Titles: Robustness encoded across essential and accessory replicons in an ecologically versatile bacterium

Running title: Genome effects on gene essentiality

Authors: George C diCenzo[1]*, Alex B Benedict[2], Marco Fondi[1], Graham C Walker[3], Turlough M Finan[4], Alessio Mengoni[1], Joel S Griffitts[2]

Affiliations: [1] Department of Biology, University of Florence, Sesto Fiorentino, FI, 50019, Italy.
[2] Department of Microbiology and Molecular Biology, Brigham Young University, Provo, UT, 84602, USA.
[3] Department of Biology, Massachusetts Institute of Technology, Cambridge, MA, 02139, USA.
[4] Department of Biology, McMaster University, Hamilton, ON, L8S 4K1, Canada.

*** Corresponding author:** George C diCenzo
Department of Biology
University of Florence
Sesto Fiorentino, FI, 50019
Italy
Email: georgecolin.dicenzo@unifi.it

Keywords: genetic interactions / synthetic lethality / systems biology / Tn-seq / constraint based modelling

ABSTRACT

Bacterial genome evolution is characterized by gains, losses, and rearrangements of functional genetic segments. The extent to which genotype-phenotype relationships are influenced by large-scale genomic alterations has not been investigated in a high-throughput manner. In the symbiotic soil bacterium *Sinorhizobium meliloti*, the genome is composed of a chromosome and two large extrachromosomal replicons (pSymA and pSymB, which together constitute 45% of the genome). Massively parallel transposon insertion sequencing (Tn-seq) was employed to evaluate contributions of chromosomal genes to fitness in both the presence and absence of these extrachromosomal replicons. Ten percent of chromosomal genes from diverse functional categories are shown to genetically interact with pSymA and pSymB. These results demonstrate the pervasive robustness provided by the extrachromosomal replicons, which is further supported by constraint-based metabolic modelling. A comprehensive picture of core *S. meliloti* metabolism was generated through a Tn-seq-guided *in silico* metabolic network reconstruction, producing a core network encompassing 726 genes. This integrated approach facilitated functional assignments for previously uncharacterized genes, while also revealing that Tn-seq alone misses over a quarter of wild type metabolism. This work highlights the strong functional dependencies and epistatic relationships that may arise between bacterial replicons and across a genome, while also demonstrating how Tn-seq and metabolic modelling can be used together to yield insights not obtainable by either method alone.

INTRODUCTION

The prediction of genotype-phenotype relationships is a fundamental goal of genetic, biomedical, and eco-evolutionary research, and this problem underpins the design of synthetic microbial systems for biotechnological applications [1]. The last decades have witnessed a shift away from the functional characterization of single genes towards whole-genome, systems-level analyses [for recent reviews, see [2,3]]. Such studies have been facilitated by the development of methods that allow for the direct interrogation of a genome to determine all genetic elements required for adaptation to a specified environment. Two primary methods are *in silico* metabolic modelling [4,5], and massively parallel sequencing of transposon insertions in bacterial mutant libraries (Tn-seq) [6,7].

The process of *in silico* genome-scale metabolic modelling consists of two stages. First, a reconstruction of all cellular metabolism is built that contains all reactions expected to be present, as well as which genes encode the enzymes performing each reaction, thereby linking genetics to metabolism [8]. Next, mathematical models such as flux balance analysis (FBA) are used to simulate the flux distribution through the reconstructed metabolic network [9], which can be used to predict how environmental perturbations or gene disruptions influence growth phenotypes. This approach allows for phenotypic predictions of all possible single, double, or higher-order gene deletion mutations within a matter of days [10,11], something that is infeasible using a direct experimental approach. However, the quality of the predictions is highly dependent on the accuracy of the metabolic reconstruction. Outside of a few model species like *Escherichia coli*, experimental genetic and biochemical data are not available at the resolution necessary to provide accurate assignment of all metabolic gene functions.

The Tn-seq approach involves the generation of a library of hundreds of thousands of mutant clones, each containing a single transposon insertion at a random genomic location [12]. The library of pooled clones is then cultured in the presence of a defined environmental challenge. Insertions resulting in altered fitness in the environment under investigation become under- or over-represented in the population, and this is monitored by deep sequencing to identify the genomic location and frequency of all transposon insertions. This approach is imperfect, as important biochemical functions may be encoded redundantly in the genome [13-15], and the loss of some essential genes can be compensated for by evolution of alternative cellular processes [16]. Moreover, fitness changes brought about by mutation in one gene may be dependent on mutation of a second gene bearing no resemblance to the first—a phenomenon known as a genetic interaction [17,18]. Such genetic interactions may cause the apparent functions of some genes to be strictly dependent on their genomic environment [19]. In other words, a gene may be essential for growth in one organism, but its orthologous counterpart in another organism may be non-essential. This significantly complicates efforts to generalize genotype-phenotype relationships [20].

Resolving the problem of genome-conditioned gene function is of broad significance in the areas of functional genomics, population genetics, and synthetic biology. For example, the ability to design and build optimized minimal cell factories on the basis of single-mutant fitness data is expected to present numerous complications [21], as evidenced by the recent effort to rationally build a functional minimal genome [22]. Tn-seq studies have suggested there is as little as 50% to 25% overlap in the essential genome of any two species [23-25]. As a striking example, 210 of the Tn-seq determined essential genes of *Pseudomonas aeruginosa* PA14 are not even present in the genome of *P. aeruginosa* PAO1 [26]. Comparison of Tn-seq data for

Shigella flexneri with the deletion analysis data for closely related *E. coli* suggested only a small number of genes were specifically essential in one species. Mutation of about 100 genes, however, appeared to result in a growth rate decrease specifically in *E. coli* [27]. Similarly, comparison of Tn-seq datasets from two *Salmonella* species revealed that mutation of nearly 40 genes had a stronger growth phenotype in one of the two species [28]. Overall, these studies suggest that the genomic environment (here defined as the genomic components that may vary from organism to organism) influences the fitness contributions of a significant proportion of an organism's genes. However, no large-scale analysis has been performed that directly illustrates how the phenotypes of individual genes are impacted when a small or large part of the genome is modified.

Here, we provide a quantitative, genome-scale evaluation of how large-scale genomic variance influences genotype-phenotype relationships. We have accomplished this in a way that minimizes the effects of laboratory-to-laboratory variation, and removes the effects of complex genome evolution. The model system used is *Sinorhizobium meliloti*, an α-proteobacterium whose 6.7-Mb genome consists of a chromosome and two additional replicons, the pSymA megaplasmid and the pSymB chromid. The pSymA and pSymB replicons constitute 45% of the *S. meliloti* genome (~2,900 genes); yet, by simply transferring only two essential genes from pSymB to the chromosome, both pSymA and pSymB can be completely removed from the genome, yielding a viable single-replicon organism [29]. We report a comparison of gene essentiality (via Tn-seq) for wild-type *S. meliloti* and the single-replicon derivative. This analysis was supplemented by an *in silico* double gene deletion analysis of a *S. meliloti* genome-scale metabolic network reconstruction. We further examine how integration of Tn-seq data with *in silico* metabolic modelling, through a Tn-seq-guided reconstruction process, overcomes the limitations of using either of these approaches in isolation to develop a consolidated view of the core metabolism of the organism. This process produced a fully referenced *core S. meliloti* metabolic reconstruction.

RESULTS

Development and validation of the Tn*5*-based transposon Tn*5*-714.

In order to interrogate the *S. meliloti* genome using a Tn-seq based approach, we first developed a new construct based on the Tn*5* transposon as described in the Materials and Methods. The resulting transposon (Figure S1) contains constitutive promoters reading out from both ends of the transposon to ensure the production of non-polar mutations. Analysis of the insertion site locations validated that the transposon performed largely as expected. Gene disruptions caused by transposon insertions were confirmed to be non-polar as illustrated by the case reported in Figure 1, and there was no strong bias in the distribution of insertions around the chromosome (Figures 2A, S2). However, there did appear to be somewhat of a bias for integration of the transposon in GC rich regions (Figure S3). Given the high GC content (62.7%) of the *S. meliloti* chromosome, it is unlikely that this moderate bias had a discernable influence on the results of this study.

Overview of the Tn-seq output.

The Tn-seq experiments reported here were undertaken with two primary aims: i) to identify the core set of genes contributing to *S. meliloti* growth in laboratory conditions, and ii) to determine the extent to which the phenotypic consequence of a gene deletion is influenced by the genomic environment (i.e. presence/absence of the secondary replicons). To accomplish this, Tn-seq libraries of two *S. meliloti* strains were prepared: a wild type strain (designated RmP3499) containing the entire genome, and a strain with both the pSymA and pSymB replicons removed (designated RmP3496 or ΔpSymAB; strains described previously in [30]). Transposon library sizes were skewed to compensate for the difference in genome sizes, resulting in nearly identical insertion site density for each library (Table S1). Both libraries were passed through selective growth regimens in either complex BRM broth (rich medium) or minimal VMM broth (defined medium) in duplicates. Following approximately nine generations of growth, the location of the transposon insertions in the population was determined, a gene essentiality index (GEI) was calculated for all chromosomal genes, and each gene was classified into one of five fitness categories (Table 1) using the procedure described in the Materials and Methods. Four genes (*pdxJ*, *fumC*, *smc01011*, *smc03995*), including two of unknown function, were independently mutated in the wild-type background, and in all cases, the mutations yielded the expected no-growth phenotype (Figure S4), supporting the accuracy of the Tn-seq output. All Tn-seq data is available as Data Set S1.

A strong correlation was observed between the number of insertions per gene in each set of duplicates (Figure S5), indicating that there was high reproducibility of the results and that differences between conditions were unlikely to reflect random fluctuations in the output. On average, insertions were found in 190,000 unique chromosomal positions with a median of 39 unique insertion positions per gene (Table S1). The similarity in the number of unique insertion positions between samples suggested that differences in the Tn-seq outputs were also unlikely to be an artefact of the quality of the libraries.

Elucidation of the core genetic components of *S. meliloti*.

There were 307 genes classified as essential independently of growth medium or strain (Figure S6). This set of 307 genes includes those encoding functions commonly understood to be essential: the DNA replication apparatus, the four RNA polymerase subunits, the housekeeping sigma factor, the general transcriptional termination factor Rho, 40 out of 55 of the annotated ribosomal protein subunits, 18 out of 20 of the annotated aminoacyl-tRNA synthetases, and 6 out of 10 of the annotated ATP synthase subunits. Considering genes classified as essential plus those genes whose mutation resulted in a large growth defect (Groups I and II in Table 1), a core growth promoting genome of 489 genes, representing ~ 15% of the chromosome, was identified (Figure 2B). This expanded list includes 51 out of 55 of the annotated ribosomal protein subunits, 19 out of 20 of the annotated aminoacyl-tRNA synthetases, and 9 out of 10 of the annotated ATP synthase subunits These 489 genes appeared to be mostly dispersed around the chromosome, although there was a bias for these genes to be found in the leading strand (Figure 2A). Based on published RNA-seq data for *S. meliloti* grown in a glucose minimal medium, these 489 genes tend to be highly expressed, with a median expression level above the 90% percentile (Figure S7). Compared to the entire chromosome (Fisher exact test, p-value < 0.05 following a Bonferroni correction for 18 tests), this set of 489 genes was enriched for genes involved in translation (5.2-fold), lipid metabolism (2.7-fold), cofactor metabolism (3.3-fold), and electron transport (2.1-fold), whereas genes involved in transport (2.1-fold), motility/attachment (9.4-fold), and hypothetical genes (2.7-fold) were under-represented (Figure 2C). Additionally, cell wall (2.2-fold) and cell division (2.3-fold) were over-represented while transcription (1.9-fold) was under-represented (Figure 2C), although these differences where not considered statistically significant.

A clear influence of the growth medium on the fitness phenotypes of gene mutations was observed. The degree to which mutant phenotypes were impacted by growth medium type is reflected in the synthetic medium index (SMI) calculated as described in the Materials and

Methods. Focusing on the wild-type strain, a core of 519 genes were identified as contributing equally to growth in both media (Figure 2D). Forty genes were identified as more important during growth in rich medium than in defined medium, and these genes had a median SMI score of 7 (values of 1 and -1 are neutral). Only translation functions (5.8-fold) displayed a statistically significant enrichment in these genes, which may reflect the faster growth rate in the rich medium (Figure S8), while there was also a non-statistically significant enrichment in signal transduction (5.1-fold) (Figure 2C). The extent of specialization for growth in the defined medium was more pronounced; 93 genes were more important during growth in the defined medium with a median SMI score of -20. These genes were enriched (statistically significant) in amino acid (9.0-fold) and nucleotide (6.7-fold) metabolism presumably due to the requirement of their biosynthesis, and carbohydrate metabolism (3.6-fold) likely as the sole carbon source was a carbohydrate (Figure 2C). The same overall pattern was observed between media for the ΔpSymAB strain (Figure S9).

Mutant fitness phenotypes are strongly influenced by their genomic environment.

The Tn-seq data sets for the wild-type and the ΔpSymAB strains were compared to evaluate the robustness of the observed fitness phenotypes in response to changes in the gene's genomic environment. Similar results were observed for both growth media, suggesting that the results were generalizable and not medium specific. Depending on the medium, either 484 or 488 genes had an equal contribution to growth in both strains, 81 or 89 genes led to stronger growth impairment when mutated in wild-type cells, and either 250 or 251 genes led to stronger growth impairment when mutated in ΔpSymAB cells (Figures 2E, 2F, and Table 2). Only minor functional bias was observed in the genes that displayed larger fitness defects in the ΔpSymAB background (Figure 2C); in both media, only electron transport (3-fold) and oxidoreductases

(9.5-fold) were over- and under-represented, respectively. Similarly, little functional bias was detected in genes with larger fitness defects in the wild-type background (Figure 2C); in both media, lipid metabolism (4.5-fold) and hypothetical genes (2-fold) were over- and under-represented, respectively, while nucleotide metabolism (5.5-fold) was also enriched in the rich medium. Overall, these results were consistent with pervasive effects of the genomic environment on the genotype-phenotype relationship that was largely independent of the biological role of the gene products.

Approximately half (9 of 16) of the genes that were independently mutated in both strains yielded the expected phenotypes on rich agar plates (Figures S10). Of the other seven genes, which were expected to be essential specifically in the ΔpSymAB strain, at least three were non-lethal but displayed obvious growth rate defects or extended lag phases during liquid culture experiments (Table S2 and Figure S11). The remaining three genes may represent false positives from the Tn-seq screen, or may reflect differences in the growth conditions, namely, competitive growth versus isogenic growth. Nevertheless, the observation that at least 75% of the selected genes were confirmed to have a genome content-dependent fitness phenotype validates that the large majority of the strain specific phenotypes observed in the Tn-seq screen represent true differences.

Level of genetic and phenotypic conservation of the essential *S. meliloti* genes.

Several recent studies have used Tn-seq to study the essential genome of *Rhizobium leguminosarum* [31-33]. We compared our Tn-seq datasets with those reported in by Perry *et al* [32] to examine the conservation of the essential genome of these two closely related N_2-fixing species. Putative orthologs for ~ 75% of all *S. meliloti* chromosomal genes were identified in *R. leguminosarum* via a Blast Bidirectional Best Hit (Blast-BBH) approach (Data Set S2). Much higher conservation of the growth promoting genome was observed; 97% of the 489 core growth promoting genes and 99% of the 307 core essential genes had a putative ortholog in *R. leguminosarum*. However, conservation of the gene did not necessarily correspond to conservation of the phenotype. Considering only the 303 conserved core essential *S. meliloti* genes (as these were the least likely to have been falsely identified as essential), 8% (25 of 303) of their orthologous genes were classified as having little contribution to growth on defined medium in *R. leguminosarum* (Figure 3A). An additional 34 genes were considered to be non-essential but growth defective when mutated (Figure 3A). Independent mutation of two genes (*fumC*, *pdxJ*) identified as specifically essential in *S. meliloti* confirmed their essentiality (Figure S4), supporting the Tn-seq data. A similar pattern is observed starting with the *R. leguminosarum* genes classified as essential in both minimal and complex medium by Perry *et al*. [32]. Of the 241 core essential *R. leguminosarum* genes with an ortholog on the *S. meliloti* chromosome, 21 (9%) of the orthologs were classified as non-essential in *S. meliloti* for growth in defined medium, while an additional 8 were considered to have a moderate growth defect (Figure 3B).

To further test the species specificity of the above-mentioned genes, the experiment was replicated *in silico*. Fifteen of the 25 orthologs specifically essential in *S. meliloti* were present both in our existing *S. meliloti* genome-scale metabolic model [34] as well as in a draft *R. leguminosarum* metabolic model (see Materials and Methods). Flux balance analysis was used to examine the *in silico* effect of deleting these 15 pairs of orthologs on growth. Three pairs of orthologs were classified as essential in both models, five were classified as non-essential in both models, and seven were classified as essential specifically in the *S. meliloti* model. Thus, at least half of the gene essentiality differences observed in the Tn-seq data are corroborated by the *in silico* metabolic simulation, despite the preliminary nature of the draft *R. leguminosarum* model.

An *in silico* analysis of the genes identified as specifically essential in *R. leguminosarum* on the basis of the Tn-seq data was not performed as only two of these genes were present in the *R. leguminosarum* model.

In silico analyses support a high potential for genetic redundancy in the S. meliloti genome.

The results of the previous two sections are consistent with a strong genomic environment effect on the phenotypic consequences of gene mutations. One possible explanation is the presence of widespread genetic redundancy, at the gene and/or pathway level. In support of this, ~ 14% of chromosomal genes had a Blast-BBH hit when the chromosomal proteome was compared against the combined pSymA/pSymB proteome (Data Set S3). Therefore, this phenomenon was further explored using a constraint-based metabolic modelling approach.

We first tested the *in silico* effect of chromosomal single gene deletions on growth rate in the presence and absence of pSymA/pSymB (Figure 4A). This analysis identified 67 genes (~ 7% of all chromosomal model genes) as having a more severely impaired growth phenotype when deleted in the absence of pSymA/pSymB genes, 38 of which were lethal. This appeared to be due to a combination of direct functional redundancy of the gene products as well as through metabolic bypasses, as deletion of 50 reactions dependent on chromosomal genes had a more severe phenotype in the absence of pSymA/pSymB, 42 of which were lethal (Figure S12).

Next, a double gene deletion analysis was performed to examine the effect on growth rate of deleting every possible pair of model genes. This analysis suggested that 49 chromosomal genes had a more significant impact on growth than expected when simultaneously deleted with a single pSymA or pSymB gene (Figure 4B). Additionally, synthetic negative phenotypes were observed for 97 chromosomal genes when simultaneously deleted with another chromosomal gene (Figure 4C). Overall, 14% of chromosomal genes were predicted to have a synthetic negative phenotype when co-deleted with a second gene, consistent with a high potential for metabolic robustness being encoded by the *S. meliloti* genome, and with a significant influence of the genomic environment on the fitness phenotype of gene mutations.

A consolidated view of core *S. meliloti* metabolism through Tn-seq-guided *in silico* metabolic reconstruction.

The results described in the previous sections made it evident that a Tn-seq approach alone is insufficient to elucidate all processes contributing to growth in a particular environment. This is especially true if also considering non-essential metabolism that is nevertheless actively present in wild type cells, such as exopolysaccharide production. Moreover, it is difficult to fully comprehend the core functions of a cell by simply examining a list of essential genes and their predicted functions. We therefore attempted to overcome these limitations by using the Tn-seq data to guide a manual *in silico* reconstruction of the core metabolic processes of *S. meliloti*. A detailed description of this process is provided in the Materials and Methods. In brief, the existing metabolic model iGD1575 was treated as a database of reactions and gene-reaction associations. Each pathway involved in central carbon metabolism or the production of essential or non-essential biomass components (Table S3) were then rebuilt in a new (initially empty) reconstruction drawing from the reactions present in iGD1575. At the same time, the genes associated with each reaction were compared to the Tn-seq data and published literature to confirm the linkage of the correct gene(s) to each reaction.

The resulting model, termed iGD726 and included as in SBML format in File S2, is summarized in Figure 5 and Table 3, and the entire model including genes, reaction formulas, and references is provided as an easy to read Excel table in Data Set S4. The process of integrating the Tn-seq data with *in silico* metabolic reconstruction resulted in a major refinement of the core metabolism compared to the existing genome-scale model: 228 new reactions were added, 115 new genes were added, and the genes associated with 135 of the 432 reactions common to both reconstructions were updated. In addition to improving the metabolic reconstruction, this process significantly expanded the view of core *S. meliloti* metabolism compared to that gained solely through the application of Tn-seq. The genes associated with approximately one third of the iGD726 reactions were not detected as growth promoting in the Tn-seq datasets (Figure 5, Table 3). While many of the additional reactions present in iGD726 are due to the inclusion of non-essential biomass components, which are part of the wild type cell but are nonetheless dispensable for growth, others are from essential metabolic pathways (Figures 5, S13). Overall, the combined approach of integrating Tn-seq data and *in silico* metabolic modelling allowed for the development of a high-quality representation of core *S. meliloti* metabolism in a way that neither approach alone was capable of accomplishing.

Tn-seq-guided *in silico* metabolic reconstruction facilitates novel gene annotation.

Over 20 of the reactions of the core metabolic reconstruction initially had no gene attributed with producing the enzyme responsible for its catalysis. Similarly, many genes with no clear biological function were found to be essential in the Tn-seq screen. By attempting to fill the gaps in the *in silico* model with the uncharacterized essential genes, we were able to assign putative functions to eight previously uncharacterized genes (Table S4). Two of these genes were chosen for further characterization, *smc01361* and *smc04042*. The *smc01361* gene was annotated as encoding a dihydroorotase, and mutation of *smc01361* resulted in pyrimidine auxotrophy (Figure S14). Given its location next to *pyrB*, and the presence of an essential PyrC dihydroorotase encoded elsewhere in the genome (Data Set S1), we propose that *smc01361* encodes an inactive dihydroorotase (PyrX) required for PyrB activity as has been observed in some other species including *Pseudomonas putida* [35,36]. The essential *smc04042* gene was annotated as an inositol-1-monphosphatase family protein. It was previously observed that rhizobia lack a gene encoding a classical L-histidinol-phosphate phosphohydrolase, and it was suggested an inositol monophosphatase family protein may fulfill this function instead [37]. Mutation of *smc04042* resulted in histidine auxotrophy (Figure S14), consistent with this enzyme fulfilling the role of a L-histidinol-phosphate phosphohydrolase. It is likely that this is true for most rhizobia, as putative orthologs of this gene were identified in all 10 of the examined Rhizobiales genomes (Data Set S4). These examples illustrate the power of the combined Tn-seq and metabolic reconstruction process in the functional annotation of bacterial genomes.

DISCUSSION

In this study, we developed a new variant of the Tn*5* transposon for construction of non-polar insertion mutations that should be readily adaptable for use with other α-proteobacteria. The Tn*5* transposon was chosen as it was expected to have low insertion site specificity in *S. meliloti* [38]. However, we observed a moderate sequence insertion bias for GC rich regions consistent with previous studies of the Tn*5* transposon [39-41]. The consensus sequence of ~ 190,000 unique insertion locations was largely consistent, but not identical, with that previously reported [41]; however, the specificity appeared to extend past the 9 base pair region that is duplicated during Tn*5* insertion (Figure S3). While this bias is unlikely to have a significant influence on the results in species with high GC content genomes, such as *S. meliloti*, accounting for this bias may be important when applying Tn*5* mutagenesis to species with low GC content genomes.

Greater than 10% of species with a sequenced genome contain a genomic architecture similar to *S. meliloti*, that is, with at least two large DNA replicons [42,43]. Several studies have revealed that, in many ways, each replicon acts as a functionally and evolutionarily distinct entity (for a review, refer to [43]); yet, there can also be regulatory cross-talk [44], as well as the exchange of genetic material between the replicons [45]. The Tn-seq analyses reported here provide new insights into the functional integration of secondary replicons into the host organism. The pSymB replicon of *S. meliloti* is known to have two essential genes (which were transferred to the chromosome for this study) [45], while pSymA has no essential genes [46]. However, the large number of chromosomal genes—across many functional groups (Figure 2)—that became conditionally essential following the removal of pSymA and pSymB indicate the presence of many genes whose products can perform essential metabolic capabilities but that remain cryptic due to inter-replicon epistatic interactions. It was also interesting to note that the strength of the correlation between duplicates (Figure S5), as determined by the size of the absolute residuals, was higher for the ΔpSymAB strain than for the wild type strain in both media (p-value $< 2.2 \times 10^{-16}$ for both media, as determined with Welch two-sample t-tests). This may be reflective of the genetic robustness encoded by the secondary replicons and the stochastic activation of these processes in the mutant population. Potentially, the high level of inter-replicon redundancy may reduce the level of purifying selection on the chromosomal copies of the genes, facilitating more rapid diversification of gene functionality and increased rates of chromosomal gene evolution. Overall, the results of these analyses suggest that secondary replicons may influence the evolution of the chromosome and play a vital role in the biology of the organism, even if these activities remain cryptic due to inter-replicon epistatic interactions.

More generally, the Tn-seq data reported here provide a unique perspective of how a gene's genomic environment influences its genotype-phenotype relationship. Previous studies have illustrated that the fitness phenotypes of orthologous genes of both distant and closely related species may differ [21,23-28,47], and even how intercellular effects within microbial communities can modify the essential genome of a species [48]. The data reported here more directly addressed the influence of the genomic environment by comparing the fitness phenotypes of mutating the exact same set of ~ 3,500 genes in two very different genomic environments. It was found that the non-essential genome had a remarkable influence on what was classified as a growth-promoting gene, with 10% of *S. meliloti* chromosomal genes exhibiting fitness-based genetic interactions with the non-essential component of the genome (Figure 2). This observation was not growth medium-dependent, was not unique to a specific gene functional class, and was not simply due to an overall reduced fitness of the ΔpSymAB strain as the findings could be largely replicated *in silico* (Figure 4).

The majority of the genes whose fitness phenotype was dependent on the genomic environment became more important for fitness following the genome reduction. In many cases, this is expected to reflect a loss of functional redundancy; the increased importance of the chromosomal cytochrome genes likely reflects a compensation for the loss of the pSymA/pSymB encoded cytochrome complexes (Figure 6). In other cases, it may reflect newly activated pathways that must compensate for the loss of a normal housekeeping pathway. The specific essentiality of proline biosynthesis, and the second half of histidine biosynthesis, in the ΔpSymAB strain during growth in rich medium presumably reflects the inability of these strains to transport these compounds and must therefore synthesize them *de novo* (Figure 6). Indeed, previous metabolomics work is consistent with the ΔpSymAB strain being unable to transport many amino acids, including proline and histidine [49]. Similarly, glycolysis appeared specifically essential in the ΔpSymAB strain in rich medium (Figure 6), likely as the reduced metabolic capacity of this strain [29] led to a greater reliance on catabolism of the abundant sucrose for energy and biosynthetic precursors. Specific gene essentiality in the ΔpSymAB background may also occur as a result of synthetic negative interactions that are not associated with metabolic redundancy, for example, synthetic effects of disrupting two independent aspects of the cell envelope. This may be reflected in the specific essentiality of the *feuNPQ* and *ndvAB* genes involved in production of periplasmic cyclic β-glucans (Figure 6) [50-53]. The cell envelope of the ΔpSymAB strain is altered compared to the wild-type, due to the loss of succinoglycan production [54] and the *bacA* gene [55], and the membrane lipid composition contains signs of increased stress [49]. The fitness of disrupting periplasmic cyclic β-glucans biosynthesis in this background, further altering the cell envelope, may therefore represent a synthetic negative interaction.

Somewhat surprisingly, approximately a quarter of the genes with a genomic environment effect had a greater fitness defect in the wild type strain. In some cases this may have been due to the reduced nutrient demand of the ΔpSymAB strain as a result of the smaller genome content. For example, mutations of genes for arginine biosynthesis and the biosynthesis of AICAR and UMP, common precursors in the synthesis of purines and pyrimidines, respectively, had fitness defects in rich medium specifically in the wild-type (Figure 6). This may reflect that in this environment, the uptake of these nutrients is growth limiting to the wild-type in the absence of their *de novo* synthesis, whereas this is not the case in the ΔpSymAB strain due to the reduced genome size, and thus lower nutrient requirement, and the already reduced growth rate (Figure S8). Another possibility is that removal of pSymAB evokes phenotypes that are epistatic to many of those brought about by chromosomal mutations. For example, the removal of pSymB is expected to have resulted in alterations of the cell membrane [49,54,55]; our observation that many mutations causing greater relative fitness defects in wild-type cells are associated with lipid metabolism, such as biosynthesis of the lipopolysaccharide core oligosaccharide (Figure 6) may be a result of those mutations being phenotypically masked in the absence of pSymB.

Our work in integrating the Tn-seq data with *in silico* metabolic modelling made it evident that Tn-seq alone is insufficient to identify the entire core metabolism of an organism; almost a third of the reactions present in the core metabolic reconstruction were not supported by Tn-seq data (Figure 5 and Table 3). Similarly, the large number of changes made in the gene-reaction relationships when producing the core model illustrated the limitations in the quality of metabolic reconstructions when high-throughput mutagenesis data are lacking. In some cases, the gaps in the Tn-seq data were due to genomic environment effects, such as genetic redundancy, in other cases it was due to the inclusion of reactions that are non-essential but that are nonetheless required for production of 'wild type' cells, and sometimes the gene associated with a reaction is simply unknown. A fourth possibility is phenotypic complementation through cross-feeding. Given that Tn-seq involves growth of a population of mutants, a mutant unable to produce an essential metabolite may still grow if the metabolite is excreted and transferred to the mutant from the rest of the population.

Regardless of the reasons why Tn-seq may have missed so many central metabolic reactions, this limitation can have a significant practical impact in the modern era of synthetic biology. The results of Tn-seq studies may be used to guide engineering of designer microbial factories with specific properties [56], or for the identification of putative new therapeutic targets

[25,57]. While Tn-seq studies undoubtedly give invaluable information to be used towards these goals, basing engineered cells solely on Tn-seq studies is insufficient, as evidenced in the recent monumental efforts to design and synthesize a minimal bacterial genome [22] . Importantly, this limitation can be overcome by combining Tn-seq with metabolic modelling. We are aware of only a few other studies making use of both Tn-seq data and metabolic reconstruction [58-62]; however, these studies almost always focus on using the Tn-seq data to refine the metabolic reconstruction. As illustrated here, combining an experimental Tn-seq approach with a ground-up *in silico* metabolic reconstruction strategy can improve not only the reconstruction but also overcome the limitations of the Tn-seq approach. A Tn-seq-guided reconstruction process forces the identification of missing essential reactions, while ensuring correct gene-reaction associations, and the integrated approach can facilitate functional annotation of genes without clear biological roles. This process allows one to obtain a very high-quality representation of the metabolism, and the underlying genetics, of the organism in the given environment. The resulting model can serve as a blueprint to simply understand the workings of the cell, or as a basis for developing new cell factories.

MATERIALS AND METHODS

Bacterial strains, media, and growth conditions.

The wild type and ΔpSymAB strains used throughout this work are the RmP3499 and RmP3496 strains, respectively, whose construction was described previously [30]. All *E. coli* or *S. meliloti* strains used in this study are described in Table S5 and were grown at 37°C or 30°C, respectively. BRM medium was used as the rich medium for growth of the *S. meliloti* strains, and it consisted of 5 g/L Bacto Tryptone, 5 g/L Bacto Yeast Extract, 50 mM NaCl, 2 mM

$MgSO_4$, 2 µM $CoCl_2$, 0.5% (w/v) sucrose, and supplemented with the following antibiotics, as appropriate: streptomycin (Sm, 200 µg/ml), neomycin (Nm, 100 µg/ml), gentamycin (Gm, 15 µg/ml). The defined medium for growth of *S. meliloti* contained 50 mM NaCl, 10 mM KH_2PO_4, 10 mM NH_4Cl, 2 mM $MgSO_4$, 0.2 mM $CaCl_2$, 0.5% (w/v) sucrose, 2.5 µM thiamine, 2 µM biotin, 10 µM EDTA, 10 µM $FeSO_4$, 3 µM $MnSO_4$, 2 µM $ZnSO_4$, 2 µM H_3BO_3, 1 µM $CoCl_2$, 0.2 µM Na_2MoO_4, 0.3 µM $CuSO_4$, 50 µg/ml streptomycin, and 30 µg/ml neomycin. *E. coli* strains were grown on Luria-Bertani (LB) supplemented with the following antibiotics as appropriate: chloramphenicol (30 mg/ml), kanamycin (Km, 30 µg/ml), gentamycin (Gm, 3 µg/ml).

Growth curves.

Overnight cultures grown in rich media with the appropriate antibiotics were pelleted, washed with a phosphate buffer (20 mM KH_2PO_4 and 100 mM NaCl), and resuspended to an OD_{600} of 0.25. Twelve µl of each cell suspension was mixed with 288 µl of growth medium, without antibiotics, in wells of a 100-well Honeycomb microplate. Plates were incubated in a Bioscreen C analyzer at 30°C with shaking, and OD_{600} recorded every hour for at least 48 hours.

***S. meliloti* mutant construction for Tn-seq validation.**

Single gene knockout mutants were generated through single cross-over plasmid integration of the suicide plasmid pJG194 [63]. Approximately 400-bp fragments homologous to the central portion of the target genes were PCR amplified using the primers listed in Table S6. PCR products as well as the pJG194 and pJG796 vectors were digested with the restriction enzymes *Eco*RI/*Hind*III, *Bam*HI/*Xba*I, or *Sal*I/*Xho*I, and each PCR fragment was ligated into the appropriately digested pJG194 or pJG796 vector using standard molecular biology techniques [64], and all recombinant plasmids verified. Recombinant plasmids were mobilized from *E. coli* to *S. meliloti* via tri-parental matings as described before [52], and transconjugants isolated on BRM Sm Nm agar plates. All *S. meliloti* mutants were verified by PCR.

Transduction of the integrated plasmids into the *S. meliloti* wild type and ΔpSymAB strains was performed using phage N3 as described elsewhere [65], with transductants recovered on BRM medium containing the appropriate antibiotics.

Construction of the transposon delivery vector pJG714

The plasmid pJG714 is a variant of the previously reported mini-Tn*5* delivery plasmid, pJG110 [63], with the primary modifications being removal of the *bla* gene and pUC origin of replication, and introduction of the *pir*-dependent R6K replication origin. A map of pJG714 is given in Figure S1A, and the complete sequence of the transposable region is provided in Figure S1B. This delivery plasmid is maintained in *E. coli* strain MFD*pir* [66], which possesses chromosomal copies of R6K *pir* and RK2 transfer functions. MFD*pir* is unable to synthesize diaminopimelic acid (DAP), thus disabling growth on rich or defined medium lacking supplemental DAP. The MFD*pir*/pJG714 strain is cultured on rich medium containing kanamycin and 12.5 μg/ml DAP.

Tn-seq experimental setup.

Transposon mutagenesis was accomplished in the wild-type and ΔpSymAB strains in parallel. Flask cultures of MFD*pir*/pJG714 and the two *S. meliloti* strains were grown overnight to saturation, and pellets were washed and suspended in BRM to a final OD_{600} value of approximately 40. Equal volumes of each suspension were mixed as bi-parental matings, to accomplish mobilization of the transposon delivery vector into the *S. meliloti* recipient strains. These cell mixtures were plated on BRM supplemented with 50 μg/ml DAP and incubated at 30°C for 6 h. Mating mixtures were collected in BRM with 10% glycerol, and cell clumps were broken up by shaking the suspended material for 30 min at 225 rpm. Aliquots were stored at -80°C. For selection of transposants, mating mixes were thawed and plated at a density of 15,000 cfu/plate (150-mm plates) on BRM supplemented with Sm and Nm. To accomplish equivalent coverage of each genome with transposon insertions, 675,000 and 360,000 colonies were selected for the wild-type and ΔpSymAB strains, respectively. For each recipient, transposon mutant colonies were collected and cell clumps were broken up as described above. The selected clone libraries were aliquoted and stored at -80°C.

For whole-population selection and massively parallel sequencing of transposon ends, 1x10^9 cells from each of the two clone libraries were transferred into 500 ml of either BRM or defined medium, allowing approximately 8-10 generations of growth at 30°C before reaching saturation. At this stage, cells were pelleted, DNA was extracted using the MoBio microbial DNA isolation kit (#12255-50), and the resulting DNA was fragmented with NEB fragmentase (#M0348S) to an average molecular weight of 1000 bp. After clean-up (Qiagen #27106), the resulting DNA fragments were appended with short 3' homopolymer (oligo-dCTP) tails using terminal deoxynucleotidyl transferase (NEB #M0315S), and this sample was used as the template for a two-round PCR that gave rise to the final Illumina-ready libraries. In the first round, a transposon end-specific primer (1TN) and oligo-G primer (1GG) were used (all primer sequences can be found in Table S6). After clean-up, a portion of the first-round product was used as the template for the second-round reaction employing a nested transposon-specific primer (2TNA-C) and a reverse index-incorporating primer (2BAR01-08). The series of three 2TN primers (A-C) were designed to incorporate base diversity in the opening cycles of Illumina sequencing, and the series of eight 2BAR primers were designed to uniquely identify each experimental condition in a single multiplexed sequencing sample. After PCR amplification of transposon-flanking sequences with concomitant incorporation of Illumina adapters and barcodes, the samples were size-selected for 200-600-bp fragments, and sequenced on an Illumina Hi-Seq instrument as 50-bp single-end reads. Raw reads were used as input into a custom-built Tn-seq analytical pipeline, which was recently described [57].

Calculation of gene and synthetic indexes.

For calculation of Gene Essentiality Index (GEI) scores, a pseudo count of one was first added to all gene read counts for each replicate. GEI were then calculated by summing the number of reads that mapped to the gene in both replicates, and dividing this number by the nucleotide length of the gene. GEI scores were calculated for each gene separately in each medium and in each strain. All GEI values are available in Data Set S1.

Synthetic Media Index (SMI) scores were calculated to represent the difference in GEI scores between the two media for the same strain. Raw SMI scores were determined by dividing the GEI of the gene in defined medium by the GEI of the gene in rich medium. Processed SMI scores, those shown throughout the manuscript, were determined as follows. If the raw value was above one, the processed SMI and the raw SMI are the same. Raw SMI scores that were below one were converted to processed SMI scores through the transformation, “1 / raw SMI score”, and presenting the value as a negative number.

Raw and processed Synthetic Rich Index (SRI) and Synthetic Defined Index (SDI) scores were calculated to represent the difference in the GEI scores of a gene between the wild-type and ΔpSymAB strains when grown in rich or defined medium, respectively. SRI and SDI indexes were calculated using the same procedure as described for the SMI scores above. All synthetic index scores are provided in Data Set S1.

Statistical analysis of the Tn-seq output.

The output of the Tn-seq analysis pipeline was used in the fitness classification of genes as follows. First, all genes with no observed insertions were classified as essential. Next, GEI scores were imported into R version 3.2.3 and log transformed. Initial clustering of the log transformed GEI scores into fitness categories was performed using the *Mclust* function of the *Mclust* package in R [67]. In short, this function attempts to explain the distribution of GEI values by fitting a series of overlapping Guassian distributions, with the number and shape of the distributions determined by *Mclust*. The data are then assigned to different categories based on the probability of the data point arising from each of the distributions. As high uncertainty in the classification of genes at the borders of groupings exists, the clusters were refined through the use of affinity propagation implemented by the *apcluster* function of the *apcluster* package of R [68]. All genes belonging to an *apcluster* grouping that contained an essential gene, as determined in any of the previous steps, were re-annotated as essential. Additionally, all genes belonging to an *apcluster* grouping that spanned the border of two *Mclust* goups were transferred to the same classification, based on which cluster the genes had a higher median probability of being derived from in the *Mclust* analysis. Finally, genes that were classified as 'essential' in one medium and 'large growth impairment' in the second medium, but that were identified as having no medium specificity based on their SMI scores, were considered as essential in both media.

Genes with GEI scores significantly different between conditions were determined as follows. The synthetic indexes (SMI, SDI, SRI) scores were imported into R and log transformed, and the following clustering performed independently for each index. The log transformed synthetic scores were clustered using *Mclust* and *apcluster* in R as described above for the GEI scores. In the case of the SMI scores, three clusters were produced 'Little to no difference', 'Moderate difference', and 'Large difference'; only genes with a SMI scores classified as 'Large difference' were considered to display a medium specificity. In the case of SDI and SRI scores, only two clusters were produced: 'Little to no difference' and 'Difference between strains'.

Gene functional enrichments.

Assignment of chromosomal genes into specific functional categories was performed largely based on the annotations provided on the *S. meliloti* Rm1021 online genome database (https://iant.toulouse.inra.fr/bacteria/annotation/cgi/rhime.cgi). This website pulls annotations from several databases including PubMed, Swissprot, trEMBL, and Interpro. Additionally, it provides enzyme codes, PubMed IDs, functional classifications, and suggested Gene Ontology (GO) terms for most genes. The numerous classifications were simplified to 18 functional categories, designed to adequately cover all core cellular processes. Occasionally, ambiguous or conflicting annotations were observed. In these cases, protein BLASTp searches through the NCBI server were performed against the non-redundant protein database. If putative domains were detected within the amino acid sequence, a combination of the best hit (lowest E-value) and consensus among domain annotations were used to categorize the gene in question. If no putative domains were detected, the functional annotation was based on the best scoring protein hits in the database. The functional annotations of all chromosomal genes are provided in Data Set S5.

Data visualization.

Tn-seq results were visualized using the Integrative Genomics Viewer v2.3.97 [69]. Scatter plots, functional enrichment plots, box plots, and line plots were generated in R using the *ggplot2* package [70]. Venn diagrams were produced in R using the *VennDiagram* package [71]. The genome map was prepared using the circos v0.67-7 software [72]; the sliding window insertion density was calculated with the *geom_histogram* function of *ggplot2*, and the GC skew was calculated using the analysis of sequence heterogeneity sliding window plots online webserver [73]. The metabolic model was visualized using the iPath v2.0 webserver [74]. The logo of the transposon insertion site specificity was generated by first extracting the nucleotides surrounding all unique insertion sites in one replicate of the wild-type grown in rich medium using Perl v5.18.2, followed by generation of a hidden Markov model with the *hmmbuild* function of HMMER v3.1b2 [75] and visualization with the Skylign webserver [76].

Blast Bidirectional Best Hit (Blast-BBH) strategy.

Putative orthologous proteins between species were identified with a Blast-BBH approach, implemented using a modified version of our in-house Shell and Perl pipeline [77]. This pipeline involved GNU bash v4.3.48(1), Perl v5.22.1, Python v2.7.12 and the Blast v2.6.0+ software [78]. Proteomes were downloaded from the National Center for Biotechnology Information repository, and the Genbank annotations were used. As a threshold to limit false positives, Blast-BBH pairs were only maintained if they displayed a minimum of 30% amino acid identify over at least 60% of the protein. To identify putative duplicate proteins in *S. meliloti*, the same Blast-BBH approach was employed to compare the *S. meliloti* chromosomal proteome with the proteins encoded by pSymA and pSymB.

In silico metabolic modeling procedures.

All simulations were performed in Matlab 2017a (Mathworks) with scripts from the Cobra Toolbox (downloaded May 12, 2017 from the openCOBRA repository) [79], and using the Gurobi 7.0.2 solver (www.gurobi.com), the SBMLToolbox 4.1.0 [80], and libSBML 5.15.0 [81]. Boundary conditions for simulation of the defined medium are given in Table S7. *In silico* analysis of redundancy in the *S. meliloti* genome was performed using the iGD1575b metabolic reconstruction, whose development is described in the following section. Single and double gene deletion analyses were performed using the *singleGeneDeletion* and *doubleGeneDeletion* functions, respectively, using the Minimization of Metabolic Adjustment (MOMA) method. All Matlab scripts used in this work are provided as File S3.

For all deletion mutants, the growth rate ratio (grRatio) was calculated (growth rate of mutant / growth rate of wild type). Single gene deletion mutants were considered to have a growth defect if the grRatio was < 0.9. For the double gene deletion analysis, if the grRatio of the double mutant was less than 0.9 the expected grRatio (based on multiplying the grRatio of the two corresponding single mutants), the double deletion was said to have a synthetic negative phenotype.

Development of iGD1575b.

For *in silico* analysis of redundancy in the *S. meliloti* genome, the previously published *S. meliloti* genome-scale metabolic model iGD1575 [34] was modified slightly. As indicated in Table S8, the biomass composition was updated to include 31 additional compounds at trace concentrations, including vitamins, coenzymes, and ions, in order to ensure the corresponding transport or biosynthetic pathways were essential. However, the original model iGD1575 was unable to produce vitamin B12 and holo-carboxylate. To rectify this, the reversibility of rxn00792_c0 was changed from ‘false’ to ‘true’, and the reactions rxn01609, rxn06864, and rxnBluB were added to the model. However, no new genes were included in the model. This updated model was termed iGD1575b and is available in SBML and Matlab format in File S2.

Simulating the removal of pSymA and pSymB *in silico*.

Several modifications to iGD1575b were required in order to produce a viable model following the deletion of all pSymA and pSymB genes. As described previously [34], succinoglycan was removed from the biomass composition, 'gapfill' GPRs (gene-protein-reaction relationships) were added to the reactions 'rxn01675_c0', 'rxn01997_c0', 'rxn02000_c0', and 'rxn02003_c0' in order to allow the continued production of the full LPS molecule, as well as to 'rxn00416_c0' to allow asparagine biosynthesis. Additionally, 'gapfill' GPRs were added to the reactions 'rxn03975_c0' and 'rxn03393_c0' so that removal of pSymA and pSymB did not prevent biosynthesis of vitamin B12 and ubiquinone-8, respectively. Finally, a glycerol export reaction via diffusion (rxnBLTPcpd00100b) was added to remove the glycerol build-up resulting from cardiolipin biosynthesis. The modified version of the model was termed iGD1575c, and is available in in SBML and Matlab format in File S2. For simulating the removal of pSymA and pSymB in Matlab, all pSymA and pSymB genes were deleted from the iGD1575b model using the *deleteModelGenes* function, followed by the removal of all constrained reactions using the *removeRxns* function.

Building the draft *R. leguminosarum* metabolic model.

A draft, fully automated model containing no manual curation for *R. leguminosarum* bv. *viciae* 3841 was built using the KBase webserver (www.kbase.us). The Genbank file (GCA_000009265.1_ASM926v1_genomic.gbff) of the *R. leguminosarum* genome [82] was uploaded to KBase and re-annotated using the 'annotate microbial genome' function, maintaining the original locus tags. An automated metabolic model was then built using the 'build metabolic model' function, with gap-filling. This model included 1537 genes, 1647 reaction, and 1731 metabolites, and is available in in SBML and Matlab format in File S2. The biomass composition was not modified from the default Gram negative biomass of Kbase. All essential model genes were determined using the Cobra Toolbox in Matlab with the *singleGeneDeletion* function and the MOMA protocol, with exchange reaction bounds set as provided in Table S7.

Building the *S. meliloti* core metabolic reconstruction, iGD726.

The iGD726 model was built from the ground-up using the existing iGD1575 model as a reaction and GPR database, and with the Tn-seq data as a guide. Each metabolic pathway included in iGD726 was rebuilt in a new file by adding individual reactions to the file. These reactions were taken from iGD1575, or were taken from other sources, primarily the Kyoto Encyclopedia of Genes and Genomes (KEGG) database [83], if an appropriate reaction was missing in iGD1575. Following the transfer of each reaction, the genes associated with the reaction were checked against the Tn-seq data, and a literature search for each associated gene was performed. The gene associations were then modified as necessary to ensure the model accurately captured the experimental data. For example, if gene was experimentally determined to be essential, but the corresponding reaction for the gene was associated with multiple alternative genes, all but the essential gene were removed from the reaction. Similarly, if a non-essential gene was associated with an essential reaction, a second gene or an Unknown was added to reflect the apparent redundancy in the genome. Where possible, unknowns in the gene associations were replaced with genes whose gene product may catalyze the reaction.

During the construction of the core model, the biomass composition was updated. This included modifying the membrane lipid composition to include lipids with different sized fatty acids based on the ratio experimentally determined [84]; the original iGD1575 model contained only one representative per each membrane lipid class. Additionally, essential vitamins, cofactors, and ions were added to the biomass composition at trace concentrations to ensure that their biosynthesis or transport was essential. The complete biomass composition is provided in Table S3.

The necessary metabolic and transport reactions to allow the model to growth with sucrose, glucose, or succinate were included in the reconstruction. Once the model was capable of producing all biomass components using any of the three carbon sources, the list of model genes was compared with the list of 489 core growth promoting genes to identify genes not included in the model but experimentally determined to contribute to growth. When possible, missing genes and their corresponding reactions were added to the core model. The final model contained 726 genes, 681 reactions, and 703 metabolites, and is provided in SBML and Matlab format in File S2, and as an Excel file in Data Set S4. The Excel file contains all necessary information for use as a *S. meliloti* metabolic resource, including the reaction name, the reaction equation using the real metabolite names, the associated genes/proteins, and references. Additionally, for each reaction, the putative orthologs of the associated genes in 10 related Rhizobiales species are included, allowing the model to provide useful information for each of these organisms.

FUNDING INFORMATION

This work was funded by National Science Foundation grant IOS-1054980 to Joel S Griffitts. George C diCenzo was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC) through a PDF fellowship. Work in the laboratory of Turlough M Finan was funded by NSERC. The funders had no role in the study design, data collection and interpretation, or the decision to submit the work for publication.

REFERENCES

1. Orgogozo V, Morizot B, Martin A. The differential view of genotype-phenotype relationships. Front Genet. 2015;6: 179. doi:10.3389/fgene.2015.00179

2. Ritchie MD, Holzinger ER, Li R, Pendergrass SA, Kim D. Methods of integrating data to uncover genotype-phenotype interactions. Nat Rev Genet. 2015;16: 85–97. doi:10.1038/nrg3868

3. Palsson B. Metabolic systems biology. FEBS Lett. 2009;583: 3900–3904. doi:10.1016/j.febslet.2009.09.031

4. Durot M, Bourguignon P-Y, Schachter V. Genome-scale models of bacterial metabolism: reconstruction and applications. FEMS Microbiol Rev. 2009;33: 164–190. doi:10.1111/j.1574-6976.2008.00146.x

5. O'Brien EJ, Monk JM, Palsson BØ. Using genome-scale models to predict biological capabilities. Cell. 2015;161: 971–987.

6. van Opijnen T, Bodi KL, Camilli A. Tn-seq: high-throughput parallel sequencing for fitness and genetic interaction studies in microorganisms. Nat Methods. 2009;6: 767–772. doi:10.1038/nmeth.1377

7. van Opijnen T, Camilli A. Transposon insertion sequencing: a new tool for systems-level analysis of microorganisms. Nature Rev Microbiol. 2013;11: 435–442. doi:10.1038/nrmicro3033

8. Thiele I, Palsson BØ. A protocol for generating a high-quality genome-scale metabolic reconstruction. Nat Protoc. 2010;5: 93–121. doi:10.1038/nprot.2009.203

9. Orth JD, Thiele I, Palsson BØ. What is flux balance analysis? Nat Biotechnol. 2010;28: 245–248. doi:10.1038/nbt.1614

10. Fowler ZL, Gikandi WW, Koffas MAG. Increased malonyl coenzyme A biosynthesis by tuning the *Escherichia coli* metabolic network and its application to flavanone production. Appl Environ Microbiol. 2009;75: 5831–5839. doi:10.1128/AEM.00270-09

11. Pratapa A, Balachandran S, Raman K. Fast-SL: an efficient algorithm to identify synthetic lethal sets in metabolic networks. Bioinformatics. 2015;31: 3299–3305. doi:10.1093/bioinformatics/btv352

12. Chao MC, Abel S, Davis BM, Waldor MK. The design and analysis of transposon insertion sequencing experiments. Nature Rev Microbiol. 2016;14: 119–128. doi:10.1038/nrmicro.2015.7

13. Thomaides HB, Davison EJ, Burston L, Johnson H, Brown DR, Hunt AC, et al. Essential bacterial functions encoded by gene pairs. J Bacteriol. 2007;189: 591–602. doi:10.1128/JB.01381-06

14. diCenzo GC, Finan TM. Genetic redundancy is prevalent within the 6.7 Mb *Sinorhizobium meliloti* genome. Mol Genet Genomics. 2015;290: 1345–1356. doi:10.1007/s00438-015-0998-6

15. Bergmiller T, Ackermann M, Silander OK. Patterns of evolutionary conservation of essential genes correlate with their compensability. PLOS Genet. 2012;8: e1002803. doi:10.1371/journal.pgen.1002803

16. Liu G, Yong MYJ, Yurieva M, Srinivasan KG, Liu J, Lim JSY, et al. Gene essentiality is a quantitative property linked to cellular evolvability. Cell. 2015;163: 1388–1399.

17. Butland G, Babu M, Díaz-Mejía JJ, Bohdana F, Phanse S, Gold B, et al. eSGA: *E. coli* synthetic genetic array analysis. Nat Methods. 2008;5: 789–795. doi:10.1038/nmeth.1239

18. Costanzo M, Baryshnikova A, Bellay J, Kim Y, Spear ED, Sevier CS, et al. The genetic

landscape of a cell. Science. 2010;327: 425–431. doi:10.1126/science.1180823

19. Phillips PC. Epistasis—the essential role of gene interactions in the structure and evolution of genetic systems. Nat Rev Genet. 2008;9: 855–867. doi:10.1038/nrg2452

20. Juhas M. On the road to synthetic life: the minimal cell and genome-scale engineering. Crit Rev Biotechnol. 2016;36: 416–423. doi:10.3109/07388551.2014.989423

21. Juhas M, Reuß DR, Zhu B, Commichau FM. *Bacillus subtilis* and *Escherichia coli* essential genes and minimal cell factories after one decade of genome engineering. Microbiology. 2014;160: 2341–2351. doi:10.1099/mic.0.079376-0

22. Hutchison CA, Chuang R-Y, Noskov VN, Assad-Garcia N, Deerinck TJ, Ellisman MH, et al. Design and synthesis of a minimal bacterial genome. Science. 2016;351: aad6253–aad6253. doi:10.1126/science.aad6253

23. Curtis PD, Brun YV. Identification of essential alphaproteobacterial genes reveals operational variability in conserved developmental and cell cycle systems. Mol Microbiol. 2014;93: 713–735. doi:10.1111/mmi.12686

24. Pechter KB, Gallagher L, Pyles H, Manoil CS, Harwood CS. Essential genome of the metabolically versatile alphaproteobacterium *Rhodopseudomonas palustris*. J Bacteriol. 2016;198: 867–876. doi:10.1128/JB.00771-15

25. Lee SA, Gallagher LA, Thongdee M, Staudinger BJ, Lippman S, Singh PK, et al. General and condition-specific essential functions of *Pseudomonas aeruginosa*. Proc Natl Acad Sci USA. 2015;112: 5189–5194. doi:10.1073/pnas.1422186112

26. Skurnik D, Roux D, Aschard H, Cattoir V, Yoder-Himes D, Lory S, et al. A comprehensive analysis of *in vitro* and *in vivo* genetic fitness of *Pseudomonas aeruginosa* using high-throughput sequencing of transposon libraries. PLOS Pathog. 2013;9:

e1003582. doi:10.1371/journal.ppat.1003582

27. Freed NE, Bumann D, Silander OK. 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. 2016;16: 203. doi:10.1186/s12866-016-0818-0

28. Canals R, Xia X-Q, Fronick C, Clifton SW, Ahmer BM, Andrews-Polymenis HL, et al. High-throughput comparison of gene fitness among related bacteria. BMC Genomics. 2012;13: 212. doi:10.1186/1471-2164-13-212

29. diCenzo GC, MacLean AM, Milunovic B, Golding GB, Finan TM. Examination of prokaryotic multipartite genome evolution through experimental genome reduction. PLOS Genet. 2014;10: e1004742. doi:10.1371/journal.pgen.1004742

30. diCenzo GC, Zamani M, Milunovic B, Finan TM. Genomic resources for identification of the minimal N_2-fixing symbiotic genome. Environ Microbiol. 2016;18: 2534–2547. doi:10.1111/1462-2920.13221

31. Perry BJ, Yost CK. Construction of a *mariner*-based transposon vector for use in insertion sequence mutagenesis in selected members of the *Rhizobiaceae*. BMC Microbiol. 2014;14: 298. doi:10.1186/s12866-014-0298-z

32. Perry BJ, Akter MS, Yost CK. The use of transposon insertion sequencing to interrogate the core functional genome of the legume symbiont *Rhizobium leguminosarum*. Front Microbiol. 2016;7: 1873. doi:10.3389/fmicb.2016.01873

33. Wheatley RM, Ramachandran VK, Geddes BA, Perry BJ, Yost CK, Poole PS. The role of O2 in the growth of *Rhizobium leguminosarum* bv. *viciae* 3841 on glucose and succinate. J Bacteriol. 2016;199: e00572–16. doi:10.1128/JB.00572-16

34. diCenzo GC, Checcucci A, Bazzicalupo M, Mengoni A, Viti C, Dziewit L, et al. Metabolic

modelling reveals the specialization of secondary replicons for niche adaptation in *Sinorhizobium meliloti*. Nat Commun. 2016;7: 12219. doi:10.1038/ncomms12219

35. Schurr MJ, Vickrey JF, Kumar AP, Campbell AL, Cunin R, Benjamin RC, et al. Aspartate transcarbamoylase genes of *Pseudomonas putida*: requirement for an inactive dihydroorotase for assembly into the dodecameric holoenzyme. J Bacteriol. 1995;177: 1751–1759.

36. Labedan B, Xu Y, Naumoff DG, Glansdorff N. Using quaternary structures to assess the evolutionary history of proteins: the case of the aspartate carbamoyltransferase. Mol Biol Evol. 2004;21: 364–373. doi:10.1093/molbev/msh024

37. Dunn MF. Key roles of microsymbiont amino acid metabolism in rhizobia-legume interactions. Critical Reviews in Microbiology. 2015;41: 411–451. doi:10.3109/1040841X.2013.856854

38. De Bruijn FJ, Lupski JR. The use of transposon Tn*5* mutagenesis in the rapid generation of correlated physical and genetic maps of DNA segments cloned into multicopy plasmids--a review. Gene. 1984;27: 131–149.

39. Berg DE, Schmandt MA, Lowe JB. Specificity of transposon Tn*5* insertion. Genetics. 1983;105: 813–828.

40. Goryshin IY, Miller JA, Kil YV, Lanzov VA, Reznikoff WS. Tn*5*/IS*50* target recognition. Proc Natl Acad Sci USA. 1998;95: 10716–10721.

41. Green B, Bouchier C, Fairhead C, Craig NL, Cormack BP. Insertion site preference of Mu, Tn*5*, and Tn*7* transposons. Mob DNA. 2012;3: 3. doi:10.1186/1759-8753-3-3

42. Harrison PW, Lower RPJ, Kim NKD, Young JPW. Introducing the bacterial “chromid”: not a chromosome, not a plasmid. Trends Microbiol. 2010;18: 141–148.

doi:10.1016/j.tim.2009.12.010

43. diCenzo GC, Finan TM. The divided bacterial genome: structure, function, and evolution. Microbiol Mol Biol Rev. 2017;81: e00019–17. doi:10.1128/MMBR.00019-17

44. Galardini M, Brilli M, Spini G, Rossi M, Roncaglia B, Bani A, et al. Evolution of intra-specific regulatory networks in a multipartite bacterial genome. PLOS Comput Biol. 2015;11: e1004478. doi:10.1371/journal.pcbi.1004478

45. diCenzo G, Milunovic B, Cheng J, Finan TM. The tRNA^{arg} gene and *engA* are essential genes on the 1.7-mb pSymB megaplasmid of *Sinorhizobium meliloti* and were translocated together from the chromosome in an ancestral strain. J Bacteriol. 2013;195: 202–212. doi:10.1128/JB.01758-12

46. Oresnik IJ, Liu SL, Yost CK, Hynes MF. Megaplasmid pRme2011a of *Sinorhizobium meliloti* is not required for viability. J Bacteriol. 2000;182: 3582–3586.

47. Koo B-M, Kritikos G, Farelli JD, Todor H, Tong K, Kimsey H, et al. Construction and analysis of two genome-scale deletion libraries for *Bacillus subtilis*. Cell Systems. 2017;4: 291–305.e7.

48. Armbruster CE, Forsyth-DeOrnellas V, Johnson AO, Smith SN, Zhao L, Wu W, et al. Genome-wide transposon mutagenesis of *Proteus mirabilis*: Essential genes, fitness factors for catheter-associated urinary tract infection, and the impact of polymicrobial infection on fitness requirements. PLOS Pathog. 2017;13: e1006434. doi:10.1371/journal.ppat.1006434

49. Fei F, diCenzo GC, Bowdish DME, McCarry BE, Finan TM. Effects of synthetic large-scale genome reduction on metabolism and metabolic preferences in a nutritionally complex environment. Metabolomics. 2016;12: 23. doi:10.1007/s11306-015-0928-y

50. Stanfield SW, Ielpi L, O'brochta D, Helinski DR, Ditta GS. The *ndvA* gene product of

Rhizobium meliloti is required for beta-(1----2)glucan production and has homology to the ATP-binding export protein HlyB. J Bacteriol. 1988;170: 3523–3530. doi:10.1128/jb.170.8.3523-3530.1988

51. Ielpi L, Dylan T, Ditta GS, Helinski DR, Stanfield SW. The *ndvB* locus of *Rhizobium meliloti* encodes a 319-kDa protein involved in the production of beta-(1----2)-glucan. J Biol Chem. 1990;265: 2843–2851.

52. Griffitts JS, Carlyon RE, Erickson JH, Moulton JL, Barnett MJ, Toman CJ, et al. A *Sinorhizobium meliloti* osmosensory two-component system required for cyclic glucan export and symbiosis. Mol Microbiol. 2008;69: 479–490.

53. Carlyon RE, Ryther JL, VanYperen RD, Griffitts JS. FeuN, a novel modulator of two-component signalling identified in *Sinorhizobium meliloti*. Mol Microbiol. 2010;77: 170–182. doi:10.1111/j.1365-2958.2010.07198.x

54. Finan TM, Kunkel B, De Vos GF, Signer ER. Second symbiotic megaplasmid in *Rhizobium meliloti* carrying exopolysaccharide and thiamine synthesis genes. J Bacteriol. 1986;167: 66–72.

55. Ferguson GP, Roop RM, Walker GC. Deficiency of a *Sinorhizobium meliloti bacA* mutant in alfalfa symbiosis correlates with alteration of the cell envelope. J Bacteriol. 2002;184: 5625–5632. doi:10.1128/JB.184.20.5625-5632.2002

56. Chan CH, Levar CE, Jiménez-Otero F, Bond DR. Genome scale mutational analysis of *Geobacter sulfurreducens* reveals distinct molecular mechanisms for respiration and sensing of poised electrodes vs. Fe(III) oxides. J Bacteriol. 2017;: JB.00340–17. doi:10.1128/JB.00340-17

57. Arnold MFF, Shabab M, Penterman J, Boehme KL, Griffitts JS, Walker GC. Genome-

wide sensitivity analysis of the microsymbiont *Sinorhizobium meliloti* to symbiotically important, defensin-like host peptides. mBio. 2017;8: e01060–17. doi:10.1128/mBio.01060-17

58. Yang H, Krumholz EW, Brutinel ED, Palani NP, Sadowsky MJ, Odlyzko AM, et al. Genome-scale metabolic network validation of *Shewanella oneidensis* using transposon insertion frequency analysis. PLOS Comput Biol. 2014;10: e1003848. doi:10.1371/journal.pcbi.1003848

59. Broddrick JT, Rubin BE, Welkie DG, Du N, Mih N, Diamond S, et al. Unique attributes of cyanobacterial metabolism revealed by improved genome-scale metabolic modeling and essential gene analysis. Proc Natl Acad Sci USA. 2016;113: E8344–E8353. doi:10.1073/pnas.1613446113

60. Bartell JA, Blazier AS, Yen P, Thøgersen JC, Jelsbak L, Goldberg JB, et al. Reconstruction of the metabolic network of *Pseudomonas aeruginosa* to interrogate virulence factor synthesis. Nat Commun. 2017;8: 14631. doi:10.1038/ncomms14631

61. Senior NJ, Sasidharan K, Saint RJ, Scott AE, Sarkar-Tyson M, Ireland PM, et al. An integrated computational-experimental approach reveals *Yersinia pestis* genes essential across a narrow or a broad range of environmental conditions. BMC Microbiol. 2017;17: 163. doi:10.1186/s12866-017-1073-8

62. Burger BT, Imam S, Scarborough MJ, Noguera DR, Donohue TJ. Combining genome-scale experimental and computational methods to identify essential genes in *Rhodobacter sphaeroides*. mSystems. 2017;2: e00015–17. doi:10.1128/mSystems.00015-17

63. Griffitts JS, Long SR. A symbiotic mutant of *Sinorhizobium meliloti* reveals a novel genetic pathway involving succinoglycan biosynthetic functions. Mol Microbiol. 2008;67:

1292–1306. doi:10.1111/j.1365-2958.2008.06123.x

64. Sambrook J, Fritsch EF, Maniatis T. Molecular cloning: a laboratory manual. New York: Cold Spring Harbor Laboratory Press; 1989.

65. Martin MO, Long SR. Generalized transduction in *Rhizobium meliloti*. J Bacteriol. 1984;159: 125–129.

66. Ferrières L, Hémery G, Nham T, Guérout A-M, Mazel D, Beloin C, et al. Silent mischief: bacteriophage Mu insertions contaminate products of *Escherichia coli* random mutagenesis performed using suicidal transposon delivery plasmids mobilized by broad-host-range RP4 conjugative machinery. J Bacteriol. 2010;192: 6418–6427. doi:10.1128/JB.00621-10

67. Fraley C, Raftery AE, Murphy TB, Scrucca L. mclust Version 4 for R: Normal mixture modeling for model-based clustering, classification, and density estimation. Washington, USA: Department of Statistics, University of Washington; 2012.

68. Bodenhofer U, Kothmeier A, Hochreiter S. APCluster: an R package for affinity propagation clustering. Bioinformatics. 2011;27: 2463–2464. doi:10.1093/bioinformatics/btr406

69. Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, et al. Integrative genomics viewer. Nat Biotechnol. 2011;29: 24–26. doi:10.1038/nbt.1754

70. Wickham H. ggplot2: elegant graphics for data analysis.[Internet]. 2009. New York, USA: Springer-Verlag; 2009.

71. Chen H, Boutros PC. VennDiagram: a package for the generation of highly-customizable Venn and Euler diagrams in R. BMC Bioinform. 2011;12: 35. doi:10.1186/1471-2105-12-35

72. Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, et al. Circos: an

information aesthetic for comparative genomics. Genome Res. 2009;19: 1639–1645. doi:10.1101/gr.092759.109

73. Mrázek J, Karlin S. Strand compositional asymmetry in bacterial and large viral genomes. Proc Natl Acad Sci USA. 1998;95: 3720–3725.

74. Yamada T, Letunic I, Okuda S, Kanehisa M, Bork P. iPath2.0: interactive pathway explorer. Nucleic Acids Res. 2011;39: W412–5. doi:10.1093/nar/gkr313

75. Eddy SR. A new generation of homology search tools based on probabilistic inference. Genome Inform. 2009;23: 205–211.

76. Wheeler TJ, Clements J, Finn RD. Skylign: a tool for creating informative, interactive logos representing sequence alignments and profile hidden Markov models. BMC Bioinform. 2014;15: 1. doi:10.1186/1471-2105-15-7

77. diCenzo GC, Zamani M, Ludwig HN, Finan TM. Heterologous complementation reveals a specialized activity for BacA in the *Medicago*-*Sinorhizobium meliloti* symbiosis. Mol Plant Microbe Interact. 2017;30: 312-324. doi:10.1094/MPMI-02-17-0030-R

78. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+: architecture and applications. BMC Bioinform. 2009;10: 421. doi:10.1186/1471-2105-10-421

79. Schellenberger J, Que R, Fleming RMT, Thiele I, Orth JD, Feist AM, et al. Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox v2.0. Nat Protoc. 2011;6: 1290–1307. doi:10.1038/nprot.2011.308

80. Keating SM, Bornstein BJ, Finney A, Hucka M. SBMLToolbox: an SBML toolbox for MATLAB users. Bioinformatics. 2006;22: 1275–1277. doi:10.1093/bioinformatics/btl111

81. Bornstein BJ, Keating SM, Jouraku A, Hucka M. LibSBML: an API library for SBML.

Bioinformatics. 2008;24: 880–881. doi:10.1093/bioinformatics/btn051

82. Young JPW, Crossman LC, Johnston AW, Thomson NR, Ghazoui ZF, Hull KH, et al. The genome of *Rhizobium leguminosarum* has recognizable core and accessory components. Genome Biol. 2006;7: R34. doi:10.1186/gb-2006-7-4-r34

83. Kanehisa M, Sato Y, Kawashima M, Furumichi M, Tanabe M. KEGG as a reference resource for gene and protein annotation. Nucleic Acids Res. 2016;44: D457–D462. doi:10.1093/nar/gkv1070

84. Basconcillo LS, Zaheer R, Finan TM, McCarry BE. A shotgun lipidomics study of a putative lysophosphatidic acid acyl transferase (PlsC) in *Sinorhizobium meliloti*. J Chromatogr B. 2009;877: 2873–2882. doi:10.1016/j.jchromb.2009.05.014

85. Thorvaldsdóttir H, Robinson JT, Mesirov JP. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Brief Bioinformatics. 2013;14: 178–192. doi:10.1093/bib/bbs017

86. Yuan Z-C, Zaheer R, Finan TM. Regulation and properties of PstSCAB, a high-affinity, high-velocity phosphate transport system of *Sinorhizobium meliloti*. J Bacteriol. 2006;188: 1089–1102. doi:10.1128/JB.188.3.1089-1102.2006

87. diCenzo GC, Sharthiya H, Nanda A, Zamani M, Finan TM. PhoU allows rapid adaptation to high phosphate concentrations by modulating PstSCAB transport rate in *Sinorhizobium meliloti*. J Bacteriol. 2017;: JB.00143–17. doi:10.1128/JB.00143-17

88. diCenzo GC, Muhammed Z, Østerås M, O'Brien SAP, Finan TM. A key regulator of the glycolytic and gluconeogenic central metabolic pathways in *Sinorhizobium meliloti*. Genetics. 2017;: [epub ahead of print]. doi:10.1534/genetics.117.300212

Table 1. Fitness classification of chromosomal genes. Genes were ranked from lowest to highest GEI, with the lowest GEI being at the 0 percentile and the highest GEI being at the 100th percentile. The approximate break points for the groupings, determined as described in the Materials and Methods, are shown for each condition.

		GEI percentile range			
Group	Description	Wild type, rich medium	ΔpSymAB, rich medium	Wild type, defined medium	ΔpSymAB, defined medium
I	Essential	0-12	0-14	0-12	0-14
II	Strong growth defect	12-17	14-23	12-18	14-24
III	Moderate growth defect	17-36	23-49	18-28	24-47
IV	Little to no growth impact	36-100	49-96	28-99	47-99
V	Growth improvement	N/A	96-100	99-100	99-100

Table 2. Sample genes showing strain specific phenotypes. The top ten genes from each of the indicated groupings, as determined based on the ratio of GEI scores of the two strains, are shown. GEI (Gene Essentiality Index) scores are shown first for the wild type (WT) followed by the scores for the ΔpSymAB (dAB) strain.

		GEI			GEI
Gene	Function	(WT ... dAB)	Gene	Function	(WT ... dAB)
More Important ΔpSymAB - Rich Medium			More Important ΔpSymAB - Defined Medium		
kpsF3	capsule expression protein	5.951 ... 0.002	*smc03782*	signal peptide protein	9.670 ... 0.001
groEL	chaperonin GroEL	2.926 ... 0.001	*amiC*	N-acetylmuramoyl-L-alanine amidase	12.273 ... 0.003
aidB	oxidoreductase	2.658 ... 0.001	*groEL*	chaperonin GroEL	2.755 ... 0.001
proA	γ-glutamyl phosphate reductase	4.505 ... 0.001	*amn*	AMP nucleosidase	2.200 ... 0.001
amiC	N-acetylmuramoyl-L-alanine amidase	2.016 ... 0.002	*ndvA*	cyclic beta-1,2-glucan ABc transporter	1.434 ... 0.001
smc03782	signal peptide protein	1.847 ... 0.001	*smc02495*	translaldolase	3.933 ... 0.003
etfA1	electron transfer flavoprotein	2.447 ... 0.002	*glnA*	glutamine synthetase	2.535 ... 0.002
smc02495	translaldolase	3.127 ... 0.003	*glmS*	glucosamine--fructose-6P aminotransferase	1.918 ... 0.002
exoN2	UTP--glucose-1P uridylyltransferase	2.292 ... 0.002	*smc00717*	ABC transporter ATP-binding protein	6.624 ... 0.006
glnA	glutamine synthetase	2.153 ... 0.002	*etfA1*	electron transfer flavoprotein	2.156 ... 0.002
More Important Wild Type - Rich Medium			More Important Wild Type - Defined Medium		
carB	carbamoyl phosphate synthase	0.001 ... 1.930	*folD2*	5,10-methylene-THF dehydrogenase	0.003 ... 1.026
argG	argininosuccinate synthase	0.002 ... 1.281	*nuoK1*	NADH dehydrogenase subunit K	0.006 ... 1.476
carA	carbamoyl phosphate synthase	0.007 ... 3.928	*prfC*	peptide chain release factor RF-3 protein	0.001 ... 0.232
purB	adenylosuccinate lyase	0.002 ... 0.799	*smc00714*	1-acyl-SN-glycerol-3P acyltransferase	0.005 ... 0.522
hrm	histone-like protein	0.011 ... 2.586	*fpr*	ferredoxin--NADP reductase	0.002 ... 0.248
nuoK1	NADH dehydrogenase subunit K	0.006 ... 1.330	*smc00532*	hypothetical protein	0.002 ... 0.203
folD2	5,10-methylene-THF dehydrogenase	0.018 ... 3.110	*ubiE*	ubiquinone biosynthesis methyltransferase	0.002 ... 0.209
coaA	pantothenate kinase	0.004 ... 0.595	*asd*	aspartate-semialdehyde dehydrogenase	0.002 ... 0.157
argF1	ornithine carbamoyltransferase	0.012 ... 1.776	*secE*	preprotein translocase subunit SecE	0.010 ... 0.796
smc00914	oxidoreductase	0.002 ... 0.240	*smc01038*	hypothetical protein	0.040 ... 2.940

Table 3. Summary of iGD726. The last column indicates reactions whose genes associations are supported by the Tn-seq data of this study. Percentage of all reactions in that category are indicated in brackets.

Pathways	Genes	Reactions	Reactions supported by Tn-seq
Overall	726	681	444 (65%)
Carbon metabolism, oxidative phosphorylation	105	54	37 (69%)
Amino acid metabolism	116	93	72 (77%)
Nucleotide metabolism	34	40	39 (98%)
Fatty acid, lipid metabolism	42	227	143 (63%)
Peptidoglycan, lipopolysaccharide, exopolysaccharide metabolism	47	43	27 (63%)
Nucleotide sugar metabolism	25	17	6 (35%)
Vitamin, cofactor, coenzyme metabolism	109	121	83 (69%)
Miscellaneous metabolism	23	15	7 (47%)
Transcription, translation, DNA replication, cell division	153	29	28 (97%)
Transport reactions	75	21	3 (14%)
Exchange reactions	0	21	N/A

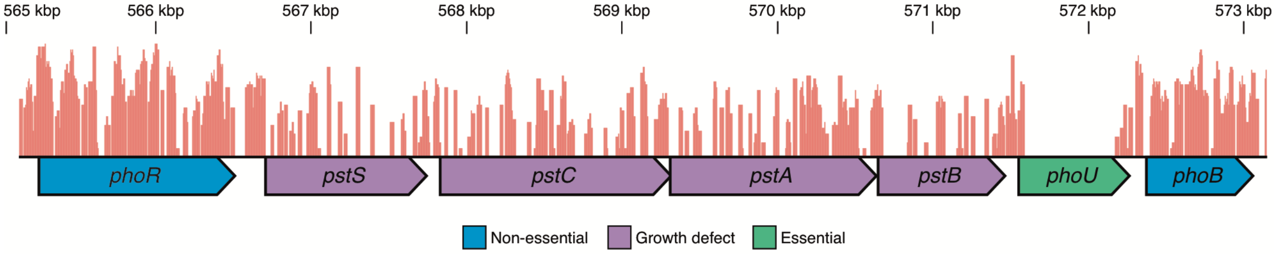

Figure 1. Visualization of the location of transposon insertion sites. An image of the *pst* locus of *S. meliloti* generated using the Integrative Genomics Viewer [85]. Chromosomal nucleotide positions are indicated along the top of the image, and the location of transposon insertions are indicated by the red bars. Non-essential genes contain a high density of transposon insertions, whereas essential genes have few to no transposon insertions. Genes are color coded based on their fitness classification, and transcripts are indicated by the arrows below the genes. The *pstS*, *pstC*, *pstA*, *pstB*, *phoU*, and *phoB* genes are co-transcribed as a single operon [86], and previous work demonstrated that polar *phoU* mutations are lethal in *S. meliloti*, whereas non-polar mutations are not lethal [87]. The lack of insertions within the *phoU* coding region is therefore consistent with the non-polar nature of the transposon.

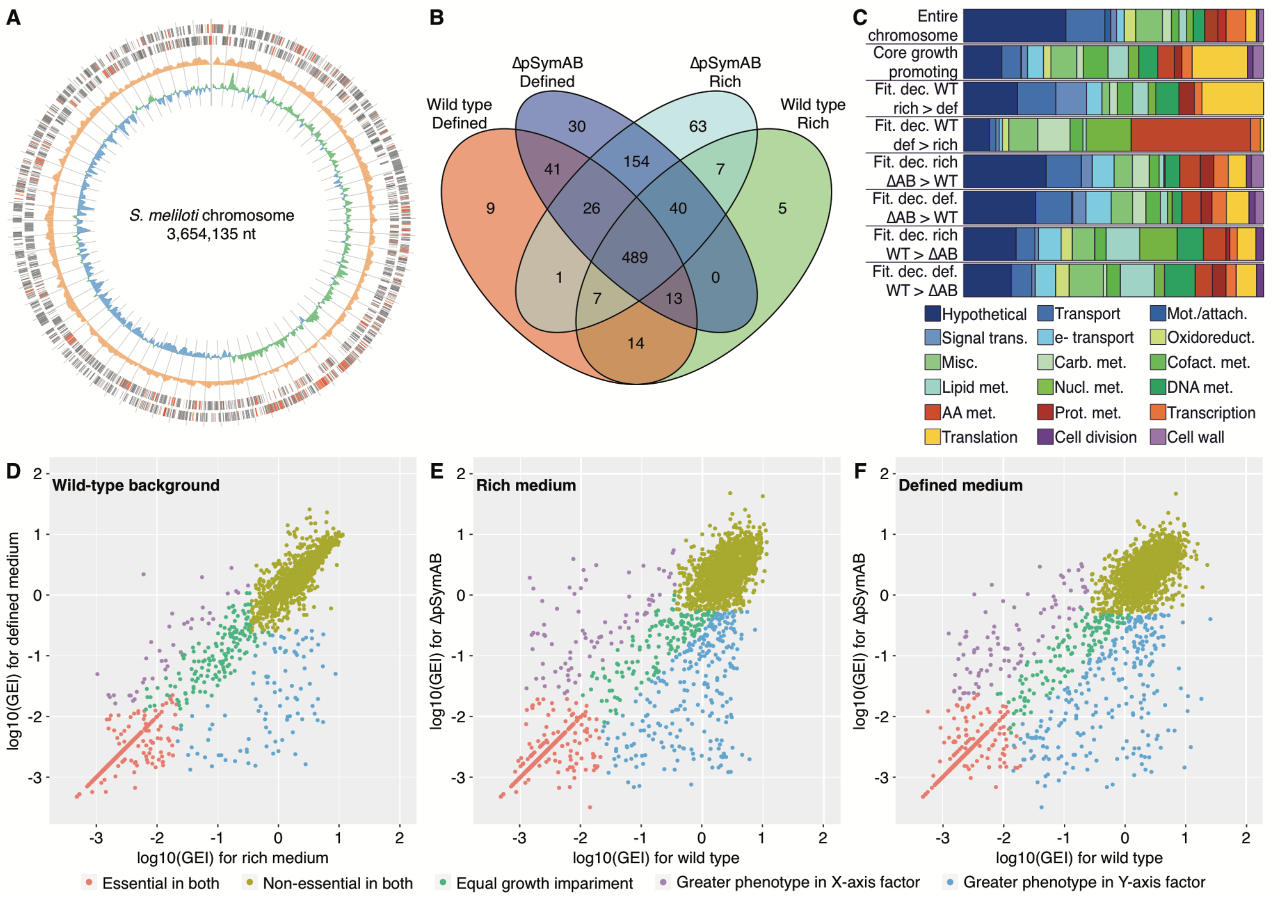

Figure 2. Characteristics of the core genetic components of *S. meliloti*. (**A**) A plot of the *S. meliloti* chromosome is shown. From the outside to inside: positive strand coding regions, negative strand coding regions, total insertion density, and GC skew. For the positive and negative strands, red lines indicate the core 489 growth promoting genes. The insertion density displays the total transposon insertions across all experiments over a 10,000-bp window. The GC skew was calculated over a 10,000-bp window, with green showing a positive skew and blue showing a negative skew. Tick marks are every 50,000 bp. (**B**) A comparison of the overlap between the growth promoting genome (Group I and II genes) of each Tn-seq data set. Each data set is labelled with the strain (wild type or ΔpSymAB) and the growth medium (defined medium or rich medium). (**C**) Functional enrichment plots for the indicated gene sets. Name abbreviations: Fit – fitness; Dec – decrease; WT – wild type; ΔAB - ΔpSymAB; Def – defined medium; Rich – rich medium. For example, 'Fit. dec. WT def > rich' means the genes with a greater fitness decrease in wild type grown in defined medium compared to rich medium. Legend abbreviations: AA – amino acid; Attach – attachment; Carb – carbohydrate; Cofact – cofactor; e- – electron; Met – metabolism; Misc – miscellaneous; Mot – motility; Nucl – nucleotide; Oxidoreduct – oxidoreductase activity; Prot – protein; Trans – transduction. (**D-F**) Scatter plots comparing the fitness phenotypes, shown as the $\log_{10}$ of the GEI scores (Gene Essentiality Index scores; i.e., number of insertions within the gene divided by gene length in nucleotides) of (**D**) wild type grown in rich medium versus wild type grown in defined medium, (**E**) wild type grown in rich medium versus ΔpSymAB grown in rich medium, and (**F**) wild type grown in defined medium versus ΔpSymAB grown in defined medium.

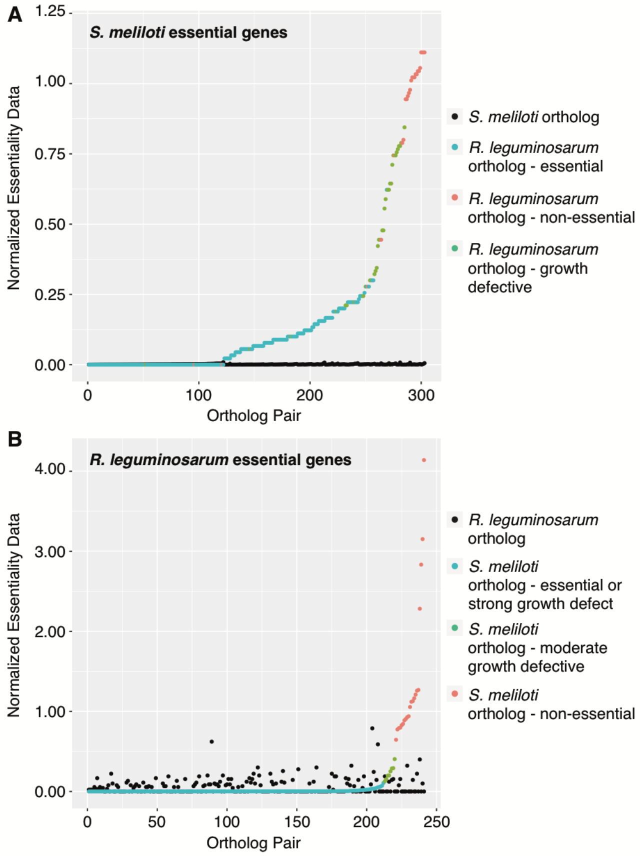

Figure 3. Comparison of *S. meliloti* and *R. leguminosarum* Tn-seq data. (**A**) The fitness phenotypes of essential *S. meliloti* genes, as determined in this study, is compared to the fitness phenotypes of the orthologous *R. leguminosarum* genes, as determined by Perry *et al*. [32]. *S. meliloti* orthologs are shown in black, while the *R. leguminosarum* orthologs are colored according to their classification by Perry *et al*. [32]. (**B**) The fitness phenotypes of essential *R. leguminosarum* genes is compared to the fitness phenotypes of the orthologous *S. meliloti* genes. *R. leguminosarum* orthologs are shown in black, while the *S. meliloti* orthologs are colored according to their classification in this study. (**A**,**B**) Normalized fitness values are used to facilitate direct comparison between the studies as different output statistics were calculated. For *S. meliloti*, the GEI score of each gene for wild type grown in minimal medium broth was divided by the median GEI for all genes under the same conditions. For *R. leguminosarum*, the insertion density of each gene during growth on minimal medium plates was divided by the median insertion density of all strains.

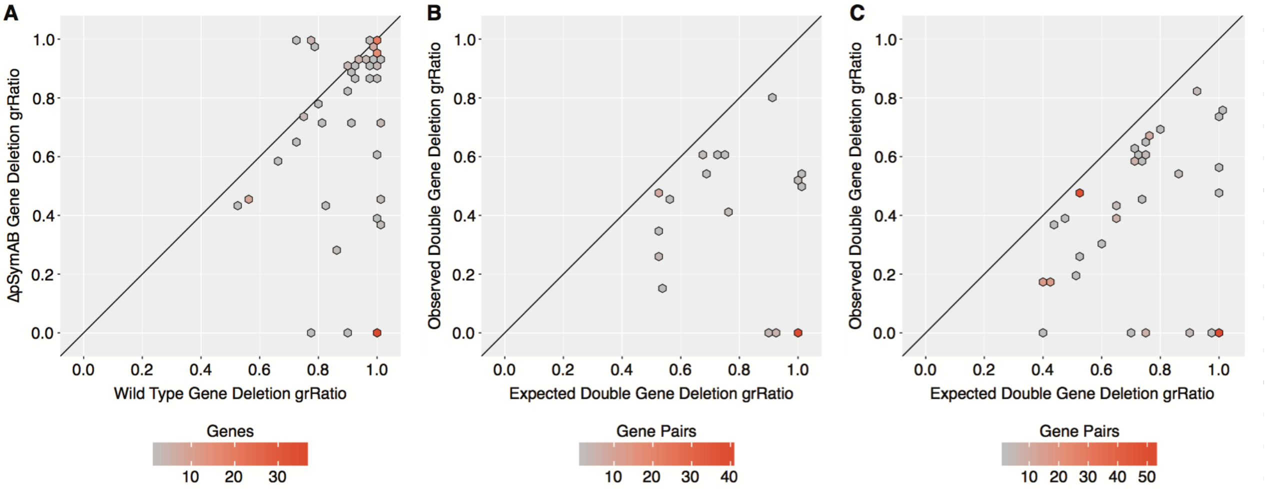

Figure 4. *In silico* analysis of genetic redundancy in *S. meliloti*. The effects of single or double gene deletion mutants were predicted *in silico* with the genome-scale *S. meliloti* metabolic model. (**A**) A scatter plot comparing the grRatio (growth rate of mutant / growth rate of non-mutant) for gene deletion mutations in the presence (wild type) versus absence (ΔpSymAB) of the pSymA/pSymB model genes. Genes whose deletion had either no effect or were lethal in both cases are not shown. (**B**) A scatter plot comparing the grRatio for each double gene deletion pair (where one gene was on the chromosome and the other on pSymA or pSymB) observed *in silico* versus the predicted grRatio based on the grRatio of the single deletions (grRatio1 * grRatio2). Only gene pairs with an observed grRatio at least 10% less than the expected are shown. (**C**) A scatter plot comparing the grRatio for each double gene deletion pair (both genes on the chromosome) observed *in silico* versus the predicted grRatio. Only gene pairs with an observed grRatio at least 10% less than the expected are shown. (**A-C**) The color of each hexagon is representative of the number of reactions plotted at that location, as illustrated by the density bar below each panel. The diagonal line serves as a reference line.

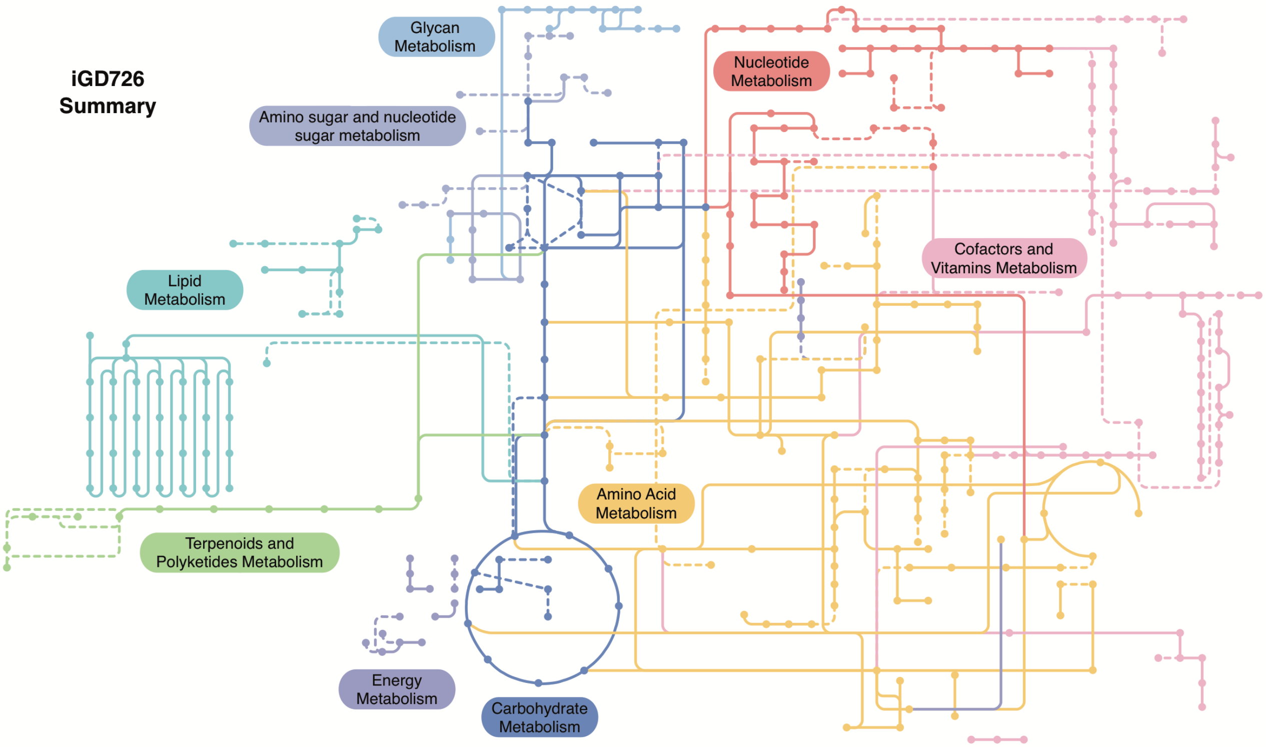

Figure 5. Summary schematic of core *S. meliloti* metabolism. The iGD726 core metabolic model was visualized using the iPath v2.0 webserver [74], which maps the reactions of the metabolic model to KEGG metabolic pathways; it therefore does not capture metabolism not present in the KEGG pathways included in iPath. Reactions and metabolites are colour coded according to their biological role, as indicated. Reactions whose associated genes were not identified as growth promoting in this study are in dashed lines.

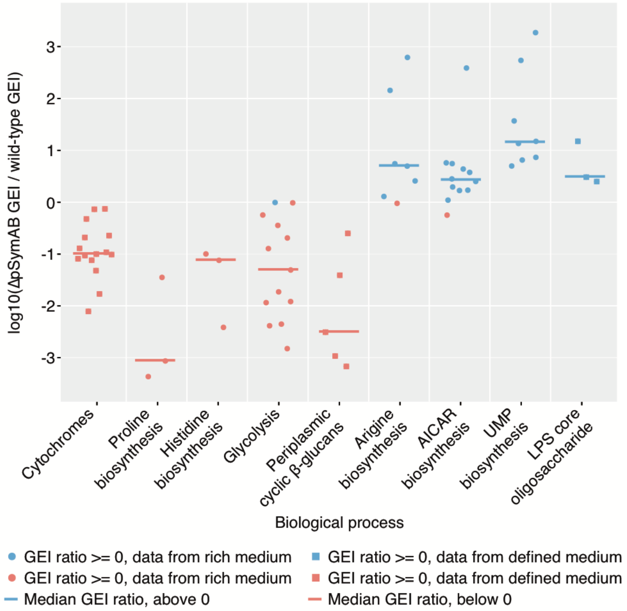

Figure 6. Gene essentiality index (GEI) changes for genes of selected biological pathways. Each circle or square represents an individual gene, and shows the $\log_{10}$ of the ratio of the GEI for that gene in the ΔpSymAB background compared to the wild-type background. Lines indicate the median value of all genes included from the biological process. The underlying data is given in Table S9. Genes included in each process are as follows: Cytochrome C oxidase related genes – *ctaB*, *ctaC*, *ctaD*, *ctaE*, *ctaG*, *ccsA*, *cycH*, *cycJ*, *cycK*, *cycL*, *ccmA*, *ccmB*, *ccmC*, *ccmD*, *ccmG*; Proline biosynthesis – *proA*, *proB1*, *proC*; Histidine biosynthesis – *hisB*, *hisD*, *smc04042*; Glycolysis and related genes – *glk*, *frk*, *pgi*, *zwf*, *pgl*, *edd*, *eda2*, *gap*, *pgk*, *gpmA*, *eno*, *pykA*, *pyc*; Periplamic cyclic β-glucan biosynthesis – *feuN*, *feuP*, *feuQ*, *ndvA*, *ndvB*; Arginine biosynthesis – *argB*, *argC*, *argD*, *argF1*, *argG*, *argH1*, *argJ*; AICAR biosynthesis – *purB*, *purC*, *purD*, *purE*, *purF*, *purH*, *purK*, *purL*, *purM*, *purN*, *purQ*, *smc00494*; UMP biosynthesis – *carA*, *carB*, *pyrB*, *pyrC*, *pyrD*, *pyrE*, *pyrF*, *smc01361*; LPS core oligosaccharide biosynthesis – *lpsC*, *lpsD*, *lpsE*.