ACS Synthetic Biology
Cite This: ACS Synth. Biol. 2019, 8, 158−169
Research Article
pubs.acs.org/synthbio

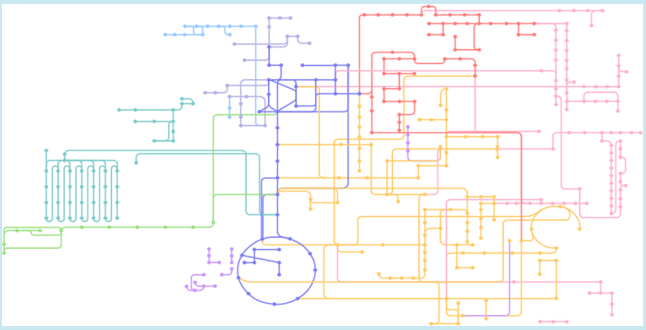
# Tn-Core: A Toolbox for Integrating Tn-seq Gene Essentiality Data and Constraint-Based Metabolic Modeling

George C. diCenzo,* Alessio Mengoni, and Marco Fondi*

Department of Biology, University of Florence, Sesto Fiorentino, Florence, 50019, Italy

Supporting Information

**ABSTRACT:** The design of synthetic cells requires a detailed understanding of the relevance of genes and gene networks underlying complex cellular phenotypes. Transposon-sequencing (Tn-seq) and constraint-based metabolic modeling can be used to probe the core genetic and metabolic networks underlying a biological process. Integrating these highly complementary experimental and *in silico* approaches has the potential to yield a highly comprehensive understanding of the core networks of a cell. Specifically, it can facilitate the interpretation of Tn-seq data sets and identify gaps in the data that could hinder the engineering of the cellular system, while also providing refined models for the accurate predictions of cellular metabolism. Here, we present Tn-Core, the first easy-to-use computational pipeline specifically designed for integrating Tn-seq data with metabolic modeling, prepared for use by both experimental and computational biologists. Tn-Core is a MATLAB toolbox that contains several custom functions, and it is built upon existing functions within the COBRA Toolbox and the TIGER Toolbox. Tn-Core takes as input a genome-scale metabolic model, Tn-seq data, and optionally RNA-seq data, and returns: (i) a context-specific core metabolic model; (ii) an evaluation of redundancies within core metabolic pathways, and optionally (iii) a refined genome-scale metabolic model. A simple, user-friendly workflow, requiring limited knowledge of metabolic modeling, is provided that allows users to run the analyses and export the data as easy-to-explore files of value to both experimental and computational biologists. We demonstrate the utility of Tn-Core using *Sinorhizobium meliloti*, *Pseudomonas aeruginosa*, and *Rhodobacter sphaeroides* genome-scale metabolic reconstructions as case studies.



**KEYWORDS:** *transposon sequencing, metabolic network reconstruction, COBRA models, systems biology, bacterial metabolism*

The use of synthetic biology to engineer cell factories through genome-wide modifications will require systems biology approaches that allow for an understanding of all cellular components and regulatory elements underlying complex biological phenotypes.[1,2] An emerging tool within this area is transposon-sequencing (Tn-seq), and related variants such as INSeq, TraDIS, and HITS.[3,4] These techniques allow for genome-wide screens of the fitness contribution of all genes to growth in a defined environment. After preparing a library of hundreds of thousands of mutants each containing a random transposon insertion, the library is pooled and passed through a selective environment. Next-generation sequencing is used to identify the location of all transposon insertions in the output cell population, and the number of insertions per gene is used as a measure of the fitness consequence of mutating the gene; insertions decreasing fitness will be lost from the population and therefore under-represented in the sequencing data. Interestingly, Tn-seq and RNA-sequencing (RNA-seq) data sets do not necessarily correspond.[5] This demonstrates that gene fitness and expression are not inherently linked and highlights how Tn-seq and RNA-seq experiments provide distinct but complementary types of information.

The results of Tn-seq studies undoubtedly provide invaluable insights into the metabolism and biology underlying a phenotype of interest, and they can serve as a basis for genome engineering. For example, Tn-seq has been used to identify genes contributing to antibiotic or phage resistance,[6,7] virulence factors,[8,9] establishment of a symbiotic interaction,[10] and motility,[11] among others. However, epistatic interactions, particularly functional redundancy, prevents Tn-seq from providing a comprehensive identification of the genes relevant to the studied phenotype. This was highlighted by the results of recent studies demonstrating that ∼10% of chromosomal genes in *Sinorhizobium meliloti* display a fitness phenotype dependent on the presence/absence of the two large secondary replicons.[12,13] Similarly, multiple studies have observed that orthologous genes of closely related species may display unique fitness patterns,[12,14,15] likely due in part to species-specific genetic interactions. Potentially pervasive genetic interactions masking phenotypes within bacterial genomes has also been demonstrated by several other studies.[8,16−18]

**Received:** October 18, 2018
**Published:** December 7, 2018






DOI: 10.1021/acssynbio.8b00432
ACS Synth. Biol. 2019, 8, 158−169

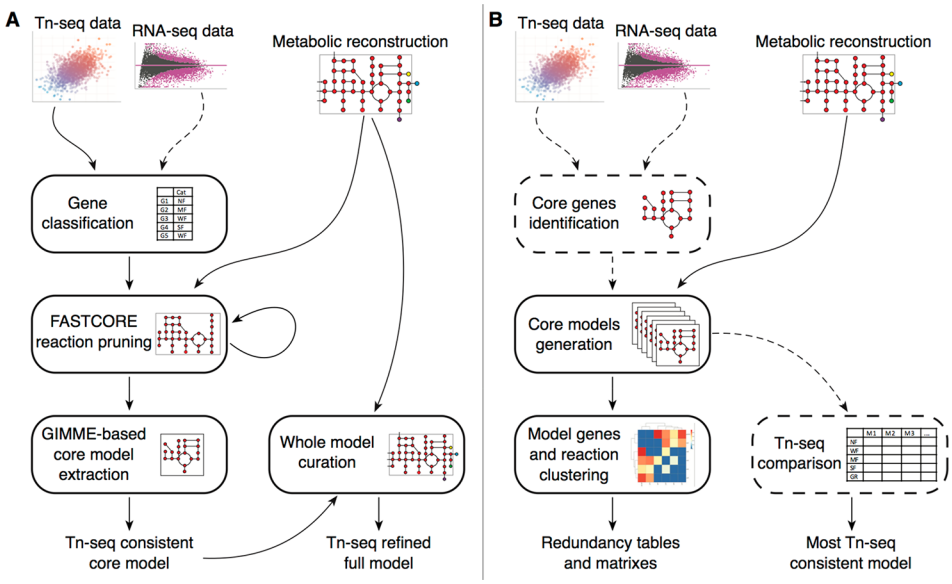



**Figure 1.** Schematic representation of the main functionalities of Tn-Core. In both panels, dashed lines indicate optional steps. (A) The main workflow for extraction of a Tn-seq-consistent core metabolic model, as well as its use in refining a genome-scale metabolic model. (B) The main workflow for evaluation of redundancy within core metabolic pathways.

The resulting knowledge gaps can hinder synthetic biology attempts at leveraging the data from Tn-seq studies in synthetic biology and biotechnology applications.[14,19] It is therefore imperative to develop methods to overcome this limitation. With respect to metabolic pathways, genome-scale metabolic network reconstruction (GSMR)[20] and constraint-based metabolic modeling (CBMM)[21] are ideally suited to fulfill this role. GSMRs are *in silico* representations of the entire metabolic potential of the cell, with each reaction linked to the corresponding enzyme(s) and its gene(s), providing a link between genetics and metabolism.[22] CBMM using optimality based approaches such as Flux Balance Analysis (FBA)[23] can then be used to predict the metabolic flux distribution through the GSMR, and comprehensively predict the phenotypes of single, double, or higher-order gene mutations or reaction perturbations.[24] Nowadays, GSMRs for a large number of species have been prepared; the BiGG database contains 85 publicly available models,[25] and many more are available elsewhere.

Over the past few years, an increasing number of studies have begun to employ both Tn-seq and metabolic modeling due to their highly complementary nature.[10,24−29] These have generally used the Tn-seq data in manual refinement of the gene-reaction associations of the GSMR,[26−28] or to compare essential gene predictions by these two approaches.[29,30] Intriguingly, recent work demonstrated that a manual Tn-seq-guided reconstruction of a core metabolic network can help illuminate and fill-in the gaps present in the Tn-seq data,[12] producing a more comprehensive understanding of core cellular metabolism than is possible with either approach alone. Several algorithms exist for the automated integration of -omics data with GSMRs with the goal of extracting a smaller, context-specific metabolic model. This most commonly involves integrating gene expression data, constraining the allowable flux across each reaction based on the expression level(s) of the corresponding gene(s).[31,32] Similar approaches exist for combining GSMRs with proteomics,[33] fluxomics,[34] and metabolomics data.[35] However, we are unaware of any algorithms or pipelines specifically developed for the automated integration of Tn-seq data with GSMRs.

Here, we report Tn-Core, a MATLAB toolbox for integration of Tn-seq (and optionally RNA-seq) data sets with GSMRs. Development of Tn-Core was motivated by a desire to produce an easy-to-use pipeline for the automated integration of Tn-seq data with metabolic models that requires limited knowledge of metabolic modeling. The Tn-Core Toolbox contains functions to (i) generate context-specific core metabolic models through integration of Tn-seq data using a gene-centric approach; (ii) evaluate redundancy within core metabolic pathways; and (iii) perform Tn-seq-guided refinement of the gene-protein-reaction (GPR) rules in a GSMR. The outputs of the pipeline can serve as a framework for interpretation of Tn-seq data, can be used to identify gaps in Tn-seq data sets, can provide a basis for metabolic engineering of a cell, and can be used in downstream metabolic modeling applications.

## RESULTS AND DISCUSSION

**Implementation of the Tn-Core Toolbox.** The current version of the Tn-Core Toolbox consists of 14 functions. The toolbox is written in MATLAB code, and is built upon existing functions within the COBRA and TIGER Toolboxes.[36,37] Tn-Core, together with a detailed manual, is available through GitHub (github.com/diCenzo-GC/Tn-Core), and future releases will be available through the same link. The functionality of the entire toolbox has been validated on two machines running different versions of MATLAB (R2015b, R2017a) and distinct COBRA toolbox setups (openCOBRA downloaded 05/2017 and 08/2017). We therefore expect that Tn-Core should work in a broad range of computing environments.

Tn-Core was developed to facilitate the integration of Tn-seq data with metabolic models, with the aim to assist in the interpretation of Tn-seq data sets, as well as to generate context-specific core metabolic models for further computational analyses. The toolbox can be largely divided into three main modules (Figure 1): (i) development of context-specific core metabolic models consistent with experimental gene fitness


DOI: 10.1021/acssynbio.8b00432
*ACS Synth. Biol.* 2019, 8, 158−169

data; (ii) evaluation of genetic and metabolic redundancy within core metabolic pathways; and (iii) refinement of genome-scale metabolic models based on gene fitness data. Sample outputs from each of the three main functions of Tn-Core are provided in Data sets S1−S3 to highlight the types of data obtained. A straightforward and customizable overall workflow function for running Tn-Core from data import to data export is included in the toolbox. This overall workflow is designed to require limited knowledge of metabolic modeling procedures; nevertheless, the user must be able to identify the exchange reactions to set the simulated medium and to identify the objective function.[22,36] We are also developing an online Web server implementing Tn-Core (at combo.dbe.unifi.it/tncore) that will allow running the tool as a web application

*Generation of Core Metabolic Models.* Several algorithms exist for the evaluation of context-specific metabolism through the integration of expression (microarray) data with GSMRs.[32,38] One of the most commonly used approaches is GIMME.[39] In the GIMME algorithm, genes expressed above a specified threshold are turned on, and the rest of the genes are turned off. If the model does not carry flux through the objective function (i.e., the model cannot produce biomass), a minimal number of genes initially turned off are reactivated based on minimizing the inconsistency; put simply, higher expressed genes are preferred over lower expressed genes. We wished to employ a GIMME-based approach for extraction of core metabolic models on the basis of Tn-seq data. However, as Tn-seq fitness data is fundamentally different than expression data, it is not possibly to simply run GIMME substituting gene fitness data for the expression data. There are at least two related features of GIMME (illustrated in Figure 2) that, while appropriate for expression data, are not valid for fitness data: (i) a short pathway or small protein complex encoded by genes with a weak fitness contribution may be preferred over a long pathway or large protein complex involving genes with a moderate fitness contribution or a mix of genes with strong and weak fitness contributions, and (ii) a pathway with all genes having moderate fitness contribution may be preferred over a pathway with a mix of genes with strong and weak fitness contribution. To overcome these issues, we developed a pipeline built upon the iterative use of FASTCORE[40] followed by GIMME. FASTCORE is an algorithm that extracts core models from a GSMR on the basis of a user-provided set of core reactions. Given a GSMR and a reaction list, FASTCORE returns a reaction list that maximizes the number of reactions from the list that can carry flux, while minimizing the number of additional reactions that must be added.[40]

The pipeline is summarized in Figure 1A and in the pseudocode of Algorithm S1. Additional details of the underlying logic are provided in the Supplementary Text. The minimum input required by Tn-Core is (i) a GSMR, and (ii) Tn-seq data for each gene, where a lower number reflects an increased fitness contribution (e.g., the number of insertions per gene, normalized by gene length). The pipeline can optionally consider RNA-seq data, as RPKM or TPM values, in addition to the Tn-seq data. Genes embedded within the model are categorized into four fitness categories (strong, moderate, weak, and no fitness contribution) based on the Tn-seq data. When RNA-seq data are provided, the no fitness contribution category is split into two groups: highly expressed, and not highly expressed. FASTCORE is then iteratively run, forcing optimal flux through the objective reaction (i.e., maximal biomass production), while successively removing reactions associated

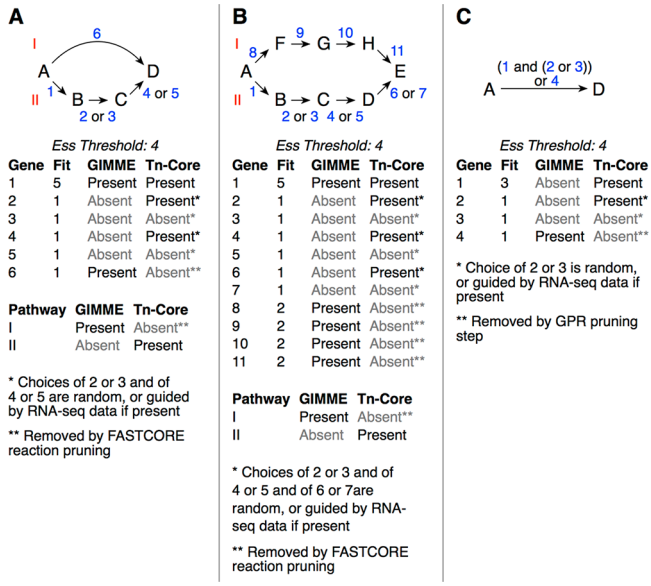



**Figure 2.** Toy pathways demonstrating key characteristics of Tn-Core during core model extraction. In each panel, metabolites are represented by the black capital letters, genes are represented by the blue numbers, and pathways are indicated by the red roman numerals. The *Ess Threshold* indicates the threshold for a gene to be essential (in these examples, a higher value indicates the gene has a greater contribution to fitness). The tables below the schematics lists the Tn-seq determined fitness value for each gene (column *Fit*), and the expected outcome when running GIMME or Tn-Core (Present, gene remains in the core model; Absent, gene is not in the core model). (A) Demonstration of how Tn-Core would favor the inclusion of a long pathway with a mix of essential and nonessential genes over the inclusion of a short, nonessential pathway. (B) Demonstration of how Tn-Core would favor the inclusion of a pathway with a mix of essential and nonessential genes over the inclusion of an equal length pathway with genes consistently categorized as having a weak or moderate fitness contribution. (C) Demonstration of how Tn-Core would favor inclusion of a multiprotein complex with a mix of genes with strong and no fitness contributions over the inclusion of an alternative protein that is nonessential.

with genes having no, weak, and moderate fitness contribution that are both (i) not required for model growth, and (ii) not required to complete a pathway associated with at least one gene above the current fitness category of interest. Once the FASTCORE-based reaction pruning is complete, gene associations are modified to remove protein complexes if an alternative complex contains at least one gene in a higher fitness category. Next, the gene-centric TIGER version of GIMME[37,39] is used with binned fitness values (Figure S1) to produce a list of genes to be included in the final core model. As the genes identified as active in GIMME are not always sufficient to produce a COBRA-formatted model capable of carrying flux through the objective function (i.e., unable to grow), a minimal set of additional genes are included in the final model, with genes with higher fitness contributions favored. At the same time, all highly expressed genes in the RNA-seq data, if provided, can be optionally reintroduced into the core model. As illustrated schematically in Figure 2, the Tn-Core workflow is capable of better selecting the pathways for inclusion in the core model on the basis of gene fitness data.

*Evaluation of Redundancy.* The function described above extracts a core metabolic model from a GSMR. This can assist in the interpretation and understanding of the core metabolism


DOI: 10.1021/acssynbio.8b00432
*ACS Synth. Biol.* 2019, 8, 158−169

underlying a specific growth condition. It does not, however, allow for an analysis of redundancy in the core metabolic network.[12] Therefore, a "brute force" implementation of GIMME was prepared that allows for analysis of gene-level and reaction-level redundancies and modules in core metabolism. The overall pipeline for analysis of redundancy is summarized in the pseudocode provided in Algorithm S2 and graphically in Figure 1B, and additional details of the underlying logic are provided in the Supplementary Text. The minimum input is a COBRA-formatted metabolic model; optionally, Tn-seq data and/or RNA-seq data can be provided. A list of genes to be "protected" during core model generation is prepared based on essential genes in the Tn-seq data and/or highly expressed genes in the RNA-seq data, plus all genes essential for sustaining growth (i.e., biomass production) of the input GSMR. A user-defined number of randomized core models are then generated as follows. For each iteration, the "non-protected" gene list is randomly shuffled. Starting from the top of the list, the genes are successively deleted from the input GSMR. If the flux through the objective function stays above the threshold, the gene is excluded from the model; otherwise, the gene is put back into the model to maintain flux through the objective function above the threshold. The order in which genes are deleted from a model can alter the content of the final model. For example, if Gene A and Gene B redundantly encode the same essential protein product, only the gene lower on the list will remain in the output model. As such, the end result of this process is a population of core metabolic models, each containing the initially protected genes and a minimal set of additional genes necessary to maintain objective function flux above the threshold. As the order genes are deleted from the model is shuffled at each iteration, each core model could potentially be unique. Additionally, if Tn-seq data are provided, the core model most consistent with the gene fitness data is recorded as described in Algorithm S2.

Functions are provided to explore the gene and reaction differences in the core metabolic models produced through the above-described process as a way of summarizing and identifying functional redundancies and functional modules. A binary presence/absence matrix is given, which indicates, for each model, which features are present or absent. A co-occurrence matrix is also provided; for each feature variably present in the core model population, a Chi-squared statistic is reported to indicate the feature pairs that are more likely than chance to appear, or not appear, in the same core models. Additionally, an easily searchable co-occurrence table is provided, indicating the prevalence of each gene and each gene-pair in the core model population, and the expected prevalence of each gene-pair. Finally, if core models are generated multiple independent times, for example, using different objective flux thresholds, a matrix can be produced comparing the percentage of models in each core model population that contains each feature (Data set S2).

*Refinement of GPRs in Genome-Scale Metabolic Models.* The preparation of high-quality GSMRs is a time and effort intensive process. Several automated metabolic network reconstruction pipelines are now available. However, they rely on automated genome annotations that can be quite error prone,[41] meaning that intensive manual refinement of the automated reconstructions remains necessary in order to correct errors.[22] One of the errors in automated reconstructions is the incorrect assignment of multiple genes to the same core metabolic reaction. In the absence of experimental data, it can be difficult to correct such errors; however, Tn-seq data is ideally suited to help guide the correction of these errors.[11,26,27,29] We therefore provide an extension of the main functionality of Tn-Core to assist in the automated curation of GSMRs using Tn-seq data.

In this process (summarized in Algorithm S3, Figure 1A, and the Supplementary Text), a core metabolic model is first extracted from the GSMR on the basis of the provided Tn-seq data. Next, for any core model reaction with a gene classified as essential based on the Tn-seq data, the GPR rules of the corresponding reaction in the original GSMR are replaced with those of the core reconstruction. The function then returns this refined GSMR, as well as a report indicating the changes that were made (see Data set S3 for example output). Tn-Core will also, in some cases (see Supplementary Text) suggest that certain GPR rules be modified to replace "or" statement with "and" statements. These suggestions are listed in the report, but they are not incorporated in the refined model by Tn-Core.

**Validation of Tn-Core Functionality.** *Analysis of the Accuracy of Tn-Core Using a S. meliloti GSMR.* To validate the accuracy and utility of core model extraction by Tn-Core, the toolbox was tested with a modified version (see Methods) of the previously prepared iGD1575 GSMR of *S. meliloti*,[42] and published Tn-seq[12] and RNA-seq[43] data sets for *S. meliloti* grown in minimal medium with sucrose as the carbon source. An advantage of working with *S. meliloti* for validation of Tn-Core is the existence of a core metabolic reconstruction, termed iGD726,[12] that was manually prepared through a Tn-seq-guided reconstruction process. The iGD726 reconstruction therefore serves as a positive control, roughly representing the expected output of Tn-Core; however, differences between iGD726 and iGD1575, including in the biomass composition, means that a core model extracted from iGD1575 will never fully resemble iGD726.

Three core models were extracted from iGD1575 using the three main implementations of Tn-Core: (i) using only Tn-seq data; (ii) using Tn-seq and RNA-seq data; and (iii) using Tn-seq and RNA-seq data, with highly expressed genes added back to the core model following core model generation. The main features of each core model are summarized in Table 1, and the

**Table 1. Summary Statistics of the *S. meliloti* GSMR and Various Core Models**

| model[a] | reactions | gene associated reactions | metabolites | genes | essential genes[b] |
|---|---|---|---|---|---|
| iGD1575 | 1828 | 1411 | 1579 | 1575 | 223 |
| iGD726 | 681 | 632 | 703 | 726 | 356 |
| Tn-Core -RNA | 530 | 464 | 524 | 486 | 309 |
| Tn-Core + RNA 1 | 532 | 466 | 527 | 493 | 309 |
| Tn-Core + RNA 2 | 550 | 484 | 539 | 517 | 298 |
| GIMME Tn-seq | 637 | 507 | 613 | 483 | 289 |
| FASTCORE | 556 | 488 | 544 | 732 | 246 |
| minNW | 501 | 430 | 525 | 695 | 256 |
| GIMME RNA-seq | 646 | 516 | 615 | 513 | 288 |

[a]A description of each model is provided in the caption of Figure 3. [b]The number of genes essential in the Tn-seq data set that are also essential in the metabolic model.



DOI: 10.1021/acssynbio.8b00432
ACS Synth. Biol. 2019, 8, 158−169

core models are provided as COBRA formatted models and Excel worksheets in Data set S1 to provide examples of the Tn-Core output. Inclusion of the RNA-seq data (without adding back highly expressed genes) had only a minor effect on the size of the core model compared to using just Tn-seq data. This was expected as RNA-seq data serves to assist in the selection of gene(s) when alternatives exist but does not directly add genes to the model. However, the RNA-seq data influenced the composition of the core model, although the majority of the core model content remained consistent; approximately 93% of the genes were common to both models, while approximately 97% of the reactions were conserved. An example difference is *tpiA* and *tpiB*, which both encode triose phosphate isomerases. These genes are functionally redundant, at least for growth with some carbon sources.[44] The core model produced without RNA-seq data included *tpiB*, whereas the model with RNA-seq data instead included *tpiA*; expression of *tpiA* was ~7.5-fold higher than that of *tpiB* in the RNA-seq data set. When Tn-Core was run with highly expressed genes added back into the model following preparation of the core model, a slightly larger model was produced than during the other two implementations of Tn-Core.

The ability of the core models produced by Tn-Core to capture context-specific core metabolism was examined by predicting the fitness contribution of central carbon metabolic genes (Figure 3). All tested genes were predicted to be nonessential in iGD1575, presumably due to network redundancy. In contrast, all but two of the genes are predicted

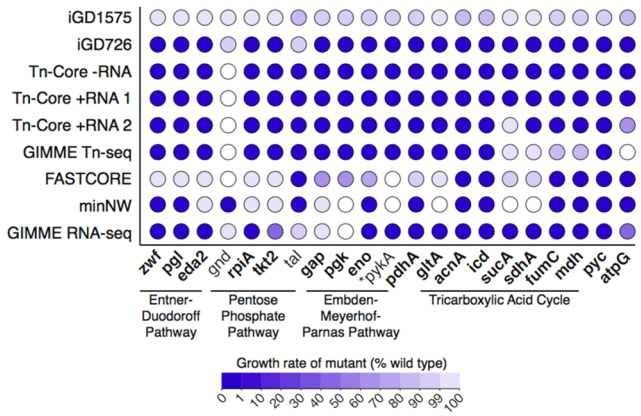



**Figure 3.** Comparison of predicted fitness scores for central carbon and energy metabolic genes in the *S. meliloti* models. Representative genes from central carbon and energy metabolism are shown; those in bold are essential based on the Tn-seq data. For each gene, an empty (white) circle is shown if the gene is absent from the model; otherwise, the circle is colored according to the effect of deleting the gene on the growth rate of the model (determined using the MOMA algorithm). A value of 100 means no growth impairment (i.e., growth at 100% the rate of wild-type), whereas a value of 0 means the gene deletion was lethal. The models included in the figure are as follows: the full *S. meliloti* genome-scale metabolic reconstruction (iGD1575); the manually produced core metabolic reconstruction (iGD726); three models produced from iGD1575 using Tn-Core (without RNA-seq [Tn-Core -RNA], with RNA-seq [Tn-Core + RNA 1], and with RNA-seq plus reintroduction of highly expressed genes [Tn-Core + RNA 2]); core models derived from iGD1575 using Tn-seq data and pipelines based on GIMME, FASTCORE, or minNW; and a core model derived from iGD1575 using RNA-seq data and a pipeline based on GIMME (*). Although the gene *pykA* was not essential, the Tn-seq data suggested that it is highly important for growth.

to be essential in the manually prepared core model iGD726, with *gnd* and *tal* correctly predicted as nonessential.[12,45] The gene fitness patterns for the two core models produced by Tn-Core with or without RNA-seq data (without adding back highly expressed genes) were the same. Both models correctly predicted 18 of 18 essential genes. The gene *gnd* was absent from both models, and therefore correctly identified as nonessential. However, like iGD726, *pykA* was essential in both core models unlike the Tn-seq data where *pykA* was nonessential but with an obvious fitness contribution. The sole difference compared to iGD726 was that *tal* was predicted as essential in the core models. As described later, this is related to the core nature of the Tn-Core generated models; when multiple, alternative pathways exist, only one will be included in the output, resulting in that pathway appearing essential. Overall, these results confirm that Tn-Core is able to accurately, and rapidly, extract a core metabolic model from an input GSMR that is highly consistent with experimental Tn-seq derived fitness data.

When Tn-Core was run with the option to reintroduce highly expressed genes into the model following the preparation of the core model, two genes became nonessential: *sucA* in the Kreb's cycle and *atpG* in the ATP synthase complex of the electron transport chain (Figure 3). We therefore generally suggest against the use of this option; however, there may be times when this option produces information of interest to users, for example, in helping identify putative redundancies.

*Comparison of S. meliloti Core Model Extraction by Tn-Core to Other Approaches.* There is currently no tool explicitly comparable to Tn-Core as none considers experimental Tn-seq data during core model generation. Nevertheless, to evaluate whether Tn-Core performs a function that is currently lacking among the available tools, we compared the effectiveness of Tn-Core to that of pipelines using existing algorithms with simple modifications to deal with Tn-seq data. These pipelines used either GIMME,[37,39] FASTCORE,[40] or minNW,[46] as described in the Methods. The GIMME pipeline is gene-centric (i.e., adds/removes genes, consequently influencing reaction content), and takes as input appropriately modified Tn-seq data. In contrast, the FASTCORE and minNW pipelines are reaction-centric (i.e., directly adds/removes reactions), and require as input a set of reactions, not genes, to be protected during core model generation. To adapt these latter two algorithms for use with Tn-seq data, the protected reactions were set as those reactions that were constrained upon deletion of the genes essential in the Tn-seq data set.

The main properties of the models are summarized in Table 1. Compared to the core model produced with Tn-Core using only Tn-seq data, the core model generated with the GIMME algorithm contained ~20% more reactions but a similar gene count. In contrast, the models produced by FASTCORE and minNW were similar in reaction count to the Tn-Core produced model (~5% more or less reactions, respectively), but were larger in terms of gene count (43% to 50% more). The higher gene count for the core models produced with FASTCORE and minNW is presumably due to the reaction-centric nature of the algorithms, and consequently, the lack of refinement of the genes associated with the included reactions. Additionally, the core models produced by these three pipelines embeded between 89% and 96% of the genes, and between 83% and 97% of the reactions, present in the Tn-Core produced model.

More importantly than the model size and content is how accurate the models capture the core metabolism as represented


DOI: 10.1021/acssynbio.8b00432
*ACS Synth. Biol.* 2019, 8, 158−169

in the Tn-seq data. In this respect, Tn-Core clearly outperformed all three of the tested pipelines when using central carbon and energy metabolism as a proxy. Whereas all 18 of the essential genes in these pathways were correctly predicted as essential by Tn-Core, 5, 12, and 8 of these genes were incorrectly predicted as nonessential by the GIMME, FASTCORE, and minNW based pipelines (Figure 3). Additionally, 309 genes essential in the Tn-seq data set were also essential in the Tn-Core derived core model, compared to 223 in the initial GSMR (Table 1). In contrast, only 289, 246, and 256 of the Tn-seq essential genes were essential in the core models derived by the GIMME, FASTCORE, and minNW pipelines, respectively (Table 1). These results confirm that Tn-Core fulfills a function that is currently lacking among the available algorithms.

We further compared the results of Tn-Core with a context-specific model produced by GIMME using only RNA-seq data, the original purpose of GIMME. As expected given the different input data, the resulting context-specific model differed from the models produced by Tn-Core and by GIMME using Tn-seq data (Table 1 and Figure 3). In terms of the fitness predictions for central carbon and energy metabolic genes, the model produced on the basis of the RNA-seq data was comparable to the model produced by GIMME using Tn-seq data, but less accurate than the models produced by Tn-Core (Figure 3). Additionally, only 288 Tn-seq essential genes were essential in the core model produced by GIMME with RNA-seq data, compared to the 309 that are present in the Tn-Core derived model (Table 1). These results demonstrate that the use of Tn-seq data or RNA-seq data during extraction of a core metabolic model may influence the final model content. These results further suggest that, when available, the use of Tn-seq data may lead to a core model that better represents the true core metabolism.

*Validation of Tn-Core Using P. aeruginosa and R. sphaeroides GSMRs.* To confirm that the functionality of Tn-Core is not limited to iGD1575, the toolbox was further validated using GSMRs of *Pseudomonas aeruginosa* and *Rhodobacter sphaeroides*. For *P. aeruginosa*, the iPae1146 GSMR[27] was used with published Tn-seq[15] and RNA-seq[47] data for *P. aeruginosa* PAO1 grown in a minimal medium with succinate as the carbon source. Tn-Core successfully produced working core metabolic models from iPae1146 that consisted of 261 genes and 390 reactions. There were 151−152 genes essential in the Tn-seq data set that were also essential in the core models, compared to 122 genes in the input GSMR (Table 2). As with iGD1575, running Tn-Core with just Tn-seq data, or with both Tn-seq and RNA-seq data sets, had little effect (Table 2), with ∼95% of the genes and 99% of the reactions common to both core models. In central carbon metabolism, seven of the eight examined genes were correctly predicted to be essential in the Tn-Core-derived models (Figure 4). The only gene incorrectly predicted to be nonessential was *gap*, which, as described later, was due to an error in the input GSMR. Additionally, the Tn-Core-derived core model predicted two genes (*tal*, *pckA*) to be essential that were not essential in the Tn-seq data set (Figure 4). We cannot determine if these represent false positives or gaps in the Tn-seq data; however, in support of the latter option, we note that both genes were essential for the growth of *R. sphaeroides* with succinate as a carbon source.[28] Finally, Tn-Core performed at least as good as the other pipelines that were tested using Tn-seq data. The core models produced with the other three pipelines contained between 10 and 127 more genes than the Tn-Core derived models, while between 132 and 151 genes were correctly predicted to be essential (Table 2 and Figure 4). Overall, the data are consistent with Tn-Core producing a core model that accurately captures the core metabolism of *P. aeruginosa* in the tested growth environment.

GIMME was used to extract a context-specific model of iPae1146 on the basis of RNA-seq data without considering Tn-seq data. The resulting model was fairly different than the models produced by Tn-Core; although 92% of reactions in the Tn-Core-derived models were present in the GIMME-derived model, only 81% of genes were shared. Nevertheless, at a global level, the core model produced by GIMME did a good job of capturing the Tn-seq fitness data, although perhaps not quite as good as Tn-Core (Table 1 and Figure 4). There were 142 genes correctly predicted to be essential in the core model produced by

**Table 2. Summary Statistics of the *P. aeruginosa* GSMR and Various Core Models**

| model[a] | reactions | gene associated reactions | metabolites | genes | essential genes[b] |
|---|---|---|---|---|---|
| iPae1146 | 1496 | 1271 | 1284 | 1146 | 122 |
| Tn-Core -RNA | 390 | 363 | 382 | 261 | 151 |
| Tn-Core + RNA | 390 | 363 | 382 | 264 | 152 |
| GIMME Tn-seq | 549 | 428 | 543 | 271 | 151 |
| FASTCORE | 429 | 400 | 421 | 388 | 138 |
| minNW | 353 | 328 | 361 | 344 | 132 |
| GIMME RNA-seq | 576 | 455 | 565 | 289 | 142 |

[a]A description of each model is provided in the caption of Figure 4. [b]The number of genes essential in the Tn-seq data set that are also essential in the metabolic model.

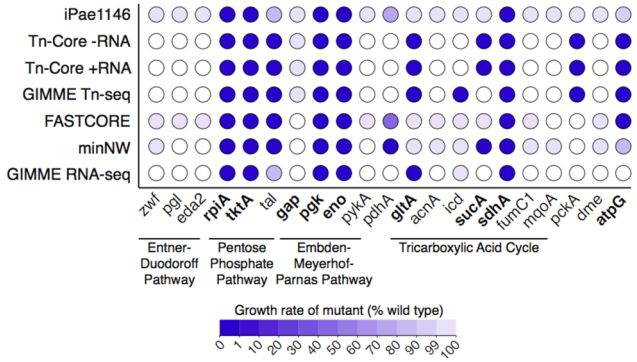



**Figure 4.** Comparison of predicted fitness scores for central carbon and energy metabolic genes in the *P. aeruginosa* models. Representative genes from central carbon and energy metabolism are shown; those in bold are essential based on the Tn-seq data. For each gene, an empty (white) circle is shown if the gene is absent from the model; otherwise, the circle is colored according to the effect of deleting the gene on the growth rate of the model (determined using the MOMA algorithm). A value of 100 means no growth impairment (i.e., growth at 100% the rate of wild-type), whereas a value of 0 means the gene deletion was lethal. The models included in the figure are as follows: the full *P. aeruginosa* genome-scale metabolic reconstruction (iPae1146); two models produced from iPae1146 using Tn-Core (without RNA-seq [Tn-Core -RNA], with RNA-seq [Tn-Core + RNA]), core models derived from iPae1146 using Tn-seq data and pipelines based on GIMME, FASTCORE, and minNW; and a core model derived from iPae1146 using RNA-seq data and a pipeline based on GIMME.



DOI: 10.1021/acssynbio.8b00432
*ACS Synth. Biol.* 2019, 8, 158−169

GIMME, compared to 151 or 152 in the core models produced by Tn-Core (Table 1). This difference included two genes (*sucA*, *atpG*) in central carbon and energy metabolism that were correctly predicted as essential in the Tn-Core-derived models but not in the GIMME-derived model (Figure 4).

Tn-Core was also effective at producing a core metabolic model from the *Rhodobacter sphaeroides* iRsp1140 GSMR,[48] using Tn-seq data[28] for *R. sphaeroides* 2.4.1 grown in a minimal medium with succinate as the carbon source. The Tn-Core-derived core model well-represented the gene fitness patterns of the examined central carbon metabolic genes, and it did so better than the core models prepared from the other tested pipelines (Figure S2). Additionally, a greater number of the Tn-seq essential genes were essential in the core model produced using Tn-Core (202 genes) than in the original GSMR (159 genes) or the core models produced with the other pipelines (167−192 genes; Table S1). Further discussion of these results is available in the Supplementary Text.

Overall, the above simulations provide strong support that Tn-Core can be applied with a broad range of GSMRs to extract core metabolic models consistent with Tn-seq-derived fitness scores. The analyses also indicate that context-specific models derived from Tn-seq or RNA-seq data differ, highlighting how integrating these distinct data types with GSMRs may provide unique insights into the metabolism of an organism.

*Limitations of Tn-Core for Extraction of Core Metabolic Models.* While the above sections demonstrate the utility of Tn-Core to extract core metabolic models from GSMRs on the basis of gene fitness data, it is important to recognize the limitations of this tool. Specifically, the quality of a core model is intrinsically linked to the quality and completeness in the input GSMR. This can be demonstrated by two results that were observed when running Tn-Core with the iPAE1146 model.

The core model derived from iPAE1146 contained ∼220 fewer genes and 140 fewer reactions than the core model produced from the *S. meliloti* iGD1575 reconstruction. The difference in size is likely due to the inclusion of a larger complement of vitamins and cofactors in the biomass composition of iGD1575 compared to iPae1146. This highlights how the completeness of the core model produced by Tn-Core (and by methods using expression data) is dependent on the completeness of the biomass composition. For example, if Tn-Core is not told that peptidoglycan is part of the biomass, a complete peptidoglycan biosynthetic pathway will not necessarily be included in the core model, regardless of the Tn-seq data.

We were surprised that the gene *gap* (*PA3001*) was not essential in any of the iPae1146 core models despite being essential in the *P. aeruginosa* Tn-seq data set. The gene *PA3001* encodes a glyceraldehyde-3-phosphate dehydrogenase that mediates the interconversion between glyceraldehyde-3-phosphate and 1,3-bisphosphoglycerate. During growth on gluconeogenic substrates such as succinate, this reaction proceeds in the reverse direction. The iPae1146 model contains two putative NADP-dependent glyceraldehyde-3-phosphate dehydrogenases (*PA3001* and *PA2323*), and one putative NAD-dependent glyceraldehyde-3-phosphate dehydrogenase (*PA3195*). Notably, only the NAD-dependent reaction was reversible in iPae1146, while the NADP-dependent reaction was only allowed to proceed in the forward direction and could therefore not function during gluconeogenesis. As a result, gluconeogenesis in iPae1146, and consequently the core models, was dependent on the NAD-dependent reaction, meaning that *PA3001*, associated with the NADP-dependent reaction, was nonessential. This example highlights how the accuracy of core models produced by Tn-Core (and by methods using expression data) is constrained by the accuracy of the input GSMRs.

*Use of Tn-Core for Identifying Redundancies within a GSMR.* In addition to extracting a Tn-seq-consistent core metabolic model, Tn-Core can be used to evaluate redundancies within core metabolism regardless of whether Tn-seq data are available. We tested this function with GSMRs for five bacterial species (*S. meliloti*,[42] *P. aeruginosa*,[27] *Escherichia coli*,[49] *Pseudoalteromonas haloplanktis*,[50] and *Acinetobacter baumannii*[51]). Similar patterns were observed for all models, indicating that the applicability of this function is not unique to a single model. All matrixes and tables produced by Tn-Core to summarize the redundancy in the *S. meliloti* iGD1575 GSMR are provided in Data set S2 as examples of the exported data. As summarized in Table S2, the running parameters influence the extent of redundancy detected. The main parameters having an effect are (i) whether or not Tn-seq and/or RNA-seq data are included, (ii) the lower growth rate limit, and (iii) whether the FBA algorithm or the MOMA algorithm is used. Generally, we recommend the use of a 50% growth threshold, the use of Tn-seq data when available, and using the FBA algorithm.

Extensive redundancy was detected within all GSMRs (Figure 5 and Figures S3−S7). The gene/reaction presence matrixes (Figure 5A, 5B, Data set S2) serve as an overview of the variability. In the case of *S. meliloti*, the core models produced with the redundancy function contained an average of 433 genes, of which 286 genes (∼66%) are invariably present and the rest are from a set of 780 genes variably present or absent from the models. In other words, a third of the core metabolic genes in the *S. meliloti* GSMR can be functionally replaced by alternative genes or pathways. This level of redundancy in *S. meliloti* is consistent with recent experimental work.[12] The variable and invariable core genes were mapped to KEGG pathways[52] using eggNOG-mapper[53] to identify functional biases. Significant redundancy was observed in a diversity of pathways, including carbon, amino acid, and nucleotide metabolism. In contrast, the most fundamental cellular processes appeared to lack redundancy, such as transcription, translation, and aminoacyl-tRNA biosynthesis.

The frequency that two genes or reactions occur in the same model relative to chance are given in the co-occurrence tables and matrixes, and the matrixes can be visualized as shown in Figure 5C,D. These outputs can be used to identify modules that work together (likely to co-occur), and genes or biochemical pathways that are functionally redundant (unlikely to co-occur). For all GSMRs used in this work, clear modules and redundant genes/pathways could be observed in the matrixes (Figure 5 and Figures S3−S7). Using these outputs, known gene and reaction redundancies could be detected in the *S. meliloti* model. For example, the two pathways for L-proline biosynthesis[54] never occurred in the same model; similarly, thiamine transport and thiamine biosynthesis were unlikely to co-occur in the same model. Additionally, the functionally redundant gene pairs *edd* and *sma0235*, *proC* and *smb20003*, and *argH1* and *argH2* were never present in the same core model.[13]

The gene *tal* was predicted to be essential in the *S. meliloti* core models produced by Tn-Core, unlike in the manually prepared iGD726 core model (Figure 3). Examining the redundancy output indicated that *tal* and *gnd* were found in the same core models less frequently than expected. We therefore examined the effect of deleting both *tal* and *gnd* from iGD726; the double



DOI: 10.1021/acssynbio.8b00432
*ACS Synth. Biol.* 2019, 8, 158−169

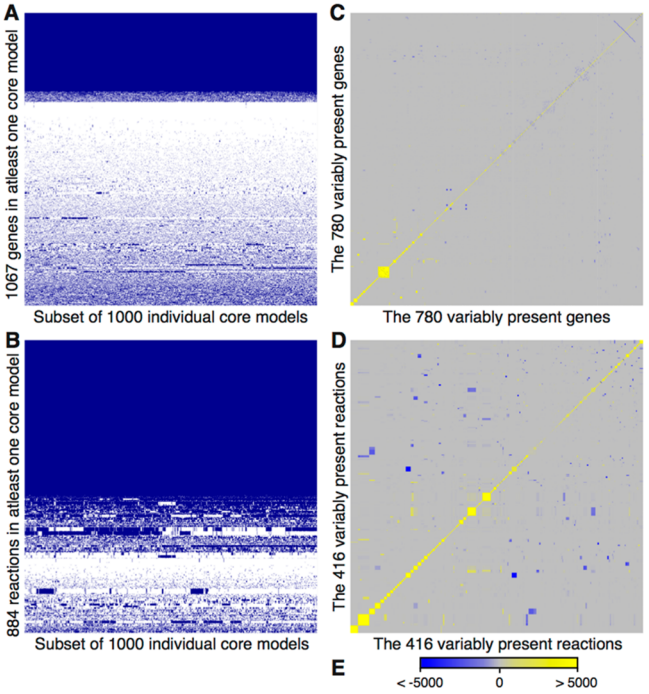



**Figure 5.** Modules and redundancies in the core metabolic pathways of *S. meliloti*. Four primary matrixes generated by Tn-Core to summarize redundancy and functional/genetic modules are shown. The Tn-Core redundancy function was run using the *S. meliloti* iGD1575 genome-scale metabolic reconstruction, with 25 000 iterations, a growth threshold of 10%, without essential genes preidentified, without RNA-seq data, and with the FBA algorithm. (A) Gene and (B) reaction presence matrixes are shown for 1000 of the randomly produced core models. Blue indicates the gene/reaction is present, white indicates the gene/reaction is absent. (C) Gene and (D) reaction co-occurrence matrixes are shown for the genes/reactions variably present in the 25 000 core models. (E) The legend for the co-occurrence matrixes is shown. The scale represents a Chi-squared statistic that summarizes if the gene or reaction pair is more (yellow) or less (blue) likely to occur in the same core model than by chance.

mutant was unable to produce biomass. Thus, the apparent essentiality of *tal* in the Tn-Core produced models is likely associated with the absence of *gnd*, and possibly other genes. Both *gnd* and *tal* encode proteins involved in the pentose-phosphate pathway.[45,55] The central carbon metabolism of *S. meliloti* is cyclic,[56] and the pentose phosphate pathway is mostly reversible. Our results are therefore consistent with mutation of *tal* being nonessential in *S. meliloti* due to a bypass of the metabolic block through a combination of the forward and reverse activities of the pentose phosphate pathway.

Similarly, the *pckA* gene was predicted to be essential in *P. aeruginosa* core models produced by Tn-Core, despite being nonessential in the Tn-seq data set (Figure 4). Examining the redundancy output revealed that *pckA* (whose product synthesizes phosphoenolpyruvate from oxaloacetate) and *ppsA* (whose product synthesizes phosphoenolpyruvate from pyruvate) are highly unlikely to co-occur in the same core model; indeed, the latter gene is absent in the core models produced by Tn-Core. Thus, the nonessentiality of *pckA* in the Tn-seq data is likely due to a bypass by the combined activities of PpsA and a malic enzyme.

Together, the results presented in the previous paragraphs confirm that the co-occurrence output of Tn-Core can be useful in detecting metabolic redundancy in core bacterial metabolism, and in understanding the biological causes underlying experimentally observed gene fitness patterns. They may also assist in identifying (and understanding) gaps in Tn-seq data sets. For example, these functions demonstrated the importance of synthesizing phosphoenolpyruvate from TCA cycle intermediates by *P. aeruginosa* during growth with succinate, a metabolic process whose essentiality was not immediately obvious based solely on the Tn-seq data set.

*Refinement of GSMRs with Tn-seq Data.* To test the utility of Tn-Core in refining GSMRs, we first validated the pipeline using the *S. meliloti* iGD1575 model described above, a draft *S. meliloti* model prepared using the KBase automated reconstruction pipeline (kbase.us), and the *P. aeruginosa* iPae1146 model. This process resulted in the modification of the GPRs of 68 reactions in iGD1575 (and suggested the replacement of "or" to "and" statements in the GPRs of 15 reactions), with 82 genes removed from the model (refined models are available in COBRA and Excel format in Data Set S1). For example, there are two annotated 2-dehydro-3-deoxyphosphogluconate aldolases in the *S. meliloti* genome (*eda1* and *eda2*),[57] and both genes are associated with the corresponding reaction (rxn03884_c0) in iG1575. Following refinement with Tn-Core, only *eda2* remained associated with this reaction, consistent with the Tn-seq data.[12] In the case of the draft *S. meliloti* model, 120 GPRs (~10% of reactions) were modified following Tn-Core mediated refinement (while the GPRs of 23 reactions were suggested to have an "or" to "and" replacement), with 78 genes deleted from the model. As with iGD1575, the reaction rxn03884_c0 was modified to leave only *eda2* as an associated gene. Using the refinement function of Tn-Core, the GPRs of 41 reactions in the *P. aeruginosa* iPae1146 model were modified, and five reactions were suggested to have an "or" to "and" replacement in their GPRs.

Previously, the iRsp1140 model of *R. sphaeroides* was manually refined based on Tn-seq data sets generated during aerobic growth with succinate as a carbon source, as well as during photosynthetic growth.[28] We performed an iterative, automated refinement of iRsp1140 with Tn-Core using the two Tn-seq data sets, and then compared the results to the manual refinement. The Tn-Core iterative refinement resulted in the modification of the GPRs of 75 reactions (and suggested "or" to "and" replacements in the GPRs of 33 reactions). Of these 75 reactions, 57 (75%) were also modified during the previously performed manual refinement;[28] similarly, the manual curation involved "or" to "and" replacements in 21 of the 33 reactions (64%) for which this was suggested by Tn-Core. While some of the modifications unique to the automated refinement may reflect correct changes that were missed during the manual refinement, others may represent changes that were incorrectly made by Tn-Core; at present, we cannot discriminate between these possibilities. However, it is important to note that automated refinement on the basis of a single Tn-seq data set has the risk of over-refinement. Specifically, genes not expressed under the given condition may be removed from the model despite the product of those genes performing the annotated function in other growth environments. This would result in condition-specific genes becoming absolutely essential. We therefore urge users to critically evaluate the results of the Tn-Core refinement process; the output table can be used to guide a manual refinement of the model keeping only those changes that the user is confident are likely correct. Additionally, only 48% of the GPRs manually modified were also modified during the automated refinement. Overall, these results demonstrate that


DOI: 10.1021/acssynbio.8b00432
*ACS Synth. Biol.* 2019, 8, 158−169

Tn-seq data and Tn-Core can play a valuable role in a metabolic model reconstruction and an automated curation pipeline. It does not, however, fully replace the need of an accurate manual curation.

## CONCLUSIONS

Tn-seq and *in silico* metabolic reconstruction with constraint-based modeling are highly complementary, systems-biology approaches to characterize the biology of an organism. Tn-Core provides automated, computational tools for integration of these data sets for the analysis of core metabolic networks of bacteria, and for the identification of gene and reaction level metabolic redundancies. Despite currently growing in popularity, CBMM still requires some computational skills to be fully exploited (e.g., a programming language like MATLAB or Python). This toolbox has been prepared in a way to maximize accessibility to both experimentalists and computational biologists, and it can provide outputs of value to all researchers. Tn-Core returns context-specific metabolic models in COBRA format for further computational evaluation, and it can also assist in automated refinement of GSMRs. At the same time, Tn-Core can export Excel-formatted metabolic models that are easily searchable, which can serve as a template for interpreting the genome-scale fitness data generated through Tn-seq experiments. Tn-Core can further return a set of tables and matrixes summarizing the redundancy present within the core metabolic pathways, assisting in identifying gaps in the gene fitness data. These outputs can also serve as a guide for synthetic biology attempts at engineering minimal cell factories. Going forward, we intend to continue to improve and enlarge the Tn-Core toolbox with additional functionality, as well as to develop an online Web server (at combo.dbe.unifi.it/tncore) to further increase the accessibility of these tools.

## METHODS

**General Metabolic Modeling Procedures.** All data presented here were generated using MATLAB 2016a (Mathworks), SBMLToolbox 4.1.0,[58] libSBML 5.13.0,[59] scripts from the COBRA Toolbox (downloaded May 12, 2017 from the openCOBRA repository),[36] the TIGER Toolbox 1.2.0-beta,[37] FASTCORE 1.0,[40] minNW (downloaded September 10, 2017 from sourceforge.net/projects/minimalnetwork),[46] and Tn-Core 1.2. The Gurobi 7.0.1 solver (gurobi.com) was used throughout; the exceptions were when running FASTCC, FASTCORE, or minNW, in which cases the iLOG CPLEX Studio 12.7.1 solver (ibm.com) was used. Effects of gene deletion on growth was determined using the *singleGeneDeletion* function and the MOMA algorithm. To ensure that core model generation with the Tn-Core redundancy function did not occasionally fail when using the MOMA algorithm, the MOMA.m script of the COBRA Toolbox was modified at line 216 to treat unbounded solutions the same as infeasible solutions. For all FBA simulations, the growth environments were set to mimic the growth conditions used in generating the Tn-seq data. The lower bounds of exchange reactions for compounds found in the growth medium were opened, setting carbon availability as the growth rate limiting factor; the lower bounds of all other exchange reactions were set to zero.

Scripts to repeat the benchmarking, as well as the output data generated in this work, are available at github.com/diCenzo-GC/Tn-Core. The complete Tn-Core toolbox, together with a reference manual, are freely available at github.com/diCenzo-GC/Tn-Core, and future releases of the toolbox will be available through the same link.

**Metabolic Models and -Omics Data Sets.** The *S. meliloti* iGD1575,[42] *P. haloplanktis* iMF721,[50] *A. baumannii* iLP844,[51] *E. coli* iJO1366,[49] *P. aeruginosa* iPae1146,[27] *R. sphaeroides* iRsp1140,[48] and *R. sphaeroides* iRsp1140_opt[28] models were previously published. Prior to using iGD1575, the model reaction content and biomass composition were modified as described recently.[12] Additionally, a $FADH_2$-dependent ubiquinone reductase reaction was added as this reaction was observed to be missing. Prior to using iLP844, the genes "Unknown1" through "Unknown160" were replaced with a single gene called "Unknown". The draft *S. meliloti* GSMR was generated on the KBase Web server (kbase.us). The RefSeq version of the *S. meliloti* genome was reannotated using the "annotate microbial genome" function, maintaining the original locus tags, following which an automated reconstruction was built using the "build metabolic model" function with gap-filling.

All Tn-seq and RNA-seq data sets were collected from published studies. The *S. meliloti* Tn-seq[12] and RNA-seq[43] data were generated for *S. meliloti* cells grown in a minimal medium with sucrose as the sole carbon source. The *P. aeruginosa* Tn-seq[15] and RNA-seq[47] data were generated for *P. aeruginosa* cells grown in a minimal medium with succinate as the sole carbon source. The *R. sphaeroides* Tn-seq[28] data were generated for *R. sphaeroides* cells grown photosynthetically or aerobically in minimal medium with succinate as the sole carbon source.

**Core Model Extraction for Benchmarking.** Four approaches were used to generate core models from the GSMR for comparison with the core models generated by Tn-Core. Each approach was based on FASTCORE, minNW, or GIMME. Each pipeline began by setting the correct boundary conditions. Reactions involving dead-end metabolites were then iteratively removed from the model, and unnecessary unknown GPRs were deleted, using code from Tn-Core. The reduced model was then used as input with each of the following pipelines.

The initial set-ups for running the FASTCORE and minNW pipelines were the same. FASTCC was used to identify the consistent set of reactions in the GSMR, and this list of reactions was used to extract the consistent model from the GSMR. The Tn-seq data was then analyzed the same way as in the Tn-Core pipeline to identify the Tn-seq essential genes. These genes were deleted from the GSMR, and the resulting list of constrained reactions was identified. The constrained reactions that were also present in the FASTCC-derived consistent model were collected as a list of protected reactions. FASTCORE was then run using the consistent model and the protected reactions that were determined in the previous step. Similarly, minNW was run with the same input model and the protected reaction list using the bigM formulation. For both algorithms, the list of reactions to be included in the core model was identified from the output, and all other reactions (and associated genes and metabolites not associated with another reaction) were removed from the GSMR to produce the final core models.

For the GIMME-based pipeline involving Tn-seq data, the Tn-seq data were log-transformed, all values multiplied by negative 1, and the minimum value (which was negative) was subtracted from all values. These transformations made the Tn-seq data suitable for GIMME by giving the genes with strong fitness contributions the highest value and setting the lowest value equal to zero. The input model was converted to TIGER format, and the Tn-Core adaptation of the TIGER version of


DOI: 10.1021/acssynbio.8b00432
ACS Synth. Biol. 2019, 8, 158−169

GIMME was run to extract a core model using the Tn-seq data; genes with a value greater than 3.5 standard deviations above the median were considered essential (i.e., the default used in Tn-Core). The genes identified as "on" in the GIMME output were used to rebuild a functional core model in COBRA format using code from Tn-Core.

The GIMME-based pipeline involving RNA-seq data was largely the same as that for the Tn-seq data. The difference was that unmodified RNA-seq data (as RPKM values) were used in place of Tn-seq data, and genes with an RPKM value equal to at least 0.02% of the sum of all RPKM values were considered highly expressed.

**KEGG Functional Mapping.** The entire *S. meliloti* Rm1021 proteome was annotated with KEGG Orthology (KO) terms using the eggNOG-mapper Web server[52] with default settings. All KO terms associated with the genes of interest were extracted from the eggNOG output file, and they were used as input for the 'KEGG Mapper − Search Pathway' function[53] that linked the KO terms to KEGG pathways. The abundances of the KEGG pathways among the gene sets of interest were then examined to identify functional biases.

## ■ ASSOCIATED CONTENT

### Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acssynbio.8b00432.

All input *S. meliloti* models, and all *S. meliloti* core or refined metabolic models generated in this work; all models are provided in both COBRA format and Excel format (ZIP)

Sample output of the redundancy function of Tn-Core, generated using the *S. meliloti* iGD1575 metabolic reconstruction (ZIP)

Sample output of the refinement function of Tn-Core, generated using the *S. meliloti* iGD1575 metabolic reconstruction (ZIP)

Supplementary text, Algorithms S1−S3, Tables S1−S2, Figures S1−S7, and the associated references (PDF)

## ■ AUTHOR INFORMATION


### Corresponding Authors

*E-mail: georgecolin.dicenzo@unifi.it.

*E-mail: marco.fondi@unifi.it.

### ORCID

Alessio Mengoni: 0000-0002-1265-8251

Marco Fondi: 0000-0001-9291-5467


### Notes

The authors declare no competing financial interest.

## ■ ACKNOWLEDGMENTS

We thank Michele Giovannini for preparing the RNA-seq datasets used in this work. G.C.D. was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC) through a PDF fellowship. M.F.'s research is supported by PNRA (Programma Nazionale di Ricerca in Antartide, grant PNRA16_00246).

## ■ ABBREVIATIONS

GSMR, genome-scale metabolic reconstruction; CBMM, constraint-based metabolic modeling; FBA, flux balance analysis; MOMA, minimization of metabolic adjustment; RPKM, reads per kilobase million; TPM, transcripts per million

## ■ REFERENCES


(1) Hutchison, C. A., Chuang, R.-Y., Noskov, V. N., Assad-Garcia, N., Deerinck, T. J., Ellisman, M. H., Gill, J., Kannan, K., Karas, B. J., Ma, L., Pelletier, J. F., Qi, Z.-Q., Richter, R. A., Strychalski, E. A., Sun, L., Suzuki, Y., Tsvetanova, B., Wise, K. S., Smith, H. O., Glass, J. I., Merryman, C., Gibson, D. G., and Venter, J. C. (2016) Design and synthesis of a minimal bacterial genome. *Science 351*, aad6253−aad6253.

(2) Nielsen, J. (2017) Systems biology of metabolism. *Annu. Rev. Biochem. 86*, 245−275.

(3) van Opijnen, T., and Camilli, A. (2013) Transposon insertion sequencing: a new tool for systems-level analysis of microorganisms. *Nat. Rev. Microbiol. 11*, 435−442.

(4) Chao, M. C., Abel, S., Davis, B. M., and Waldor, M. K. (2016) The design and analysis of transposon insertion sequencing experiments. *Nat. Rev. Microbiol. 14*, 119−128.

(5) Jensen, P. A., Zhu, Z., and van Opijnen, T. (2017) Antibiotics disrupt coordination between transcriptional and phenotypic stress responses in pathogenic bacteria. *Cell Rep. 20*, 1705−1716.

(6) Gallagher, L. A., Shendure, J., and Manoil, C. (2011) Genome-scale identification of resistance functions in *Pseudomonas aeruginosa* using Tn-seq. *mBio 2*, e00315−10.

(7) Christen, M., Beusch, C., Bösch, Y., Cerletti, D., Flores-Tinoco, C. E., Del Medico, L., Tschan, F., and Christen, B. (2016) Quantitative Selection Analysis of Bacteriophage φCbK Susceptibility in *Caulobacter crescentus*. *J. Mol. Biol. 428*, 419−430.

(8) van Opijnen, T., and Camilli, A. (2012) A fine scale phenotype-genotype virulence map of a bacterial pathogen. *Genome Res. 22*, 2541−2551.

(9) Gawronski, J. D., Wong, S. M. S., Giannoukos, G., Ward, D. V., and Akerley, B. J. (2009) Tracking insertion mutants within libraries by deep sequencing and a genome-wide screen for *Haemophilus* genes required in the lung. *Proc. Natl. Acad. Sci. U. S. A. 106*, 16422−16427.

(10) Goodman, A. L., McNulty, N. P., Zhao, Y., Leip, D., Mitra, R. D., Lozupone, C. A., Knight, R., and Gordon, J. I. (2009) Identifying genetic determinants needed to establish a human gut symbiont in its habitat. *Cell Host Microbe 6*, 279−289.

(11) Cameron, D. E., Urbach, J. M., and Mekalanos, J. J. (2008) A defined transposon mutant library and its use in identifying motility genes in *Vibrio cholerae*. *Proc. Natl. Acad. Sci. U. S. A. 105*, 8736−8741.

(12) diCenzo, G. C., Benedict, A. B., Fondi, M., Walker, G. C., Finan, T. M., Mengoni, A., and Griffitts, J. S. (2018) Robustness encoded across essential and accessory replicons of the ecologically versatile bacterium *Sinorhizobium meliloti*. *PLoS Genet. 14*, e1007357.

(13) diCenzo, G. C., and Finan, T. M. (2015) Genetic redundancy is prevalent within the 6.7 Mb *Sinorhizobium meliloti* genome. *Mol. Genet. Genomics 290*, 1345−1356.

(14) Canals, R., Xia, X.-Q., Fronick, C., Clifton, S. W., Ahmer, B. M., Andrews-Polymenis, H. L., Porwollik, S., and McClelland, M. (2012) High-throughput comparison of gene fitness among related bacteria. *BMC Genomics 13*, 212.

(15) Turner, K. H., Wessel, A. K., Palmer, G. C., Murray, J. L., and Whiteley, M. (2015) Essential genome of *Pseudomonas aeruginosa* in cystic fibrosis sputum. *Proc. Natl. Acad. Sci. U. S. A. 112*, 4110−4115.

(16) Koo, B.-M., Kritikos, G., Farelli, J. D., Todor, H., Tong, K., Kimsey, H., Wapinski, I., Galardini, M., Cabal, A., Peters, J. M., Hachmann, A.-B., Rudner, D. Z., Allen, K. N., Typas, A., and Gross, C. A. (2017) Construction and analysis of two genome-scale deletion libraries for *Bacillus subtilis*. *Cell Syst 4*, 291−305.

(17) Freed, N. E., Bumann, D., and Silander, O. K. (2016) Combining *Shigella* Tn-seq data with gold-standard *E. coli* gene deletion data suggests rare transitions between essential and non-essential gene functionality. *BMC Microbiol. 16*, 203.

(18) Joshi, S. M., Pandey, A. K., Capite, N., Fortune, S. M., Rubin, E. J., and Sassetti, C. M. (2006) Characterization of mycobacterial virulence genes through genetic interaction mapping. *Proc. Natl. Acad. Sci. U. S. A. 103*, 11760−11765.




DOI: 10.1021/acssynbio.8b00432
*ACS Synth. Biol.* 2019, 8, 158−169

(19) Juhas, M., Reuß, D. R., Zhu, B., and Commichau, F. M. (2014) *Bacillus subtilis* and *Escherichia coli* essential genes and minimal cell factories after one decade of genome engineering. *Microbiology 160*, 2341−2351.
(20) Varma, A., and Palsson, B. O. (1994) Stoichiometric flux balance models quantitatively predict growth and metabolic by-product secretion in wild-type *Escherichia coli* W3110. *Appl. Environ. Microbiol. 60*, 3724−3731.
(21) Bordbar, A., Monk, J. M., King, Z. A., and Palsson, B. Ø. (2014) Constraint-based models predict metabolic and associated cellular functions. *Nat. Rev. Genet. 15*, 107−120.
(22) Thiele, I., and Palsson, B. Ø. (2010) A protocol for generating a high-quality genome-scale metabolic reconstruction. *Nat. Protoc. 5*, 93−121.
(23) Orth, J. D., Thiele, I., and Palsson, B. Ø. (2010) What is flux balance analysis? *Nat. Biotechnol. 28*, 245−248.
(24) Pratapa, A., Balachandran, S., and Raman, K. (2015) Fast-SL: an efficient algorithm to identify synthetic lethal sets in metabolic networks. *Bioinformatics 31*, 3299−3305.
(25) Schellenberger, J., Park, J. O., Conrad, T. M., and Palsson, B. Ø. (2010) BiGG: a Biochemical Genetic and Genomic knowledgebase of large scale metabolic reconstructions. *BMC Bioinf. 11*, 213.
(26) Broddrick, J. T., Rubin, B. E., Welkie, D. G., Du, N., Mih, N., Diamond, S., Lee, J. J., Golden, S. S., and Palsson, B. Ø. (2016) Unique attributes of cyanobacterial metabolism revealed by improved genome-scale metabolic modeling and essential gene analysis. *Proc. Natl. Acad. Sci. U. S. A. 113*, E8344−E8353.
(27) Bartell, J. A., Blazier, A. S., Yen, P., Thøgersen, J. C., Jelsbak, L., Goldberg, J. B., and Papin, J. A. (2017) Reconstruction of the metabolic network of *Pseudomonas aeruginosa* to interrogate virulence factor synthesis. *Nat. Commun. 8*, 14631.
(28) Burger, B. T., Imam, S., Scarborough, M. J., Noguera, D. R., and Donohue, T. J. (2017) Combining genome-scale experimental and computational methods to identify essential genes in *Rhodobacter sphaeroides*. *mSystems 2*, e00015−17.
(29) Yang, H., Krumholz, E. W., Brutinel, E. D., Palani, N. P., Sadowsky, M. J., Odlyzko, A. M., Gralnick, J. A., and Libourel, I. G. L. (2014) Genome-scale metabolic network validation of *Shewanella oneidensis* using transposon insertion frequency analysis. *PLoS Comput. Biol. 10*, e1003848.
(30) Senior, N. J., Sasidharan, K., Saint, R. J., Scott, A. E., Sarkar-Tyson, M., Ireland, P. M., Bullifent, H. L., Rong Yang, Z., Moore, K., Oyston, P. C. F., Atkins, T. P., Atkins, H. S., Soyer, O. S., and Titball, R. W. (2017) An integrated computational-experimental approach reveals *Yersinia pestis* genes essential across a narrow or a broad range of environmental conditions. *BMC Microbiol. 17*, 163.
(31) Blazier, A. S., and Papin, J. A. (2012) Integration of expression data in genome-scale metabolic network reconstructions. *Front. Physiol. 3*, 299.
(32) Machado, D., and Herrgård, M. (2014) Systematic evaluation of methods for integration of transcriptomic data into constraint-based models of metabolism. *PLoS Comput. Biol. 10*, e1003580.
(33) Großeholz, R., Koh, C.-C., Veith, N., Fiedler, T., Strauss, M., Olivier, B., Collins, B. C., Schubert, O. T., Bergmann, F., Kreikemeyer, B., Aebersold, R., and Kummer, U. (2016) Integrating highly quantitative proteomics and genome-scale metabolic modeling to study pH adaptation in the human pathogen *Enterococcus faecalis*. *NPJ. Syst. Biol. Appl. 2*, 16017.
(34) Zamboni, N., Fischer, E., and Sauer, U. (2005) FiatFlux–a software for metabolic flux analysis from 13C-glucose experiments. *BMC Bioinf. 6*, 209.
(35) Yizhak, K., Benyamini, T., Liebermeister, W., Ruppin, E., and Shlomi, T. (2010) Integrating quantitative proteomics and metabolomics with a genome-scale metabolic network model. *Bioinformatics 26*, i255−60.
(36) Schellenberger, J., Que, R., Fleming, R. M. T., Thiele, I., Orth, J. D., Feist, A. M., Zielinski, D. C., Bordbar, A., Lewis, N. E., Rahmanian, S., Kang, J., Hyduke, D. R., and Palsson, B. Ø. (2011) Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox v2.0. *Nat. Protoc. 6*, 1290−1307.
(37) Jensen, P. A., Lutz, K. A., and Papin, J. A. (2011) TIGER: Toolbox for integrating genome-scale metabolic models, expression data, and transcriptional regulatory networks. *BMC Syst. Biol. 5*, 147.
(38) Pacheco, M. P., Pfau, T., and Sauter, T. (2015) Benchmarking Procedures for High-Throughput Context Specific Reconstruction Algorithms. *Front. Physiol. 6*, 410.
(39) Becker, S. A., and Palsson, B. Ø. (2008) Context-specific metabolic networks are consistent with experiments. *PLoS Comput. Biol. 4*, e1000082.
(40) Vlassis, N., Pacheco, M. P., and Sauter, T. (2014) Fast reconstruction of compact context-specific metabolic network models. *PLoS Comput. Biol. 10*, e1003424.
(41) Schnoes, A. M., Brown, S. D., Dodevski, I., and Babbitt, P. C. (2009) Annotation error in public databases: misannotation of molecular function in enzyme superfamilies. *PLoS Comput. Biol. 5*, e1000605.
(42) diCenzo, G. C., Checcucci, A., Bazzicalupo, M., Mengoni, A., Viti, C., Dziewit, L., Finan, T. M., Galardini, M., and Fondi, M. (2016) Metabolic modelling reveals the specialization of secondary replicons for niche adaptation in *Sinorhizobium meliloti*. *Nat. Commun. 7*, 12219.
(43) diCenzo, G. C., Muhammed, Z., Østerås, M., O'Brien, S. A. P., and Finan, T. M. (2017) A key regulator of the glycolytic and gluconeogenic central metabolic pathways in *Sinorhizobium meliloti*. *Genetics 207*, 961−974.
(44) Poysti, N. J., and Oresnik, I. J. (2007) Characterization of *Sinorhizobium meliloti* triose phosphate isomerase genes. *J. Bacteriol. 189*, 3445−3451.
(45) Hawkins, J. P., Ordonez, P. A., and Oresnik, I. J. (2018) Characterization of mutants that affect the non-oxidative pentose phosphate pathway in *Sinorhizobium meliloti*. *J. Bacteriol. 200*, e00436−17.
(46) Röhl, A., and Bockmayr, A. (2017) A mixed-integer linear programming approach to the reduction of genome-scale metabolic networks. *BMC Bioinf. 18*, 2.
(47) Turner, K. H., Everett, J., Trivedi, U., Rumbaugh, K. P., and Whiteley, M. (2014) Requirements for *Pseudomonas aeruginosa* acute burn and chronic surgical wound infection. *PLoS Genet. 10*, e1004518.
(48) Imam, S., Noguera, D. R., and Donohue, T. J. (2013) Global insights into energetic and metabolic networks in *Rhodobacter sphaeroides*. *BMC Syst. Biol. 7*, 89.
(49) Orth, J. D., Conrad, T. M., Na, J., Lerman, J. A., Nam, H., Feist, A. M., and Palsson, B. Ø. (2011) A comprehensive genome-scale reconstruction of *Escherichia coli* metabolism–2011. *Mol. Syst. Biol. 7*, 535−535.
(50) Fondi, M., Maida, I., Perrin, E., Mellera, A., Mocali, S., Parrilli, E., Tutino, M. L., Liò, P., and Fani, R. (2015) Genome-scale metabolic reconstruction and constraint-based modelling of the Antarctic bacterium *Pseudoalteromonas haloplanktis* TAC125. *Environ. Microbiol. 17*, 751−766.
(51) Presta, L., Bosi, E., Mansouri, L., Dijkshoorn, L., Fani, R., and Fondi, M. (2017) Constraint-based modeling identifies new putative targets to fight colistin-resistant *A. baumannii* infections. *Sci. Rep. 7*, 3706.
(52) Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M., and Tanabe, M. (2016) KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res. 44*, D457−D462.
(53) Huerta-Cepas, J., Forslund, K., Coelho, L. P., Szklarczyk, D., Jensen, L. J., von Mering, C., and Bork, P. (2017) Fast genome-wide functional annotation through orthology assignment by eggNOG-mapper. *Mol. Biol. Evol. 34*, 2115−2122.
(54) diCenzo, G. C., Zamani, M., Cowie, A., and Finan, T. M. (2015) Proline auxotrophy in *Sinorhizobium meliloti* results in a plant-specific symbiotic phenotype. *Microbiology 161*, 2341−2351.
(55) Geddes, B. A., and Oresnik, I. J. (2014) Physiology, genetics, and biochemistry of carbon metabolism in the alphaproteobacterium *Sinorhizobium meliloti*. *Can. J. Microbiol. 60*, 491−507.


DOI: 10.1021/acssynbio.8b00432
*ACS Synth. Biol.* 2019, 8, 158−169

(56) Fuhrer, T., Fischer, E., and Sauer, U. (2005) Experimental identification and quantification of glucose metabolism in seven bacterial species. *J. Bacteriol. 187*, 1581−1590.
(57) Galibert, F., Finan, T. M., Long, S. R., Puhler, A., Abola, P., Ampe, F., Barloy-Hubler, F., Barnett, M. J., Becker, A., Boistard, P., Bothe, G., Boutry, M., Bowser, L., Buhrmester, J., Cadieu, E., Capela, D., Chain, P., Cowie, A., Davis, R. W., Dreano, S., Federspiel, N. A., Fisher, R. F., Gloux, S., Godrie, T., Goffeau, A., Golding, B., Gouzy, J., Gurjal, M., Hernández-Lucas, I., Hong, A., Huizar, L., Hyman, R. W., Jones, T., Kahn, D., Kahn, M. L., Kalman, S., Keating, D. H., Kiss, E., Komp, C., Lelaure, V., Masuy, D., Palm, C., Peck, M. C., Pohl, T. M., Portetelle, D., Purnelle, B., Ramsperger, U., Surzycki, R., Thébault, P., Vandenbol, M., Vorhölter, F. J., Weidner, S., Wells, D. H., Wong, K., Yeh, K. C., and Batut, J. (2001) The composite genome of the legume symbiont Sinorhizobium meliloti. *Science 293*, 668−672.
(58) Keating, S. M., Bornstein, B. J., Finney, A., and Hucka, M. (2006) SBMLToolbox: an SBML toolbox for MATLAB users. *Bioinformatics 22*, 1275−1277.
(59) Bornstein, B. J., Keating, S. M., Jouraku, A., and Hucka, M. (2008) LibSBML: an API library for SBML. *Bioinformatics 24*, 880−881.



DOI: 10.1021/acssynbio.8b00432
*ACS Synth. Biol.* 2019, 8, 158−169