Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in Biotechnology

Rapidly moving new bacteria to model-organism status

Hualan Liu and Adam M Deutschbauer

The paradigm of large research communities collectively working on a small number of model bacteria such as *Escherichia coli* and *Bacillus subtilis* is changing. While these classic model bacteria will continue to be important for advanced systems biology and new technology development, we envision that increasingly small research teams will be deeply investigating their own favorite strains, for example as new hosts for metabolic engineering or as key members of a complex microbiome. Given the lack of a research community and the sheer number of possible bacteria to interrogate, the development and application of technologies to rapidly and inexpensively advance these unstudied strains to 'model-organism' status is imperative. Here, we discuss the minimal information and tools necessary to develop a new model bacterium and how existing approaches can bring this power into the hands of a single investigator.

Address
Environmental Genomics and Systems Biology Division, Lawrence Berkeley National Laboratory, 1 Cyclotron Road, MS 977-250, Berkeley, CA 94720, United States

Corresponding author: Deutschbauer, Adam M (AMDeutschbauer@lbl.gov)

Current Opinion in Biotechnology 2018, **51**:116–122

This review comes from a themed issue on **Systems biology**

Edited by **Nathan Price** and **Eran Segal**

https://doi.org/10.1016/j.copbio.2017.12.006

Introduction

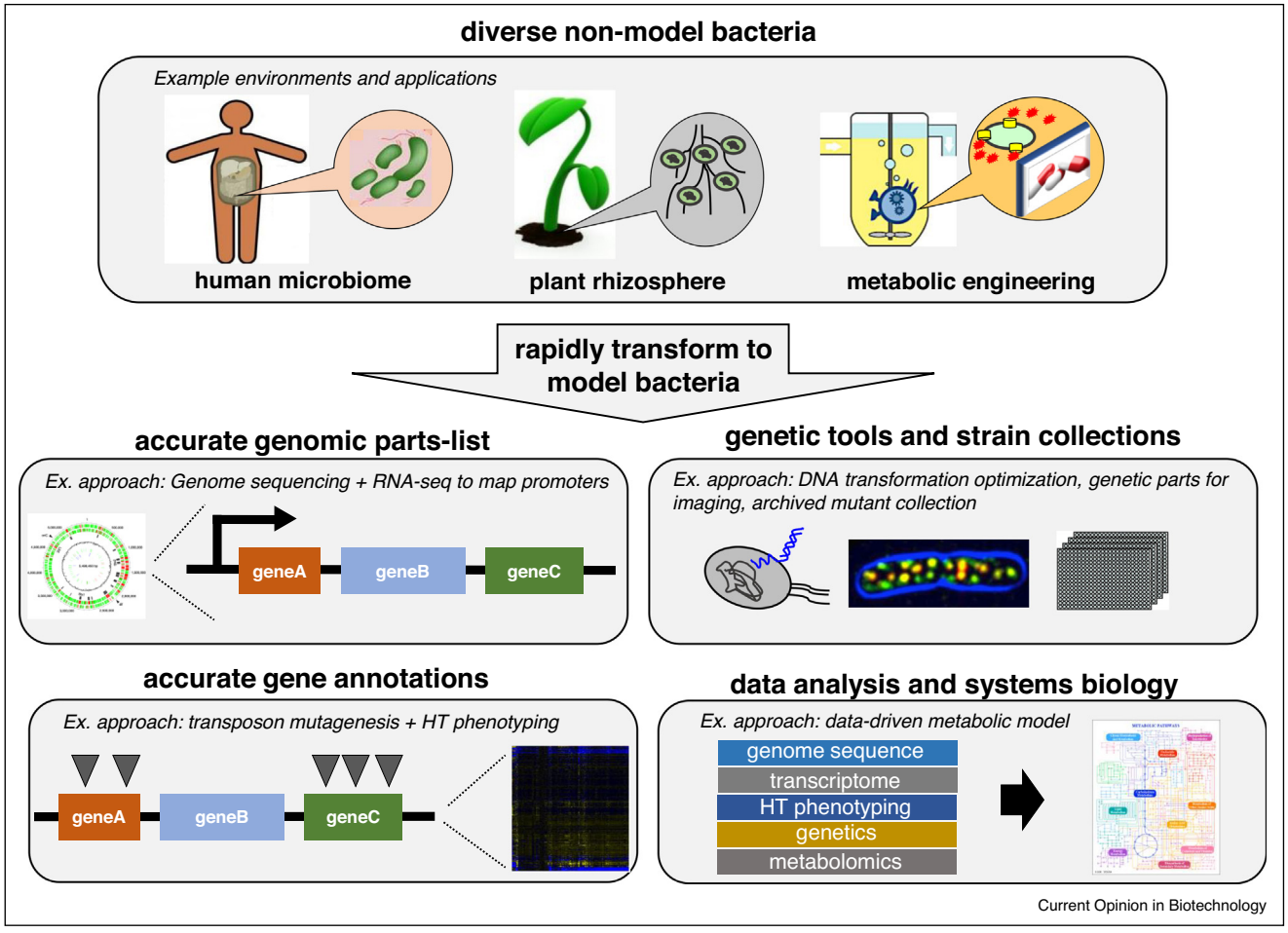
Microbiology-based solutions have been proposed for many of our most pressing planetary challenges including human health, sustainable agriculture, biomanufacturing, and environmental stewardship [1]. Meeting these challenges will require the molecular genetic investigation of a significant number of bacteria for which no literature exists, for example the newly isolated members of a complex microbiome [2] or the development a new autotrophic host for metabolic engineering of advanced chemicals [3]. The investigation of a new bacterium poses unique challenges, the most obvious being the lack of a research community with available genetic tools, strain collections, and a wealth of accumulated knowledge, such as those that have been developed in *E. coli* and *B. subtilis*. Therefore, translating the potential of these molecularly unstudied bacteria into applications will be hindered unless we accelerate the development of new model bacteria.

In this short review, we present some of the minimum criteria necessary to move a new bacterium to 'model-organism' status: (1) an accurate parts-list of genes, proteins, and promoters, (2) a genetic toolbox, (3) accurate, data-driven gene annotations, and (4) computational platforms for data integration and systems-level analyses (Figure 1). We describe existing technologies and resources available to meet each of these four criteria with a focus on approaches that can be applied by a single investigator rapidly and at a relatively low cost.

An accurate genomic parts-list

A complete and accurately assembled genome, with precise annotation of the structure and function of its features, is a crucial foundation for the study of any bacterium. The development of next-generation sequencing (NGS) technologies has made bacterial genome sequencing routine. Though these assemblies are often unclosed draft genomes, due to the short read lengths of the Illumina sequencing platform and repeat elements in bacterial genomes, third-generation sequencing technologies that produce longer read lengths, such as PacBio [4], should help to seal the gaps and generate complete genome sequences. This hybrid strategy combining both short-read and long-read sequencing has shown its power for high quality genome assembly [5]. Despite its higher error rate relative to Illumina sequencing, PacBio sequencing data has improved to the point that complete bacterial genome assemblies can be achieved using this platform alone [6], although the cost is still significantly higher than Illumina sequencing. An added benefit of the PacBio platform is the utility of these data to identify methylated bases and thus define all methylated motifs (the methylome) in the genome [7]. In a recent study, the methylome of 230 bacteria and archaea was determined using PacBio sequencing data, and revealed 834 different methylated motifs [8•]. As described below, there are downstream benefits to both a complete genome sequence (e.g., for normalization of transposon site sequencing data) and identification of methylated motifs (for designing strategies to increase transformation efficiency), therefore we recommend generating PacBio data for any bacteria being advanced to model-organism status.

Figure 1



Developing new model bacteria. High-throughput tools can be rapidly applied to transition diverse bacteria to new model systems. In the panels at the bottom, we highlight the criteria for a new model bacterium and provide example approaches that can be applied to meet each of these criteria.

There are a number of annotation pipelines for calling genes and assigning putative functions including the Integrated Microbial Genome (IMG) [9] and the Rapid Annotation using Subsystem Technology (RAST) systems [10]. Nevertheless, the accuracy of the gene calls and the protein function predictions are less than ideal, in particular for a strain under more in-depth investigation. Furthermore, these annotation servers do not typically identify gene structures beyond the coding sequence, leaving critical information such as promoters, transcriptional start sites (TSSs) antisense transcription, and small regulatory RNAs (sRNA) unknown. Now well-established transcriptomic and proteomic approaches can be applied at low cost to rapidly define the genomic parts-list of an organism. Differential RNA-seq (dRNA-seq) has become the preferred approach for precise mapping of 5′ TSSs in bacterial genomes as this approach can differentiate between primary and degraded/processed transcripts [11]. dRNA-seq requires the preparation and sequencing of two libraries, one treated with terminator exonuclease and one without, and has been used to identify thousands of TSSs in *E. coli* including over 5000 antisense transcripts [12]. Putative sRNAs can also be readily mapped using RNA-seq as illustrated in *Synechocystis* sp. PCC 6803 [13] and *Acinetobacter baumannii* [14].

Another established approach, termed proteogenomics, uses mass spectrometry-based proteomic data, to search against pre-existing protein databases, to identify novel proteins missed by the primary annotation and to modify gene models [15]. Application of this approach to *Mycobacterium smegmatis* validated hundreds of predicted gene models and identified 63 new proteins [16]. Proteogenomics can be coupled to transcriptomics to further refine gene models. For example, in the sulfate-reducing bacterium *Desulfovibrio vulgaris*, a combination of proteomics, tiling microarrays, and RNA-seq was used to identify over 1000 transcriptional start sites and revise 505 protein