

The Use of In Silico Genome-Scale Models for the Rational Design of Minimal Cells



Jean-Christophe Lachance, Sébastien Rodrigue, and Bernhard O. Palsson

Abstract Organism-specific genome-scale metabolic models (GEMs) can be reconstructed using genome annotation and biochemical data available in literature. The systematic inclusion of biochemical reactions into a coherent metabolic network combined with the formulation of appropriate constraints reveals the set of metabolic capabilities harbored by an organism, hereby allowing the computation of growth phenotypes from genotype information. GEMs have been used thoroughly to assess growth capabilities under varying conditions and determine gene essentiality. This simulation process can rapidly generate testable hypotheses that can be applied for the systematic evaluation of growth capabilities in genome reduction efforts and the definition of a minimal cell. Here we review the most recent computational methods and protocols available for the reconstruction of genome-scale models, the formulation of objective functions, and the applications of models in the prediction of gene essentiality. These methods and applications are suited to the emerging field of genome reduction and the development of minimal cells as biological factories.

Keywords Computational modeling · Metabolic modeling · Constraint-based modeling · Gene essentiality prediction

J.-C. Lachance · S. Rodrigue

Département de Biologie, Université de Sherbrooke, Sherbrooke, QC, Canada

B. O. Palsson (✉)

Department of Bioengineering, University of California, San Diego, CA, USA

Bioinformatics and Systems Biology Program, University of California, San Diego, CA, USA

Department of Pediatrics, University of California, San Diego, CA, USA

Novo Nordisk Foundation Center for Biosustainability, Technical University of Denmark, Lyngby, Denmark

e-mail: palsson@ucsd.edu

© Springer Nature Switzerland AG 2019

A. R. Lara, G. Gosset (eds.), *Minimal Cells: Design, Construction,*

Biotechnological Applications, https://doi.org/10.1007/978-3-030-31897-0_6

1 Introduction

1.1 Engineering Biology

Biologists of the past 200 years have provided a breadth of knowledge on the fundamentals of life on Earth. Current theories and dogmas emerged from a maze of suppositions and hypotheses through the succession of key findings and incremental advances. Nowadays, few molecular functions necessary to support life remain unknown. While biology has considerably matured as a science discipline, we will discuss here how the exhaustive characterization of organisms along with proper modeling frameworks should drive a new era, in which cell engineering will develop into an independent discipline. Because of their lower complexity, microorganisms—particularly minimal bacteria—are expected to play a very important role in this endeavor.

Scientists investigate that which already is; Engineers create that which has never been.

—Albert Einstein

We discuss here the historical context and key steps leading to the birth of biological engineering. This historical recap should highlight the importance of minimal cell models while providing readers with a perspective on the entire field of biology. We divided it in four stages: *classical biology*, *molecular biology*, *genomics*, and finally *synthetic biology* (Fig. 1).

1.1.1 Classical Biology

In 1859, Darwin published his work entitled “On the Origin of Species by Means of Natural Selection, or, the Preservation of Favoured Races in the Struggle for Life.” Less than a decade later, in 1865, Mendel proposed mechanisms for heredity. Both theories used observable phenotypes at the organism level to infer potential mechanisms driving their evolution. While Darwin’s work explained the driving forces underlying the emergence of phenotypes and speciation, Mendel’s work was focused on a mechanistic explanation of the basic principles of genetics. While not specifically described by Mendel, his conclusions gave birth to the concept of gene. Understanding the chemical basis of the gene and heredity then became the main endeavor of this first era of biology, defined here as the *classical biology* era (Fig. 1). This objective remained one of the grand challenges of biology until, in 1953, Watson and Crick published the structure of deoxyribonucleic acid (DNA) (Watson and Crick 1953). This historically significant finding allowed scientists to ask more intricate questions on the molecular functions sustaining life, marking the beginning of the *molecular biology* era (Waddington 1961).

nucleotides lacking the 3'-OH group necessary for chain elongation (Sanger et al. 1977a). This method enabled sequencing the complete 5375 bp genome of phage ϕ X174 (Sanger et al. 1977b). While Sanger's dideoxy termination method generated sequences ranging from 15 to 200 nucleotides, massive scaling up was necessary to allow more ambitious sequencing efforts.

The automated DNA sequencer (Smith et al. 1986) and the advent of shotgun sequencing (Anderson 1981) considerably increased the capacity of DNA sequencing, resulting in longer whole-genome sequences (WGS) (Heather and Chain 2016). Following the Santa Cruz workshop in 1985 (Sinsheimer 1989), the Human Genome Project (HGP) was initiated and reached completion in 2001 (Venter et al. 2001; Lander et al. 2001). Taking advantage of the technologies developed for the HGP, smaller-scale WGS projects were completed before the new millennium (Fig. 1). In a historic reference to the first type II restriction enzyme isolated, the first WGS of a free-living organism, *Haemophilus influenzae*, was reported in 1995 (Fleischmann et al. 1995). The genome of *Mycoplasma genitalium*, the smallest free-living organism, was published shortly after (Fraser et al. 1995). The more complex model organisms *Saccharomyces cerevisiae* and *Escherichia coli* followed in 1996 and 1997, respectively (Goffeau et al. 1996; Blattner et al. 1997).

1.1.3 Genomic

We arbitrarily defined the beginning of the genomic era with the completion of the first WGS of a free-living organism, the ~1.9 million bp genome of *Haemophilus influenzae* (Fleischmann et al. 1995) (Fig. 1). The number and size of WGS made available following this first sequencing effort steadily increased, eventually including the ~3.2 billion base pairs (bp) haploid human genome (Venter et al. 2001; Lander et al. 2001). Automation and computational tools were further improved to expand the capacity of Sanger sequencing. Nevertheless, the advent of next-generation sequencing (NGS) technologies developed by private companies upon the completion of the Human Genome Project represented a major breakthrough. While the sequencing by synthesis paradigm was preserved between Sanger sequencing and the NGS methods, the ability to parallelize the sequencing within one reaction massively increased the throughput (Heather and Chain 2016).

NGS allowed the elaboration of new initiatives such as the 1000 Genomes Project (Spencer 2008) and the Human Microbiome Project (McGuire et al. 2008), both initiated in 2008. The power of NGS technologies could not be exemplified any better than by considering that, in spite of their much greater scale, these projects reached their primary goals within 4 years (1000 Genomes Project Consortium 2012; Human Microbiome Project Consortium 2012), merely one third of the time required for the HGP. The accessibility of sequencing now contributes to an unprecedented breadth of knowledge that is meant to continue. Recently, the development by Oxford Nanopore (minION) of a portable, benchtop, real-time sequencer (Lu et al. 2016) further expands the applications of NGS for fundamental discovery.

Obtaining the genome sequences of a wide number of species is of paramount importance for understanding their phylogenetic relationships and the potential

functions they encode. However, the genetic information encrypted in the DNA of a cell is essentially static and does not reveal the dynamic nature of molecular phenotypes. This reality became evident shortly after the completion of the HGP, when the predicted number of genes in human was found to have been grossly overestimated (Brower 2001). Gladly, the efforts for high-throughput interrogation of other important cellular components started early with the development of untargeted approaches for the sequencing of proteins (Mørtz et al. 1996). More than a decade later, the elaboration of a protocol for high-throughput RNA sequencing using NGS technologies revealed the full transcriptomic profile of yeast (Nagalakshmi et al. 2008). From that point on, the three main macromolecules of the *central dogma* of biology (Crick 1970) could be sequenced at a genome-scale in an untargeted way.

The remaining components of the cell are less ubiquitous, and the application of untargeted methods for organism-wide identification is more complex. The identification of all water-soluble components is termed metabolomic, whereas the hydrophobic content is generally referred to as lipidomic (Riekeberg and Powers 2017). Liquid chromatography followed by mass spectrometry (LC-MS) allows for both metabolomic and lipidomic determination (Riekeberg and Powers 2017; Yang and Han 2016) with the difference in extraction method reflecting the polarity of the compounds. These methods along with others (Ingolia et al. 2009; Lahner et al. 2003; Zamboni et al. 2009) allow the characterization of a dynamic state of the cell that can be leveraged in systems biology (Haas et al. 2017).

1.1.4 Synthetic Biology

The term synthetic biology is closely associated with the application of engineering principles to biological systems. DNA synthesis enabled the generation and assembly of synthetic DNA parts. In turn, these capabilities allowed creating “that which did not exist,” hence defining synthetic biology as a field (Andrianantoandro et al. 2006; Heinemann and Panke 2006; Hughes and Ellington 2017). The first attempt at synthesizing DNA happened shortly after the elucidation of its structure. In 1957, Bessman and colleagues used the DNA polymerase from *E. coli* to produce DNA fragments. They noted that the presence of polymerized DNA is necessary for the reaction. This concept was later reused both by Sanger for DNA sequencing (Sanger et al. 1977a) and later for the famous polymerase chain reaction (PCR) (Saiki et al. 1985). The DNA oligonucleotide primers used for the development of PCR were produced using the phosphoramidite method (Matteucci and Caruthers 1981; Beaucage and Caruthers 1981). This chemistry is still currently used in most modern DNA synthesis platforms (LeProust 2016) but is limited by the oligonucleotide length that can be obtained without accumulating undesired mutations. This problem was circumvented by Stemmer in 1995, who first reported a technique to generate a long synthetic DNA fragment (>1000 bp) by assembling oligonucleotides (Stemmer et al. 1995). While the cost of DNA synthesis stayed more or less the same in the last 10 years (Hughes and Ellington 2017), recent progress toward high-throughput DNA synthesis strategies using microarrays may soon overcome this issue (LeProust 2016) and promise to make the synthesis of large DNA fragments an affordable

solution for routine molecular biology experiments or industrial strain design (Hughes and Ellington 2017; Bassalo et al. 2016).

The utmost objective of DNA synthesis is the conception and assembly of entire genomes. To reach this goal, the development of robust methods to assemble DNA fragments into larger sequences was necessary. This goal was met in 2008 when a team at the John Craig Venter Institute (JCVI) realized the complete synthesis and assembly of the *Mycoplasma genitalium* genome (Gibson et al. 2008). This achievement was made possible by a hierarchical strategy relying on in vitro recombination of DNA cassettes (Gibson et al. 2009). The assembly of overlapping DNA oligonucleotides to create larger fragments was later shown to be even more effective in vivo using yeast (Gibson 2009). The development of whole-genome synthesis and assembly methods together with that of whole-genome transplantation (Lartigue et al. 2007) enabled the creation of the first cell living with an entirely synthetic genome (Gibson et al. 2010).

Recent years have seen groundbreaking synthetic biology efforts that will undoubtedly have an impact on the future of this field. In 2014, Romesberg and colleagues created a bacteria functioning with an altered DNA containing six different bases (Malyshev et al. 2014), thereby offering an additional base pairing combination. No known living organism contains these synthetic nucleobases, resulting in a new life form on Earth. Following the path of the first free-living organism containing a synthetic genome, the team at JCVI designed and assembled a cell with a greatly reduced gene content, resulting in a working approximation of a minimal cell (Hutchison et al. 2016). Finally, the Sc2.0 project was initiated and in 2017 an international consortium reported the complete de novo synthesis of five entire chromosomes of the yeast *Saccharomyces cerevisiae* (Richardson et al. 2017).

With the advent of NGS, multiple omics methods for the dynamic characterization of the cell, targeted genome editing methods (Qi et al. 2013), and the development of high-throughput DNA synthesis and assembly methods, synthetic biology is now poised to create life forms that will revolutionize many industrial research fields such as microbial drug synthesis, biofuel production, or alternative approaches for disease treatment (Smolke et al. 2018).

1.2 The Minimal Cell Concept

The hydrogen atom of biology

—Harold J. Morowitz

The idea of a minimal cell was approached by biophysicist Harold J. Morowitz in a guest lecture in 1984 (Morowitz 1984) where he reasoned that a free-living organism would have a lower limit on the number of atoms from which it is composed. Below this number, the necessary functions to support life would not be met. This logical deduction somewhat resembles that of Schrödinger in his famous book *What is life?* (Schrödinger 1967), where the famous physicist questioned the material support of the gene and applied limitations imposed by quantum physics to correctly predict that it would be a molecule that could form a

crystal. In his lecture, Morowitz proposed that the mollicutes, a phylogenetic group of bacterium deprived of a cell wall, would be the best candidates to match the constraint and the endeavor of generating what he then defined as a “minimal cell.” The choice was made firstly according to their size with the idea that the smaller cell, much like the hydrogen atom in physics, would be the simplest system to study and hence yield fundamental understanding applicable to other, more complex biological systems. The prediction was accurate. The mollicute *Mycoplasma genitalium*, second entirely sequenced free-living organism (Fig. 1) (Fraser et al. 1995), has the smallest gene content of any known naturally occurring organisms. The purpose of studying minimal cells was then clearly stated: defining the basic principles of life (Glass et al. 2017).

As soon as more than one whole-genome sequence was generated, Mushegian and Koonin sought to compare the two phylogenetically distant species in the hope of finding orthologous genes that would be a working approximation of a minimal gene set (Mushegian and Koonin 1996). The initial proposition was that 256 genes would be sufficient to support life. This proposition was later experimentally shown to be a relatively low estimate. Gene essentiality in genome-reduced bacteria probed with random transposon insertion estimated that the number of genes would be between 265 and 350 (Hutchison et al. 1999). With the increasing number of whole-genome sequences available, comparative genomics allowed to deepen the understanding of the concept of minimal gene set. When comparing the eukaryote *Saccharomyces cerevisiae* to its initial proposition, Koonin realized that very few genes were conserved (only 40%) (Koonin 2000). The suggested explanation for this was that non-orthologous gene displacement (Koonin et al. 1996) (NOD) would have a higher frequency than originally anticipated. The definition of NOD states that genes with similar functions can evolve independently. This induced a paradigm shift in the concept of minimal gene set, where the identity of the genes themselves was deferred to a second level, with the functional activity they provide becoming more important. From an engineering standpoint, the minimal set of functions is indeed more interesting than the set of genes (Danchin and Fang 2016). In this context, the various genes become interchangeable parts to accomplish a given function (Fig. 2).

The many progresses in synthetic biology realized by scientists at the JCVI led to the design and synthesis of the first working approximation of a minimal cell: JCVI-syn3.0 (Hutchison et al. 2016). The 463 genes encoded in the chromosome of this cell is a lower number than any other known free-living organism (Glass et al. 2017) but is substantially higher than computationally and experimentally determined minimal gene sets (Mushegian and Koonin 1996; Hutchison et al. 1999; Koonin 2000; Glass et al. 2006). Although essential for cell growth, a significant fraction (149/463, ~30%) of the JCVI-syn3.0 gene set had no proposed function (Hutchison et al. 2016; Glass et al. 2017; Danchin and Fang 2016). Danchin and Fang extensively reviewed these genes in search for a molecular mechanism that would need to be fulfilled (Danchin and Fang 2016) and provided potential functions based on known or projected necessities for 32 of those 84 generic and 65 “unknown unknowns.” The validity of these hypotheses has yet to be determined, and therefore the original question raised by Morowitz, seeking for the completeness of molecular biology, remains unanswered.

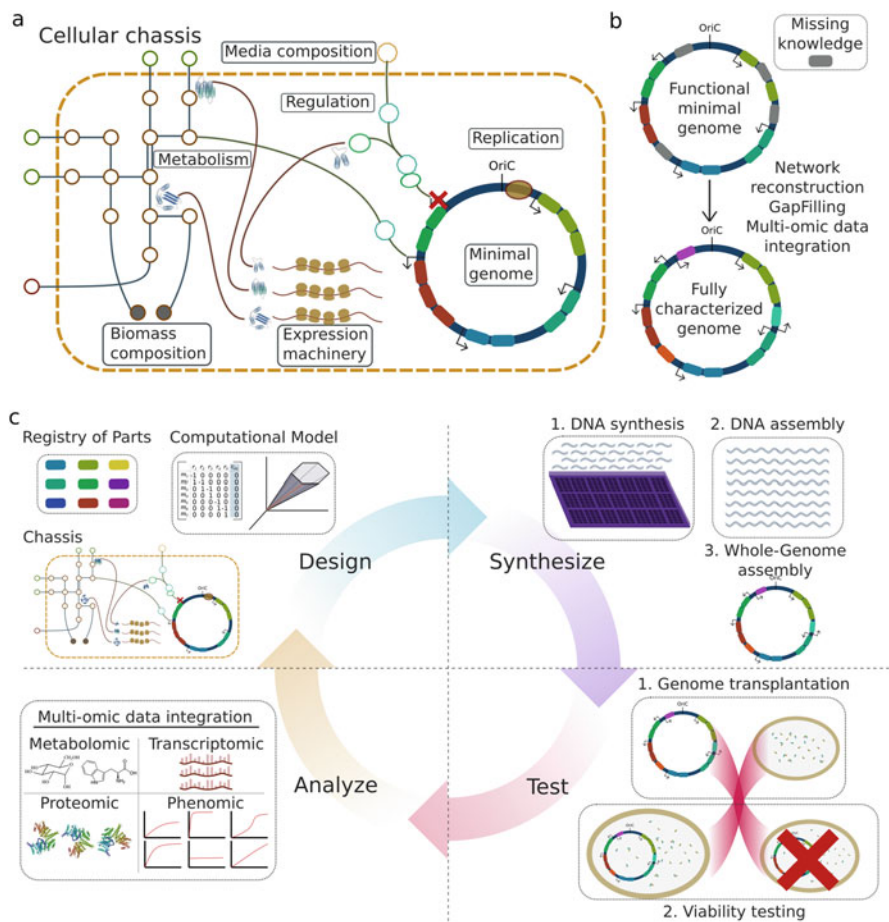


Fig. 2 Design of cells using a computer model. **(a)** Naive representation of a cellular chassis in which all mandatory cellular functions and their interactions are understood and characterized. **(b)** The generation of computer models for minimal cells can accelerate the identification of missing knowledge and facilitate the generation of hypothesis for essential uncharacterized cellular functions. **(c)** A design-build-test-analyze loop for the generation of minimal cells and their improvement toward production strains. Mathematical models are used to predict functional genotypes, and the current DNA synthesis technologies mentioned in the text are used to generate the proposed genome. Cloning of entire genomes in living cells allows to test for viability, and multiple omic datasets are used to characterize the synthetic organism

While absolute minimal cells should inform on the first principles of life, heavily reduced cells entail a strong interest from an engineering perspective. Cho and colleagues reviewed some of the potential advantages of reduced bacteria for strain design (Choe et al. 2016). As mentioned, high-throughput characterization of cellular phenotypes through omic data generation and increase in throughput of DNA synthesis should allow for in vitro fabrication of designed genomes (Fig. 2). We list here some of the advantages that were pointed out.

The first bacterium harboring a synthetic chromosome, JCVI-syn1.0 (Gibson et al. 2010), was reported to represent a 40-million-dollar endeavor (Sleator 2010). A smaller genome obviously results in a reduced DNA synthesis cost. While a decrease in the price per base was announced by companies such as Twist Bioscience, others do not foresee a price reduction that would defeat Moore's law (Smolke et al. 2018). Hence, the economic impact of generating several small genomes would remain significant. From a systems biology or design perspective, a reduced number of genes translate into a lower probability of negative interactions that could affect the desired outcomes of the initiated design. The development of high-throughput and untargeted approaches in the *genomic* era has allowed the rapid characterization of cells, but the outcome of genetic modifications is still not entirely reliable. The idea here is that genome-scale modeling of minimal cells could lead to more accurate model predictions. For instance, efforts have already been invested in reducing the complexity of metabolic models in the attempt of making the generated solutions more human readable (Ataman and Hatzimanikatis 2017). Genome reduction and minimization also allows for the design of biocontainment strategies. These include auxotrophy(ies) or programmed cell death, which will be highly beneficial as synthetic biology becomes more common in commercial applications. Finally, for more complex organisms, the deletion of genomic sections could accelerate genome replication while potentially increasing genomic stability through the removal of duplicated elements.

2 Constraint-Based Modeling

In the last section, we reviewed how biology developed from a pure science discipline at its inception to a mechanistic and engineering discipline in our times. The advent of high-throughput characterization methods for organisms together with biological reductionism that entails mechanistic description of processes sustaining life led to the birth of synthetic biology as a field. In this context, we reviewed the idea of a minimal cell, which should be a functional chassis for the design of production strains or a platform for fundamental understanding of biology (Danchin 2012). As mentioned, the current status of minimal cell research, with the 149 genes with no function associated with the synthetic organism JCVI-syn3.0 (Hutchison et al. 2016), requires further characterization of molecular functions to reach a complete understanding of every molecular functions necessary to sustain life. This biological reductionism approach should feed into a computational framework geared toward integrative analysis where *in silico* simulations based on mathematical models take advantage of the high-throughput methods to generate predictions. In this section we describe flux balance analysis (FBA) (Orth et al. 2010), a mathematical approach that allowed the generation of genome-scale models from whole-genome sequences around the new millennium (Edwards and Palsson 1999, 2000). This modeling approach is a solid basis on which minimal cells can be designed *in silico*.

2.1 Concept of Constraints in Metabolism

Flux balance analysis (FBA) arose from an attempt at generating simple coarse-grained models for the fermentation of the chemical industry's feedstocks in bacterial hosts (Papoutsakis 1984). One initial model suggested by Papoutsakis relied on the assumption that the fermentative process could be resumed in a single stoichiometric equation where elemental balance is conserved. Interestingly, the definition of the so-called fermentation equation utilized the known stoichiometry of reactions involved in the fermentation of butyric acid. The stoichiometry of biochemical reactions in a metabolic network was later used by Majewski and Domach in an attempt to establish a theoretical understanding for acetate overflow metabolism in *Escherichia coli* cells grown under aerobic conditions (Majewski and Domach 1990). The model presented for the acetate overflow entailed many key elements of FBA. The proposed hypothesis was that a flow network with a given objective could represent and explain the shift in metabolic state of *E. coli* responsible for the excretion of acetate.

The problem was summarized as a linear optimization problem on which network constraints would apply. Fixing the objective as to maximize the production of ATP and applying two constraints, (1) limiting the amount of reducing equivalents that can be produced by the electron transport chain and (2) assuming that a given enzyme of the Krebs cycle is limiting hereby limits the flux through a given reaction, the authors demonstrated that linear programming could correctly predict a bacterial metabolic state.

The use of a metabolic flux network optimized with linear programming served as a basis for the development of mathematical formalism for FBA (Savinell and Palsson 1992a, b). The concept was extended with the definition of a stoichiometric matrix (S). In this matrix, each column represents a reaction in the metabolic network, and each row is a different metabolite (Fig. 3). The mathematical formulation of the metabolite concentration over time using the S matrix then becomes:

$$\frac{dX}{dt} = S \cdot v \quad (1)$$

where X is the vector of metabolites and v is the vector of fluxes. FBA assumes that the metabolic network will reach a steady state. In this case, the concentration of metabolites over time should be in equilibrium where the inputs are equal to the outputs so that:

$$0 = S \cdot v \quad (2)$$

FBA has the advantage of requiring only the stoichiometry of the reactions to operate. The details of thermodynamics for each reaction are not necessary. Nevertheless, reaction directionality can be obtained from thermodynamics, hereby adding another set of constraint on the system. A physiologically meaningful objective (Z) can be defined in order to simulate the desired metabolic phenotype.

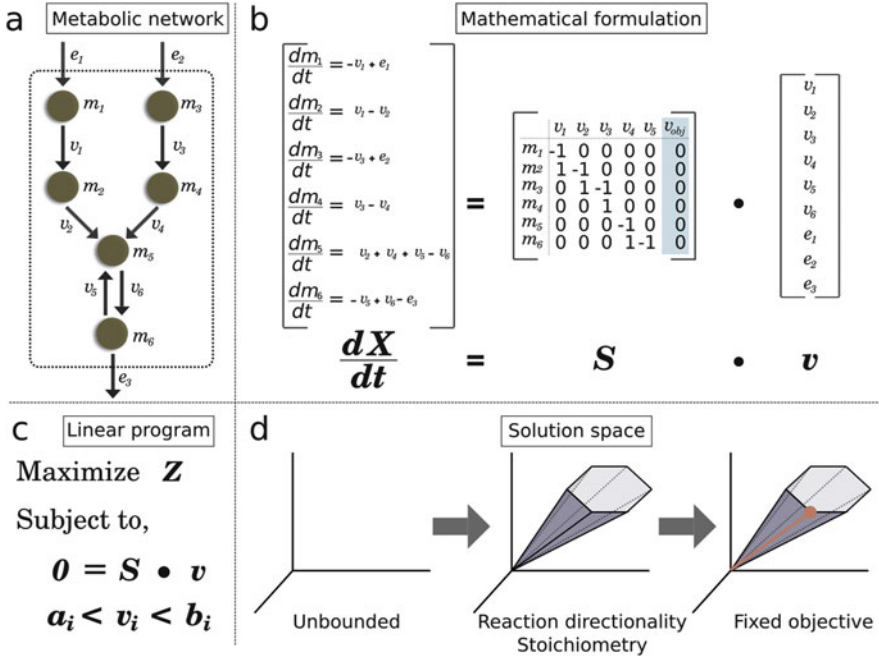


Fig. 3 Constraint-based modeling using linear programming. **(a)** A given metabolic network composed of metabolites (nodes) and reactions (links) can be represented in the form of a stoichiometric matrix S . **(b)** In this matrix, each row represents a metabolite while each column is associated with a reaction. The variation of metabolite concentration over time ($\frac{dX}{dt}$) can then be represented as the matrix-vector product of S by v , the vector of fluxes for each reaction in the network. **(c)** Defining a physiologically meaningful objective Z , the optimal solution for the metabolic network can be represented as a linear optimization problem with given flux constraints on metabolic reactions, and, at steady state, the variation in metabolite concentration is equal to 0. **(d)** The application of constraints on the optimization problem limits the solution space, while applying a proper objective allows finding the line of optimality within that bounded convex solution space

$$\begin{aligned} &\text{maximize } Z, \\ &0 = S \cdot v \\ &a_i < v_i < b_i \end{aligned} \tag{3}$$

This mathematical formulation can be solved using linear programming and allows finding the optimal solution of a given metabolic network at steady state. We will now review how this formulation allows for the generation of genome-scale models and how the objective function can be tailored to represent specific physiological states.

2.2 *Metabolic Network Reconstruction*

The completion of whole-genome sequences in the genomic era (Fig. 1) allowed the generation of genome-scale metabolic models (GEMs). In most cases, genome annotation yields the predicted function of proteins encoded by an organism. For the metabolic enzymes, the annotation together with extensive literature research can link a DNA sequence to a biochemical reaction in the metabolic network. The process of extracting a maximum number of reactions from the genome is termed reconstruction and has been reviewed in detail (Thiele and Palsson 2010). We explain the key steps in the reconstruction of a stoichiometric matrix at the genome scale (Fig. 4).

First, a draft reconstruction must be generated. The process of building this draft can be performed manually or automatically. The automated methods for draft reconstruction of metabolic networks are reviewed in Sect. 3. The generation of a draft reconstruction process consists in extracting biochemical reactions from genome annotation. Through this process, the stoichiometry of every reaction in the network is obtained. The reactions can be fetched from annotated EC numbers or gene names, and the candidate metabolic genes are linked to a reaction of the S matrix. The association between a gene and its reaction is key for future predictions generated by the model and should therefore be evaluated carefully.

Second, the initial draft is examined more closely through a refinement step. The key elements of this step are the examination of the mass-balance conservation for each reaction; that is the number of atoms in the reactants should be equal to the number of atoms in the products. The same rationale goes for the charge of the reactions. The balanced equations should have a neutral charge. These assumptions are linked to the fundamental principles of chemistry, hereby ensuring that no mass or energy is created in any reaction of the metabolic network. The gene-protein-reaction (GPR) association is then verified for all reactions, and a confidence score is attributed that facilitates further evaluation of the results once the model simulations are compared to experimental data.

Non-gene-associated reactions are then added. Spontaneous reactions are reactions for which no gene is associated and represent the natural occurrence of a reaction that is thermodynamically favorable without the need for a gene-encoded catalyst (enzyme). Other non-gene-associated reactions are exchange, sinks, and demands. These reactions represent the environment/culture media of the cell. They are not mass-balanced or charge-balanced by default since they represent the uptake/dumping of metabolites from/to the media. They are nevertheless necessary for the simulation of growth phenotypes under a given environment. Finally, a biomass reaction and ATP maintenance (ATPM) are added. The idea of a biomass reaction is to force the model to produce metabolites necessary for the growth of the organism, and its potential in simulating growth will be discussed later. The ATPM reaction is an ATP hydrolysis reaction that allows modelers to set a certain rate of ATP consumption for a growing cell. Knowing the experimental energy requirements hereby allows for more precise growth rate predictions.

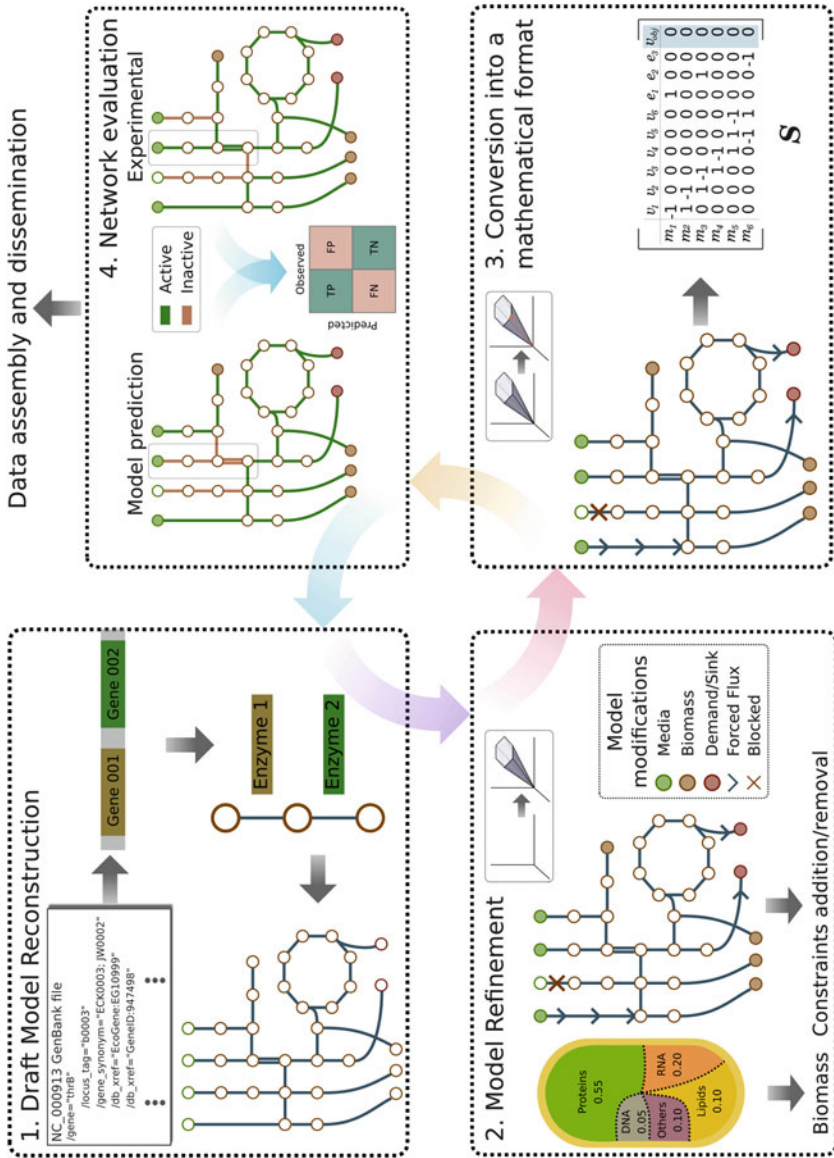


Fig. 4 The four main stages of metabolic network reconstruction and simulation. (1) A draft reconstruction is generated from genome annotation. (2) The draft reconstruction is refined by generating a biomass equation and applying constraints. (3) The reconstruction is converted into an optimization problem allowing the simulation of cellular phenotype. (4) Model predictions are compared to experimental observations. The discrepancies between predictions and observations are used to improve the model in an iterative process

Finally, the reconstructed network is ready for validation and simulation. Setting simulation constraints together with a defined objective allows the formulation of predictions that can be tested. The model may or may not yield a feasible solution. In the latter case, extensive unit testing may be required to address issues in the formulated reconstruction (e.g., metabolite accumulation). Iteratively fixing those issues allows the generation of a functional model that can be used for simulation. The comparison of the formulated predictions with readily available experimental data can either confirm how the system is expected to function or reveal potential gaps in knowledge.

2.3 Objective Function

In linear programming, the objective function is the numerical value to maximize or minimize. The significance of the value to be optimized is dependent on the situation that the modeler wishes to simulate. For instance, Papoutsakis (Papoutsakis 1984) generated a model for butyrate production. In that case, the numerical value is the amount of butyrate (a feedstock of the chemical industry) that can be produced over time. To simulate and explain acetate overflow, Majewski and Domach (Majewski and Domach 1990) maximized the ATP production by the network. Finally, the FBA red blood cell (RBC) model (Bordbar et al. 2011) maximizes the flux through the Na⁺/K⁺ ATPase pump. The choice of the objective function hence reflects the physiological situation and is key for the predictions generated by the model. Since the RBC cannot duplicate itself, it is assumed that the actual biological objective of the cell is to maintain a proper gradient of sodium and potassium, a task that requires the production of energy in the form of ATP. This proper objective definition along with the integration of high-throughput experimental data allowed the identification of biomarkers for RBC degradation upon storage (Yurkovich et al. 2017).

A common objective for modelers is to predict a growth phenotype. In this case, a biomass objective function (BOF) is defined that contains every metabolite necessary for the doubling of the cell (Feist and Palsson 2010). The BOF is modeled through the addition of an extra reaction (column) of the stoichiometric matrix (S). The proportion of each element within the cell is given as stoichiometric coefficients in the reaction. In order to provide an estimation of growth rate, a basis is given (Varma and Palsson 1993) such that the product of cell weight by time is equal to 1 g of cellular dry weight per hour (gDW/h). While the metabolite composition of the BOF may vary from a species to another, many components are shared across prokaryotes that are necessary for growth (Xavier et al. 2017). The proper integration of biomass components effectively present along with the stoichiometric coefficients that reflect the experimental composition of the cell (Beck et al. 2018) in species changes the accuracy of the model predictions (Lachance et al. 2018). The definition of the BOF is therefore crucial to generate gene essentiality predictions, a key topic for the endeavor of generating minimal cells *in silico* through genome-scale modeling with FBA.

2.4 Conversion into a Mathematical Format and Evaluation

Genome-scale models have the power to simulate the organism's metabolic capabilities. Converting the reconstruction into a mathematical format through the establishment of proper objective (e.g., precise definition of the biomass objective function) and constraints (e.g., media definition, internal flux bounds, uptake and secretion rates, etc.) provides the model with that potential. The model can then be used to formulate predictions of the metabolic state of the organism. The predictions formulated by the model and the datasets used to validate them vary based on the scientific objective of the conducted research.

A common objective used to enhance the quality of the model is growth (maximize flux through the biomass reaction). The direct prediction is the growth rate, which can be matched with the experimentally determined value. Getting a correct doubling time is dependent on correctly determining the cellular energy expenses and the stoichiometric coefficients of biomass precursors included in the biomass reaction. Optimizing for biomass production can also be used to determine gene essentiality by iteratively removing single genes and solving the model, a common measure of a model's quality, and will be covered in more detail later (see Sect. 4). Finally, FBA provides a flux state with the given solution. While FBA finds a unique optimal solution for the given objective function, many flux states may lead to it. Different methods have been developed that study the variability of the flux states that will be covered later (Gudmundsson and Thiele 2010). Modelers can then sample and study the variability of the flux state to identify fluxes that are out of biologically feasible ranges and apply supplementary constraints that improve the model's quality.

Compliance with experimental data can then be assessed. As mentioned, the gene essentiality prediction of the model is commonly used as a reference for a model's general quality since it accounts for the quality of the assigned GPRs together with the network topology, biomass, and media composition. A Punnett matrix is often used to visualize the predictions formulated by the model with all four combinations of true/false positive/negative represented. A metric such as accuracy or Matthews correlation coefficient can be used to quantify the quality of the model's prediction in a single number.

3 Computational Methods for Genome-Scale Reconstruction

With whole-genome sequences available for a greater number of species, the number of genome-scale metabolic models (GEMs) developed over the last two decades increased steadily (Monk et al. 2014). The number of computational tools tailored for the reconstruction of biological metabolic networks as well as the analysis and integration of omics data in these models has been developed accordingly (Lewis et al. 2012). In this section we review the methods and databases used for the

reconstruction of the stoichiometric matrix (S), filling reaction gaps in the network and objective definition.

3.1 *Tools for Network Reconstruction*

The reconstruction of a GEM begins with the reconstruction of the stoichiometric matrix of reactions and metabolites (Fig. 4). Careful inspection of the genome annotation allows to link a gene and its sequence to a particular function in the network. In order to connect these elements together, modelers can use the many publicly available databases of pathways and biochemical reactions that are specifically designed to provide the association between genes, biochemical reactions, and/or the metabolic pathways (Kanehisa et al. 2017; Artimo et al. 2012; Placzek et al. 2017; Wattam et al. 2017; Aziz et al. 2008; Devoid et al. 2013; Fabregat et al. 2018; King et al. 2016; Caspi et al. 2008).

The identification of metabolic candidates in the reference genome is the first step of genome-scale reconstruction. To do so, modelers can either obtain enzyme commission numbers (EC) from specialized software (Nursimulu et al. 2018) or extract the information contained in the publicly available databases. In both cases, the standardization of metabolite and reaction identifier is key for the consistency and readability of the model. Since these identifiers vary considerably from one database to another, draft reconstructions may not be readable in another format. This type of issue has been addressed and can potentially be overcome by the use of MetaNetX (Moretti et al. 2016) or BiGG (King et al. 2016). MetaNetX is a web-based platform that attempts to centralize the identification of metabolite and reactions while also providing methods for automated genome-scale reconstructions. The main focus of the BiGG database is to list GEMs formulated in the BiGG nomenclature. Nevertheless, reactions and metabolites stored on BiGG are linked to other commonly used databases such as Reactome, KEGG, SEED, CHEBI, BioCyc, and MetaNetX. Choosing an identification system and ensuring the conversion from an annotation system to another is therefore key for the establishment of the draft reconstruction of the model.

The network reconstruction can be executed in different frameworks based on modeler's preferences. The SEED (Devoid et al. 2013) and Merlin (Dias et al. 2015) both allow for the automated generation of GEMs. While these functional models provide predictions, exhaustive literature search and model fine-tuning are usually necessary before a model is released (Thiele and Palsson 2010). The Open COBRA (Constraint-Based Reconstruction and Analysis) suite is designed to include every step of the process and is currently available under three different programming languages: Python (COBRAPy, Ebrahim et al. 2013), MatLab (COBRA Toolbox 3.0, Schellenberger et al. 2011), and Julia (COBRA.jl, Heirendt et al. 2017). Implemented in MatLab, the RAVEN toolbox (Agren et al. 2013) is another option for reconstruction that also entails the visualization of the metabolic networks. The sybil toolbox allows R users to operate FBA, MOMA (Segrè et al. 2002), and ROOM (Shlomi et al. 2005) in their preferred language (Gelius-Dietrich et al. 2013). While Open COBRA

does not specifically entail visualization, the generated network can be visualized by building a metabolic map with Escher (King et al. 2015).

3.2 *Tools for Network Analysis*

The main functionalities of the reconstruction toolboxes mentioned above are to allow the creation of foundational elements of models (metabolite, reaction, and gene objects) and store them into a model object that can be saved or imported in the desired format(s). These toolboxes also include basic model simulation functionalities such as the definition of objective and a bridge to the solver interface necessary to optimize the model. These preliminary simulation functionalities are useful for the conversion of the model into a mathematical format which can later be used for more intensive simulation processes and the evaluation of the organism's metabolic capabilities. We cover here some of the algorithms that have been developed to increase the quality of models before they are used for simulation.

3.2.1 **Gaps in Network**

In order to reveal biological capabilities, the network needs to be maximally functional, that is, flux can go through as many reactions as possible. As discussed, the mathematical formulation of FBA (steady-state assumption) does not allow for the accumulation of metabolites. This means that for a given linear pathway, a single missing reaction would block flux through every upstream and downstream reactions. The entire pathway would then be considered unfunctional, a hypothesis of debatable biological value that should be handled with care by modelers.

Several algorithms have been developed that aim at identifying problematic metabolites and reactions, solving gaps in the biological network, finding reactions that could fill those gaps and eventually proposing genes that could catalyze the suggested reaction(s) (Orth and Palsson 2010; Pan and Reed 2018). As mentioned, the general framework of these algorithms first identifies dead-end metabolites, that is, metabolites that cannot be produced or consumed in the metabolic network. Solving a gap in the network may be accomplished by adding one or many reactions. To find candidate reactions, these algorithms usually query larger reaction databases such as those contained in KEGG (Kanehisa et al. 2017) or MetaCyc (Caspi et al. 2008). The value of adding a given specific set of reactions can only be measured by the relatedness of this proposed mechanism to the actual species being studied. Therefore, the third step aims at identifying the best possible genes that can associate with those reactions.

The first gap filling algorithm (Satish Kumar et al. 2007) did not include this third task, but subsequent versions incorporated different ways to input experimental data along with the suggested reactions. GlobalFit (Hartleb et al. 2016) and ProbAnnoPy (King et al. 2018) are good examples of gap filling methods attempting to improve a metabolic model based on experimental data. For a deeper coverage of the available methods, interested readers can consult this review by Pan and Reed (Pan and Reed

2018). GlobalFit was used to increase the quality of two GEMs, *Escherichia coli* iJO1366 and *Mycoplasma genitalium* iPS189. It uses a bi-level optimization problem to minimize the gap between predicted gene essentiality and the experimental data and allows the incorporation of new reactions within the model or new exchange reactions (media components) as well as biomass precursors (metabolites of the BOF). ProbAnno (Web and Py) attributes a probability based on BLASTp search e-value to rank the reactions used to fill the gaps in the network.

Such approaches are relevant in the current context of minimal cell research and design. While a minimal cell has already been generated experimentally, the number of genes it contains for which a precise function could not be attributed is a significant portion of the complete genome (149/463). An ideal cellular chassis should have no unknown properties (Danchin 2012) since it would serve as a blueprint for further design. Hence, reconstructing metabolic networks and using gap filling algorithms that provide functional annotation are systematic ways to address the fulfillment of missing knowledge.

3.2.2 Objective Functions

The metabolic objectives of the cells can be summarized in a reaction of the stoichiometric matrix and set as an objective: the biomass objective function (BOF). The identification of key components necessary for a cell to grow is nevertheless a daunting task. This process can be accomplished in a biased way, which attempts at incorporating as much of the current knowledge of the organism's composition as possible or in an unbiased way where experimental data is utilized to infer the cellular objectives. A worthy effort at summarizing the current knowledge on prokaryotic biomass composition was accomplished by Rocha and colleagues (Xavier et al. 2017). In this extensive study, the biomass composition of 71 manually curated models available in the BiGG database (King et al. 2016) was compared along with the phylogenetic distance of the species that they represent. Swapping the BOF from a model to another showed that reaction essentiality prediction is sensitive to the BOF composition. Further studying the impact of biomass composition on gene essentiality predictions of several species, the authors found a set of universally essential cofactors in prokaryotes. This foundational knowledge highlights the importance of accurate BOFs for gene essentiality prediction by GEMs and provides an important resource for future modeling work.

Using previously established essential cell components, modelers can partly define the BOF for their organism of interest. Nevertheless, the remaining part of the BOF is species-specific and can be completed using an unbiased approach. Much like gap filling, cellular objective search can be performed algorithmically. Historically, most algorithms developed for that purpose have used metabolic flux analysis (MFA) data together with various optimization methods (Burgard and Maranas 2003; Gianchandani et al. 2008; Zhao et al. 2016). While MFA is a particularly well-suited type of data for flux models, the state-of-the-art number of fluxes generated by the method does not scale to the number of reactions included in GEMs. Recently developed algorithms attempt to use other types of data to find

cellular objectives. BOFdat (Lachance et al. 2019b) uses a genetic algorithm to find the biomass compositions that provide the best match between predicted and experimental gene essentiality. The metabolites identified by the algorithm are then clustered based on their relative distance in the metabolic network to form clusters of metabolic objectives that can be interpreted by modelers. Another approach called BIG-BOSS integrates multiple omics data types to formulate the cellular objective by using a proteome constrained model, with a bi-level optimization problem similar to BOSS (Gianchandani et al. 2008). By augmenting MFA for a subset of fluxes with proteomics, the biomass composition was recovered more accurately than using just one data type alone.

4 Data Integration and Phenotypic Predictions

Once a GEM is reconstructed, converted into a mathematical format, and validated with experimental data, systematic model-driven hypothesis generation can take place that will guide the design of the desired strain. Much like the design of a production strain, the realization of a minimal cell requires in-depth knowledge of the organism that can be acquired through the generation of extensive high-throughput data. The integration of such data is made possible by GEMs, and a plethora of software has been written that helps modelers in this task. Here we cover available methods for the integration of high-throughput data as well as strain design algorithms that can be leveraged for the design of synthetic minimal cells (Fig. 6).

4.1 Cellular Objectives and Gene Essentiality Prediction

A key concept for the design of minimal cells is the identification of removable content. That is: “what genes are non-essential under laboratory conditions?”. To formulate such prediction in silico, one must first determine the requirements for growth (Fig. 6). As we mentioned, those are represented by the BOF in GEMs. The definition of the BOF is tightly linked to the evolutionary pressure applied on the strain, which is in turn function of its growth environment. For instance, given an *E. coli* cell suddenly shifted from aerobic to anaerobic conditions, the instantaneous modification in phenotype is the result of chemical and physical properties, i.e., utilization of new substrate, shift in metabolic state, changes in gene expression, etc. This rapid adaptation can be termed proximal causation (Palsson 2015). Its counterpart, termed distal causation, happens over time and is the result of evolutive adaptation. Distal causation is proper to biological systems and entails a modification of the genotype to fit the constraints imposed by the environment under which the species is grown. Since the biomass composition of a cell is a result of its evolution, each species entails different metabolite requirements for growth with some essential components shared across a wide range of organisms (Xavier et al. 2017) (Fig. 6).

4.1.1 Gene Essentiality Prediction

GEMs can be used to formulate reaction or gene essentiality predictions (Fig. 6). To formulate that prediction, all reactions in the model are individually removed, and every time, the model is optimized for growth (Suthers et al. 2009a; Joyce and Palsson 2008). An appropriate threshold is necessary to discriminate between viable and nonviable phenotypes. This allows determining which reactions are essential to carry flux through the biomass reaction in the model. The proper definition of the BOF is therefore critical for the accurate prediction of gene essentiality. The qualitative definition of the BOF defines the growth requirements of the organism, and the pathways that lead to the production of these metabolites are then activated. Other constraints such as the growth media, uptake rates of the main carbon sources, and/or oxygen also impact gene essentiality predictions.

The added value of GEMs is that most reactions within the framework are associated with one or many genes. This association between gene and reactions is termed GPR and accounts for reactions catalyzed by a single gene or multiple genes in a complex, symbolized by an “and” rule, as well as isozymes, symbolized by an “or” rule (Fig. 6). Whole-model gene essentiality can be generated easily using the reconstruction toolboxes mentioned previously since they include an implementation of this function.

It is noteworthy that GEMs are very efficient at predicting gene essentiality. Highly curated models like that of *E. coli* have achieved essentiality predictions on different growth conditions with accuracies up to 93.4% (Monk et al. 2017). The quality of the prediction relies both on the high level of biochemical information included in the *E. coli* reconstruction and the precise knowledge of the growth conditions. These limitations will be discussed later.

4.1.2 Beyond Single Gene Deletion

An advantage of GEMs is the ability to formulate predictions of synthetic lethality (SL) (Fig. 6). The phenomenon was reported early in the classical era of biology in an attempt to describe the observation that the combination of observable traits did not yield viable descendants (Bridges 1922). At the gene level, SL is known as the observation that simultaneously knocking out two genes yields a lethal phenotype when their independent individual knockout provided a viable phenotype (Fig. 6). Experimentally studying SL at the systems level is complex since it involves screening several combinations of gene knockouts. For an organism containing N individually nonessential genes, the number of combinations is the binomial coefficient: $\frac{n!}{k!(n-k)!}$. Obtaining all possible SL combinations for an organism implies generating a library of knockouts on top of a knockout library. This task has been accomplished for heavily studied organisms such as *Saccharomyces cerevisiae* in which gene editing methods are commonplace (Goodson et al. 1996; Deutscher et al. 2006) but is generally too demanding to be generated for most species.

The computation of SL genes in FBA models is computationally expensive but still orders of magnitude faster than generating the data experimentally. Using this approach can guide the design of minimal genomes since it adds a level of information that could not be otherwise fetched from the genome or single knockout libraries. GEMs also provide the possibility to expand the SL study to more than gene pairs and include triple or quadruple knockouts (Suthers et al. 2009a), an undeniable advantage over the strictly experimental approach.

An interesting usage of SL analysis is the MinGenome algorithm written by Wang and Maranas (2018). This algorithm takes as input the genome sequence of the organism of interest, a GEM, genome-scale in vivo essentiality data, operon and promoter sites, and transcription factor information. Using this information, MinGenome iteratively finds the largest section of DNA that can be removed without killing the cell. The operon structure along with the promoters and transcription factor information are used to keep regulatory elements in place which should increase the probability that the suggested minimal genome is functional in vivo.

4.2 Multiple Omics Dataset Integration

As previously mentioned, the genomic era has enabled the generation of high-throughput data (“omic”) for many different types of molecules. The integration of these sizeable datasets into comprehensive biological knowledge requires a proper framework. Metabolic models have been shown to provide a systematic way for the integration of multiple omic datasets for mechanistic understanding (Monk et al. 2014; Bordbar et al. 2014). Ralser and colleagues discussed the integration of seven types of omic datasets: genomic, transcriptomic, proteomic, lipidomic, metabolomic, ionic, and phenomic (Haas et al. 2017). The approach used to incorporate this multi-omics information in GEMs will be discussed below.

GEMs use genomic information to extract biological functions of metabolic genes. While the regulation of gene expression is not accounted in metabolic models, the transcriptomic and proteomic datasets can be used to apply supplementary constraints on the model. The flux bound can be limited based on the level of expression or simply shut down when the genes are not expressed so that the reaction(s) associated with these genes cannot carry flux (Fig. 6). The concept of a minimal cell assumes a very specialized cell with reduced metabolic capabilities. Integrating the gene expression datasets in models hence has the potential to generate context-specific models that meet the expectation of highly specialized minimal cells.

Other datasets characterize molecules outside the central *dogma* of biology (Crick 1970). Metabolite concentrations themselves are not included in standard FBA, but a variant termed uFBA (Bordbar et al. 2017) allows the incorporation of time-course metabolomics into GEM resulting in more accurate predictions of the metabolic state

of the cell. Lipidomic and ionic results are useful to determine the composition of the cell, valuable information for the definition of the BOF.

The integration of multiple omic datasets with genome-scale models provides mechanistic explanation of the organism's phenotype under different environments (Lewis et al. 2010). Using multiple omic datasets, Lewis et al. showed that *E. coli* strains evolved under different conditions modify their pattern of gene expression in a manner that is consistent with a variant of FBA termed parsimonious FBA (pFBA). pFBA uses a bi-level linear programming approach to minimize the enzyme-associated flux while maximizing biomass production. The flux state generated using pFBA was consistent with the differential gene expression across conditions. These findings provided support for the biological relevance of FBA. The implication for the design of minimal cells is that generating an FBA-based model for such a cell would allow to design its optimal state ahead of conception.

5 Systems Biology of Minimal Cells

Since the proposition by Morowitz that minimal cells would allow understanding the basic principles of life (Morowitz 1984), many efforts have been driven toward the identification of theoretical minimal gene sets through comparative genomics (Mushegian and Koonin 1996), gene-wide essentiality probing (Glass et al. 2006), and a combination of these approaches (Baby et al. 2018). Genome reduction in complex bacteria has also been attempted experimentally for several complex bacteria (Choe et al. 2016), and ultimately, nearly 10 years of groundbreaking efforts led to the realization of a working approximation of a minimal cell in vitro (JCVI-syn3.0) (Hutchison et al. 2016; Sleator 2010).

We covered how the use of GEMs, which are mathematically structured knowledge bases of metabolism, provides phenotypic predictions from genomic information and thus can be leveraged for the rational design of minimal cells (Wang and Maranas 2018). We will now review GEMs for some naturally occurring near-minimal bacteria from the class of mollicutes and then cover expansion of modeling methods beyond metabolism.

5.1 Available GEMs for Naturally Occurring Minimal Organisms

Mollicutes have been the object of much research since they were proposed as the smallest free-living organisms (Morowitz and Tourtellotte 1962). Extensive knowledge of the particular metabolism (Miles 1992) of these species allowed the generation of GEMs for the most studied of them. The first GEM for a mollicute was reconstructed for the human urogenital pathogen *Mycoplasma genitalium* (Suthers et al. 2009b). This model includes 189 genes, 168 gene-associated reactions, and

274 metabolites. Using the experimental essentiality data (Glass et al. 2006), the model was consistent with 87% of essential genes and 89% of nonessentials. While this model prediction may be accurate, several approximations were used for the reconstruction. The biomass composition and the growth and non-growth-associated maintenance costs that can be calculated from substrate uptake rate and secretion rates were estimated from *E. coli*. Since there is no defined media for *M. genitalium*, the growth media was also estimated.

Formerly known as Eaton's agent, *Mycoplasma pneumoniae* is associated with atypical pneumonia in humans (Dajani 1965; Lind 1966). Multiple efforts at characterizing *M. pneumoniae* have been undertaken providing genome re-annotation (Dandekar et al. 2000), and the transcriptome (Güell et al. 2009), proteome (Kühner et al. 2009), and metabolism (Yus et al. 2009) have been studied in-depth. This allowed the generation of a quantitative model for *M. pneumoniae* (Wodke et al. 2013). The amount of experimental data available allowed modelers to compare predicted sugar utilization and obtain the energy utilization throughout the growth phases. Constraining the model with this data allowed dissecting the pathway usages at different growth stages.

The predictions formulated by the *M. pneumoniae* model revealed that a substantial amount of ATP is not directed toward biomass production but rather toward cell maintenance functions such as chaperone-assisted protein folding, DNA maintenance, and posttranslational modifications. Strikingly, the ATPase was responsible for most of the energy usage (57–80%) in order to maintain intracellular pH and a favorable proton gradient across the membrane. The authors suggested that four factors may impact the overall energy usage: the topology of the metabolic network, the growth rate, the environmental conditions, and the cell size. These findings are particularly interesting as they show that using a systems biology approach such as GEMs for the design of bacteria can go beyond gene essentiality prediction and reveal intrinsic properties affecting cellular energetics. These factors could hardly be predicted without the integration of experimental data into a mathematically structured knowledge base.

5.2 Genome-Scale Modeling of Synthetic Minimal Organisms

Recently, modeling efforts were dedicated to JCVI-syn3.0, a synthetic working approximation of a minimal cell (Breuer et al. 2019). The metabolic reconstruction was generated using the gene annotation of the parental strain JCVI-syn1.0 (*Mycoplasma mycoides*) for which much information is available. Collecting the breadth of knowledge into a single computational format is a significant step forward in order to define the functional metabolic requirements of a minimal cell. As discussed, GEMs can be used to formulate phenotypic predictions such as gene essentiality and integrate high-throughput data such as gene expression (see Sect. 4). Breuer et al. recently provided a dataset of high-density transposon mutagenesis operated on JCVI-syn3.0 as well as a quantitative proteomic dataset. The gene essentiality data allowed identifying discrepancies between model predictions and observations.

Together with the reconstruction process, the authors were able to formulate several hypotheses on the remaining gene functions that could not be removed but nevertheless unknown.

Using proteomic data allowed contextualizing the activity of the expressed proteins in JCVI-syn3.0, but the analysis is nevertheless limited. Indeed, while the resulting GEM for this organism is the first and closest representation of a synthetic minimal cell, more accurate model predictions would have required the detailed biomass composition of this bacterium along with a chemically defined medium. Including these parameters within the model should expand its predictive capabilities.

6 Perspectives on the Use of Models for Minimal Cell Design

A key objective of minimal cell research is to gather exhaustive understanding of the cell. The FBA framework allows to generate multiple predictions on the metabolic state of the cell, but the scope is limited to metabolism. Other approaches have been developed that allow including constraints from various cellular functions such as the expression machinery, regulatory network, enzyme kinetics, and thermodynamics. We propose here to extend the definition originally proposed by Morowitz for “the completeness of molecular biology” which entailed that every element in the cell should be characterized.

What I do not understand I cannot create

—Richard Feynman

6.1 *Expanding the Scope of Models Beyond Metabolism*

6.1.1 Modeling Gene Expression

Using the constraint imposed by the stoichiometry of reactions was key for the development of flux balance analysis (Kauffman et al. 2003) and later to genome-scale models of metabolism. In an attempt to expand the scope of models beyond metabolism, Thiele et al. reconstructed the expression matrix for *E. coli* (Thiele et al. 2009). The reconstruction of this matrix, named E-matrix by opposition to the M-matrix for metabolism, was executed using the same protocol that was mentioned above (Thiele and Palsson 2010). All reactions necessary for RNA transcription and protein translation are included in the E-matrix. Interestingly, every element necessary for the synthesis of proteins is considered as a metabolite in the network. For instance, transfer RNA (tRNA) and ribosomal RNA (rRNA) are both metabolites that can be produced from the transcription reactions. The tRNA are then charged and used in another reaction which synthesizes proteins. While the number of genes

included in the E-matrix (423 genes) was smaller than that of the M-matrix [1515 genes (Monk et al. 2017)], the number of reactions is significantly higher (13,694 reactions in E vs 2719 reactions in M). The large size of the E-matrix is due to the high number of similar reactions catalyzed by the expression machinery.

Much like the M-matrix, the stoichiometry imposed by the E-matrix can be used as a constraint, and the reconstruction can be converted into a mathematical format by applying reaction bounds and fixing an objective. In this case, the uptake rates of amino acids and nucleotides need to be fixed as they are the necessary metabolites for the production of every downstream metabolites. The production of ribosome by the model can then be optimized for different growth rates since ribosome production is key for cell growth. Refining the constraints allowed the model to generate a number of ribosomes matching the experimental data. This work demonstrated the applicability of FBA to systems other than metabolism.

In order to couple the machinery of gene expression to the metabolism of the cell and generate a unified model for cellular growth, additional constraints were needed. Termed “coupling constraints,” these equations are a function of the organism’s doubling time and account for the dilution of material in doubling cells while providing upper limits on enzyme expression (Lerman et al. 2012; Lloyd et al. 2018; Thiele et al. 2012; O’Brien et al. 2013). These new constraints are both integer and linear and therefore define a mixed integer linear programming (MILP) problem. This type of problem is computationally more intensive than the regular linear programming problem solved in FBA and also requires more specific solvers (Yang et al. 2016).

6.1.2 Simulating with ME Models

An ME-model links metabolism to gene expression and can be used to generate experimentally testable predictions such as growth rate, substrate uptake and secretion rates, metabolic fluxes, and gene product expression levels (O’Brien et al. 2013). This last property is important as it simplifies comparison with experimental gene expression levels, which can now be routinely generated under many different environments. The ease of integration of multiple omics data in ME models has allowed the identification of key biological regularities (Ebrahim et al. 2016). Experimental proteomic data can provide absolute protein counts within a cell, which can be used to constrain the amount of protein in the ME model. Fluxomic data can also be used as a constraint since it provides the flux through a certain number of reactions. Combining these two types of data into ME-model simulations allowed to generate turnover rates (k_{eff}) for enzymes in the model, an example of model-driven generation of knowledge.

Simulating ME models over 333 different environmental conditions, Yang et al. identified genes consistently essential for optimal growth in *E. coli* (Yang et al. 2015). The formulated model-driven prediction of the core proteome was also found to be consistent with non-differentially expressed genes. Obviously, the functional incorporation of expression subsystems provided by the ME matrix allows the

identification of more functional categories [COG (Koonin et al. 2004)] when determining a minimal gene set. This was further exemplified by the fact that DNA replication and repair mechanisms, functional categories absent from the ME model, were not represented in the core proteome. The further expansion of the ME-model to include other cellular systems such as a constraint-based approach of the regulatory machinery would provide a working approximation of a whole-cell model requiring fewer experimental parameters than what has been previously generated (Karr et al. 2012).

Potentially because of the size of the E-matrix, the reconstruction of entire ME-models has been contained to only two species so far, namely, *Thermotoga maritima* and *E. coli* (Lerman et al. 2012; O'Brien et al. 2013). Much like the generation of M-models is eased by the existence of toolboxes, the reconstruction of ME-models could be widespread by the recent publication of COBRAME, a Python framework for the reconstruction of ME-models (Lloyd et al. 2018).

6.2 *Perspectives on the Use of Models to Design Minimal Cells*

We delved into the historical evolution of biology and highlighted the possibility that a part of the discipline could turn into a field of engineering, in which the concept of a minimal cell would play a central role. The main idea surrounding this minimal cell concept is that of biological reductionism (Glass et al. 2017), which entails the complete description of every molecular functions harbored by a free-living cell (Lachance et al. 2019a). Reaching this level of knowledge is of paramount importance for the establishment of key design rules for organisms. With the advent of DNA synthesis techniques and whole-genome assembly, the creation of entirely new organisms is within reach. Such an example has been achieved with JCVI-syn3.0 (Hutchison et al. 2016), completing the first functional in vitro approximation of a minimal cell.

JCVI-syn3.0 reveals the state of the art in the design of minimal cells. Cutting-edge methods together with extensive work over many years have been put in place in order to produce this framework. The amount of labor necessary is met with the high-throughput capabilities of our days and age, both in terms of DNA synthesis and cloning and assembly, but the limiting factor remains the predictability of a given design. This struggle, relevant for both academic and industrial researchers, is one of the grand challenges that lays ahead in synthetic biology, and it is understood that laboratories which possess the best predictive power may outcompete those with high-throughput production and analysis capabilities.

In this context, the development of models for minimal cells is of paramount importance. We have reviewed the standard FBA approach for the genome-scale modeling of metabolism (Figs. 3, 4, and 5) and its applications for high-throughput data integration and the formulation of phenotypic predictions such as the flux

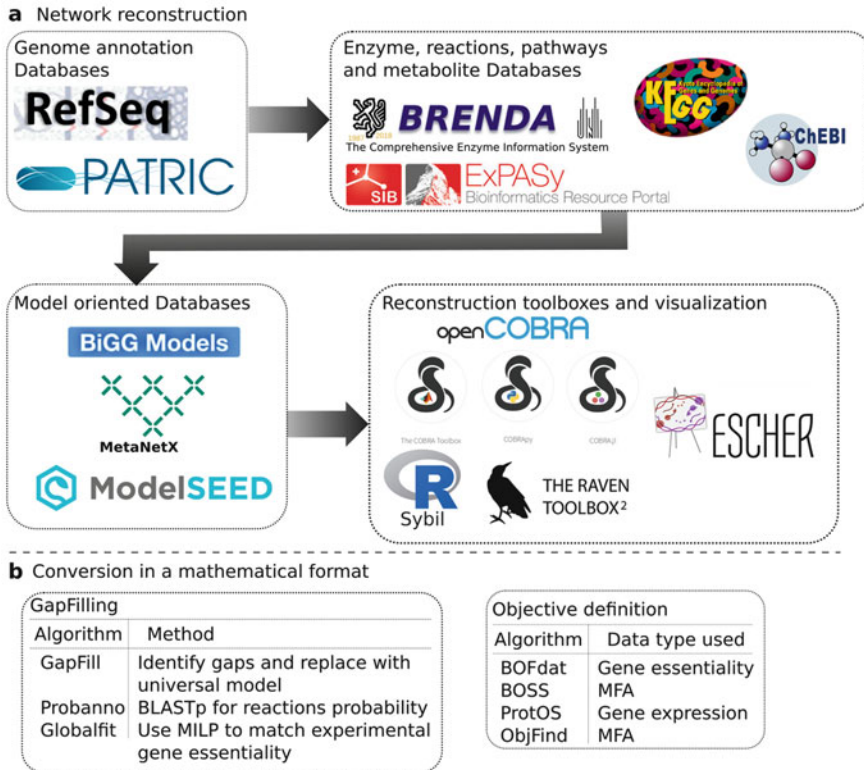


Fig. 5 Tools for genome-scale reconstruction and analysis. **(a)** Non-exhaustive list of computational tools and databases for the reconstruction of metabolic networks. First, querying annotation databases allows the identification of metabolic gene candidates (RefSeq, PATRIC). These genes can be associated to reactions by consulting reaction databases (KEGG, Brenda, ExPaSy, Chebi). The reactions and metabolites are associated to model specific identifiers using model-oriented databases (BiGG, MetaNetX, ModelSEED). The reconstruction toolboxes are designed to facilitate the creation of reaction, metabolite, gene and model objects programmatically (Open COBRA, Sybil, Raven). **(b)** Non-exhaustive list of computational tools to facilitate the identification of gaps in the network and cellular objectives

through metabolic reactions and gene essentiality (Fig. 6) (Suthers et al. 2009a; Zomorodi and Maranas 2010). Integrating this knowledge into a single framework is important to offer a systematic way of addressing knowledge gaps (Orth and Palsson 2010; Pan and Reed 2018) as demonstrated by Breuer et al. in their GEM of JCVI-syn3.0 (Breuer et al. 2019).

What lays ahead is up for debate. Further development of models for mollicutes will require more exhaustive biomass and growth media definition to impose relevant constraints on the system. Given their small genomes, the number of biochemical studies needed before exhaustive characterization is reached is reduced and, with the help of models, could be addressed rather quickly (Danchin and Fang 2016). A recently developed algorithmic method allows to generate a minimal genome sequence from transcriptional architecture and an ME-model (Wang and Maranas 2018) which could help in reducing genomes of more elaborate organisms

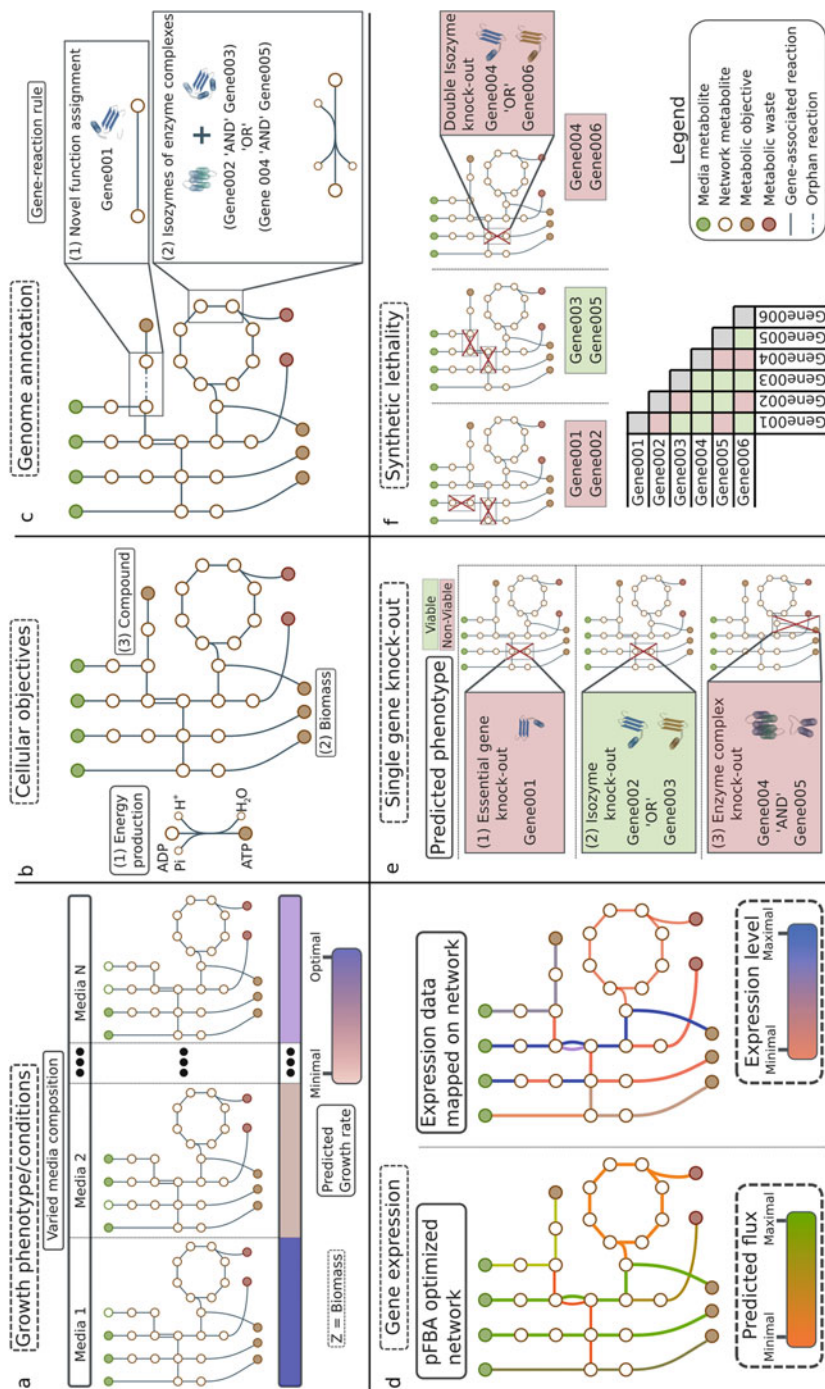


Fig. 6 The multiple uses of genome-scale models. (a) Growth media composition. The opened exchange reactions determine the *in silico* media composition. The impact of modifying the media composition is twofold. First, the necessary metabolites for growth are identified, hereby defining the minimal media. Second, the growth capabilities of the organism are explored, that is, the identification of nitrogen and carbon sources allowing growth. (b) Cellular objectives.

that are already used as production strains such as *E. coli* and *S. cerevisiae*. As we just discussed, constraint-based approaches can be expanded beyond metabolism, allowing the generation of models of metabolism and expression, ME-models. These models have already been employed to generate an in silico prediction of the core proteome by simulating on a wide array of different environments (Yang et al. 2015). With one of the main conclusions of this study being that the inclusion of more cellular systems be important for accurate predictions of a minimal gene set, it is interesting to consider that the expansion of modeling methods beyond metabolism and expression may be key for the rational design of minimal cells.

Finally, in silico writing of functional genome should be the following step. The integration of software tools for the conception of genomes is underway with the “Autocad” for genome recently published (Bates et al. 2017) as well as a genetic circuit compiler (Waites et al. 2018). Such tools are inspired by the experience acquired in the field of engineering, and the interest spurred by the community suggests a widespread application for the future of biology. For now, no organism is fully characterized, and hence the proposed completeness of biology (Morowitz 1984) is yet to be achieved. The use of genome-scale models together with genome writing tools might accelerate this process, and once a well-understood minimal cell chassis is described, strain design will reach a new paradigm.

References

- 1000 Genomes Project Consortium, Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM et al (2012) An integrated map of genetic variation from 1,092 human genomes. *Nature* 491:56–65
- Agren R, Liu L, Shoaie S, Vongsangnak W, Nookaew I, Nielsen J (2013) The RAVEN toolbox and its use for generating a genome-scale metabolic model for *Penicillium chrysogenum*. *PLoS Comput Biol* 9:e1002980
- Anderson S (1981) Shotgun DNA sequencing using cloned DNase I-generated fragments. *Nucleic Acids Res* 9:3015–3027
- Andrianantoandro E, Basu S, Karig DK, Weiss R (2006) Synthetic biology: new engineering rules for an emerging discipline. *Mol Syst Biol* 2:2006.0028



Fig. 6 (continued) The cellular objective can be modified to suit modeling needs. The production of energy, of biomass, or of a relevant metabolite can be studied. **(c)** Genome annotation. The reactions in the network are associated with a gene via the gene-reaction rule. The metabolic model provides a dynamic context in which to situate the gene annotation and review a gene’s function. **(d)** Gene expression. Parsimonious FBA (pFBA) is used to generate optimal flux state assuming an optimal enzyme usage by the cell. The flux through each reaction can then be compared to the gene expression level. **(e)** Gene essentiality. Genes and reactions are associated together through the gene-reaction rule. A gene is deemed essential if its deletion identifies it as the single determinant for the production of a biomass precursor. **(f)** Synthetic lethality. Performing simultaneous knockouts is possible with GEMs, an important asset for genome reduction and the production of synthetic minimal cells

- Artimo P, Jonnalagedda M, Arnold K, Baratin D, Csardi G, de Castro E et al (2012) ExPASy: SIB bioinformatics resource portal. *Nucleic Acids Res* 40:W597–W603
- Ataman M, Hatzimanikatis V (2017) lumpGEM: systematic generation of subnetworks and elementally balanced lumped reactions for the biosynthesis of target metabolites. *PLoS Comput Biol* 13:e1005513
- Aziz RK, Bartels D, Best AA, DeJongh M, Disz T, Edwards RA et al (2008) The RAST Server: rapid annotations using subsystems technology. *BMC Genomics* 9:75
- Baby V, Lachance J-C, Gagnon J, Lucier J-F, Matteau D, Knight T et al (2018) Inferring the minimal genome of *Mesoplasma florum* by comparative genomics and transposon mutagenesis. *mSystems* 3. <https://doi.org/10.1128/mSystems.00198-17>
- Bassalo MC, Garst AD, Halweg-Edwards AL, Grau WC, Domaille DW, Mutalik VK et al (2016) Rapid and efficient one-step metabolic pathway integration in *E. coli*. *ACS Synth Biol* 5:561–568
- Bates M, Lachoff J, Meech D, Zulkower V, Moisy A, Luo Y et al (2017) Genetic constructor: an online DNA design platform. *ACS Synth Biol* 6:2362–2365
- Beaucage SL, Caruthers MH (1981) Deoxynucleoside phosphoramidites—a new class of key intermediates for deoxypolynucleotide synthesis. *Tetrahedron Lett* 22:1859–1862
- Beck AE, Hunt KA, Carlson RP (2018) Measuring cellular biomass composition for computational biology applications. *Processes* 6:38
- Blattner FR, Plunkett G III, Bloch CA, Perna NT, Burland V, Riley M et al (1997) The complete genome sequence of *Escherichia coli* K-12. *Science* 277:1453–1462
- Bordbar A, Jamshidi N, Palsson BO (2011) iAB-RBC-283: a proteomically derived knowledge-base of erythrocyte metabolism that can be used to simulate its physiological and pathophysiological states. *BMC Syst Biol* 5:110
- Bordbar A, Monk JM, King ZA, Palsson BO (2014) Constraint-based models predict metabolic and associated cellular functions. *Nat Rev Genet* 15:107–120
- Bordbar A, Yurkovich JT, Paglia G, Rolfsson O, Sigurjónsson ÓE, Palsson BO (2017) Elucidating dynamic metabolic physiology through network integration of quantitative time-course metabolomics. *Sci Rep* 7:46249
- Breuer M, Earnest TM, Merryman C, Wise KS, Sun L, Lynott MR et al (2019) Essential metabolism for a minimal cell. *Elife*:8. <https://doi.org/10.7554/eLife.36842>
- Bridges CB (1922) The origin of variations in sexual and sex-limited characters. *Am Nat* 56:51–63
- Brower V (2001) Proteomics: biology in the post-genomic era: companies all over the world rush to lead the way in the new post-genomics race. *EMBO Rep* 2:558–560
- Burgard AP, Maranas CD (2003) Optimization-based framework for inferring and testing hypothesized metabolic objective functions. *Biotechnol Bioeng* 82:670–677
- Caspi R, Foerster H, Fulcher CA, Kaipa P, Krummenacker M, Latendresse M et al (2008) The MetaCyc Database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. *Nucleic Acids Res* 36:D623–D631
- Choe D, Cho S, Kim SC, Cho B-K (2016) Minimal genome: worthwhile or worthless efforts toward being smaller? *Biotechnol J* 11:199–211
- Crick F (1970) Central dogma of molecular biology. *Nature* 227:561–563
- Dajani AS (1965) Experimental infection with mycoplasma pneumoniae (Eaton's agent). *J Exp Med* 121:1071–1086
- Danchin A (2012) Scaling up synthetic biology: do not forget the chassis. *FEBS Lett* 586:2129–2137
- Danchin A, Fang G (2016) Unknown unknowns: essential genes in quest for function. *Microb Biotechnol* 9:530–540
- Dandekar T, Huynen M, Regula JT, Ueberle B, Zimmermann CU, Andrade MA et al (2000) Re-annotating the *Mycoplasma pneumoniae* genome sequence: adding value, function and reading frames. *Nucleic Acids Res* 28:3278–3288
- Danna K, Nathans D (1971) Specific cleavage of simian virus 40 DNA by restriction endonuclease of *Hemophilus influenzae*. *Proc Natl Acad Sci U S A* 68:2913–2917
- Deutscher D, Meilijson I, Kupiec M, Ruppin E (2006) Multiple knockout analysis of genetic robustness in the yeast metabolic network. *Nat Genet* 38:993–998

- Devoid S, Overbeek R, DeJongh M, Vonstein V, Best AA, Henry C (2013) Automated genome annotation and metabolic model reconstruction in the SEED and Model SEED. *Methods Mol Biol* 985:17–45
- Dias O, Rocha M, Ferreira EC, Rocha I (2015) Reconstructing genome-scale metabolic models with merlin. *Nucleic Acids Res* 43:3899–3910
- Ebrahim A, Lerman JA, Palsson BO, Hyduke DR (2013) COBRApy: COntstraints-based reconstruction and analysis for python. *BMC Syst Biol* 7:74
- Ebrahim A, Brunk E, Tan J, O'Brien EJ, Kim D, Szubin R et al (2016) Multi-omic data integration enables discovery of hidden biological regularities. *Nat Commun* 7:13091
- Edwards JS, Palsson BO (1999) Systems properties of the *Haemophilus influenzae* Rd metabolic genotype. *J Biol Chem* 274:17410–17416
- Edwards JS, Palsson BO (2000) The *Escherichia coli* MG1655 in silico metabolic genotype: its definition, characteristics, and capabilities. *Proc Natl Acad Sci U S A* 97:5528–5533
- Fabregat A, Jupe S, Matthews L, Sidiropoulos K, Gillespie M, Garapati P et al (2018) The reactome pathway knowledgebase. *Nucleic Acids Res* 46(D1):D649–D655. <https://doi.org/10.1093/nar/gkx1132>
- Feist AM, Palsson BO (2010) The biomass objective function. *Curr Opin Microbiol* 13:344–349
- Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, Kerlavage AR et al (1995) Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* 269:496–512
- Fraser CM, Gocayne JD, White O, Adams MD, Clayton RA, Fleischmann RD et al (1995) The minimal gene complement of *Mycoplasma genitalium*. *Science* 270:397–403
- Gelius-Dietrich G, Desouki AA, Fritzscheier CJ, Lercher MJ (2013) Sybil—efficient constraint-based modelling in R. *BMC Syst Biol* 7:125
- Gianchandani EP, Oberhardt MA, Burgard AP, Maranas CD, Papin JA (2008) Predicting biological system objectives de novo from internal state measurements. *BMC Bioinf* 9:43
- Gibson DG (2009) Synthesis of DNA fragments in yeast by one-step assembly of overlapping oligonucleotides. *Nucleic Acids Res* 37:6984–6990
- Gibson DG, Benders GA, Andrews-Pfannkoch C, Denisova EA, Baden-Tillson H, Zaveri J et al (2008) Complete chemical synthesis, assembly, and cloning of a *Mycoplasma genitalium* genome. *Science* 319:1215–1220
- Gibson DG, Young L, Chuang R-Y, Venter JC, Hutchison CA III, Smith HO (2009) Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nat Methods* 6:343–345
- Gibson DG, Glass JI, Lartigue C, Noskov VN, Chuang R-Y, Algire MA et al (2010) Creation of a bacterial cell controlled by a chemically synthesized genome. *Science* 329:52–56
- Glass JI, Assad-Garcia N, Alperovich N, Yooseph S, Lewis MR, Maruf M et al (2006) Essential genes of a minimal bacterium. *Proc Natl Acad Sci U S A* 103:425–430
- Glass JI, Merryman C, Wise KS, Hutchison CA, Smith HO (2017) Minimal cells—real and imagined. *Cold Spring Harb Perspect Biol* 9(12). <https://doi.org/10.1101/cshperspect.a023861>
- Goffeau A, Barrell BG, Bussey H, Davis RW, Dujon B, Feldmann H et al (1996) Life with 6000 genes. *Science* 274(546):563–567
- Goodson HV, Anderson BL, Warrick HM, Pon LA, Spudich JA (1996) Synthetic lethality screen identifies a novel yeast myosin I gene (MYO5): myosin I proteins are required for polarization of the actin cytoskeleton. *J Cell Biol* 133:1277–1291
- Gudmundsson S, Thiele I (2010) Computationally efficient flux variability analysis. *BMC Bioinf* 11:489
- Güell M, van Noort V, Yus E, Chen W-H, Leigh-Bell J, Michalodimitrakis K et al (2009) Transcriptome complexity in a genome-reduced bacterium. *Science* 326:1268–1271
- Haas R, Zelezniak A, Iacovacci J, Kamrad S, Townsend S, Ralser M (2017) Designing and interpreting “multi-omic” experiments that may change our understanding of biology. *Curr Opin Syst Biol* 6:37–45
- Hartleb D, Jarre F, Lercher MJ (2016) Improved metabolic models for *E. coli* and *Mycoplasma genitalium* from GlobalFit, an algorithm that simultaneously matches growth and non-growth data sets. *PLoS Comput Biol* 12:e1005036

- Heather JM, Chain B (2016) The sequence of sequencers: the history of sequencing DNA. *Genomics* 107:1–8
- Heinemann M, Panke S (2006) Synthetic biology—putting engineering into biology. *Bioinformatics* 22:2790–2799
- Heirendt L, Thiele I, Fleming RMT (2017) DistributedFBA.jl: high-level, high-performance flux balance analysis in Julia. *Bioinformatics* 33:1421–1423
- Holley RW (1965) Structure of an alanine transfer ribonucleic acid. *JAMA* 194:868–871
- Hughes RA, Ellington AD (2017) Synthetic DNA synthesis and assembly: putting the synthetic in synthetic biology. *Cold Spring Harb Perspect Biol* 9. <https://doi.org/10.1101/cshperspect.a023812>
- Human Microbiome Project Consortium (2012) Structure, function and diversity of the healthy human microbiome. *Nature* 486:207–214
- Hutchison CA III, Chuang R-Y, Noskov VN, Assad-Garcia N, Deerinck TJ, Ellisman MH et al (2016) Design and synthesis of a minimal bacterial genome. *Science* 351:aad6253
- Hutchison CA, Peterson SN, Gill SR, Cline RT, White O, Fraser CM et al (1999) Global transposon mutagenesis and a minimal *Mycoplasma* genome. *Science* 286:2165–2169
- Ingolia NT, Ghaemmaghami S, Newman JRS, Weissman JS (2009) Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science* 324:218–223
- Jacob F, Perrin D, Sanchez C, Monod J (1960) The operon: a group of genes whose expression is coordinated by an operator. *C R Seances Acad Sci* 250:1727–1729
- Joyce AR, Palsson BØ (2008) Predicting gene essentiality using genome-scale in silico models. *Methods Mol Biol* 416:433–457
- Kanehisa M, Furumichi M, Tanabe M, Sato Y, Morishima K (2017) KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res* 45:D353–D361
- Karr JR, Sanghvi JC, Macklin DN, Gutschow MV, Jacobs JM, Bolival B Jr et al (2012) A whole-cell computational model predicts phenotype from genotype. *Cell* 150:389–401
- Kauffman KJ, Prakash P, Edwards JS (2003) Advances in flux balance analysis. *Curr Opin Biotechnol* 14:491–496
- King ZA, Dräger A, Ebrahim A, Sonnenschein N, Lewis NE, Palsson BO (2015) Escher: a web application for building, sharing, and embedding data-rich visualizations of biological pathways. *PLoS Comput Biol* 11:e1004321
- King ZA, Lu J, Dräger A, Miller P, Federowicz S, Lerman JA et al (2016) BiGG Models: a platform for integrating, standardizing and sharing genome-scale models. *Nucleic Acids Res* 44:D515–D522
- King B, Farrah T, Richards MA, Mundy M, Simeonidis E, Price ND (2018) ProbAnnoWeb and ProbAnnoPy: probabilistic annotation and gap-filling of metabolic reconstructions. *Bioinformatics* 34:1594–1596
- Koonin EV (2000) How many genes can make a cell: the minimal-gene-set concept. *Annu Rev Genomics Hum Genet* 1:99–116
- Koonin EV, Mushegian AR, Bork P (1996) Non-orthologous gene displacement. *Trends Genet* 12:334–336
- Koonin EV, Fedorova ND, Jackson JD, Jacobs AR, Krylov DM, Makarova KS et al (2004) A comprehensive evolutionary classification of proteins encoded in complete eukaryotic genomes. *Genome Biol* 5:R7
- Kühner S, van Noort V, Betts MJ, Leo-Macias A, Batisse C, Rode M et al (2009) Proteome organization in a genome-reduced bacterium. *Science* 326:1235–1240
- Lachance J-C, Monk JM, Lloyd CJ, Seif Y, Palsson BO, Rodrigue S et al (2018) BOFdat: generating biomass objective function stoichiometric coefficients from experimental data [Internet]. [bioRxiv:243881. https://doi.org/10.1101/243881](https://doi.org/10.1101/243881)
- Lachance J-C, Rodrigue S, Palsson BO (2019a) Minimal cells, maximal knowledge. *Elife* 8. <https://doi.org/10.7554/eLife.45379>
- Lachance J-C, Lloyd CJ, Monk JM, Yang L, Sastry AV, Seif Y et al (2019b) BOFdat: generating biomass objective functions for genome-scale metabolic models from experimental data [Internet]. *PLoS Comput Biol*:e1006971. <https://doi.org/10.1371/journal.pcbi.1006971>

- Lahner B, Gong J, Mahmoudian M, Smith EL, Abid KB, Rogers EE et al (2003) Genomic scale profiling of nutrient and trace elements in *Arabidopsis thaliana*. *Nat Biotechnol* 21:1215–1221
- Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J et al (2001) Initial sequencing and analysis of the human genome. *Nature* 409:860–921
- Lartigue C, Glass JI, Alperovich N, Pieper R, Parmar PP, Hutchison CA III et al (2007) Genome transplantation in bacteria: changing one species to another. *Science* 317:632–638
- LeProust EM (2016) Rewriting DNA synthesis. *Chem Eng Prog* 2016:30–35
- Lerman JA, Hyduke DR, Latif H, Portnoy VA, Lewis NE, Orth JD et al (2012) In silico method for modelling metabolism and gene product expression at genome scale. *Nat Commun* 3:929
- Lewis NE, Hixson KK, Conrad TM, Lerman JA, Charusanti P, Polpitiya AD et al (2010) Omic data from evolved *E. coli* are consistent with computed optimal growth from genome-scale models. *Mol Syst Biol* 6:390
- Lewis NE, Nagarajan H, Palsson BO (2012) Constraining the metabolic genotype-phenotype relationship using a phylogeny of in silico methods. *Nat Rev Microbiol* 10:291–305
- Lind K (1966) Isolation of mycoplasma pneumoniae (Eaton agent) from patients with primary atypical pneumonia. *Acta Pathol Microbiol Scand* 66:124–134
- Lloyd CJ, Ebrahim A, Yang L, King ZA, Catoiu E, O'Brien EJ et al (2018) COBRAME: a computational framework for genome-scale models of metabolism and gene expression. *PLoS Comput Biol* 14:e1006302
- Lu H, Giordano F, Ning Z (2016) Oxford Nanopore MinION sequencing and genome assembly. *Genomics Proteomics Bioinformatics* 14:265–279
- Majewski RA, Domach MM (1990) Simple constrained-optimization view of acetate overflow in *E. coli*. *Biotechnol Bioeng* 35:732–738
- Malyshev DA, Dhami K, Lavergne T, Chen T, Dai N, Foster JM et al (2014) A semi-synthetic organism with an expanded genetic alphabet. *Nature* 509:385–388
- Matteucci MD, Caruthers MH (1981) Synthesis of deoxyoligonucleotides on a polymer support. *J Am Chem Soc* 103:3185–3191
- McGuire AL, Colgrove J, Whitney SN, Diaz CM, Bustillos D, Versalovic J (2008) Ethical, legal, and social considerations in conducting the Human Microbiome Project. *Genome Res* 18:1861–1864
- Miles RJ (1992) Catabolism in mollicutes. *J Gen Microbiol* 138:1773–1783
- Monk J, Nogales J, Palsson BO (2014) Optimizing genome-scale network reconstructions. *Nat Biotechnol* 32:447–452
- Monk JM, Lloyd CJ, Brunk E, Mih N, Sastry A, King Z et al (2017) iML1515, a knowledgebase that computes *Escherichia coli* traits. *Nat Biotechnol* 35:904–908
- Moretti S, Martin O, Van Du Tran T, Bridge A, Morgat A, Pagni M. MetaNetX/MNXref--reconciliation of metabolites and biochemical reactions to bring together genome-scale metabolic networks. *Nucleic Acids Res* 2016;44: D523–D526.
- Morowitz HJ (1984) Special guest lecture the completeness of molecular biology. *Isr J Med Sci* 2
- Morowitz HJ, Tourtellotte ME (1962) The smallest living cells. *Sci Am* 206:117–126
- Mørtz E, O'Connor PB, Roepstorff P, Kelleher NL, Wood TD, McLafferty FW et al (1996) Sequence tag identification of intact proteins by matching tandem mass spectral data against sequence data bases. *Proc Natl Acad Sci U S A* 93:8264–8267
- Mushegian AR, Koonin EV (1996) A minimal gene set for cellular life derived by comparison of complete bacterial genomes. *Proc Natl Acad Sci U S A* 93:10268–10273
- Nagalakshmi U, Wang Z, Waern K, Shou C, Raha D, Gerstein M et al (2008) The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science* 320:1344–1349
- Nirenberg MW, Jones OW, Leder P, Clark BFC, Sly WS, Pestka S (1963) On the coding of genetic information. *Cold Spring Harb Symp Quant Biol* 28:549–557
- Nirenberg M, Leder P, Bernfield M, Brimacombe R, Trupin J, Rottman F et al (1965) RNA codewords and protein synthesis, VII. On the general nature of the RNA code. *Proc Natl Acad Sci U S A* 53:1161–1168
- Nursimulu N, Xu LL, Wasmuth JD, Krukov I, Parkinson J (2018) Improved enzyme annotation with EC-specific cutoffs using DETECT v2. *Bioinformatics* 34:3393–3395

- O'Brien EJ, Lerman JA, Chang RL, Hyduke DR, Palsson BØ (2013) Genome-scale models of metabolism and gene expression extend and refine growth phenotype prediction. *Mol Syst Biol* 9:693
- Orth JD, Palsson BØ (2010) Systematizing the generation of missing metabolic knowledge. *Biotechnol Bioeng* 107:403–412
- Orth JD, Thiele I, Palsson BØ (2010) What is flux balance analysis? *Nat Biotechnol* 28:245–248
- Palsson BØ (2015) *Systems biology: constraint-based reconstruction and analysis*. Cambridge University Press, Cambridge
- Pan S, Reed JL (2018) Advances in gap-filling genome-scale metabolic models and model-driven experiments lead to novel metabolic discoveries. *Curr Opin Biotechnol* 51:103–108
- Papoutsakis ET (1984) Equations and calculations for fermentations of butyric acid bacteria. *Biotechnol Bioeng* 26:174–187
- Placzek S, Schomburg I, Chang A, Jeske L, Ulbrich M, Tillack J et al (2017) BRENDA in 2017: new perspectives and new tools in BRENDA. *Nucleic Acids Res* 45:D380–D388
- Qi LS, Larson MH, Gilbert LA, Doudna JA, Weissman JS, Arkin AP et al (2013) Repurposing CRISPR as an RNA-guided platform for sequence-specific control of gene expression. *Cell* 152:1173–1183
- Richardson SM, Mitchell LA, Stracquadanio G, Yang K, Dymond JS, DiCarlo JE et al (2017) Design of a synthetic yeast genome. *Science* 355:1040–1044
- Riekeberg E, Powers R (2017) New frontiers in metabolomics: from measurement to insight. *F1000Res* 6:1148
- Roberts RJ (2005) How restriction enzymes became the workhorses of molecular biology. *Proc Natl Acad Sci U S A* 102:5905–5908
- Saiki RK, Scharf S, Faloona F, Mullis KB, Horn GT, Erlich HA et al (1985) Enzymatic amplification of beta-globin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia. *Science* 230:1350–1354
- Sanger F, Nicklen S, Coulson AR (1977a) DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A* 74:5463–5467
- Sanger F, Air GM, Barrell BG, Brown NL, Coulson AR, Fiddes JC et al (1977b) Nucleotide sequence of bacteriophage ϕ X174 DNA. *Nature* 265:687
- Satish Kumar V, Dasika MS, Maranas CD (2007) Optimization based automated curation of metabolic reconstructions. *BMC Bioinf* 8:212
- Savinell JM, Palsson BO (1992a) Optimal selection of metabolic fluxes for in vivo measurement. I. Development of mathematical methods. *J Theor Biol* 155:201–214
- Savinell JM, Palsson BO (1992b) Optimal selection of metabolic fluxes for in vivo measurement. II. Application to *Escherichia coli* and hybridoma cell metabolism. *J Theor Biol* 155:215–242
- Schellenberger J, Que R, Fleming RMT, Thiele I, Orth JD, Feist AM et al (2011) Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox v2.0. *Nat Protoc* 6:1290–1307
- Schrodinger E (1967) *What is life? The physical aspect of the living cell and mind and matter; mind and matter*. Cambridge University Press, Cambridge
- Segrè D, Vitkup D, Church GM (2002) Analysis of optimality in natural and perturbed metabolic networks. *Proc Natl Acad Sci U S A* 99:15112–15117
- Shlomi T, Berkman O, Ruppin E (2005) Regulatory on/off minimization of metabolic flux changes after genetic perturbations. *Proc Natl Acad Sci U S A* 102:7695–7700
- Sinsheimer RL (1989) The Santa Cruz Workshop—May 1985. *Genomics* 5:954–956
- Sleator RD (2010) The story of *Mycoplasma mycoides* JCVI-syn1.0: the forty million dollar microbe. *Bioeng Bugs* 1:229–230
- Smith HO, Wilcox KW (1970) A restriction enzyme from *Hemophilus influenzae*. I. Purification and general properties. *J Mol Biol* 51:379–391
- Smith LM, Sanders JZ, Kaiser RJ, Hughes P, Dodd C, Connell CR et al (1986) Fluorescence detection in automated DNA sequence analysis. *Nature* 321:674–679
- Smolke C, Lee SY, Nielsen J, Stephanopoulos G (2018) *Synthetic biology: parts, devices and applications*. Wiley

- Spencer G (2008) International consortium announces the 1000 Genomes project. See http://www.1000genomes.org/bcms/1000_genomes/Documents/1000Genomes-NewsRelease.pdf
- Stemmer WP, Cramer A, Ha KD, Brennan TM, Heyneker HL (1995) Single-step assembly of a gene and entire plasmid from large numbers of oligodeoxyribonucleotides. *Gene* 164:49–53
- Suthers PF, Zomorodi A, Maranas CD (2009a) Genome-scale gene/reaction essentiality and synthetic lethality analysis. *Mol Syst Biol* 5:301
- Suthers PF, Dasika MS, Kumar VS, Denisov G, Glass JI, Maranas CDA (2009b) Genome-scale metabolic reconstruction of *Mycoplasma genitalium*, iPS189. *PLoS Comput Biol* 5(2): e1000285. <https://doi.org/10.1371/journal.pcbi.1000285>
- Thiele I, Palsson BØ (2010) A protocol for generating a high-quality genome-scale metabolic reconstruction. *Nat Protoc* 5:93–121
- Thiele I, Jamshidi N, Fleming RMT, Palsson BØ (2009) Genome-scale reconstruction of *Escherichia coli*'s transcriptional and translational machinery: a knowledge base, its mathematical formulation, and its functional characterization. *PLoS Comput Biol* 5:e1000312
- Thiele I, Fleming RMT, Que R, Bordbar A, Diep D, Palsson BO (2012) Multiscale modeling of metabolism and macromolecular synthesis in *E. coli* and its application to the evolution of codon usage. *PLoS One* 7:e45635
- Varma A, Palsson BO (1993) Metabolic capabilities of *Escherichia coli*: I. synthesis of biosynthetic precursors and cofactors. *J Theor Biol* 165:477–502
- Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG et al (2001) The sequence of the human genome. *Science* 291:1304–1351
- Waddington CH (1961) Molecular biology or ultrastructural biology? *Nature* 190:184
- Waites W, Mısırlı G, Cavaliere M, Danos V, Wipat A (2018) A genetic circuit compiler: generating combinatorial genetic circuits with web semantics and inference. *ACS Synth Biol*. <https://doi.org/10.1021/acssynbio.8b00201>
- Wang L, Maranas CD (2018) MinGenome: an in silico top-down approach for the synthesis of minimized genomes. *ACS Synth Biol* 7:462–473
- Watson JD, Crick FHC (1953) Others. Molecular structure of nucleic acids. *Nature* 171:737–738
- Wattam AR, Davis JJ, Assaf R, Boisvert S, Brettin T, Bun C et al (2017) Improvements to PATRIC, the all-bacterial bioinformatics database and analysis resource center. *Nucleic Acids Res* 45: D535–D542
- Wodke JAH, Puchałka J, Lluch-Senar M, Marcos J, Yus E, Godinho M et al (2013) Dissecting the energy metabolism in *Mycoplasma pneumoniae* through genome-scale metabolic modeling. *Mol Syst Biol* 9:653
- Xavier JC, Patil KR, Rocha I (2017) Integration of biomass formulations of genome-scale metabolic models with experimental data reveals universally essential cofactors in prokaryotes. *Metab Eng* 39:200–208
- Yang K, Han X (2016) Lipidomics: techniques, applications, and outcomes related to biomedical sciences. *Trends Biochem Sci* 41:954–969
- Yang L, Tan J, O'Brien EJ, Monk JM, Kim D, Li HJ et al (2015) Systems biology definition of the core proteome of metabolism and expression is consistent with high-throughput data. *Proc Natl Acad Sci U S A* 112:10810–10815
- Yang L, Ma D, Ebrahim A, Lloyd CJ, Saunders MA, Palsson BO (2016) solveME: fast and reliable solution of nonlinear ME models. *BMC Bioinf* 17:391
- Yurkovich JT, Yang L, Palsson BO (2017) Biomarkers are used to predict quantitative metabolite concentration profiles in human red blood cells. *PLoS Comput Biol* 13:e1005424
- Yus E, Maier T, Michalodimitrakis K, van Noort V, Yamada T, Chen W-H et al (2009) Impact of genome reduction on bacterial metabolism and its regulation. *Science* 326:1263–1268
- Zamboni N, Fendt S-M, Rühl M, Sauer U (2009) 13C-based metabolic flux analysis. *Nat Protoc* 4:878
- Zhao Q, Stettner AI, Reznik E, Paschalidis IC, Segrè D (2016) Mapping the landscape of metabolic goals of a cell. *Genome Biol* 17:109
- Zomorodi AR, Maranas CD (2010) Improving the iMM904 *S. cerevisiae* metabolic model using essentiality and synthetic lethality data. *BMC Syst Biol* 4:178