

Building the Blueprint of Life

Christopher S. Henry,^{1,2} Ross Overbeek,³ and Rick L Stevens^{1,2}

¹Mathematics and Computer Science Department, Argonne National Laboratory, 9700 S. Cass Avenue, Argonne, IL 60439, USA

²Computation Institute, The University of Chicago, 5640 S. Ellis Avenue, Chicago, IL 60637, USA

³Fellowship for Interpretation of Genomes, 15W155 81st Street, Burr Ridge, IL 60527, USA

Corresponding author:
Dr. Christopher Henry
Computation Institute
The University of Chicago
5640 S. Ellis Avenue
Chicago, IL 60637, USA
Phone: 847-757-4377

Email addresses:
CSH: cshenry@uchicago.edu
RO: rossoverbeek@gmail.com
RLS: stevens@anl.gov

Keywords:
Synthetic biology
Minimal organism
Bacillus subtilis
Escherichia coli
Mycoplasma genitalium

Abstract

With recent breakthroughs in experimental microbiology making it possible to synthesize and implant an entire genome to create a living cell, the challenge of constructing a working blueprint for the first truly minimal synthetic organism is more important than ever. Here we review the significant progress made in the design and creation of a minimal organism. We discuss how comparative genomes, gene essentiality data, naturally small genomes, and metabolic modelling are all being applied to produce a catalogue of the biological functions essential for life. We compare the minimal gene sets from three published sources with functions identified in 13 existing gene essentiality datasets. We examine how genome-scale metabolic models have been applied to design a minimal metabolism for growth in simple and complex media. Additionally, we survey the progress of efforts to construct a minimal organism, either through implementation of combinatorial deletions in *B. subtilis* and *E. coli* or through the synthesis and implantation of synthetic genomes.

Introduction

The recent breakthrough involving the de novo synthesis and implantation of an entire prokaryotic genome to create a living synthetic cell [1] represents a major triumph integrating decades of research in virtually every area of biology. This proof-of-principle breakthrough has the potential to dramatically change how metabolic engineering is done: rather than modify existing organisms in small ways to fit them to our needs, entire genomes can be designed, synthesized, and implanted to produce living synthetic cells. This is far from being a new idea in biology [2-4], but the work of Venter and colleagues changes the consideration of this idea from

a hypothetical “what if” into something that is technologically achievable in the near future if a working blueprint for the synthetic cell can be created.

If we are to produce this blueprint for a living, dividing, and robust synthetic cell, we must know which parts to include in the blueprint, we must understand the function of each part, and we must ensure that enough regulatory machinery exists for the parts to function as a harmonious whole [2]. It is precisely this need for a minimal catalogue of the essential biological functions required for life that drives the pursuit for the design and creation of a minimal organism [5-9]. Here, we define a minimal organism as a living, dividing cell where every gene in the genome is indispensable for viability in a defined chemical environment. This is similar to existing definitions [2, 5, 6, 10] except we propose that the chemical environment of the cell be strictly defined so the nutrients required by the cell for survival are completely understood. A minimal organism satisfying this definition will most likely not possess the smallest gene set required for life, as the organism would include additional genes required to survive in the defined chemical environment. Instead, this definition seeks to strike a balance between a number of closely related goals relating to minimality, level of understanding for each gene, and the ability to engineer a functional system. Over the next decade, we anticipate the creation of numerous synthetic microorganisms that will fit this definition of a “minimal cell”, and while differences between specific incarnations will certainly exist, these will reflect the scientific or industrial goal driving the synthesis of the organism.

Efforts to develop a minimal organism, both theoretically and experimentally, have been a driving force in basic biological science, requiring that we understand in detail every essential biological function [2]. These efforts are already helping to improve genome annotations [5, 6, 11], correct models [12-14], and gain insights into global cell regulation [8, 15].

The goal of creating a minimal organism also has practical applications beyond basic science. A minimal organism will lack mobile DNA elements (e.g., insertion elements, transposases, phages, integrases, and site-specific recombinases) that plague efforts to control strains used in the industrial setting over the long term [15]. A minimal organism will also lack competing metabolic pathways that drive raw materials away from desired end products and toward useless or even toxic by-products [2]. Moreover, a minimal organism will lack the complex layer of transcriptional regulatory interactions that make natural microorganisms resistant to engineering efforts [16]. A minimal organism will have applications in medicine, commodity chemicals, pharmaceuticals, fuel production, carbon sequestration, and waste cleanup.

Researchers have been contemplating the possible content of a minimum genome since the first complete genome sequence emerged 15 years ago [17, 18]. Here we examine efforts to compile the minimal gene set using various methods, including analysis of naturally small genomes [19], gene essentiality studies [18, 20], comparative genomics [11], *in vitro* implantation of biological subsystems [2], or a combination of all approaches [6]. We compile the minimal gene lists proposed in several sources; and using SEED functional roles and sequence homology [21], we compare the lists to identify those elements that are universal among all lists or appear in a subset of lists. We also examine efforts to synthesize a minimal genome in the wet-lab. These efforts usually involve either the reduction of nonminimal genomes by combinatorial deletions (the top-down approach) or the de novo synthesis and implantation of synthetic minimal genomes (the bottom-up approach). We discuss the progress, strengths, and challenges associated with each effort.

The Natural Minimal Organism: Mycoplasma genitalium

Mycoplasma genitalium is a natural first stop in the pursuit of the minimal gene set [19]. With 0.58 MB encoding 482 genes [19], it is the smallest culturable bacterium sequenced to date. Because *M. genitalium* meets most of the definition of a minimal organism (except for a defined growth media), its 482 genes establish an upper bound on the number of genes in the minimal set. Unfortunately, functions have been assigned to only 326 (68%) of the genes in *M. genitalium*; hence, much is still not understood about how this organism survives (Table I and Supplementary Table S1).

To identify which genes in *M. genitalium* might be expendable, researchers have compared *M. genitalium* with a close relative: *Mycoplasma pneumonia* with 0.82 MB encoding 677 genes [22]. The comparison of these two “minimal” genomes revealed that all 482 open reading frames found in the *M. genitalium* genome are also included in the *M. pneumonia* genome [23]. While the comparison with *M. pneumonia* did not help identify expendable genes in *M. genitalium*, the 110 genes found to be unique to *M. pneumonia* did help explain why *M. genitalium* is much more fastidious than *M. pneumonia* [23].

Another method used to identify the expendable genes in *M. genitalium* was global transposon mutagenesis [20]. In this technique, transposons were targeted at every gene in the *M. genitalium* genome, but strains with successful insertions were observed only when a nonessential gene was disrupted. In this study, 100 genes in *M. genitalium* were found to be individually dispensable [20]. It is unlikely, however, that all 100 of these dispensable genes can be simultaneously removed without disrupting cell viability. Thus, this study placed only a potential lower bound on the size of a minimal gene set: ~382. Based on this analysis alone, the predicted minimal genome consists of between 382 and 482 genes, with biological functions

assigned to 272 (71%) of the 382 essential genes and 54 (54%) of the 100 nonessential genes (Table I and Supplementary Table S1). This constitutes one possible hypothesis for the minimal gene set for life.

Comparison of Published Hypothetical Minimal Gene Sets

These types of analyses of the *M. genitalium* genome have formed for the foundation for most hypothetical minimal gene sets proposed in the literature. However, the proposed minimal gene sets employ a variety of additional filters in an effort to remove all dispensable genes while still preserving the functions required for life. One of the first minimal gene sets was generated through a comparative genomics analysis of the *M. genitalium* and *H. Influenza* genomes [11]. This analysis proposed that any genes found to be conserved in these two taxonomically distant organisms must be essential for life; a set of 240 such conserved genes was identified. The analysis also examined the “completeness” of the set of biological functions encoded by this set of 240 and, through this analysis, identified another 16 genes required to fill clear gaps in the functional set, bringing the total set of genes in the minimal set to 256 (Table I and Supplementary Table S3). This set of 256 was later revisited by Koonin once more complete genome sequences were available for comparison. In his study, Koonin found that only 60 of the original 256 genes were conserved across the larger set of genomes [24]. However, he proposed that the functions encoded by the remaining unconserved 196 genes were nevertheless present in other genomes as *nonorthologous gene displacements*. Remarkably, the conclusion of this revisit was to leave the original set of 256 unchanged [24]. As might be expected, every gene in the Koonin minimal set overlaps with the *M. genitalium* genome. (Fig. 2b).

Another recent minimal gene set was derived by Gill and colleagues using a combination of comparative genomics, analysis of small genomes, and analysis/comparison of essential gene

sets [6]. The approach yielded a list of 206 essential genes (Table I and Supplementary Table S3). This list overlapped significantly with the list generated by Koonin, with several key exceptions (Fig. 2a). The Gill list included additional genes associated with DNA replication, metabolism, and translation. This list also lacked many of the genes associated with metabolism in the Koonin minimal set. The significant differences in the metabolic genes included in the Gill and Koonin sets highlights the flexibility possible in the design of a minimal metabolic network. Like the Koonin list, this list showed good overlap with both *M. genitalium* and our derived universal essential gene set (Fig. 2c). However, unlike the Koonin set, the Gill set also included 40 genes not found in *M. genitalium* genome. These 40 genes are primarily related to metabolism, translation, and poorly characterized functions.

Most recently, Church and Forster proposed a new minimal gene set consisting of only 115 genes and 37 RNA [2] (Table I and Supplementary Table S3). The derivation of this set is unique in that it did not depend largely on comparative genomics or gene essentiality datasets. Instead, this set exploited knowledge gained from the in vitro or biochemical implementation of numerous biological subsystems, including translation [25], membrane formation and fission [3], DNA replication [26, 27], and RNA processing [28, 29]. As such, the set primarily consists of simple machinery with well-understood functions required for protein synthesis, mRNA expression, DNA replication, cell division, and extremely simple metabolism. Highly conserved and often essential functions such as DNA repair were left out. Highly simplified alternative biochemical systems were applied that do not resemble the primary mechanisms used in living cells in order to produce a minimal cell where every component is completely understood [3, 26-29]. Comparison of the Church list with the Koonin and Gill revealed significant overlap in the translation machinery, which appears to be inescapable in the design of a minimal cell (Fig. 2a).

Many of the unique genes in the Church list relate to synthesis of RNA components used for DNA replication of other functions. Similarly, the overlap with *M. genitalium* primarily involved functions associated with translation (Fig. 2d).

Gaining More from Gene Essentiality Data

One potential problem with using only the essentiality data from *M. genitalium* in the analysis, formation, and validation of minimal genes sets is the accuracy of the experiments used to identify the 382 essential *M. genitalium* genes. These experiments are often plagued by “false essentials” due to three problems: (i) some genes are never mutated in the screen, (ii) deletion of some genes causes formation of toxic intermediates or incomplete protein complexes that kill the cell and give the illusion of essentiality, and (iii) disruption of some genes has an inadvertent effect on neighbouring genes in the genome, leading to an essential multi-knockout [2]. Additionally, some of the 382 genes identified as essential may be essential only in *M. genitalium* because of complex interdependencies with other *M. genitalium* functions or because of the specific culture conditions in which *M. genitalium* must be grown. These problems can be overcome by integrating many taxonomically diverse gene essentiality datasets into the analysis with a focus on the biological functions that are conserved across multiple datasets. This approach has been applied previously using essentiality datasets from five different organisms [30], but now we expand the study to the 13 currently available datasets [20, 31-44].

We began our analysis by comparing the lists of SEED functional roles associated with the essential genes in each of the 13 datasets to identify the biological functions that are conserved across multiple datasets. We found 2,206 distinct SEED functional roles associated with the 5,252 unique genes represented in the 13 datasets (Supplementary Table S2). Interestingly, no functional roles are conserved across all thirteen datasets, only four are

conserved across twelve datasets, and approximately 40 functions each are conserved across 11, 10, 9, 8, 7, 6, and 5 datasets (Fig. 2a). The number of functions conserved in four and fewer datasets grows rapidly, indicating that the essential functions at that level of conservation are diverging into organism specific functions and not universal functions. To further support this conclusion, we determined the fraction of the functions conserved across 12, 11..., 2, and 1 datasets that were included in each of the proposed minimal gene sets: the *M. genitalium* genome, the essential genes in *M. genitalium*, the Koonin set, the Gill set, and the Church set (Fig. 2b). Nearly 100% of the functions conserved across 12, 11, 10, and 9 datasets are included in nearly all proposed minimal gene sets. The Church set is the key exception, which is expected as this dataset is significantly smaller than the others and employs some mechanisms for translation, transcription, and DNA replication that are not typically used in bacteria. The degree of overlap between the conserved essential functions and the minimal gene sets steadily declines as the number of essentiality datasets the functions are conserved in decreases. This observation strongly reinforces the conclusion that the more universally essential biological functions should be included in the minimal gene sets, but as universality increases, organism and environment specific functions begin to appear among the conserved essential functions.

We selected a subset of the 280 most universal conserved essential functions for deeper analysis, using conservation in at least five essentiality datasets as our cut-off (list shown in Supplementary Table S3). The cut-off of five is somewhat arbitrary as it was selected to produce a manageable list of functions for deeper analysis; we by no means propose this list as a new minimal gene set. Each function in our list of 280 was associated with a biological subsystem (e.g. Protein biosynthesis, Carbohydrate metabolism) similar to those used in the Koonin, Gill, and Church datasets (subsystems shown in Supplementary Table S3). We then analyzed how the

subsystem classification of the conserved essential functions changes as the universality of the conserved functions declines (Table 2). Nearly all functions conserved in nine or more essentiality datasets are associated with non-metabolic subsystems including Protein biosynthesis, tRNA synthesis, Transcription machinery, and DNA replication. Among functions conserved in fewer than nine essentiality data sets, metabolic functions become progressively more heavily represented, with functions being associated with Cofactor biosynthesis. Many essential cofactor molecules (e.g. folate) are difficult to transport and lack long term stability in the environment. For these reasons, many organisms synthesize these compounds internally rather than importing them, which causes these biosynthesis pathways to appear in many essential gene sets.

We also examine the subsystems associated with the conserved essential functions that are not included Koonin, Gill, or Church minimal gene sets (Table 2). As expected, nearly all functions with conserved essentiality across 12, 11, 10, and 9 datasets are included in at least one of the Koonin, Gill, or Church datasets. The exceptions are of interest however, as these represent functions that may be missing from the current minimal gene sets. These include *GTP-binding protein EngA* in Protein biosynthesis; *Glutamyl-tRNA(Gln) amidotransferase subunit B (EC 6.3.5.7)* and *Aspartyl-tRNA(Asn) amidotransferase subunit B (EC 6.3.5.6)* in tRNA synthesis, *Phosphoglucosamine mutase (EC 5.4.2.10)* in Carbohydrate metabolism; *UDP-N-acetylglucosamin 1-carboxy-vinyltransferase (EC 2.5.1.7)*, *UDP-N-acetylmuramoylalanine-D-glutamate ligase (EC 6.3.2.9)*, and *N-acetylglucosamine transferase (EC 2.4.1.227)* in Cofactor biosynthesis, and one uncharacterized protein (MG464 in *M. genitalium*). There are two additional uncharacterized functions conserved in eight essentiality datasets that are also potential targets for additional study (*MG046* and *MG208* in *M. genitalium*) and inclusion in

proposed minimal gene sets. The functions that are conserved in fewer than nine datasets involve far more metabolic functions (most in cofactor biosynthesis) that are not included in the Koonin, Gill, or Church datasets. It is unlikely that these functions are good candidates for inclusion in the minimal gene sets as they are probably essential due to specific biological needs and growth conditions of their host organisms.

Model-driven Design of a Minimal Metabolism

As the comparison of our minimal gene sets confirms, metabolism is one of the more flexible elements of the hypothetical minimal organism. Many alternative metabolic pathways exist that can achieve the minimal metabolic goals needed for life. In particular, biochemical energy can be synthesized in the form of ATP by using a wide variety of methods.

Fortunately, genome-scale metabolic models exist that can be used to predict the set of metabolic functions required for a minimal organism to be viable in a specified chemical environment. Metabolic modelling has been applied to analyze the connectivity and behaviour of the simple metabolic network included in Gill's minimal set of 206 genes [45]. This work found the simple network to function successfully as a concerted whole and behave similarly to natural metabolic networks.

In other important work, the *iJR904* [46] genome-scale metabolic model of *E. coli* was applied with mixed-integer linear optimization to predict the minimal set of metabolic reactions needed for *E. coli* viability in minimal and complex media [47]; the study found that 122 metabolic reactions are required for growth in complex media, with an additional 102 reactions required for growth in minimal media. This is the first work to quantify approximately how many metabolic genes must be added to the minimal organism in order to obtain growth on defined minimal media instead of undefined complex media (~102 additional genes). This also

provides a mechanism for using metabolic models to select exactly which metabolic genes must be included in a minimal organism to ensure viability in desired media conditions.

Ideally this analysis should be repeated using a pan-genome metabolic network rather than constraining the solution space to *E. coli* metabolism only. The biomass reaction used in this analysis should also be adjusted to reflect the reduced biological needs of a minimal organism. Both these modifications would be useful in the development of a metabolic blueprint for a minimal organism. Given the utility of linear optimization and genome-scale metabolic modelling as a mechanism for designing, understanding, and checking the consistency of our knowledge of the minimal metabolism of an organism, we propose that the metabolic model construct could be useful for design of entire minimal genomes if models could be expanded to integrate the non-metabolic genes required for life. Many of these genes can be integrated into the same stoichiometric representation used for metabolism as done in the E-matrix approach [48]. Non-metabolic genes can also be integrated into a logical boolean network like those used for integration of regulatory constraints in metabolic models [49].

Bringing the Blueprint to Life with the Creation of a Minimal Organism

Efforts are under way in many labs throughout the world to produce a living strain with a minimal genome [1, 7, 8, 15, 50-52]. These efforts are applying various approaches, depending on the organisms being used as a starting point and the specific scientific or industrial objectives motivating the effort. All the approaches can be classified as either top down or bottom up [4]. Top-down approaches involve starting with the genome of an existing (often far from minimal) organism and combining deletions to produce progressively smaller genomes [7, 8, 15, 50-52]. Bottom-up approaches involve starting with a very small genome and engineering a reduced version of the entire genome for implantation and viability testing [1]. An even more

fundamental bottom-up approach is being pursued in which no natural genome is used as the starting point. This approach essentially involves assembling various self-replicating biochemical subsystems together in vitro and integrating them in a simple lipomembrane cell [2, 3].

Knocking Out Complexity with the Top-Down Approach

Most efforts to produce a minimal organism fit into the top-down paradigm, where chromosomal regions are systematically deleted to produce progressively smaller strains while preserving viability in set culture conditions; this process continues until no further deletions are possible without loss of viability. Thus far, two organisms have been used in top-down studies: *E. coli* [7, 15, 50, 51] and *B. subtilis* [8, 52].

One primary disadvantage of this approach is that the genomes used as a starting point are typically far from being minimal. *E. coli* contains 4.64 MB encoding 4,312 genes, and *B. subtilis* contains 4.21 MB encoding 4,114 genes. Thus, over 3.6 MB and 3,629 genes must be deleted from each of these organisms just to obtain a genome equal to *M. genitalium* (0.58 MB and 482 genes) in size. Then 150-250 additional genes must be deleted to reach the hypothetical minimal genome. Another disadvantage is that these model organisms likely contain co-dependent infrastructure that is technically dispensable for life but results in unviable strains when deleted in the wrong sequence [7, 15]. These co-dependencies must be disentangled before this infrastructure can be removed from the cell.

The primary advantage of this approach is that both *E. coli* and *B. subtilis* have highly effective genetic transformation mechanisms, making execution of knockouts technically straightforward and relatively fast. In *B. subtilis*, the native natural competence mechanism of these cells is exploited for the uptake of computationally designed primers and antibiotic resistance cassette. These primers and cassette are integrated into the genome by homologous

recombination, at which point the target chromosomal region is spontaneously snipped out. Transformed strains are identified by the antibiotic resistance conferred on them by the input cassette. This cassette is then popped out so the process can be repeated for the knockout of a second chromosomal region [53]. The procedure is similar in *E. coli*, but the cassette used for selecting transformed strains is different, and electro-competence must be used for inserting primers because *E. coli* cells are not naturally competent [7].

Another significant advantage of the top-down approach is that *E. coli* and *B. subtilis* are both versatile organisms that grow rapidly even on minimal media. Hence, a defined (even a minimal) medium may be used to test for viability throughout the genome minimization process and may be selected as the targeted culture condition for the minimal strain. This has significant implications for application of the minimal strain as an industrial or scientific platform. Minimal organisms that inherit the fastidiousness and slow growth of *M. genitalium* will most likely be impractical for use in industry or science. Additionally, at the end of the minimization process, one is left with a catalogue of the *biological subsystems* that were removed during the process. These parts may be reintegrated into the minimal strain to either ascertain their function or confer new desired capabilities on the strain.

The top-down approach also has the advantage of improving our understanding of the organism on which it is used. As chromosomal regions are progressively removed, the phenotypes of intervening strains may be tested and compared with predictions from available genome-scale models [13, 36]. When predictions are incorrect, models are adjusted to remove errors and reveal new insights into biology of the organism being reduced. In current genome reduction efforts in *B. subtilis*, new essential and coessential genes have been identified, new metabolic pathways have been revealed, and essential metabolic cofactors have been identified

[12]. Our understanding of the genome-wide regulation of both *B. subtilis* and *E. coli* has been enhanced by the top-down projects involving these organisms.

Currently, top-down approaches have produced a *B. subtilis* strain reduced by ~1.4 MB (33%) [12] and an *E. coli* strain reduced by 1.38 MB (30%) [7]. Both these efforts are approximately halfway to producing strains of *B. subtilis* or *E. coli* that are smaller than that of *M. genitalium*.

Rewriting the Operating System of Life from the Bottom Up

The bottom-up approach to the creation of a minimal organism is fundamentally different from and far more technologically challenging than the top-down approach. In the pure bottom-up approach, the minimal genome is designed computationally, synthesized in its entirety, and implanted in a living cell to produce a viable minimal organism. Every experimental step involved in implementing the bottom-up approach has now been successfully demonstrated with the genome of *Mycoplasma mycoides* as a template [1]. First, the *M. mycoides* genome was resequenced and computationally disassembled into 1,078 overlapping cassettes, each 1,080 BP long [1]. These cassettes were chemically synthesized and implanted in yeast, where the chromosome repair machinery of yeast was used to assemble these strands into a complete chromosome [9, 54, 55]. Next, the complete chromosome was injected into a *Mycoplasma capricolum* cell, effectively rewriting the operating system of that cell with the instruction set from the injected *M. mycoides* genome [1]. With this proof of principle complete, efforts are now beginning on the design and synthesis of reduced versions of the *M. mycoides* genome. These efforts will continue until the synthetic genomes cannot be reduced further without loss of viability upon implantation.

The most significant advantage of this approach is its lack of reliance on any native cellular machinery for the transformation of the genome. Additionally, there are no intervening strains in this approach, preventing extremely fastidious or slow growing intermediate strains from disrupting efforts to further reduce the genome. Such strains produced during the top-down approach would have to be abandoned because further genome transformations would no longer be practical. Primarily because of this advantage, a minimal strain produced by the bottom-up approach is expected to be smaller than a minimal strain produced by the top-down approach.

Because extremely fastidious organisms can be used with the bottom-up approach, the starting point for this approach is a much smaller genome. The native *M. mycoides* genome includes only 1.08 MB encoding 1,021 genes (much smaller than the 4+ MB and 4,000+ genes used as starting points in the top-down approach). As a result, far fewer portions of the chromosome need to be removed in order to reach a minimal organism.

The primary disadvantage of this approach is the technical difficulty, time, and expense associated with it. Fifteen years were spent just developing the technologies required to enable each step of the currently implemented process, and every attempt to further reduce the *M. mycoides* genome will require the genome assembly and implantation processes to be repeated. Currently this work is just beginning, now that the necessary experimental methods are in place.

Conclusions

Comparative genomics, gene essentiality experiments, genome annotation, and metabolic modelling each have an important role to play in the continued efforts to design the hypothetical minimal genome. The comparison of the three published minimal gene sets, the gene set in *M. genitalium*, and the set of 280 universally essential genes reveals more differences than expected. These results clearly demonstrate that there are likely to be multiple solutions to the minimal

genome challenge. While only one solution is likely to satisfy the strict condition of including the *smallest set of distinct genes required for life*, many solutions probably exist that satisfy the weaker condition of *containing no dispensable genes*. Additionally, each distinct minimal gene set generated, synthesized, and validated is likely to require significantly different growth conditions.

Another important point is that the function of many of the genes included in these minimal sets remain unclear or, in some cases, unknown. Clearly more work is needed to characterize these vital biological functions before successful design of a blueprint for a minimal cell will be possible. One important area for future focus would be the highly conserved essential genes that are not currently included in the published minimal gene sets (specific examples listed in that discussion). Another important area of focus would be the highly conserved essential genes and genes in the published minimal gene sets for which no clear function is known.

Analysis of metabolism revealed a wide range of possibly essential metabolic genes depending on the culture conditions targeted. Approximately 123 genes are need for growth in complex media, while an additional 75 genes are required for growth in minimal media. This study reveals the strength of genome-scale metabolic model and flux balance analysis as a means of designing a minimal organism capable of surviving in a specific chemical environment.

Producing a complete hypothetical minimal gene set in which the function of every gene is well understood is likely an essential prerequisite to the successful design of a functional minimal genome from the ground up. The top-down approach being used to produce minimal strains of *E. coli* and *B. subtilis* is generating data on gene functions and functional interdependencies that will be essential to the completion of this minimal blueprint. The experimental techniques being developed in the bottom-up approach will be essential to

converting this blueprint into a living, metabolizing, and dying synthetic minimal organism. Clearly a synergy exists between these two approaches that will result in a faster path to the successful design and creation of a minimal synthetic organism.

Acknowledgments

This work was supported in part by the U.S. Department of Energy under contract DE-ACO2-06CH11357. We thank the entire SEED development team for advice and assistance in using the SEED annotation system. We thank Kosei Tanaka and Philippe Noirot for data on the top-down *B. subtilis* minimization project.

Conflict of interest statement

The authors have declared no conflict of interest.

References

- [1] Gibson, D. G., Glass, J. I., Lartigue, C., Noskov, V. N., *et al.*, Creation of a bacterial cell controlled by a chemically synthesized genome. *Science* 2010.
- [2] Forster, A. C., Church, G. M., Towards synthesis of a minimal cell. *Mol Syst Biol* 2006, 2, 45.
- [3] Szostak, J. W., Bartel, D. P., Luisi, P. L., Synthesizing life. *Nature* 2001, 409, 387-390.
- [4] Luisi, P. L., Toward the engineering of minimal living cells. *Anat Rec* 2002, 268, 208-214.
- [5] Koonin, E. V., How many genes can make a cell: the minimal-gene-set concept. *Annu Rev Genomics Hum Genet* 2000, 1, 99-116.
- [6] Gil, R., Silva, F. J., Pereto, J., Moya, A., Determination of the core of a minimal bacterial gene set. *Microbiol Mol Biol R* 2004, 68, 518-537.
- [7] Hashimoto, M., Ichimura, T., Mizoguchi, H., Tanaka, K., *et al.*, Cell size and nucleoid organization of engineered *Escherichia coli* cells with a reduced genome. *Mol Microbiol* 2005, 55, 137-149.
- [8] Morimoto, T., Kadoya, R., Endo, K., Tohata, M., *et al.*, Enhanced recombinant protein productivity by genome reduction in *Bacillus subtilis*. *DNA Res* 2008, 15, 73-81.
- [9] Gibson, D. G., Benders, G. A., Andrews-Pfannkoch, C., Denisova, E. A., *et al.*, Complete chemical synthesis, assembly, and cloning of a *Mycoplasma genitalium* genome. *Science* 2008, 319, 1215-1220.
- [10] Peterson, S. N., Fraser, C. M., The complexity of simplicity. *Genome Biol* 2001, 2, 1-7.
- [11] Mushegian, A. R., Koonin, E. V., A minimal gene set for cellular life derived by comparison of complete bacterial genomes. *Proc Natl Acad Sci U S A* 1996, 93, 10268-10273.

- [12] Tanaka, K., Henry, C., Jolivet, E., Zinner, J. F., *et al.*, unpublished results. 2010.
- [13] Henry, C. S., Zinner, J., Cohoon, M., Stevens, R., iBsu1103: a new genome scale metabolic model of *B. subtilis* based on SEED annotations. *Genome Biol* 2009, *10*, R69.
- [14] Suthers, P. F., Dasika, M. S., Kumar, V. S., Denisov, G., *et al.*, A genome-scale metabolic reconstruction of *Mycoplasma genitalium*, iPS189. *PLoS Comput Biol* 2009, *5*, e1000285.
- [15] Posfai, G., Plunkett, G., 3rd, Feher, T., Frisch, D., *et al.*, Emergent properties of reduced-genome *Escherichia coli*. *Science* 2006, *312*, 1044-1046.
- [16] Fischer, E., Sauer, U., Large-scale in vivo flux analysis shows rigidity and suboptimal performance of *Bacillus subtilis* metabolism. *Nat Genet* 2005, *37*, 636-640.
- [17] Fleischmann, R. D., Adams, M. D., White, O., Clayton, R. A., *et al.*, Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* 1995, *269*, 496-512.
- [18] Itaya, M., An estimation of minimal genome size required for life. *FEBS Lett* 1995, *362*, 257-260.
- [19] Fraser, C. M., Gocayne, J. D., White, O., Adams, M. D., *et al.*, The minimal gene complement of *Mycoplasma genitalium*. *Science* 1995, *270*, 397-403.
- [20] Glass, J. I., Assad-Garcia, N., Alperovich, N., Yooseph, S., *et al.*, Essential genes of a minimal bacterium. *Proc Natl Acad Sci U S A* 2006, *103*, 425-430.
- [21] Overbeek, R., Disz, T., Stevens, R., The SEED: A peer-to-peer environment for genome annotation. *Communications of the ACM* 2004, *47*, 46-51.
- [22] Himmelreich, R., Hilbert, H., Plagens, H., Pirkel, E., *et al.*, Complete sequence analysis of the genome of the bacterium *Mycoplasma pneumoniae*. *Nucleic Acids Res* 1996, *24*, 4420-4449.
- [23] Himmelreich, R., Plagens, H., Hilbert, H., Reiner, B., Herrmann, R., Comparative analysis of the genomes of the bacteria *Mycoplasma pneumoniae* and *Mycoplasma genitalium*. *Nucleic Acids Res* 1997, *25*, 701-712.
- [24] Koonin, E. V., Comparative genomics, minimal gene-sets and the last universal common ancestor. *Nat Rev Microbiol* 2003, *1*, 127-136.
- [25] Kung, H. F., Chu, F., Caldwell, P., Spears, C., *et al.*, The mRNA-directed synthesis of the alpha0peptide of beta-galactosidase, ribosomal proteins L12 and L10, and elongation factor Tu, using purified translational factors. *Arch Biochem Biophys* 1978, *187*, 457-463.
- [26] Zhong, X. B., Lizardi, P. M., Huang, X. H., Bray-Ward, P. L., Ward, D. C., Visualization of oligonucleotide probes and point mutations in interphase nuclei and DNA fibers using rolling circle DNA amplification. *Proc Natl Acad Sci U S A* 2001, *98*, 3940-3945.
- [27] Sauer, B., Cre/lox: one more step in the taming of the genome. *Endocrine* 2002, *19*, 221-228.
- [28] Forster, A. C., Altman, S., External guide sequences for an RNA enzyme. *Science* 1990, *249*, 783-786.
- [29] Forster, A. C., Symons, R. H., Self-cleavage of virusoid RNA is performed by the proposed 55-nucleotide active site. *Cell* 1987, *50*, 9-16.
- [30] Gil, R., Silva, F. J., Pereto, J., Moya, A., Determination of the core of a minimal bacterial gene set. *Microbiol Mol Biol Rev* 2004, *68*, 518-537.
- [31] Gerdes, S., Edwards, R., Kubal, M., Fonstein, M., *et al.*, Essential genes on metabolic maps. *Curr Opin Biotechnol* 2006, *17*, 448-456.
- [32] Zhang, R., Ou, H. Y., Zhang, C. T., DEG: a database of essential genes. *Nucleic Acids Res* 2004, *32*, D271-272.

- [33] Durot, M., Le Fevre, F., de Berardinis, V., Kreimeyer, A., *et al.*, Iterative reconstruction of a global metabolic model of *Acinetobacter baylyi* ADP1 using high-throughput growth phenotype and gene essentiality data. *BMC Syst Biol* 2008, 2, 85.
- [34] Akerley, B. J., Rubin, E. J., Novick, V. L., Amaya, K., *et al.*, A genome-scale analysis for identification of genes required for growth or survival of *Haemophilus influenzae*. *Proc Natl Acad Sci U S A* 2002, 99, 966-971.
- [35] Sasseti, C. M., Boyd, D. H., Rubin, E. J., Genes required for mycobacterial growth defined by high density mutagenesis. *Mol Microbiol* 2003, 48, 77-84.
- [36] Feist, A. M., Henry, C. S., Reed, J. L., Krummenacker, M., *et al.*, A genome-scale metabolic reconstruction for *Escherichia coli* K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information. *Mol Syst Biol* 2007, 3, 121.
- [37] Salama, N. R., Shepherd, B., Falkow, S., Global transposon mutagenesis and essential gene analysis of *Helicobacter pylori*. *J Bacteriol* 2004, 186, 7926-7935.
- [38] Knuth, K., Niesalla, H., Hueck, C. J., Fuchs, T. M., Large-scale identification of essential *Salmonella* genes by trapping lethal insertions. *Mol Microbiol* 2004, 51, 1729-1744.
- [39] Ji, Y., Zhang, B., Van, S. F., Horn, *et al.*, Identification of critical staphylococcal genes using conditional phenotypes generated by antisense RNA. *Science* 2001, 293, 2266-2269.
- [40] Thanassi, J. A., Hartman-Neumann, S. L., Dougherty, T. J., Dougherty, B. A., Pucci, M. J., Identification of 113 conserved essential genes using a high-throughput gene disruption system in *Streptococcus pneumoniae*. *Nucleic Acids Res* 2002, 30, 3152-3162.
- [41] Jacobs, M. A., Alwood, A., Thaipisuttikul, I., Spencer, D., *et al.*, Comprehensive transposon mutant library of *Pseudomonas aeruginosa*. *Proc Natl Acad Sci U S A* 2003, 100, 14339-14344.
- [42] Kobayashi, K., Ehrlich, S. D., Albertini, A., Amati, G., *et al.*, Essential *Bacillus subtilis* genes. *Proc Natl Acad Sci U S A* 2003, 100, 4678-4683.
- [43] French, C. T., Lao, P., Loraine, A. E., Matthews, B. T., *et al.*, Large-scale transposon mutagenesis of *Mycoplasma pulmonis*. *Mol Microbiol* 2008, 69, 67-76.
- [44] Gallagher, L. A., Ramage, E., Jacobs, M. A., Kaul, R., *et al.*, A comprehensive transposon mutant library of *Francisella novicida*, a bioweapon surrogate. *Proc Natl Acad Sci U S A* 2007, 104, 1009-1014.
- [45] Gabaldon, T., Pereto, J., Montero, F., Gil, R., *et al.*, Structural analyses of a hypothetical minimal metabolism. *Philos Trans R Soc Lond B Biol Sci* 2007, 362, 1751-1762.
- [46] Reed, J. L., Vo, T. D., Schilling, C. H., Palsson, B. O., An expanded genome-scale model of *Escherichia coli* K-12 (iJR904 GSM/GPR). *Genome Biol* 2003, 4, 1-12.
- [47] Burgard, A. P., Vaidyaraman, S., Maranas, C. D., Minimal reaction sets for *Escherichia coli* metabolism under different growth requirements and uptake environments. *Biotechnology Progress* 2001, 17, 791-797.
- [48] Thiele, I., Jamshidi, N., Fleming, R. M., Palsson, B. O., Genome-scale reconstruction of *Escherichia coli*'s transcriptional and translational machinery: a knowledge base, its mathematical formulation, and its functional characterization. *PLoS Comput Biol* 2009, 5, e1000312.
- [49] Covert, M. W., Palsson, B. O., Transcriptional regulation in constraints-based metabolic models of *Escherichia coli*. *J Biol Chem* 2002, 277, 28058-28064.
- [50] Ussery, D. W., Leaner and meaner genomes in *Escherichia coli*. *Genome Biol* 2006, 7, 237.
- [51] Kolisnychenko, V., Plunkett, G., 3rd, Herring, C. D., Feher, T., *et al.*, Engineering a reduced *Escherichia coli* genome. *Genome Res* 2002, 12, 640-647.

- [52] Tanaka, K., Henry, C., Jolivet, E., Zinner, J. F., *et al.*, Model assisted design, implementation, and analysis of large scale deletions in *B. subtilis*. *In preparation* 2010.
- [53] Fabret, C., Ehrlich, S. D., Noirot, P., A new mutation delivery system for genome-scale approaches in *Bacillus subtilis*. *Mol Microbiol* 2002, 46, 25-36.
- [54] Lartigue, C., Vashee, S., Algire, M. A., Chuang, R. Y., *et al.*, Creating bacterial strains from genomes that have been cloned and engineered in yeast. *Science* 2009, 325, 1693-1696.
- [55] Benders, G. A., Noskov, V. N., Denisova, E. A., Lartigue, C., *et al.*, Cloning whole bacterial genomes in yeast. *Nucleic Acids Res*, 38, 2558-2569.

Figure Legends

Figure 1. Comparison of Minimal Gene Sets. Here we show the extent to which the Gill, Church and Koonin minimal gene sets overlap (a). We also show how the Gill (b), Church (c) and Koonin (d) sets each overlap with the *M. genitalium* genome and the 280 genes derived from the comparison of the available gene essentiality data.

Figure 2. Identification of Universal Essential Functions. Here we show how the number of universal functional roles conserved in 12, 11, 10..., 2, and 1 gene essentiality datasets (a). The large number of functional roles found in only one genome (1421) is likely due in part to poorly annotated genes or functional roles with inconsistent names in these genomes. We also determined the fraction of essential functional roles conserved in 12, 11, 10..., 2, and 1 datasets that overlap with the proposed minimal gene sets including (b): the *M. genitalium* genome (black); the essential genes in *M. genitalium* (red); a combination of the Koonin, Gill, and Church gene sets (blue), and the individual Koonin (purple), Gill (light blue), and Church (orange) gene sets.

The following government license should be removed before publication.

The submitted manuscript has been created by UChicago Argonne, LLC, Operator of Argonne National Laboratory ("Argonne"). Argonne, a U.S. Department of Energy Office of Science laboratory, is operated under Contract No. DE-AC02-06CH11357. The U.S. Government retains for itself, and others acting on its behalf, a paid-up

nonexclusive, irrevocable worldwide license in said article to reproduce, prepare derivative works, distribute copies to the public, and perform publicly and display publicly, by or on behalf of the Government.