Briefings in Functional Genomics, 16(2), 2017, 70–79

doi: 10.1093/bfgp/elv063
Advance Access Publication Date: 8 February 2016
Review paper

Advances and perspectives in computational prediction of microbial gene essentiality

Fredrick M. Mobegi, Aldert Zomer, Marien I. de Jonge, and Sacha A. F. T. van Hijum

Corresponding authors: Fredrick M. Mobegi, Laboratory of Pediatric Infectious Diseases and Centre for Molecular and Biomolecular Informatics, Radboud Institute for Molecular Life Sciences, Radboud University Medical Centre, Nijmegen 6500 HB, the Netherlands. Tel: +31243651545; E-mail: Fredrick.mobegi@radboudumc.nl. Sacha A.F.T. van Hijum, Bacterial Genomics Group, Center for Molecular and Biomolecular Informatics; Radboud Institute of Molecular Life Sciences, Radboud University Medical Centre, Nijmegen 6500 HB, the Netherlands. Tel: +31243619389; E-mail: Sacha.vanhijum@radboudumc.nl

Abstract

The minimal subset of genes required for cellular growth, survival and viability of an organism are classified as essential genes. Knowledge of essential genes gives insight into the core structure and functioning of a cell. This might lead to more efficient antimicrobial drug discovery, to elucidation of the correlations between genotype and phenotype, and a better understanding of the minimal requirements for a (synthetic) cell. Traditionally, constructing a catalog of essential genes for a given microbe involved costly and time-consuming laboratory experiments. While experimental methods have produced abundant gene essentiality data for model organisms like *Escherichia coli* and *Bacillus subtilis*, the knowledge generated cannot automatically be extrapolated to predict essential genes in all bacteria. In addition, essential genes identified in the laboratory are by definition 'conditionally essential', as they are essential under the specified experimental conditions: these might not resemble conditions in the microorganisms' natural habitat(s). Also, large-scale experimental assaying for essential genes is not always feasible because of the time investment required to setup these assays. The ability to rapidly and precisely identify essential genes *in silico* is therefore important and has great potential for applications in medicine, biotechnology and basic biological research. Here, we review the advances made in the use of computational methods to predict microbial gene essentiality, perspectives for the future of these techniques and the possible practical applications of essential genes.

Key words: gene essentiality prediction; computational methods; homology; transposons; next-generation sequencing

Introduction

Inactivation of essential genes in an otherwise wild type organism results in lethality. These genes therefore represent the foundation of cellular life [1, 2]. Identifying essential genes is therefore valuable and important in biology, industrial bioprocessing and medicine. For example, it could aid in comprehending the basic principles behind how cells function [3], and the complex relations between genotype and phenotype [4], which are fundamental questions in biology and genetics. Understanding the function of essential genes is prerequisite to

Fredrick M. Mobegi is a PhD fellow in Bioinformatics at the Laboratory of Pediatric Infectious Diseases, and the Centre for Molecular and Biomolecular Informatics (Radboud University Medical Centre). His research involves developing and applying bioinformatics tools to study the genetics behind the etiology, molecular epidemiology, clinical manifestation of disease and antibiotic resistance in bacterial pathogens.
Aldert Zomer is an assistant professor at the Faculty of Veterinary Medicine (Utrecht University). His research focuses on analysis of bacterial (meta)genomes for comparative genomics, molecular epidemiology and linking the bacterial genotype to clinical manifestation of disease.
Marien I. de Jonge is leader of the Laboratory of Pediatric Infectious Diseases at Radboud University Medical Centre. Research in his group focuses on the interaction of the developing immune system with pathogens causing pediatric infections, to understand mucosal and systemic immune responses in the context of infection and vaccination.
Sacha A. F. T. van Hijum is a principal scientist bioinformatics at NIZO food research, and an associate professor and leader of the bacterial (meta)genomics group at the Centre for Molecular and Biomolecular Informatics (Radboud University Medical Centre). Research in his group focuses on establishing the relation between microbes and health.

discovering the core components of a minimal cell [5], potentially facilitating reengineering of microorganisms [6] with desired phenotypical traits for research and biotechnology. Additionally, because essential genes confer lethal phenotypes to microorganisms when deleted or inactivated, they form promising drug targets on which potent antibiotics could be developed [7, 8]. Knowledge of gene essentiality has been applied in discovering candidate 'human disease genes' (genes with disease-associated alleles), their mode of inheritance and contribution to developmental abnormalities or disease [9].

Essential genes have been identified in a number of model organisms [10–15]. However, as recently reviewed [2], several studies querying gene essentiality in the same organisms under similar experimental conditions have produced different catalogs of essential genes. This lack of consensus makes it challenging to determine gene essentiality in model organisms, let alone in non-model or poorly researched organisms. The differences are possibly a result of 'conditional or contextual' essentiality: the essentiality of a gene depends on its context, which might be a defined growth media or conditions, genetic context or a particular developmental stage of a microorganism [16]. Moreover, the longer timespans required for conducting experiments also give enough time for isozymes to be upregulated, significantly affecting essentiality prediction. Most studies have consistently deciphered essential genes under rich media conditions (Supplementary Table 1); in other words, in the richness of a full complement of vital nutrients and devoid of environmental stress [8, 11, 13, 14, 17]. Although laboratory rich media conditions are undoubtedly not a proxy of conditions in a microorganism's natural niche, essential genes determined under these conditions provide a near-complete representation of genes needed in most *in situ* niches [11]. Therefore, for the purpose of this review, we define the 'essentiality' of a gene as its indispensability under rich media conditions.

Gene essentiality studies have advanced significantly in the past few years owing to a plethora of *in vitro*, *in vivo* (laboratory) and *in silico* methods. Laboratory methods assess gene essentiality by observing lethal phenotypes ensuing from random or systematic gene inactivation using transposon mutagenesis [12], gene knockouts [11, 18], genetic complementation [19] and RNA interference [20]. However, genomic-scale discovery of essential genes using laboratory techniques is often complex, costly, time-consuming and is contextual because it can be influenced by growth conditions as well as genetic context [16]. Therefore, to establish accurate results, a consensus of predicted essential genes across multiple laboratories is required. To circumvent these complexities, *in silico* techniques have been developed to predict essential genes [21–23]. Computational methods have gained popularity over the past years for numerous reasons. First, computational methods are less time-consuming, and they benefit from knowledge obtained from other organisms. The essential genes identified from several microorganisms provide seed information for training gene essentiality predictors for less-researched organisms. Second, the abundance of 'omics' data from genomic sequencing projects provides opportunities for microbial functional genomics. Finally, bioinformatics has greatly developed over recent years, significantly advancing tools available to discover essential genes in sequenced genomes. It is noteworthy that computational methods cannot (yet) predict conditional essentiality but rather predict whether a gene is essential.

In this review, we focus on advances made in genome-wide microbial gene essentiality prediction, particularly using computational methods. We discuss the fundamental principles of computational gene essentiality prediction tools, and provide an opinion on the choice of method. We also explore the possible practical applications of essential genes and give a perspective into the future of computational methods in predicting gene essentiality.

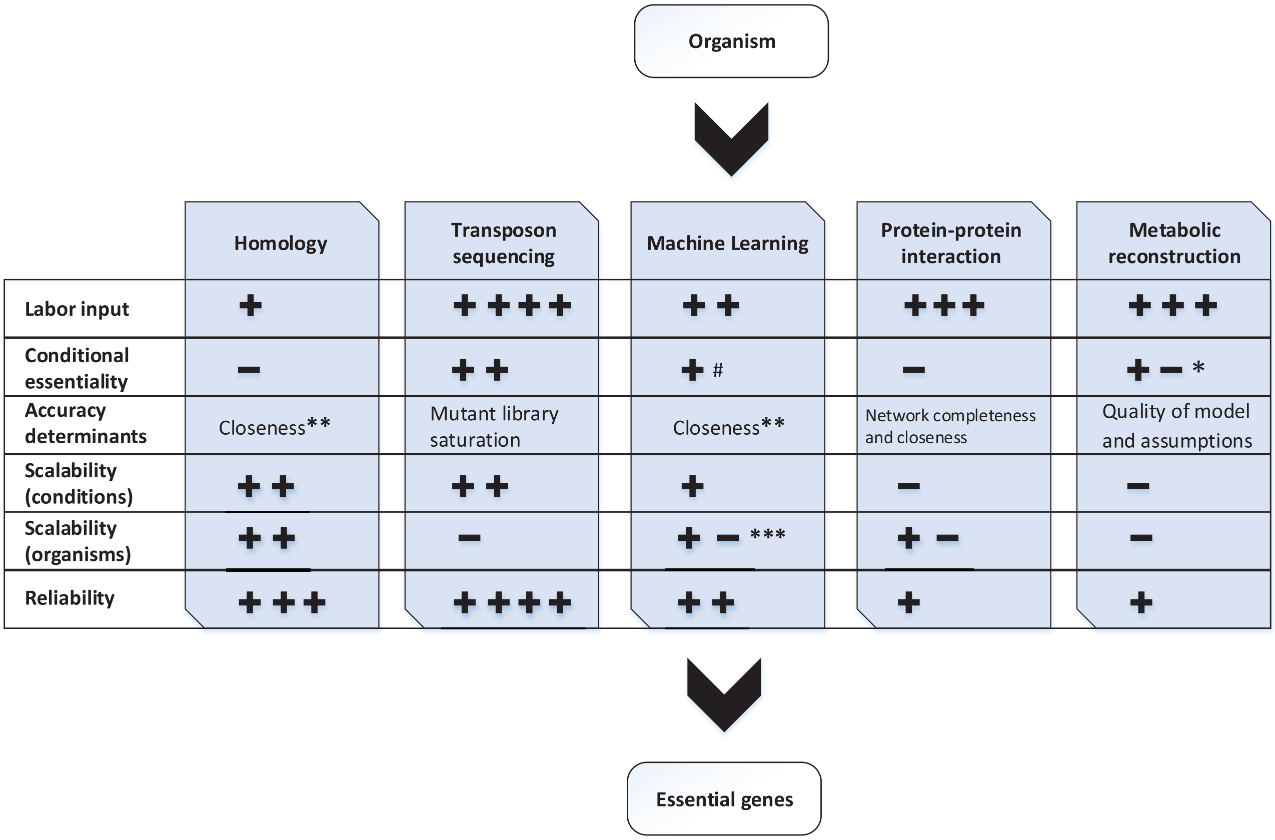
Computational techniques for gene essentiality prediction

Many *in silico* prediction methods have been established to aid in *post hoc* analysis of experimental readouts, or mining 'omics' data for encoded signatures to identify essential genes. Below, we discuss approaches commonly used to predict gene essentiality (Figure 1). They commonly analyze intrinsic genomic features, such as localization signals, codon adaptation indices, guanine cytosine (GC) content, gene orthologs, rate of gene evolution and phyletic gene retention [21, 22, 24]. Other integrated approaches such as network analysis [3, 25] and machine learning (ML) on combinations of features and approaches [24, 26] are also discussed.

Transposon sequencing methods

Transposons have been widely used in techniques like signature-tagged mutagenesis [27] to manipulate genes in various microorganisms [12, 28, 29], albeit with low resolution. Recently however, various high-throughput techniques including Tn-seq [15], INSeq [30], HITS [31], TraDIS [32] and variants thereof have harnessed the power of traditional transposon mutagenesis, next-generation sequencing (NGS) and *post hoc in silico* tracking of the insertions, to explore gene function and higher-order genome organization [14].

Transposon sequencing and analysis (TSA) techniques commonly rely on the construction of transposon mutant libraries in which nonessential genes contain transposon insertions, followed by growth of the mutant libraries in defined *in vitro* or *in vivo* (e.g. host infection models) conditions. The relative frequency of each mutant in the population at the beginning and the end of the experiment is then determined by means of NGS at the transposon junctions. Genes that are essential for growth under a particular condition will not accumulate transposon insertions. From these data, the fitness of every gene to the experimental conditions to which the transposon libraries were subjected is quantified [15, 30–32]. The relatedness and differences between various TSA techniques have comprehensively been reviewed [33, 34]. Their main advantages are the high levels of accuracy and sensitivity in predicting gene essentiality, and their ability to be adapted for analyses in a wide range of species. By using certain regimes to store mutant libraries, it is also possible to obtain strains with desired gene knockout(s). In addition, the sequencing protocols used generate short sequence reads of millions of DNA molecules simultaneously, allowing whole genomes to be investigated in a single experiment. Nonetheless, TSA techniques are dependent on strong molecular amenability of an organism to allow creation of saturated mutant libraries and accurate deep sequencing [8], making them expensive for routine use. For this reason, computational approaches like homology mapping and ML, which may rely solely on computer-mined essentiality determinants, would be desirable (Figure 1).



Organism

	Homology	Transposon sequencing	Machine Learning	Protein-protein interaction	Metabolic reconstruction
Labor input	+	++++	++	+++	+++
Conditional essentiality	−	++	+ #	−	+ − *
Accuracy determinants	Closeness**	Mutant library saturation	Closeness**	Network completeness and closeness	Quality of model and assumptions
Scalability (conditions)	++	++	+	−	−
Scalability (organisms)	++	−	+ − ***	+ −	−
Reliability	+++	++++	++	+	+

Essential genes

Figure 1. Summary of the computational methods used in predicting essential genes.

Notes. #The ability to identify conditionally essential genes is usually specific for the training data set. *Prediction of conditional essentiality is influenced by the quality of the input model and objective function among other factors. **In homology models, the query and subject sequences should have maintained enough closeness throughout evolution. In ML models, the study subjects should be close enough (e.g. same or close species) to the training set, hence disadvantageous while dealing with distantly related species.

Homology mapping models

The term ‘homologs’ refers to two or more genes related by descent from a common ancestral DNA sequence. This relationship may arise between genes separated by speciation (orthologs) or genetic duplication (paralogs). Prediction of essential genes based on sequence homology, especially to known essential genes, is arguably the simplest and earliest used method in the genomic era. Shortly after the availability of the first two completely sequenced bacterial genomes, homology models were used to predict gene essentiality [35], and to establish the minimal genome [17] in *Haemophilus influenzae* and *Mycoplasma genitalium*. These models rely on heuristic algorithms embedded in sequence alignment programs like Muscle [36], Clustal [37–39], T-Coffee [40] and database search tools like BLAST, to compare query sequences with a library or database of subject sequences whose essentiality is known. Sequences that are similar above defined percentage identity, ‘e-value’ threshold, and length coverage are grouped as homologs. Homology models show high confidence levels owing to these metric thresholds.

Essential genes evolve slowly and tend to be more conserved than nonessential genes in bacteria [41]. Selection on essential genes is more stringent than on nonessential genes, increasing the average likelihood that orthologs of essential genes are conserved in bacteria [1, 42] and almost certainly essential. This allows extrapolation of essentiality from one member to an entire group of homologous genes. Many bacterial genomes are publicly available from genome sequencing projects (Figure 2). Various homology prediction tools and databases that collate homologous proteins (Supplementary Table 2) are also publicly available. This has greatly simplified determining genes that share ancestry, making homology models attractive for predicting gene essentiality based solely on genomic sequences.

However, they have various limitations. First, they are limited to conserved orthologs between species, which often account for a small portion of the genome [43]. Moreover, because the model only considers computationally determined orthologous genes based on sequence similarity, highly evolving genes may be overlooked, consequently leading to underestimation of essential genes in a genome [5]. Second, orthologs, especially in distantly related species, often show variations in gene regulations, posttranslational protein modification, divergence in cellular pathways, redundancies in processes, gene duplications and other niche specializations [44], leading to potential multiplicity in relative gene essentiality. For example, Hutchison and colleagues [28] successfully used transposon insertions to disrupt some of the 256 essential genes predicted using homology [35], suggesting that they are possibly nonessential. Yu *et al.* also identified 787 nonessential *Streptococcus sanguinis* genes, which had orthologs in all 48 *Streptococcus* genomes they analyzed [45].

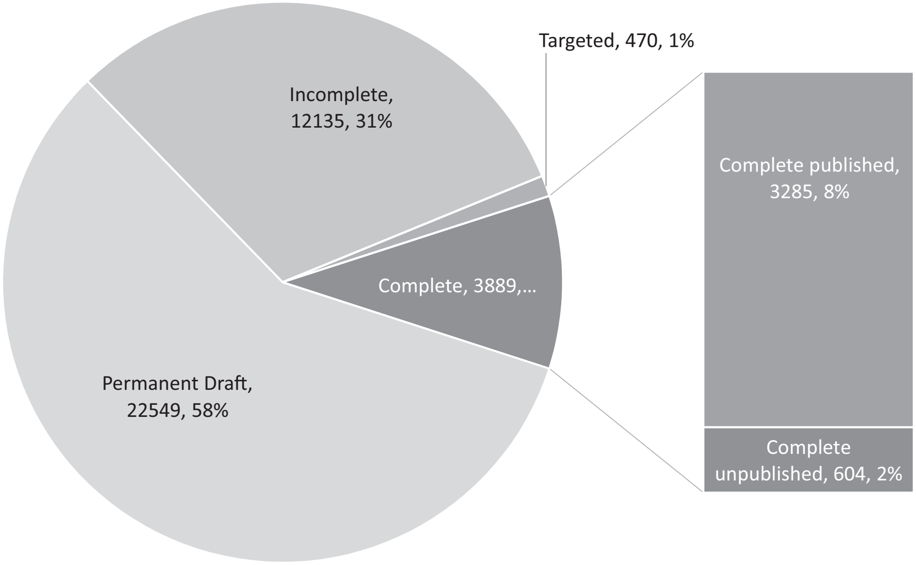


Figure 2. The genome project coverage for bacteria.

Approximately 39 442 bacterial genome projects are documented according to Genome OnLine Database (http://www.genomesonline.org) as of 30 June 2015.

Additionally, the gene encoding alanyl-tRNA synthetase (*alaS*) is essential in *Escherichia coli* but not in *Pseudomonas aeruginosa*: probably because of functional redundancy caused by a paralog, PA2106, in *Pseudomonas* [26]. For this reason, absence of paralogs is generally thought to be a strong indicator of essentiality in cross-species homology analyses [8, 46]. Although it is more straightforward to predict essentiality for single copy genes, paralogous genes could still be essential: deletion of all copies of a duplicated gene that encodes an essential function should lead to lethality. Finally, notwithstanding the tendency of essential genes to be highly conserved, genes conserved across species are not always essential. Indeed, in model organisms, only <25% of all conserved genes have experimentally been validated to be essential [29, 47], indicating that similarity of sequences does not always warrant extrapolation of the essentiality annotation among homologs.

Protein network topology models

Two or more proteins can establish physical contact (protein–protein interactions; PPIs) as a result of biochemical events and/or electrostatic forces. These interactions are mainly determined from quantum chemistry, molecular dynamics, signal transduction and biochemistry assays among others [48–51] and in certain cases can be used to predict essentiality. The abundance of experimental data generated from small-scale analyses and high-throughput procedures has assisted in defining PPIs within the interactome (all possible molecular interactions within a cell). Large-scale exploration of the topological properties of these networks is important in understanding the organizational and functional principles of individual proteins in biological pathways [25], and consequently their essentiality. Various descriptors for centrality of a node in a network have been effectively applied to identify essential proteins in PPIs [25, 52, 53]. For example, deletion of a hub (highly linked) protein is more likely to lethally perturb the network than deletion of a non-hub (peripheral) protein [54]. It is therefore widely agreed that hub proteins evolve slowly and are most likely to be essential [53]. However, prediction of gene essentiality in less-studied genomes using protein networks is expensive and arduous, primarily because of the limitations of experimental data (containing missing values, false positives, false negatives or differing between replicates—suggesting that the data are erroneous, incomplete or both) necessary to build and characterize PPIs [55], and the complexities in computational inference of PPI networks [56]. Their accuracy also depends on the completeness of the network, and they cannot be used to predict conditional essentiality. Moreover, proteins that lack known interactions with other proteins are completely disregarded in PPI models.

Metabolic network reconstruction and simulation models

While studying PPI networks gives a basic understanding of gene and protein interactions, they are limited in elucidating the complex and dynamic interactions among molecular components of cellular networks at genome scale. Whole-genome metabolic network reconstruction under constraint-based reconstruction and analysis scaffold [57, 58] allows for an in-depth insight into the metabolic capabilities of an organism, particularly in correlating the genome with molecular physiology [59]. These methods are based on the following fundamental concepts: (1) the burden of physicochemical constraints to limit quantifiable phenotypes, (2) identification and algebraic account of evolutionary selective pressures and (3) a genome-scale perception of cell metabolism that accounts for all cellular metabolic gene products [60]. Metabolic networks for many microorganisms have been reconstructed [61] or can be reconstructed and, to some extent, curated using automated systems such as model SEED [62], RAST [63] and BiGG [64]. Also, by integrating homology modeling, genome-scale models that show substantial predictive power in auxotrophy and essentiality predictions have been created for multiple strains of lesser studied organisms by starting with the genome-scale model of a well-studied organism like *Escherichia coli* K12 MG1655 [65]. Nonetheless, significant efforts are required to manually curate and ascertain the reliability of automatically generated metabolic models for improving their reliability for gene essentiality prediction. It is noteworthy that following the immense success in metabolic networks reconstruction, significant efforts are being made to model transcriptional networks [66], signaling networks [67] and computing of protein expressions needed to perform metabolism and proteome synthesis [68]. These efforts are providing crucial input to extend gene essentiality prediction using network reconstruction modeling from metabolism to incorporate other non-metabolic cellular processes.

Flux balance analysis (FBA) is the most commonly used constrained-based approach for *in silico* prediction of microbial phenotypes from metabolic models [69, 70]. It integrates biochemical constraints with the stoichiometry of metabolic and transport reactions, their reversibility and subcellular localization, thereby reducing the intricacy of potential dynamic states to predict metabolite fluxes at steady state. FBA has been used to simulate gene knockout and evaluate the associated lethality on the system, enabling the identification of essential genes [70]. For a given objective function and each *in silico* gene deletion, essentiality is evaluated by calculating the optimal production of defined biosynthetic precursors: identified auxotrophic requirements and impaired functions (indicating simulated gene knockouts, which inhibit *in silico* production of precursors contained within the objective function, e.g. alanine production) are classified as essential for the objective function. FBA relies on stoichiometric characteristics and does not require kinetic parameters (which are often difficult to obtain), allowing it to be used on any fully sequenced and annotated organism [69]. Nevertheless, FBA has important limitations: first, while FBA could be integrated with modal analyses at steady state, it cannot be used to investigate genome-scale metabolic reactions under transient dynamic states without including data on enzyme kinetics [71]. Second, FBA cannot be used to directly predict immediate suboptimal flux states and metabolite concentration following a genetic perturbation. Organisms naturally adapt to perturbations by readjusting various regulatory mechanisms, enzyme expressions and fluxes to bypass the effects. Such immediate changes and the effect of regulatory mechanisms cannot be explicitly specified in FBA. Some of these limitations have been addressed in variants of FBA such as MOMA [72], ROOM [73], MEA [74] and dynamic FBA [71]. Finally, FBA sometimes disagrees with experimental data; these discrepancies could be addressed by the addition of enzyme reactions through 'gap filling' [69, 75].

Overall, given the substantial input required and the inability to provide direct readout for conditionally essential genes, metabolic network reconstruction models are undesirable first-choice methods for exploring gene essentiality in novel genomes.

Integrated features ML models

Integrative ML models rely on constructing and training a classifier for predicting gene essentiality. They integrate multiple characteristics or features encoded in an organism's genomic sequence, which are known to be associated with essentiality [26, 76]. The classifiers are trained and tested using well-annotated genomes, then applied to identify putatively essential genes in other (novel) genomes [23–25]. The ever-increasing number of experimentally determined essential genes has improved the understanding of distinguishing properties of essential genes. As a result, it is possible to easily select features toward improving the predictive accuracy of ML models, making them less laborious. Predictive accuracy of ML classifiers resulting from a combination of different features may vary, but no specific combinations have been confirmed to be optimally robust. The reliability of ML models however depends on the closeness of the training data set to the study data set. Normally, ML models may be prone to overfitting, potentially allowing irrelevant information or noise to be presented as valid predictions. Domingos and colleagues reviewed overfitting as well as other sources of errors in ML and the possible methods of combating them [77]. These models may not be suitable for predicting conditional essentiality. Various experimental, genomic and protein features have been used to train and build classifier for genes essentiality prediction in different studies (Table 1). However, no single study has reported use of all the features in a single predictive model to predict gene essentiality. The features are often used selectively based on their accuracy and whether they can patently be determined for the organism under study.

Applications of essential genes

Discovering potential drug targets

Several diseases are becoming increasingly difficult to control because of the emergence of drug-resistant pathogenic strains, necessitating a search for new antimicrobials. Identification and prioritization of drug targets in novel pathogens is the initial and one of the most important steps during drug discovery (Figure 3). Understanding the functions of the target proteins, and consequently the mechanisms of action (MOA) are important to design putative inhibitors. As such, drug target discovery sets a foundation for developing drugs with desired therapeutic properties. Inhibiting essential proteins will confer bacteriostatic or bactericidal effects. They therefore form promising targets for discovering potent antibiotics against novel pathogens [7]. In our recent study, using a high-throughput genome-wide screening approach, we identified essential genes in bacterial respiratory pathogens [8]. From these essential genes, additional criteria were applied to prioritize, and experimentally validate some potential target proteins and pathways that can be modulated by bioactive agents.

Despite the intensifying research efforts, adoption of the biomedical discoveries into developmental stages of drug discovery, and subsequently into marketable products has been dismal. Indeed, over 80% of all potential products going into the drug development pipeline never make it to the market [85]. The problem might partly be because of the lack of comprehensive biochemical knowledge of the drug targets and the MOA of their 'lead compound' inhibitors, such that unexpected biological effects are not fully assessed before clinical trials [86]. Moreover, taking a drug through research and development to clinical approval requires immense investments in both time and cost further exacerbating the problem. In fact, only a handful of therapeutic molecules have been approved in recent years by the regulatory agencies in the USA and Europe [85].

Food microbiology and industrial bioprocessing

Numerous food products, including ripened cheese, pickles, wine, beer, bread, yoghurt and other fermented foods, owe their production and characteristics to microorganisms [87]. These foods are to some extent naturally preserved because of the fermentation process. Concurrently, their shelf life is prolonged significantly over that of the raw materials from which they are

Table 1. Features that can be used for *in silico* prediction of gene essentiality

Feature	Rationale	Reference
Gene expression profile[a]	Genes that are not expressed under given conditions are less likely to be essential. Co-expressed genes are often involved in the same pathway or similar cellular function. Interacting proteins are frequently co-expressed.	[78]
Protein localization and biological processes [(enrichment of Gene Ontology (GO)][b]	Essential proteins are enriched in, but not exclusive to the cytoplasm: compared with essential genes, significantly higher proportions of nonessential genes are located in the cytoplasmic membrane, periplasm, outer membrane, cell wall and extracellularly. GO term transcriptional regulation annotations are enriched in essential genes.	[21, 79]
Functional domains	The functional units of proteins are domains, most of which are highly conserved in diverse genera.	[80, 81]
Total upstream gene size[c]	Genes with larger upstream sizes (promoter regions) are significantly underrepresented in indispensable genes. Genes regulated by multiple transcriptional regulators are likely to have larger upstream regions to house the various *cis*-regulatory elements. Genes with more complex regulation are generally dispensable.	[3, 21, 42, 82]
Phyletic retention measure[d]	Essentiality of a gene is extrapolated if the annotated function of that gene can be detected in different genera as opposed to sequence similarity. Specificity increases with inclusion of diverse genera in the analysis.	[21, 45]
GC content	Commonly used to identify genes that are essential under high temperature selection. The DNA double helix is stabilized primarily by hydrogen bonds between nucleotides and base-stacking interactions among aromatic nucleobases: the GC pair contains three hydrogen bonds, whereas the adenine thymine pairs contain two. DNA with high GC content is believed to be more robust and stable.	[83]
Codon usage	The probability of a deleterious substitution in essential proteins is expected to be negligible, resulting in lower nonsynonymous substitution rates.	[41]
Orthology and paralogy	In bacteria, essential genes are generally more conserved across species (orthologs) than nonessential genes. Duplicated genes within a genome (paralogs) are also less likely to be essential because the duplicate gene serves as a backup and can replace the original copy. Inactivation of both copies may however result to lethality.	[41, 46]
Protein connectivity	Highly connected proteins in a network evolve slowly and are more likely to be essential; see PPI networks.	[53, 54]
Strand bias	Essential genes tend to be encoded on the leading strand of the circular chromosome.	[84]

[a] An organism's genetic code is interpreted by gene expression into functional gene products: Properties of gene expression give rise to a phenotype, often expressed by synthesis of proteins that act as catalytic enzymes in specific metabolic pathways, or control the organism s physical traits [102].

[b] Essential functional proteins domains have also been identified and used to predict gene essentiality [80, 81].

[c] The connection between regulation complexity, intergenic distance, and gene essentiality has been shown in *Drosophila melanogaster* and *Caenorrhabditis elegans* [82]. Since transcription factor binding sites in the promoter region are discovered using laborious experimental methods, the possibility of using easy-to-determine upstream region size, as a representation for regulatory complexity in integrative models, is advantageous.

[d] Often confused with conservation; a measure of substitution rate, phyletic gene retention is a measure of the number of organism in which an ortholog is present [21, 45]. It is therefore assumed that most essential genes could be predicted based on the genome annotations. However, the number of essential genes is likely to decrease with increased diversity in the genera, subsequently leading to under-prediction [45].

manufactured. Several biopolymers produced by microorganisms are also used in the food industry [88]. In addition, 'generally recognized as safe' bacteria or probiotics are becoming increasingly vital in the food industry [89]. However, microorganisms also constitute potential food and process contaminants, and foodborne pathogens [90]. Understanding essential genes is therefore invaluable in optimizing various facets of industrial fermentation and bioprocessing.

First, given a growth medium containing a defined carbon source (raw product), metabolic reconstruction can be used to predict the best processing conditions required for a given microorganism to maximize production. The ability to evaluate interactions between microorganisms and relevant metabolic capabilities in new microbes also provides leads for novel and safe starter cultures. Moreover, the knowledge gives an insight into all (conditionally) essential metabolic pathways involved in producing a given product and co-expressed nonessential pathways, which if feasible, could be inactivated to improve efficiency. For example, attempts have been made in using microbes to produce alginate, a product conventionally isolated from farmed brown seaweed [91]. Knowledge of (conditional) gene essentiality was used to pinpoint the interactions between multiple microorganisms during the course of the fermentation. To facilitate product optimization strategies, catalogs of (conditionally) essential genes for specific microorganisms in specific food products can aid in establishing biobanks and biorepositories. The biobanks can be screened for new 'safe' microbes with desired phenotypic traits and predicted interactions for a given fermentation, or to create novel fermented products.

Second, while most foodborne pathogens or food spoilage bacteria and industrial contaminants are cleared by standard sterilization, cooking and preservatives, including bacterial toxins like nisin, studies have shown that some potentially harmful microorganisms can survive these conventional food processing methods [92, 93]. Food samples could be tested by polymerase chain reaction for the presence of general spoiler marker genes, indicating a possible contamination. Such tracking tests are significantly faster than conventional techniques.

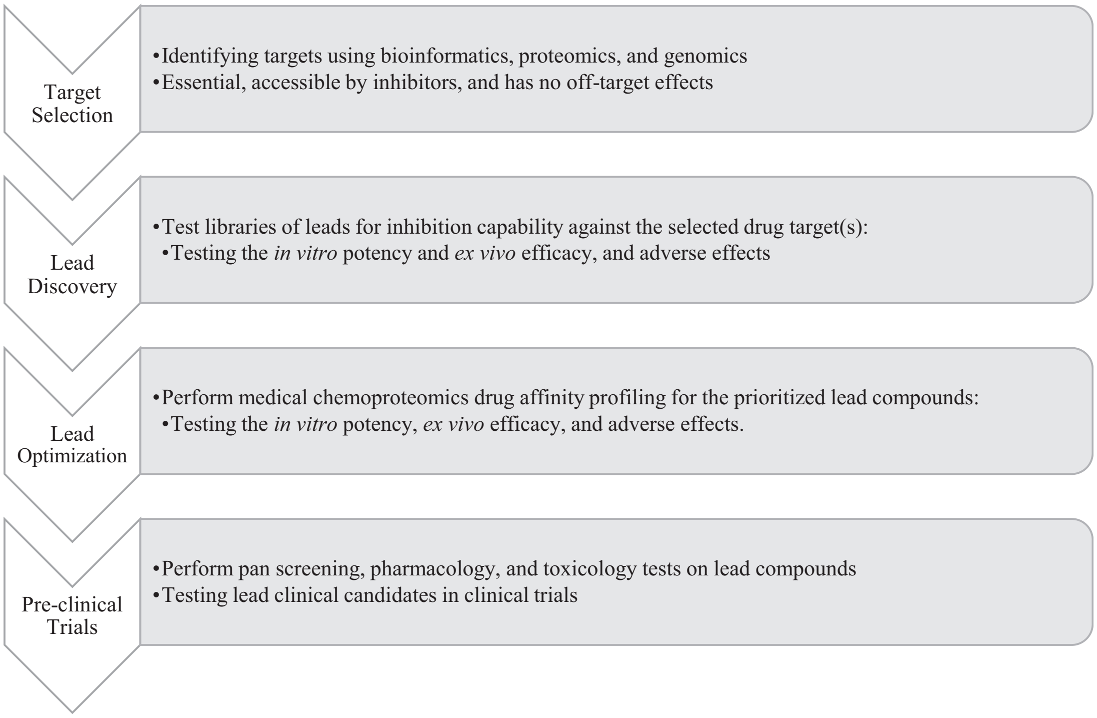

Target Selection
•Identifying targets using bioinformatics, proteomics, and genomics
•Essential, accessible by inhibitors, and has no off-target effects
Lead Discovery
•Test libraries of leads for inhibition capability against the selected drug target(s):
•Testing the *in vitro* potency and *ex vivo* efficacy, and adverse effects
Lead Optimization
•Perform medical chemoproteomics drug affinity profiling for the prioritized lead compounds:
•Testing the *in vitro* potency, *ex vivo* efficacy, and adverse effects.
Pre-clinical Trials
•Perform pan screening, pharmacology, and toxicology tests on lead compounds
•Testing lead clinical candidates in clinical trials

Figure 3. Schematic representation of the drug discovery pipeline.

Moreover, unique and essential spoiler genes could be considered as prime targets for bespoke decontamination strategies without inversely altering the production pipeline.

Finally, genomics-based determination of gene essentiality also generates valuable knowledge that can be used for metabolic engineering, optimizing cell factories and development of novel preservation methods, provided that these solutions are ethically acceptable.

Bioremediation

Compared with conventional physicochemical strategies, microbes provide a safe and cheap alternative for environmental remediation, pollution prevention and waste treatment [94]. Although highly diverse and specialized microbial populations present in the environment efficiently eliminate many pollutants, the process is normally slow, potentially permitting pollutants to accumulate above hazardous levels. For example, bioremediation of the 'Exxon Valdez' oil spill in Alaska using indigenous microflora was cost-effective and scientifically rational [95]. However, fertilizers had to be applied to accelerate the process. The fertilizers present a separate environmental imbalance, albeit minimal compared with the oil spill. Unlike oil whose constituent hydrocarbons are largely biodegradable, most recalcitrant compounds, especially heavy metals, contain structural elements or substituents that seldom occur in nature. Because of the rarity of these compounds, currently known microorganisms have probably not evolved appropriate pathways to bioaccumulate them. While some xenobiotics are inefficiently or incompletely biotransformed, or their complex mixtures inhibit degradation by existing pathways, for others, derivative pathways have not been described [96]. Knowledge of (conditional) gene essentiality can therefore aid in identifying novel biodegradation pathways in (new) microorganisms. Additionally, the knowledge could facilitate genetic modification of microbes to broaden their substrates range, successfully enhancing cleanup while producing specialized (end- or by-) products with less ecological harm.

Genotype–phenotype correlation

Mendel's classical observations of varied phenotypes in peas conjured a paradigm of distinct alterations in an organism's DNA (genotype) that cause disruptions in gene function and characteristic phenotype. Ever since, phenotypes have been used to systematically discover their plausible genetic background. However, the phenotype of a given strain is not only a product of its gene content but also its cellular regulatory mechanisms [97] and environmental factors [98]. Although genotype–phenotype association studies do not factor in the effects of regulatory mechanisms, they allow for straightforward screening of candidate genotype to phenotype relationships. Additionally, the natural diversity and adaptive responses of microbial strains to environmental changes could also be investigated using knowledge of conditional gene essentiality. For example, using transcriptomic diversity between strains of *Lactococcus lactis* isolated from diary and nondairy niches, the basis of phenotypic differences observed in fermented food products at the level of acidification properties has been investigated [99].

Perspectives

Gene essentiality prediction using computational methods will become more important with the ongoing advances in biology. Expanding computational methods to predict conditionally essential genes, which are currently predicted exclusively using laboratory techniques, may soon be realized. Evidently, predicting conditional essentiality requires many experimentally determined features that cannot be determined computationally yet. *In silico* reconstruction of microbial genomes with preferred phenotypic traits also stands to benefit. In fact, the *ab initio* assembly of a synthetic cell [6], and genome transplantation [100] has been accomplished in *Mycoplasma mycoides*. Fabricating viable cells that harbor housekeeping functions and only genes encoding desired phenotypes is therefore achievable and can be perfected in the future. There have been genome engineering attempts to improve *de novo* biosynthesis of vanillin [101]. With the global demand for natural food ingredients, flavors, fragrances, biopolymers and drugs increasing rapidly, specialized fabricated microbes that perform 'natural-like' bioconversions more efficiently might be desirable. Such projects will undoubtedly revolutionize processes beyond current technologies when they are scaled up to industrial size production. Additionally, by creating, testing and optimizing specialized genetic circuits, our understanding of cell biology will also advance significantly. In conclusion, it is our believe that future studies could build on the knowledge reviewed here, and expand it to improve accuracy and dependability of *in silico* tools in predicting essential genes.

Key Points

- Essential genes are fundamental for cellular growth and viability of an organism.
- They form attractive drug targets and essential components of a minimum cell for biotechnology and basic biological research.
- Supplementing or complementing traditional laboratory gene essentiality prediction methods with high-throughput computational approaches is gaining interest.
- Currently, transposon insertion sequencing is the most reliable but expensive method that combines wet laboratory and computational tracking to predict gene essentiality.
- Solely computational methods, including homology models, ML models, metabolic network reconstruction and protein–protein interaction models, are reliable but largely influenced by the quality of data and evolutionary distance between subjects.

Supplementary Data

Supplementary data are available online at http://bib.oxfordjournals.org/.

References

1. Kobayashi K, Ehrlich SD, Albertini A, *et al*. Essential *Bacillus subtilis* genes. *Proc Natl Acad Sci* 2003;**100**:4678–83.
2. Juhas M, Eberl L, Glass JI. Essence of life: essential genes of minimal genomes. *Trends Cell Biol* 2011;**21**:562–8.
3. Yu H, Greenbaum D, Xin Lu H, *et al.* Genomic analysis of essentiality within protein networks. *Trends Genet* 2004;**20**:227–31.
4. Dowell RD, Ryan O, Jansen A, *et al*. Genotype to phenotype: a complex problem. *Science* 2010;**328**:469.
5. Gil R, Silva FJ, Peretó J, *et al*. Determination of the core of a minimal bacterial gene set. *Microbiol Mol Biol Rev* 2004;**68**:518–37.
6. Gibson DG, Glass JI, Lartigue C, *et al*. Creation of a bacterial cell controlled by a chemically synthesized genome. *Science* 2010;**329**:52–6.
7. Chung BKS, Dick T, Lee DY. In silico analyses for the discovery of tuberculosis drug targets. *J Antimicrob Chemother* 2013;**68**:2701–9.
8. Mobegi FM, van Hijum SA, Burghout P, *et al*. From microbial gene essentiality to novel antimicrobial drug targets. *BMC Genomics* 2014;**15**:958.
9. Dickerson JE, Zhu A, Robertson DL, *et al*. Defining the role of essential genes in human disease. *PLoS One* 2011;**6**:e27368.
10. Akerley BJ, Rubin EJ, Novick VL, *et al*. A genome-scale analysis for identification of genes required for growth or survival of *Haemophilus influenzae*. *Proc Natl Acad Sci USA* 2002;**99**:966–71.
11. Giaever G, Chu AM, Ni L, *et al*. Functional profiling of the *Saccharomyces cerevisiae* genome. *Nature* 2002;**418**:387–91.
12. Salama NR, Shepherd B, Falkow S. Global transposon mutagenesis and essential gene analysis of *Helicobacter pylori*. *J Bacteriol* 2004;**186**:7926–35.
13. Glass JI, Assad-Garcia N, Alperovich N, *et al*. Essential genes of a minimal bacterium. *Proc Natl Acad Sci USA* 2006;**103**:425–30.
14. Christen B, Abeliuk E, Collier JM, *et al*. The essential genome of a bacterium. *Mol Syst Biol* 2011;**7**:528.
15. van Opijnen T, Bodi KL, Camilli A. Tn-seq: high-throughput parallel sequencing for fitness and genetic interaction studies in microorganisms. *Nat Methods* 2009;**6**:767–72.
16. D'Elia MA, Pereira MP, Brown ED. Are essential genes really essential? *Trends Microbiol* 2009;**17**:433–8.
17. Mushegian A. The minimal genome concept. *Curr Opin Genet Dev* 1999;**9**:709–14.
18. Roemer T, Jiang B, Davison J, *et al*. Large-scale essential gene identification in *Candida albicans* and applications to antifungal drug discovery. *Mol Microbiol* 2003;**50**:167–81.
19. Andreadaki M, Morgan RN, Deligianni E, *et al*. Genetic crosses and complementation reveal essential functions for the Plasmodium stage-specific actin2 in sporogonic development. *Cell Microbiol* 2014;**16**:751–67.
20. Cullen LM, Arndt GM. Genome-wide screening for gene function using RNAi in mammalian cells. *Immunol Cell Biol* 2005;**83**:217–23.
21. Gustafson A, Snitkin E, Parker S, *et al*. Towards the identification of essential genes using targeted genome sequencing and comparative analysis. *BMC Genomics* 2006;**7**:265.
22. Seringhaus M, Paccanaro A, Borneman A, *et al*. Predicting essential genes in fungal genomes. *Genome Res* 2006;**16**:1126–35.
23. Chen Y, Xu D. Understanding protein dispensability through machine-learning analysis of high-throughput data. *Bioinformatics* 2005;**21**:575–81.
24. Deng J. An integrated machine-learning model to predict prokaryotic essential genes. In: Lu LJ (ed). *Gene Essentiality*. New York: Springer, 2015, 137–51.
25. Plaimas K, Eils R, Konig R. Identifying essential genes in bacterial metabolic networks with machine learning methods. *BMC Syst Biol* 2010;**4**:56.
26. Deng JY, Deng L, Su SC, *et al*. Investigating the predictability of essential genes across distantly related organisms using an integrative approach. *Nucleic Acids Res* 2011;**39**:795–807.
27. Hensel M, Shea J, Gleeson C, *et al*. Simultaneous identification of bacterial virulence genes by negative selection. *Science* 1995;**269**:400–3.
28. Hutchison CA, Peterson SN, Gill SR, *et al.* global transposon mutagenesis and a minimal *Mycoplasma* genome. *Science* 1999;**286**:2165–9.
29. Song JH, Ko KS, Lee JY, *et al*. Identification of essential genes in *Streptococcus pneumoniae* by allelic replacement mutagenesis. *Mol Cells* 2005;**19**:365–74.

30. Goodman AL, McNulty NP, Zhao Y, *et al*. Identifying genetic determinants needed to establish a human gut symbiont in its habitat. *Cell Host Microbe* 2009;**6**:279–89.
31. Gawronski JD, Wong SM, Giannoukos G, *et al*. Tracking insertion mutants within libraries by deep sequencing and a genome-wide screen for *Haemophilus* genes required in the lung. *Proc Natl Acad Sci USA* 2009;**106**:16422–7.
32. Langridge GC, Phan MD, Turner DJ, *et al*. Simultaneous assay of every *Salmonella Typhi* gene using one million transposon mutants. *Genome Res* 2009;**19**:2308–16.
33. van Opijnen T, Camilli A. Transposon insertion sequencing: a new tool for systems-level analysis of microorganisms. *Nat Rev Microbiol* 2013;**11**:435–42.
34. Barquist L, Boinett CJ, Cain AK. Approaches to querying bacterial genomes with transposon-insertion sequencing. *RNA Biol* 2013;**10**:1161–9.
35. Mushegian AR, Koonin EV. A minimal gene set for cellular life derived by comparison of complete bacterial genomes. *Proc Natl Acad Sci USA* 1996;**93**:10268–73.
36. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* 2004;**32**:1792–7.
37. Thompson JD, Gibson TJ, Plewniak F, *et al*. The CLUSTAL_X Windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res* 1997;**25**:4876–82.
38. Thompson JD, Higgins DG, Gibson TJ. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* 1994;**22**:4673–80.
39. Larkin MA, Blackshields G, Brown NP, *et al*. Clustal W and Clustal X version 2.0. *Bioinformatics* 2007;**23**:2947–8.
40. Notredame C, Higgins DG, Heringa J. T-coffee: a novel method for fast and accurate multiple sequence alignment1. *J Mol Biol* 2000;**302**:205–17.
41. Jordan IK, Rogozin IB, Wolf YI, *et al*. essential genes are more evolutionarily conserved than are nonessential genes in bacteria. *Genome Res* 2002;**12**:962–8.
42. Fang G, Rocha E, Danchin A. How essential are nonessential genes? *Molr Biol Evol* 2005;**22**:2147–56.
43. Bruccoleri RE, Dougherty TJ, Davison DB. Concordance analysis of microbial genomes. *Nucleic Acids Res* 1998;**26**:4482–6.
44. Kim DU, Hayles J, Kim D, *et al*. Analysis of a genome-wide set of gene deletions in the fission yeast *Schizosaccharomyces pombe*. *Nat Biotech* 2010;**28**:617–23.
45. Xu P, Ge X, Chen L, *et al*. Genome-wide essential gene identification in *Streptococcus sanguinis*. *Sci Rep* 2011;**1**:125.
46. Doyle M, Gasser R, Woodcroft B, *et al*. Drug target prediction and prioritization: using orthology to predict essentiality in parasite genomes. *BMC Genomics* 2010;**11**:222.
47. Zalacain M, Biswas S, Ingraham KA, *et al*. A global approach to identify novel broad-spectrum antibacterial targets among proteins of unknown function. *J Mol Microbiol Biotechnol* 2003;**6**:109–26.
48. Herce HD, Deng W, Helma J, *et al*. Visualization and targeted disruption of protein interactions in living cells. *Nat Commun* 2013;**4**:2660.
49. Fields S, Song OK. A novel genetic system to detect protein-protein interactions. *Nature* 1989;**340**:245–6.
50. Cremazy FGE, Manders EMM, Bastiaens PIH, *et al*. Imaging in situ protein–DNA interactions in the cell nucleus using FRET–FLIM. *Exp Cell Res* 2005;**309**:390–6.
51. Tsuganezawa K, Nakagawa Y, Kato M, *et al*. A fluorescent-based high-throughput screening assay for small molecules that inhibit the interaction of MdmX with p53. *J Biomol Screen* 2013;**18**:191–8.
52. Acencio ML, Lemke N. Towards the prediction of essential genes by integration of network topology, cellular localization and biological process information. *BMC Bioinformatics* 2009;**10**:290.
53. Hahn MW, Kern AD. comparative genomics of centrality and essentiality in three eukaryotic protein-interaction networks. *Mol Biol Evol* 2005;**22**:803–6.
54. He X, Zhang J. Why do hubs tend to be essential in protein networks? *PLoS Genet* 2006;**2**:e88.
55. Marcotte EM, Pellegrini M, Thompson MJ, *et al*. A combined algorithm for genome-wide prediction of protein function. *Nature* 1999;**402**:83–6.
56. Browne F, Zheng H, Wang H, *et al*. From experimental approaches to computational techniques: a review on the prediction of protein-protein interactions. *Adv Artif Intell* 2010;**2010**:1–15.
57. Ebrahim A, Lerman J, Palsson B, *et al*. COBRApy: constraints-based reconstruction and analysis for Python. *BMC Syst Biol* 2013;**7**:74.
58. Schellenberger J, Que R, Fleming RMT, *et al*. Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox v2.0. *Nat Protoc* 2011;**6**:1290–307.
59. Francke C, Siezen RJ, Teusink B. Reconstructing the metabolic network of a bacterium from its genome. *Trends Microbiol* 2005;**13**:550–8.
60. Lewis NE, Nagarajan H, Palsson BO. Constraining the metabolic genotype–phenotype relationship using a phylogeny of in silico methods. *Nat Rev Micro* 2012;**10**:291–305.
61. Feist AM, Herrgard MJ, Thiele I, *et al*. Reconstruction of biochemical networks in microorganisms. *Nat Rev Micro* 2009;**7**:129–43.
62. Henry CS, DeJongh M, Best AA, *et al*. High-throughput generation, optimization and analysis of genome-scale metabolic models. *Nat Biotech* 2010;**28**:977–82.
63. Aziz RK, Bartels D, Best AA, *et al*. The RAST Server: rapid annotations using subsystems technology. *BMC Genomics* 2008;**9**:75.
64. Schellenberger J, Park J, Conrad T, *et al*. BiGG: a biochemical genetic and genomic knowledgebase of large scale metabolic reconstructions. *BMC Bioinformatics* 2010;**11**:213.
65. Monk JM, Charusanti P, Aziz RK, *et al*. Genome-scale metabolic reconstructions of multiple *Escherichia coli* strains highlight strain-specific adaptations to nutritional environments. *Proc Natl Acad Sci USA* 2013;**110**:20338–43.
66. Thiele I, Fleming RMT, Bordbar A, *et al*. Functional characterization of alternate optimal solutions of *Escherichia coli*'s transcriptional and translational machinery. *Biophys J* 2010;**98**:2072–81.
67. Hyduke DR, Palsson BØ. Towards genome-scale signalling-network reconstructions. *Nat Rev Genet* 2010;**11**:297–307.
68. O'Brien EJ, Monk JM, Palsson BO. Using genome-scale models to predict biological capabilities. *Cell* 2015;**161**:971–87.
69. Orth JD, Thiele I, Palsson BO. What is flux balance analysis? *Nat Biotech* 2010;**28**:245–8.
70. Basler G. Computational prediction of essential metabolic genes using constraint-based approaches. In: Lu LJ (ed). *Gene Essentiality*. New York: Springer, 2015, 183–204.

71. Mahadevan R, Edwards JS, Doyle FJ, III. Dynamic flux balance analysis of diauxic growth in *Escherichia coli*. *Biophys J* 2002;**83**:1331–40.
72. Segrè D, Vitkup D, Church GM. Analysis of optimality in natural and perturbed metabolic networks. *Proc Natl Acad Sci USA* 2002;**99**:15112–17.
73. Shlomi T, Berkman O, Ruppin E. Regulatory on/off minimization of metabolic flux changes after genetic perturbations. *Proc Natl Acad Sci USA* 2005;**102**:7695–700.
74. Kim PJ, Lee DY, Kim TY, *et al*. Metabolite essentiality elucidates robustness of *Escherichia coli* metabolism. *Proc Natl Acad Sci USA* 2007;**104**:13638–42.
75. Zomorrodi AR, Suthers PF, Ranganathan S, *et al*. Mathematical optimization applications in metabolic networks. *Metab Eng* 2012;**14**:672–86.
76. Cheng J, Xu Z, Wu W, *et al*. Training set selection for the prediction of essential genes. *PLoS One* 2014;**9**:e86805.
77. Domingos P. A few useful things to know about machine learning. *Commun ACM* 2012;**55**:78–87.
78. Jansen R, Greenbaum D, Gerstein M. Relating whole-genome expression data with protein-protein interactions. *Genome Res* 2002;**12**:37–46.
79. Peng C, Gao F. Protein localization analysis of essential genes in prokaryotes. *Sci Rep* 2014;**4**:6001.
80. Goodacre NF, Gerloff DL, Uetz P. Protein domains of unknown function are essential in bacteria. *MBio* 2013;**5**:e00744–13.
81. Lu Y, Lu Y, Deng J, *et al*. Discovering essential domains in essential genes. In: Lu LJ (ed). *Gene Essentiality*. New York, NY: Springer, 2015, 235–45.
82. Nelson C, Hersh B, Carroll S. The regulatory content of intergenic DNA shapes genome architecture. *Genome Biol* 2004;**5**:R25.
83. Yakovchuk P, Protozanova E, Frank-Kamenetskii MD. Base-stacking and base-pairing contributions into thermal stability of the DNA double helix. *Nucleic Acids Res* 2006;**34**:564–74.
84. Rocha EPC, Danchin A. Essentiality, not expressiveness, drives gene-strand bias in bacteria. *Nat Genet* 2003;**34**:377–78.
85. Lesko LJ, Woodcock J. Translation of pharmacogenomics and pharmacogenetics: a regulatory perspective. *Nat Rev Drug Discov* 2004;**3**:763–9.
86. Chan JNY, Nislow C, Emili A. Recent advances and method development for drug target identification. *Trends Pharmacol Sci* 2010;**31**:82–8.
87. Jay J. Fermented foods and related products of fermentation. In: *Modern Food Microbiology*. Netherlands: Springer, 1992, 371–409.
88. Breuer U. Book reviews: microbial production of biopolymers and polymer precursors: applications and perspectives. Edited by Bernd H. A. Rehm, *Clean Soil Air Water* 2009;**37**:414.
89. Marshall VM. Probiotics and prebiotics: scientific aspects (2005). *Int J Dairy Technol* 2007;**60**:63–4.
90. Balter S. Foodborne pathogens: microbiology and molecular biology, emerging infectious. *Diseases* 2006;**12**:2003.
91. Hay ID, Rehman ZU, Moradali MF, *et al*. Microbial alginate production, modification and its applications. *Microb Biotechnol* 2013;**6**:637–50.
92. Lima LJR, Kamphuis HJ, Nout MJR, *et al*. Microbiota of cocoa powder with particular reference to aerobic thermoresistant spore-formers. *Food Microbiol* 2011;**28**:573–82.
93. Scheldeman P, Herman L, Foster S, *et al*. *Bacillus sporothermodurans* and other highly heat-resistant spore formers in milk. *J Appl Microbiol* 2006;**101**:542–55.
94. Singh SN, Tripathi RD. *Environmental Bioremediation Technologies*. Berlin; New York: Springer, 2007.
95. Pritchard PH, Mueller JG, Rogers JC, *et al*. Oil spill bioremediation: experiences, lessons and results from the Exxon Valdez oil spill in Alaska. *Biodegradation* 1992;**3**:315–335.
96. Pieper DH, Reineke W. Engineering bacteria for bioremediation. *Curr Opin Biotechnol* 2000;**11**:262–70.
97. Dressaire C, Gitton C, Loubière P, *et al*. Transcriptome and proteome exploration to model translation efficiency and protein stability in *Lactococcus lactis*. *PLoS Comput Biol* 2009;**5**:e1000606.
98. Alberch P. From genes to phenotype: dynamical systems and evolvability. *Genetica* 1991;**84**:5–11.
99. Tan-a-ram P, Cardoso T, Daveran-Mingot ML, *et al*. Assessment of the diversity of dairy *Lactococcus lactis* subsp. lactis isolates by an integrated approach combining phenotypic, genomic, and transcriptomic analyses. *Appl Environ Microbiol* 2011;**77**:739–48.
100. Lartigue C, Glass JI, Alperovich N, *et al*. Genome transplantation in bacteria: changing one species to another. *Science* 2007;**317**:632–8.
101. Kaur B, Chakraborty D. Biotechnological and molecular approaches for vanillin production: a review. *Appl Biochem Biotechnol* 2013;**169**:1353–72.
102. Baba T, Ara T, Hasegawa M, *et al*. Construction of *Escherichia coli* K-12 in-frame, single-gene knockout mutants: the Keio collection. *Mol Syst Biol* 2006;**2**:1–11.