# *In silico* method for modelling metabolism and gene product expression at genome scale

**Joshua A. Lerman**[1,*], **Daniel R. Hyduke**[1,*], **Haythem Latif**[1], **Vasiliy A. Portnoy**[1], **nathan E. Lewis**[1], **Jeffrey D. Orth**[1], **Alexandra C. Schrimpe-Rutledge**[2], **Richard D. Smith**[2], **Joshua n. Adkins**[2], **Karsten Zengler**[1], and **Bernhard O. Palsson**[1]

Bernhard O. Palsson: palsson@ucsd.edu

[1]Department of Bioengineering, University of California–San Diego, PFBH Room 419, 9500 Gliman Drive, La Jolla, California 92093-0412, USA

[2]Biological Sciences Division, Pacifc Northwest National Laboratory, Richland, Washington 99352, USA

## Abstract

Transcription and translation use raw materials and energy generated metabolically to create the macromolecular machinery responsible for all cellular functions, including metabolism. A biochemically accurate model of molecular biology and metabolism will facilitate comprehensive and quantitative computations of an organism's molecular constitution as a function of genetic and environmental parameters. Here we formulate a model of metabolism and macromolecular expression. Prototyping it using the simple microorganism *Thermotoga maritima*, we show our model accurately simulates variations in cellular composition and gene expression. Moreover, through *in silico* comparative transcriptomics, the model allows the discovery of new regulons and improving the genome and transcription unit annotations. Our method presents a framework for investigating molecular biology and cellular physiology *in silico* and may allow quantitative interpretation of multi-omics data sets in the context of an integrated biochemical description of an organism.

A goal of systems biology is to provide comprehensive biochemical descriptions of organisms that are amenable to mathematical enquiry[1]. These models may then be used to investigate fundamental biological questions[1], guide industrial strain design[2] and provide a systems perspective for analysis of the expanding ocean of omics data[3]. Over the past decade, there has been steady progress in developing genome-scale models of metabolism (M-Models) for basic research and industrial applications[4–6]. M-Models are stoichiometric representations of the enzymatic and spontaneous biochemical reactions associated with an

organism's metabolic network at the genome scale; however, M-Models do not quantitatively describe gene expression (Fig. 1a). The lack of an explicit representation for enzyme production precludes quantitative interpretation of omics data and can result in biologically implausible predictions[7,8]. Because M-Models do not contain chemical representations of transcription and translation, to date, it has only been possible to use omics data as *ad hoc* constraints for enzyme activities[9–12].

A modelling approach that accounts for the production and degradation of a cell's macromolecular machinery will provide a full genetic basis for every computed molecular phenotype (Fig. 1b). Such computations in turn enable the direct comparison of simulation to omics data and the simulation of variable expression and enzyme activity[13,14]. In other words, an integrated model of metabolism and macromolecular expression (ME-Model) affords a genetically consistent description of a self-propagating organism at the molecular level and moves us substantially closer to establishing a systems-level quantitative basis for biology.

Here, we developed an ME-Modelling approach for the relatively simplistic microorganism, *Thermotoga maritima*, which metabolizes a variety of feedstocks into valuable products including $H_2$ (ref. 15). *T. maritima* possess a number of characteristics conducive to systems-level investigations of the genotype–phenotype relationship: a compact 1.8-Mb genome[16], wealth of structural proteome data[17], a limited repertoire of transcription factors (TFs)[18] and reduced genome organizational complexity compared with other microbes (H.L. *et al.*, Unpublished data). Taken together, *T. maritima*'s small set of TFs and reduced genome complexity impose a narrowed range of viable regulatory and functional states (H.L. *et al.*, unpublished data). The existence of few regulatory states may simplify the addition of synthetic capabilities and facilitate metabolic engineering efforts by reducing unexpected and irremediable side-effects arising from genetic manipulation[19]. A combination of metabolic versatility and genomic simplicity make *T. maritima* a promising candidate for investigating fundamental relationships between molecular and cellular physiology, both *in silico* and *in vivo*, and for the creation of a minimal chassis for chemical synthesis[20]. Our *T. maritima* ME-Model simulates changes in cellular composition with growth rate, in agreement with previously reported experimental findings[21,22]. We observed positive correlations between *in silico* and *in vivo* transcriptomes and proteomes for the 651 genes in our ME-Model with statistically significant ($P < 1 \times 10^{-15}$ *t*-test) Pearson correlation coefficients (PCC) of 0.54 and 0.57, respectively. And, when we used our ME-Model as an exploratory platform for an *in silico* comparative transcriptomics study, we discovered putative TF-binding motifs and regulons associated with L-arabinose (L-Arab) and cellobiose metabolism, and improved functional and transcription unit (TU) architecture annotation. Overall, ME-Models provide a chemically and genetically consistent description of an organism, thus they begin to bridge the gap currently separating molecular biology and cellular physiology.

## Results

### Genome-scale modelling of metabolism and expression

We developed a network reconstruction and modelling method that includes macromolecular synthesis and post-transcriptional modifications in addition to metabolism (Fig. 1c; Supplementary Methods). Specifically, our method accounts for the production of TUs, functional RNAs (that is, transfer RNAs (tRNAs), ribosomal RNAs (rRNAs) and so on) and peptide chains, as well as the assembly of multimeric proteins and dilution of macromolecules to daughter cells during growth. Based on available genomic, structural proteomic and biochemical literature we constructed an ME-Model for T. maritima that

accounts for the functional activities of 50% of the annotated gene products and, more importantly, mechanistically links these enzyme activities to the genome.

To accurately model self-replicating cells at the molecular level, it is necessary to account for material dilution during cell division as a result of volume doubling, and to provide limits on the number of proteins that may be translated from an messenger RNA before the mRNA decays or is transmitted to a daughter cell. To approximate dilution of transcripts and proteins to daughter cells and prevent infinite translation of peptides from an mRNA, we devised a series of coupling constraints (Fig. 1d; Supplementary Methods). These constraints effectively provide upper limits on enzyme expression and activity and are a function of the organism's doubling time ($T_d$). These coupling constraints may be tuned for specific mRNAs or enzymes if their, respective, degradation rates or catalytic turnover constants ($k_{cat}$) are known.

Applications of M-Models often involve simulating log-phase cellular growth using flux balance analysis (FBA)[23,24]. The organism's gross lipid, nucleotide, amino acid (AA) and cofactors, as well as growth-associated and maintenance ATP usage, are experimentally measured. Then, these measurements are integrated with the organism's $T_d$ to define a biomass reaction that approximates the dilution of cellular materials during formation of daughter cells. However, cellular composition is known to vary as a function of $T_d$ and medium[21]—with Schaechter *et al.* indicating that $T_d$ is more influential than growth medium.

Our ME-Model explicitly describes transcription, translation and the dilution of gene products to daughter cells, thus it is unnecessary to use a gross biomass production reaction when simulating growth. Instead, ME-Models contain a structural reaction that accounts for the dilution of structural materials (that is, DNA, cell wall, lipids and so on) during division and the energy cost associated with cellular maintenance of the structure (Supplementary Table S1). Conceptually, this structural reaction approximates the production of a cell whose composition varies as a function of environment and growth rate (Fig. 2a).

## Molecularly efficient simulation of cellular physiology

The RNA-to-protein mass ratio ($r$) has been observed to increase as a function of specific growth rate ($\mu$)[21,22] and decreases as a function of translation efficiency[22]. Schaechter *et al.* also observed an increase in the number of ribonucleoprotein particles with increasing $\mu$, whereas the translation rate per ribonucleoprotein particle was relatively constant[21]. The increase in $r$ and ribonucleoproteins may be due to the reduced number of translation events mediated by a ribosome as $T_d$ decreases.

To ascertain whether our ME-Model recapitulated the observed increases in $r$, ribosomal RNA and proteins with increasing $\mu$ we simulated a range of growth rates in a defined minimal medium[25] (Supplementary Table S2). To simulate the molecular physiology of *T. maritima* for a particular $\mu$, we used FBA[24] subject to linear programming optimization[26] to identify the minimum ribosome production rate required to support a given $\mu$ (Fig. 2b). Ribosome production has been shown to be linearly correlated with growth rate in *Escherichia coli*[22,27,28]. Assuming that efficient use of enzymes contributes to the fitness of an evolutionarily adapted lineage[29], we would expect a successful organism to produce the minimal amount of ribosomes required to support expression of the proteome.

Consistent with experimental observations[21,22], our ME-Model simulated an increase in $r$ with increasing $\mu$ and with decreasing translation efficiency (Fig. 3a). We observed that the fraction of the transcriptome associated with ribosomal RNA *in silico* increased with $\mu$ (Fig. 3b). In addition, the ribosomal proteins account for a larger proportion of the total proteome

as $\mu$ increases (Fig. 3c). These results indicate that it is possible to mechanistically model changes in cellular physiology that have only recently yielded to phenomenological modelling[22].

With M-Models, the cellular macromolecular composition is constant, ergo they cannot reproduce the observed increases in $r$ or ribosomes with increasing $\mu$. Although it is possible to empirically determine a relationship between gross biomass composition and $\mu$ and then use this relationship to study variable composition in M-Models[30], the M-Models will compute a solution space where the range of activity for a number of enzymes may be rather broad and even infinite[7], if not specifically constrained. The biologically implausible sections of the M-Model solution space are due, in large part, to unconstrained thermodynamically infeasible internal loops that can operate at an arbitrary flux level[8]. These arbitrary activities contradict previous observations that efficient organisms should maintain a minimal total flux through their biochemical network[29,31].

By explicitly accounting for enzyme expression and activity, ME-Model simulations should identify the set of proteins that will result in optimally efficient conversion of growth substrates into cells. To determine whether our ME-Model was more economic in terms of enzyme usage than the M-Model, we compared our ME-Model simulation to a random sampling of the M-Model solution space[7]. After we fit a normal distribution to the sampled M-Model space, we found that there is a small ($2.1 \times 10^{-5}$) probability of finding an M-Model solution as efficient as the ME-Model solution (Fig. 3d). Because ME-Models explicitly account for the costs of enzyme expression and dilution to daughter cells, the most efficient growth simulations will minimize the materials required to assemble the cell; that is, ME-Models will efficiently use enzymes when simulating a $\mu$.

To compare the range of permissible, that is, computationally feasible, activity for each metabolic enzyme in the ME-Model versus the M-Model, we performed flux variability analysis. Flux variability analysis identifies the flux range that each reaction may carry given that the model must also simulate the specified objective value, such as $\mu$, with a set tolerance. The permissible enzyme activities for simulating efficient growth with a 1% tolerance tended to have smaller ranges in the ME-Model compared with the M-Model (Fig. 3e; Supplementary Data 1), highlighting the sharply reduced flexibility in the ME-Model solution space when simulating optimal growth.

Our ME-Model contains gene products that carry out 142 of the 206 functions estimated as essential for a minimal organism[32], whereas the M-Model contains only 65 of these core functions. With the ME-Model, 120 of the 142 functions were essential for ribosome production, whereas only 23 of the 65 functions in the M-Model were essential for biomass production (Supplementary Data 2). This broader coverage of cellular functions means that ME-Models may be used for *in silico* investigations of phenotypic states that are inaccessible to M-Models.

## Gene product production and turnover alters pathway activity

In addition to simulating variable cellular composition and effectively eliminating the infinite catalysis problem, there are a number of metabolic activities that are required for optimally efficient growth with the ME-Model but not with the M-Model (Fig. 4). These differences are due to the ME-Model producing small metabolites as by-products of gene expression and explicitly accounting for the material and energy costs of macromolecule production and turnover. The ME-Model includes metabolic activities for recycling *S*-adenosylhomocysteine, which is a by-product of rRNA and tRNA methylation, and guanine, which is a by-product of queuosine modification of various tRNAs (Fig. 4a). The ME-Model, also, produces CTP from CMP that is produced during mRNA degradation (Fig. 4b).

Interestingly, the M-Model does not require CDP production to simulate growth, whereas CDP production is essential in the ME-Model. The ME-Model exhibits frugality with respect to central metabolic reactions (Fig. 4c) and proposes the canonical gylcolytic pathway during efficient growth, whereas the M-Model indicates that alternate pathways are as efficient. When the efficiency requirement is relaxed these less-efficient pathways may be active in the ME-Model solution space (Supplementary Data 1). The genes associated with optimal activities tended to be strongly expressed (approximately 60th–90th percentile) in transcriptome data.

These differences highlight the interplay between macromolecular synthesis and degradation, metabolism and salvage, and optimal use of the proteome. The ME-Models allow a fine resolution view of these processes and their simultaneous reconciliation. Not only can one analyse specific pathways in isolation, such as the three examples given above, but it is now possible to investigate in detail the coordination of functions within an organism's biochemical repertoire.

## Simulation of systems-level molecular phenotypes

To assess our ME-Model's ability to simulate systems-level molecular phenotypes, we compared model predictions to substrate consumption, product secretion, AA composition, transcriptome and proteome measurements. With the only external constraints for the ME-Model being the experimentally determined $\mu$ during log-phase growth in maltose minimal medium at 80 °C, our model accurately predicted maltose consumption and acetate and $H_2$ secretion (Fig. 5a; Supplementary Table S3). Predicted AA incorporation was linearly correlated (0.79 PCC; $P < 4.1 \times 10^{-5}$ $t$-test) with measured AA composition (Fig. 5b).

FBA simulates reaction fluxes, whereas transcraiptomics and proteomics technologies provide semiquantitative measurements of expressed gene product abundance. Thus, the simulated fluxes through the transcriptome and proteome do not directly approximate the respective omics measurements; however, for macromolecules there should be a positive correlation between gene and protein synthesis fluxes and the respective gene product abundances during log-phase growth. In other words, proteins and genes are relatively stable and when an organism is growing at steady state a relative increase in expression rate for a protein will effectively increase the quantity of that protein.

Interestingly, when we compared the simulated transcriptome and proteome fluxes to transcriptome and proteome measurements, respectively, there were statistically significant ($P < 2.2 \times 10^{-16}$ $t$-test) positive correlations for both the transcriptome (0.54 PCC; Fig. 5c) and the proteome (0.57 PCC; Fig. 5d). This degree of concordance was unexpected because the model does not account for transcriptional regulation or transcript-specific RNA degradation rates. However, this concordance may be the result of our simulation objective being aligned with *T. maritima*'s regulatory programme, whereas a decreased concordance would be expected if the regulatory network was responding to a stress. We have previously observed a tendency to increase the expression of metabolically efficient pathways, and decrease inefficient alternatives, by *E. coli* After adaptive evolution under growth selection pressure[31]. Also, we have observed that *T. maritima*'s genome is highly active with > 89% of the protein-coding genes expressed in diverse conditions (H.L. *et al.*, Unpublished data), which could indicate a general eschewal of complex and expensive circuitry within the global regulatory strategy.

Approximately 30% of *T. maritima*'s genome is not functionally annotated and 50% of the functionally annotated genes fall outside of the scope of our ME-Model. A number of genes not accounted in our model were expressed *in vivo* (Supplementary Fig. S1), and the costs of their expression as well as their functional activities may contribute to the differences

between simulation and measurement. In addition, unknown regulatory features might be responsible for irregularities observed when comparing simulation to the measurement. For instance, ribosomal RNAs and proteins are expected to be expressed at stoichiometric ratios, as occurs with the simulation, yet there is sizable variability in their measured values (Fig. 5c,d, blue colouring). These results illustrate that it is possible to sketch a molecular description of a replicating organism solely from simple, but stoichiometrically accurate, chemical equations represented on a genome scale.

### *In silico* gene expression profiling drives discovery

With our ME-Model it is now possible to compute the gene expression profile associated with growth in a specific condition or for a specific mutant. These gene expression profiles may then be compared to identify genes that are likely differentially regulated. The set of differentially expressed *in silico* genes may then be used to drive biological discovery or improve our model (Fig. 6).

Towards this end, we computed the transcriptome profiles for *T. maritima* grown in a minimal medium with either L-Arab or cellobiose as the sole carbon source (Fig. 6a). Our computations identified genes that were exclusively expressed and essential for growth with each carbon source. Because these genes are essential for growth on the respective substrate they are conditionally essential genes. Conditionally essential genes are often subject to transcriptional regulation, however, they may be constitutively expressed. To assess whether the genes were differentially expressed *in vivo*, we measured the transcriptome of *T. maritima* growing in minimal medium with L-Arab or cellobiose as the carbon source. The genes with the strongest differential expression *in vivo* were among the set of differentially expressed genes *in silico* (Fig. 6b) providing supporting evidence for the presence of transcriptional regulation.

Conditionally expressed genes may be regulated by the same TF[33]. Thepresence of a common motif in the promoter regions of a set of genes may indicate regulation by a common TF. To identify potential TF-binding motifs, we scanned the promoter and upstream regions of the *in silico* differentially expressed genes with MEME (Multiple Expectation Maximum for Motif Elicitation)[34]. Surprisingly, there was a high-scoring motif for the genes essential for growth on L-Arab and a high-scoring motif for the genes essential for growth on cellobiose (Fig. 6c). The motif found upstream of the L-Arab upregulated genes is similar to the AraR motif from *Bacillus subtilis*[35] (Supplementary Fig. S2). Also, the motif upstream of the cellobiose upregulated genes bears resemblance to catabolite-responsive elements (*cres*), known to have an important global role in catabolite repression through the binding of the CcpA protein in *B. subtilis*[36]. Here, we term the motif the CelR motif, as the regulated genes are involved in cellobiose metabolism. These discoveries highlight how ME-Model simulations can guide discovery of new regulons.

After identifying the putative AraR and CelR motifs, we scanned *T. maritima*'s genome for the presence of other members of the putative regulons. For the nondegenerate AraR motif 5-GTACG TAC-3, we identified a single additional instance in an intergenic region upstream of the TU-containing genes TM0277, TM0278 and TM0279 (Fig. 6d). These genes were induced when L-Arab was the carbon source, but not when cellobiose or maltose serves as the carbon source (Supplementary Fig. S3). L-Arab transport is an orphaned activity in our model, which means that *T. maritima* may import L-Arab, however, the responsible loci are not known. When we examined these genes using the SEED RAST server[37], TM0278 and TM0279 were classified as permeases of an ABC transporter putatively involved in L-Arab utilization, whereas TM0277 was not classified because it was annotated as containing an authentic frameshift[38]. Recent resequencing of *T. maritima*'s genome (H.L. *et al.*, Unpublished data) refute the initial annotation that TM0277 contains a

frameshift mutation; and the SEED RAST annotation for TM0277 is a predicted sugar-binding protein for an arabinoside ABC transporter. Interestingly, the TUs containing ABC transporters for maltose and chitobiose are organized in the same manner: a binding protein followed by two permeases. The presence of the AraR motif, the strong upregulation of the TM0277/TM0278/TM0279 TU in response to L-Arab *in vivo*, the SEED RAST classification and resequenced genome strongly suggest that we have identified a functional L-Arab transport system in this organism. This discovery illustrates how *in silico* molecular biology at the genome scale can be used to expand regulons and improve genome annotation.

When we scanned *T. maritima*'s genome for matches to the degenerate CelR motif TGWAAAYRTTTWCA, the promoter regions of TUs associated with cellobiose metabolism were identified. Interestingly, the promoter region of the TU-containing TM1222, TM1221, TM1220, TM1219 and TM1218 did not contain a CelR motif (Fig. 6c,d). TM1222, TM1221, TM1220 and TM1219 encode for a cellobiose ABC transporter, while TM1218 is annotated as a LacI family transcription regulator. However, the promoter region of the TU for TM1233, which is directly upstream of TM1222, contains the CelR motif. TM1233 encodes for the cellobiose-binding protein that facilitates cellobiose transport. In the TU architecture of our model, there was a predicted Rho-independent terminator following TM1223 that resulted in a new TU starting with TM1222. However, no promoter was detected in the intergenic region between TM1223 and TM1222 using PromBase[39]. Thisresult leads us to believe that the initial assignment of TM1223 and TM1222 to separate TUs was incorrect (Fig. 6d). The presence of the cellobiose transport system in the updated TU, the strong CelR motif and the annotation of TM1218 as a TF suggest that TM1218 may encode for CelR.

## Discussion

Our ME-Modelling approach represents a fundamental advance in the evolution of genome-scale biochemical models of life and significantly broadens the scope of microbial systems biology. It is now possible to ask systems-level questions *in silico* beyond metabolism and quantitatively analyse, in a bottom-up and mechanistic manner, a variety of omics data in the context of a growing organism. For instance, we can use a systems perspective to identify the minimal number of genes required to support homeostasis and replication— 120 of the 142 of the proposed minimal bacterial genome[32] were essential for ribosome production in maltose minimal medium (Supplementary Data 2).

Not only can ME-Models predict global phenotypes that are traditionally employed with M-Models, such as maximal growth rate in a defined medium, but they can also be used to calculate whether the system has any material and energy reserves available for ancillary functions. For example, the measured maltose consumption rate was greater than the one that we calculated for economically efficient growth (Fig. 5a). This discrepancy between measurement and simulation could indicate that *T. maritima* does not strive for economic efficiency or represent the portion of sugar used to support the activities of the unannotated genes or regulatory circuitry. Given that the expression levels for the gene products associated with the more efficient pathways were highly expressed (Fig. 4c), we are disposed towards the latter. Although the ME-Model does not account for regulatory events, the presence of a strong discordance between simulation and measurement would indicate that factors other than economic efficiency are influencing the expressome, thus informing hypothesis generation. For example, if a more expensive isozyme was expressed *in vivo* than *in silico*, then it would be possible to estimate the improvement in $k_{cat}$ required for the expensive isozyme to offset its higher materials and energy costs.

Technological advances have contributed to an expanding ocean of omics data that has been under-explored[3]. Omics data have been under-analysed, in part, due to the lack of a mechanistic systems-level framework for analysing myriad molecular components in the context of cellular physiology. To date, with the notable exception of C13 metabolic flux analysis, it has only been possible to perform indirect comparative analysis between omics data and M-Models[31] or to neglect the complexity of the genotype–phenotype relationship and use omics data as *ad hoc* constraints for M-Model enzyme activities[9–12]. Because ME-Models explicitly represent gene expression, directly investigating omics data in the context of the whole is now feasible.

Viewing multi-omics data in the context of biochemically and genomically consistent ME-Models may allow us to extract more value from legacy and future omics data. Comparing *in silico* and *in vitro* transcriptomes, or proteomes, can highlight under-explored areas of molecular biology. For example, a set of genes highly expressed *in silico* but not expressed *in vivo* may indicate the presence of transcriptional regulation. Differential expression of a class of genes may indicate incompleteness in our knowledge of how those gene products interact or allude to, heretofore unknown, moonlighting functions. For instance, in the case of ribosomal proteins (Fig. 5c,d, blue) the model predicts uniform expression, whereas omics measurements exhibit variability. The model was designed based on evidence that ribosomal protein synthesis is highly coordinated[40], and does not account for feedback circuits affecting degradation rates that have yet to be fully elucidated[40,41].

Although there is a positive correlation between the simulated transcriptome fluxes and semiquantitative transcriptome data there was still a substantial amount of dispersion (Fig. 5c). When comparing *in silico* and *in vivo* transcriptome measurements it is important to realize that both are approximations of the transcript levels in an organism, and that omics technologies have been inherently noisy to date[42]. Incomplete knowledge, such as a lack of specific translation efficacy for each protein and degradation rates for each mRNA, and lack of signalling and regulatory circuitry will contribute to deviations from reality by ME-Model simulations. Similarly, probe-binding and sample-labelling efficacies, as well as other technical issues, serve as barriers to absolute quantitative transcriptome measurements[43].

Although it is a non-trivial endeavour to identify the source of all variation between the simulated and measured transcriptomes, it is possible to use the ME-Model for comparative transcriptomics approaches similar to two-channel DNA microarray studies. Despite the early technological limitations of DNA microarrays, biological discovery was enabled by performing comparative transcriptomics[44–47]. Transcriptome profiling has been used extensively to identify genes that are differentially regulated as a function of genetics and environment[44]. Analysis of differentially expressed genes has contributed to the identification of gene products responsible for unannotated enzymatic activities[45]. In combination with sequence analysis, differential gene expression data can be used to investigate transcriptional regulation[46,47].

We devised and implemented a workflow for *in silico* comparative transcriptomics, which resulted in the discovery of new regulons and improved both genome and TU annotation (Fig. 6a–d). The similarities between the comparative transcriptomics *in silico* (Fig. 6a) and *in vivo* (Fig. 6b) studies are striking, given the variation observed between the simulated and measured transcriptomes (Fig. 5c)—this emphasizes that, in spite of its shortcomings, the ME-Modelling framework is a powerful tool for biological research.

Finally, ME-Models enable integrated molecular biology on a genome scale while accounting for the metabolic requirements, which partially fulfills the challenge of Project K[48] and moves us one step closer to a molecular representation of CellMap[1].

## Methods

### Network reconstruction procedure

The detailed procedure and formalism are described in detail in the Supplementary Methods. Our method accounts for biochemical reactions associated with transcription of TUs, TU degradation, translation, protein maturation, RNA processing, protein complex formation, ribosomal assembly, rRNA modification, tRNA modification, tRNA charging, aminoacyl-tRNA synthetase charging, charging EF-TU, cleavage of polycistronic TUs to release stable RNA products, sources, sinks and tRNA activation (EF-TU) as well as metabolism. In our formalism, metabolic reactions are represented as multi-step processes including substrate binding by the enzyme and dissociation of substrate–enzyme complex to enzyme and products. The metabolic content for our reconstruction was based on the previously published model[17], with updates to correct errors and incorporate new data (Supplementary Data 3).

Te molecular machinery (for example, proteins, genes, RNAs) involved in macromolecular synthesis were identified from the genome annotation[16], SEED subsystem analysis[49], comparative genomics analysis of the *E. coli* model[28] and KEGG[38]. Thefunctions of each of the 159 proteins associated with macromolecular synthesis in *T. maritima* were determined by primary literature when available. When no primary literature was available, the Uniprot[50] and SEED[49] databases were used to infer function by homology. All proteins currently believed to be used for macromolecular synthesis by *T. maritima* are enumerated in Supplementary Data 4, and 93% of these genes are mechanistically linked in our ME-Model.

Te reactions associated with transcription and translation, including initiation, biopolymerization and termination, were generated from the genome sequence and a set of *T. maritima* template reactions (Supplementary Methods). In our modelling formalism, reversible reactions were represented as two unique reactions: one for the forward direction and one for the reverse.

### Protein complexes

For each functional protein, we used primary literature and the RCSB Protein Data Bank51 to determine whether the machine was a monomer or oligomer. The Protein Data Bank entries provided an opportunity to integrate 3D structural data into our reconstruction (this model includes structures for 32 additional open reading frames compared with Zhang et al.). When data for multimeric state were unavailable for a protein of interest, state data for orthologs from closely related organisms were used; otherwise, the Uniprot database50 was consulted. In the absence of data providing insight into the multimeric state of the protein, we assumed that the functional protein was a monomer.

### Genetic code determination

From inspection of tRNA sequences and structures downloaded from the transfer RNA database[52], we determined that T. maritima uses uniform-GUC decoding with only 46 tRNA genes (see Supplementary Data 5). In both Archaea and Bacteria, but not in Eukarya, the conversion of C34 of a CAU-anticodon to lysidine (k2C) or analogue generates an anticodon for isoleucine[53]. TMtRNA-Met-2 was assigned this role based on a strong sequence alignment to E. coli tRNAs containing k2C.TheT. maritima genome encodes two additional tRNA genes with CAU anticodons, TMtRNA-Met-1 and TMtRNA-Met-3. Based on structural similarity[54] to those found in a crystal structure of E. coli's formyl-methionyl-tRNAfMet[55], TMtRNA-Met-1 may be involved in translation initiation, therefore, TMtRNA-Met-3 was designated to participate in translation elongation.

**TU architecture determination**

We assembled a draft TU architecture (Supplementary Data 6) for T. maritima based on a series of rules (Supplementary Methods). In short, we assumed all TUs start with a gene start and proceed until one of the following conditions is met: (1) two genes are found in convergent orientation on different strands, (2) two genes are found in divergent orientation on different strands, (3) a high-confidence Rho-independent transcription terminator is found separating two genes oriented in series on the same strand, (4) more than 55 base pairs separate two genes in series on the same strand or (5) experimental evidence indicates a TU boundary. Finally, to reflect the possibility of internal transcription start sites in TUs reconstructed using the rules above, we added an additional TU in cases where a high-confidence promoter was found in the region separating two genes oriented in series on the same strand.

***In silico* molecular biology**

Log-phase growth simulations were performed using FBA24. Linear programming was used to identify the maximum $\mu$ or minimum ribosome production flux supporting a particular $\mu$ from the components of the *in silico* minimal media. Because of the presence of fast (metabolic) and slow (macromolecular synthesis) timescale reactions, the parameters in the ME-Model span a wide range that can result in inaccurate simulations due to floating point limitations of currently available floating point linear programming software (Supplementary Methods). To remove the possibility of simulation results being artefacts arising from floating point limitations, we used the exact simplex routines available in the QSopt_ex package[26], with default parameter settings for ME-Model simulations. The predicted transcription level of a gene was determined by summing across the sink fluxes of TUs containing the gene, which is equivalent to the transcription fluxes less the TU degradation fluxes. Translation levels were reported as the sum across the relevant translation initiation fluxes, as many TUs can contribute to the production of a given protein. These values were compared with each other in the case of simulated nutrient shifts or to the abundances reported experimentally.

***In vivo* methods**

*T. maritima* MSB8 (ATCC: 43589) was grown in 500 ml serum bottles containing 200 ml of anoxic minimal media with 10 mM maltose, l-arabinose or cellobiose as the sole carbon source at 80 °C. All samples were collected during log-phase growth. Substrate uptake and by-product secretion rates, compositional analyses, and transcriptome and proteome measurements were performed as described in the Supplementary Methods. Transcriptome data have been submitted to the NCBI Gene Expression Omnibus (accession ID: GSE28822) and processed values are in Supplementary Data 7. Proteomics data are available through Pacific Northwest National Laboratory (http://omics.pnl.gov) and processed values are in Supplementary Data 8.

**RNA modifications**

A variety of post-transcriptional modifications of rRNAs are represented in our model. For 16S rRNA, there was experimental evidence for ten modifications[56] in this organism (Supplementary Table S4). The locations of pseudouridines, which are mass silent, were not available, but an 11th modification, U to Y at position 516, was included in the reconstruction based on the fact that it is well conserved in bacteria and the alignment (Supplementary Data 9) supports its inclusion. An unusual derivative of cytidine-designated N-330 has been sequenced to position 1,404 (ref. 56) in the decoding region of the 16S rRNA. This modified nucleoside was excluded from the reconstruction as the exact chemical composition of the modification is unknown. We were unable to find organism-specific

literature supporting modifications to the 5S and the 23S rRNA. Modifications to 5S rRNA are infrequent in bacteria[57]. Attempting to extrapolate 23S rRNA modifications from *E. coli* was relatively unsuccessful as alignment via ClustalW2[58] showed significant differences near many of the putative modification sites (Supplementary Data 10). The alignment reveals that the 23S rRNA of *T. maritima* is significantly longer (> 100 bp) than that of *E. coli*. Only three proteins with annotated roles in modifying the 23S rRNA were added to the model for a total of six modifications (Supplementary Table S5). Those were TM0940, TM0462 and TM1715.

Post-transcriptional modification of tRNA also requires a significant investment in genes, enzymes, substrates and energy[59]. We included a variety of modifications (Supplementary Table S6) in our model based on bioinformatics predictions and literature evidence (Supplementary Table S7).

### Sensitivity analysis

To explore the influence of some of the newly introduced parameters on model output, the bulk parameters used for the coupling constraints (Supplementary Methods) were varied (two-, four- and eight-fold increases and decreases away from the parameter set used). The results are summarized in Supplementary Fig. S4.

### File formats

Our final model is available as a Systems Biology Markup Language (SBML) XML file (Supplementary Data 11). The model is also available as an LP file (Supplementary Data 12) for use with linear programming solvers.

## Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

## Acknowledgments

## References

1. Brenner S. Sequences and consequences. Philos Trans R Soc Lond B Biol Sci. 2010; 365:207–212. [PubMed: 20008397]

2. Otero JM, Nielsen J. Industrial systems biology. Biotechnol Bioeng. 2010; 105:439–460. [PubMed: 19891008]

3. Palsson B, Zengler K. The challenges of integrating multi-omic data sets. Nat Chem Biol. 2010; 6:787–789. [PubMed: 20976870]

4. Mahadevan R, Palsson BO, Lovley DR. *In situ* to *in silico* and back: elucidating the physiology and ecology of *Geobacter* spp. using genome-scale modelling. Nat Rev Microbiol. 2011; 9:39–50. [PubMed: 21132020]

5. Feist AM, Palsson BO. The growing scope of applications of genome-scale metabolic reconstructions using *Escherichia coli*. Nat Biotechnol. 2008; 26:659–667. [PubMed: 18536691]

6. Oberhardt MA, Palsson BO, Papin JA. Applications of genome-scale metabolic reconstructions. Mol Syst Biol. 2009; 5:320. [PubMed: 19888215]

7. Reed JL, Palsson BO. Genome-scale *in silico* models of *E. coli* have multiple equivalent phenotypic states: assessment of correlated reaction subsets that comprise network states. Genome Res. 2004; 14:1797–1805. [PubMed: 15342562]

8. Schellenberger J, Lewis NE, Palsson BO. Elimination of thermodynamically infeasible loops in steady-state metabolic models. Biophys J. 2011; 100:544–553. [PubMed: 21281568]

9. Akesson M, Forster J, Nielsen J. Integration of gene expression data into genome-scale metabolic models. Metab Eng. 2004; 6:285–293. [PubMed: 15491858]

10. Becker SA, Palsson BO. Context-specific metabolic networks are consistent with experiments. PLoS Comput Biol. 2008; 4:e1000082. [PubMed: 18483554]

11. Colijn C, et al. Interpreting expression data with metabolic flux models: predicting Mycobacterium tuberculosis mycolic acid production. PLoS Comput Biol. 2009; 5:e1000489. [PubMed: 19714220]

12. Shlomi T, Cabili MN, Herrgard MJ, Palsson BO, Ruppin E. Network-based prediction of human tissue-specific metabolism. Nat Biotechnol. 2008; 26:1003–1010. [PubMed: 18711341]

13. Allen TE, Palsson BO. Sequence-based analysis of metabolic demands for protein synthesis in prokaryotes. J Teor Biol. 2003; 220:1–18.

14. Tiele, I. Dissertation: A Stoichiometric Model of Escherichia coli's Macromolecular Synthesis Machinery and its Integration with Metabolism. ProQuest; Ann Arbor, MI: 2008.

15. Schröder C, Selig M, Schönheit P. Glucose fermentation to acetate, CO2, and H2 in the anaerobic hyperthermophilic eubacterium Termotoga maritima: involvement of the Embden-Meyerhof pathway. Arch Microbiol. 1994; 161:460–470.

16. Nelson KE, et al. Evidence for lateral gene transfer between Archaea and bacteria from genome sequence of *Termotoga maritima*. Nature. 1999; 399:323–329. [PubMed: 10360571]

17. Zhang Y, et al. Tree-dimensional structural view of the central metabolic network of *Termotoga maritima*. Science. 2009; 325:1544–1549. [PubMed: 19762644]

18. Kummerfeld SK, Teichmann SA. DBD: a transcription factor prediction database. Nucleic Acids Res. 2006; 34:D74–81. [PubMed: 16381970]

19. Andrianantoandro E, Basu S, Karig DK, Weiss R. Synthetic biology: new engineering rules for an emerging discipline. Mol Syst Biol. 2006; 2 2006.0028.

20. Vickers CE, Blank LM, Kromer JO. Grand challenge commentary: Chassis cells for industrial biochemical production. Nat Chem Biol. 2010; 6:875–877. [PubMed: 21079595]

21. Schaechter M, Maaloe O, Kjeldgaard NO. Dependency on medium and temperature of cell size and chemical composition during balanced grown of *Salmonella typhimurium*. J Gen Microbiol. 1958; 19:592–606. [PubMed: 13611202]

22. Scott M, Gunderson CW, Mateescu EM, Zhang Z, Hwa T. Interdependence of cell growth and gene expression: origins and consequences. Science. 2010; 330:1099–1102. [PubMed: 21097934]

23. Feist AM, Palsson BO. The biomass objective function. Curr Opin Microbiol. 2010; 13:344–349. [PubMed: 20430689]

24. Orth JD, Tiele I, Palsson BO. What is flux balance analysis? Nat Biotechnol. 2010; 28:245–248. [PubMed: 20212490]

25. Rinker KD, Kelly RM. Growth physiology of the hyperthermophilic Archaeon *Thermococcus litoralis*: development of a sulfur-free defined medium, characterization of an exopolysaccharide, and evidence of biofilm formation. Appl Environ Microbiol. 1996; 62:4478–4485. [PubMed: 16535464]

26. Applegate DL, Cook W, Dash S, Espinoza DG. Exact solutions to linear programming problems. Operations Res Lett. 2007; 35:693–699.

27. Gupta RS, Schlessinger D. Coupling of rates of transcription, translation, and messenger ribonucleic acid degradation in streptomycin-dependent mutants of *Escherichia coli*. J Bacteriol. 1976; 125:84–93. [PubMed: 1107319]

28. Thiele I, Jamshidi N, Fleming RM, Palsson BO. Genome-scale reconstruction of *Escherichia coli*'s transcriptional and translational machinery: a knowledge base, its mathematical formulation, and its functional characterization. PLoS Comput Biol. 2009; 5:e1000312. [PubMed: 19282977]

29. Holzhutter HG. The principle of flux minimization and its application to estimate stationary fluxes in metabolic networks. Eur J Biochem. 2004; 271:2905–2922. [PubMed: 15233787]

30. Pramanik J, Keasling JD. Stoichiometric model of *Escherichia coli* metabolism: incorporation of growth-rate dependent biomass composition and mechanistic energy requirements. Biotechnol Bioeng. 1997; 56:398–421. [PubMed: 18642243]

31. Lewis NE, et al. Omic data from evolved E coli are consistent with computed optimal growth from genome-scale models. Mol Syst Biol. 2010; 6:390. [PubMed: 20664636]

32. Gil R, Silva FJ, Pereto J, Moya A. Determination of the core of a minimal bacterial gene set. Microbiol Mol Biol Rev. 2004; 68:518–537. [PubMed: 15353568]

33. Browning DF, Busby SJ. The regulation of bacterial transcription initiation. Nat Rev Microbiol. 2004; 2:57–65. [PubMed: 15035009]

34. Bailey TL, et al. MEME SUITE: tools for motif discovery and searching. Nucleic Acids Res. 2009; 37:W202–8. [PubMed: 19458158]

35. Franco IS, Mota LJ, Soares CM, de Sa-Nogueira I. Probing key DNA contacts in AraR-mediated transcriptional repression of the *Bacillus subtilis* arabinose regulon. Nucleic Acids Res. 2007; 35:4755–4766. [PubMed: 17617643]

36. Miwa Y, Nakata A, Ogiwara A, Yamamoto M, Fujita Y. Evaluation and characterization of catabolite-responsive elements (cre) of *Bacillus subtilis*. Nucleic Acids Res. 2000; 28:1206–1210. [PubMed: 10666464]

37. Aziz RK, et al. The RAST Server: rapid annotations using subsystems technology. BMC Genom. 2008; 9:75.

38. Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 2000; 28:27–30. [PubMed: 10592173]

39. Rangannan V, Bansal M. PromBase: a web resource for various genomic features and predicted promoters in prokaryotic genomes. BMC Res Notes. 2011; 4:257. [PubMed: 21781326]

40. Dennis PP. *In vivo* stability, maturation and relative differential synthesis rates of individual ribosomal proteins in *Escherichia coli* B/r. J Mol Biol. 1974; 88:25–41. [PubMed: 4613843]

41. Singer P, Nomura M. Stability of ribosomal protein mRNA and translational feedback regulation in *Escherichia coli*. Mol Gen Genet. 1985; 199:543–546. [PubMed: 3897787]

42. Ji H, Liu XS. Analyzing 'omics data using hierarchical models. Nat Biotechnol. 2010; 28:337–340. [PubMed: 20379180]

43. Canales RD, et al. Evaluation of DNA microarray results with quantitative gene expression platforms. Nat Biotechnol. 2006; 24:1115–1122. [PubMed: 16964225]

44. Schena M, Shalon D, Davis RW, Brown PO. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science. 1995; 270:467–470. [PubMed: 7569999]

45. Kharchenko P, Vitkup D, Church GM. Filling gaps in a metabolic network using expression information. Bioinformatics. 2004; 20(Suppl 1):i178–85. [PubMed: 15262797]

46. Sabatti C, Rohlin L, Oh MK, Liao JC. Co-expression pattern from DNA microarray experiments as a tool for operon prediction. Nucleic Acids Res. 2002; 30:2886–2893. [PubMed: 12087173]

47. Rhodius VA, LaRossa RA. Uses and pitfalls of microarrays for studying transcriptional regulation. Curr Opin Microbiol. 2003; 6:114–119. [PubMed: 12732299]

48. Crick F. Project K: The Complete Solution of *E. coli*. Perspect Biol Med. 1973; 17:67–70.

49. Overbeek R, et al. The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. Nucleic Acids Res. 2005; 33:5691–5702. [PubMed: 16214803]

50. Wu CH, et al. The Universal Protein Resource (UniProt): an expanding universe of protein information. Nucleic Acids Res. 2006; 34:D187–91. [PubMed: 16381842]

51. Rose PW, et al. The RCSB Protein Data Bank: redesigned web site and web services. Nucleic Acids Res. 2011; 39:D392–401. [PubMed: 21036868]

52. Juhling F, et al. tRNAdb 2009: compilation of tRNA sequences and tRNA genes. Nucleic Acids Res. 2009; 37:D159–62. [PubMed: 18957446]

53. Tong KL, Wong JT. Anticodon and wobble evolution. Gene. 2004; 333:169–177. [PubMed: 15177692]

54. Mandal N, Mangroo D, Dalluge JJ, McCloskey JA, Rajbhandary UL. Role of the three consecutive G:C base pairs conserved in the anticodon stem of initiator tRNAs in initiation of protein synthesis in *Escherichia coli*. RNA. 1996; 2:473–482. [PubMed: 8665414]

55. Schmitt E, Panvert M, Blanquet S, Mechulam Y. Crystal structure of methionyl-tRNAfMet transformylase complexed with the initiator formyl-methionyl-tRNAfMet. EMBO J. 1998; 17:6819–6826. [PubMed: 9843487]

56. Guymon R, Pomerantz SC, Ison JN, Crain PF, McCloskey JA. Post-transcriptional modifications in the small subunit ribosomalRNAfrom Thermotoga maritima, including presence of a novel modifed cytidine. RNA. 2007; 13:396–403. [PubMed: 17255199]

57. Szymanski M, Barciszewska MZ, Erdmann VA, Barciszewski J. 5S Ribosomal RNA Database. Nucleic Acids Res. 2002; 30:176–178. [PubMed: 11752286]

58. Larkin MA, et al. Clustal W and Clustal X version 2.0. Bioinformatics. 2007; 23:2947–2948. [PubMed: 17846036]

59. Gustilo EM, Vendeix FA, Agris PF. tRNA's modifications bring order to gene expression. Curr Opin Microbiol. 2008; 11:134–140. [PubMed: 18378185]
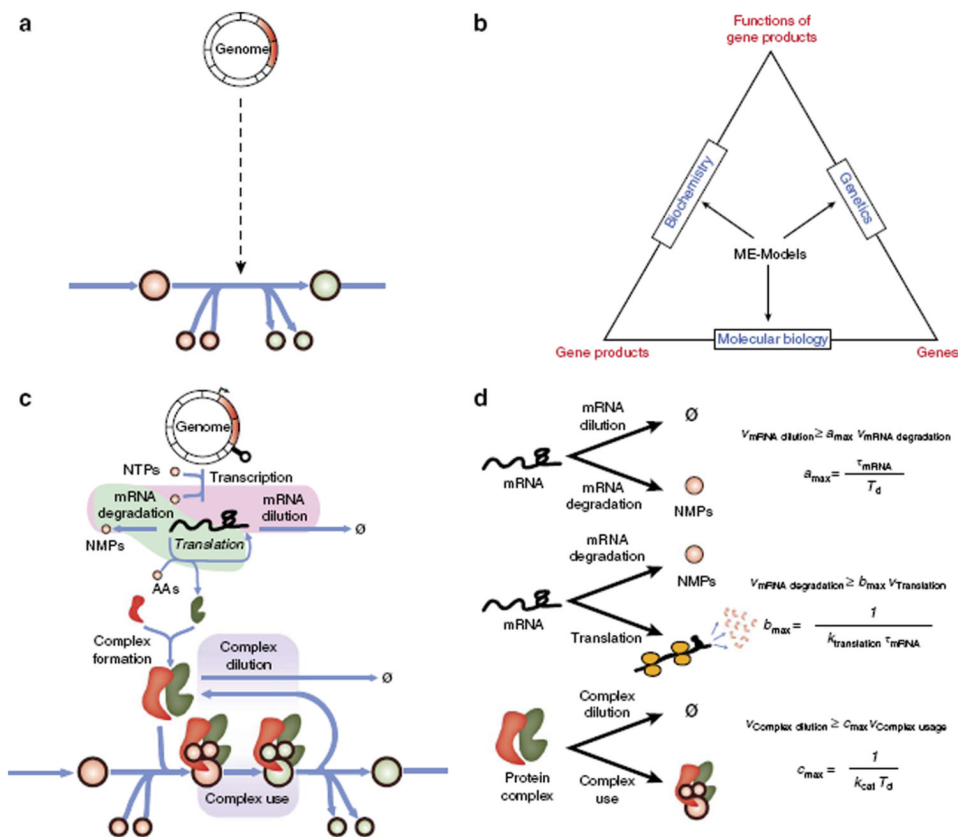
**Figure 1. Genome-scale modelling of metabolism and expression**

(**a**) Modern stoichiometric models of metabolism (M-Models) relate genetic loci to their encoded functions through causal Boolean relationships. The gene and its functions are either present or absent. The dashed arrow signifies incomplete and/or uncertain causal knowledge, whereas blue arrows signify mechanistic coverage. (**b**) ME-Models provide links between the biological sciences. With an integrated model of metabolism and macromolecular expression, it is possible to explore the relationships between gene products, genetic perturbations and gene functions in the context of cellular physiology. (**c**) models of metabolism and expression (ME-Models) explicitly account for the genotype–phenotype relationship with biochemical representations of transcriptional and translational processes. This facilitates quantitative modelling of the relation between genome content, gene expression and cellular physiology. (**d**) When simulating cellular physiology, the transcriptional, translational and enzymatic activities are coupled to doubling time ($T_d$) using constraints that limit transcription and translation rates as well as enzyme efficiency. $\tau_{mRnA}$, mRnA half-life; $k_{cat}$, catalytic turnover constant; $k_{translation}$, translation rate; $v$, reaction flux.
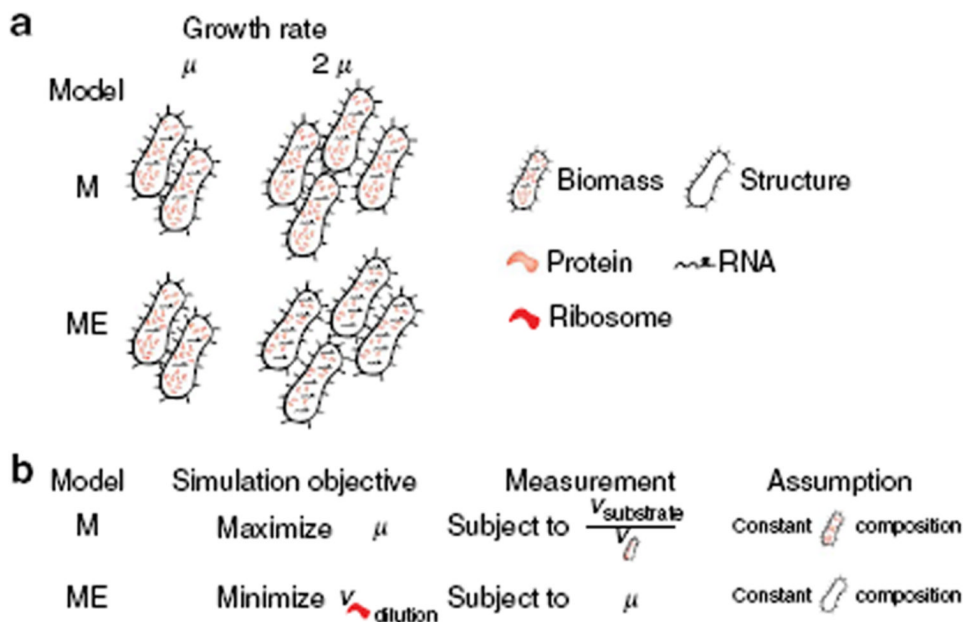
**Figure 2. Comparison of M- and ME-Models objective functions and assumptions**
(**a**) M-Models simulate constant cellular composition (biomass) as a function of specific growth rate ($\mu$), whereas ME-Models simulate constant structural composition with variable composition of proteins and transcripts. (**b**) Linear programming simulations with M-Models are designed to identify the maximum $\mu$ that is subject to experimentally measured substrate uptake rates. only biomass yields are predicted as $\mu$ enters indirectly as an input through the supplied substrate uptake rate (see the measurement column for M-Models). Importantly, the substrate uptake rate is derived by normalizing to biomass production. Linear programming simulations with ME-Models aim to identify the minimum ribosome production rate required to support an experimentally determined $\mu$. $\mu$ enters into the coupling constraints and so it must be supplied (or sampled) as the problem would otherwise be a nonlinear Program (nLP). As all M-Models reactions are contained within the ME-Models, ME-Models can simulate all M-Models objectives in addition to the broad range of objectives associated with macromolecular expression.
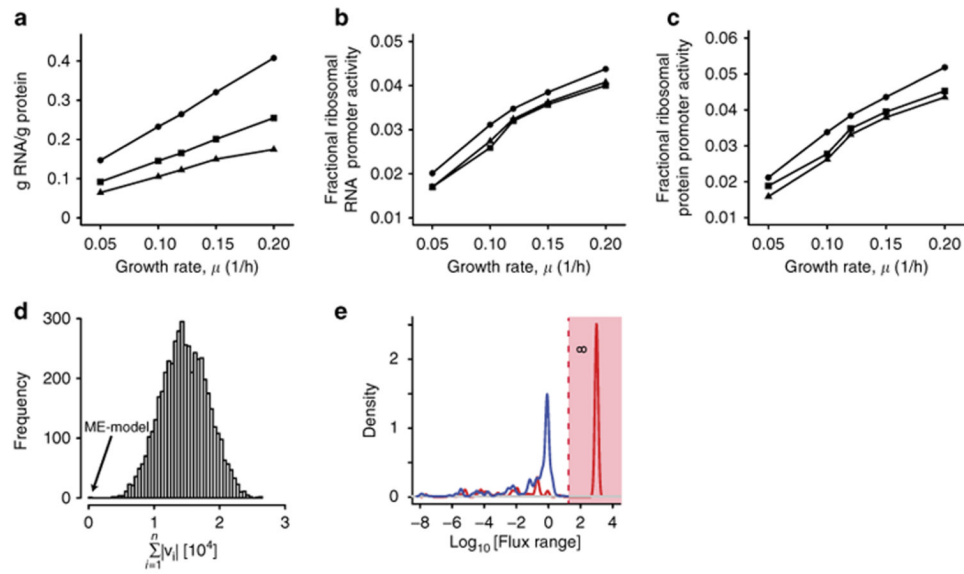
**Figure 3. Simulation of variable cellular composition and efficient use of enzymes**
(**a**) With our ME-Model, the RnA/protein ratio increases linearly with growth rate and with a slope proportional to translational capacity in amino acids per second (circles: 5 AA/s, squares: 10 AA/s, triangles: 20 AA/s). (**b**) Ribosomal RNA (rRnA) synthesis increases, relative to total RNA synthesis, with growth rate (symbols as in **a**). (**c**) Ribosomal protein promoter activity increases, relative to total RNA synthesis, with growth rate (symbols as in **a**). (**d**) Random sampling of the M-Model solution space indicates that the M-Model solution space contains numerous internal solutions with a broad range of total network flux. The probability of finding an M-Model solution as efficient as an ME-Model simulation is $2.1 \times 10^{-5}$; the probability was calculated from a normal distribution constructed from the M-Model sample space. The M-Model sample contains 5,000 flux vectors randomly sampled from the M-Model solution space. (**e**) smooth estimate of the density of the flux ranges for the metabolic enzymes that may be simulated while maintaining the objective for efficient growth with a 1% tolerance (M-Model: red line, ME-Model: blue line). The shaded area denotes biologically unrealistic flux values. All simulations were performed with an *in silico* minimal medium with maltose as the sole carbon source.
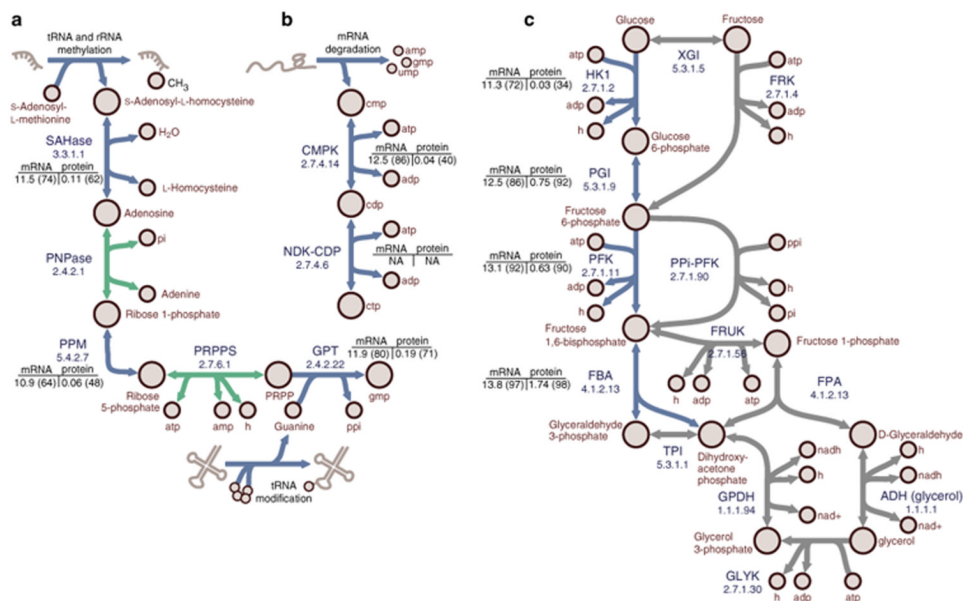
**Figure 4. Metabolic reactions required for efficient growth with the ME-Model but not the M-Model**

(**a**) Recycling of by-products of RNA modifications. Adenosylhomocysteinease (SAHase) hydrolyses *S*-adenosylhomocysteine (SAH) to L-homocysteine (L-HCys) and adenosine. Purine nucleoside phosphorylase (PNP) phosphorylases adenosine to adenine and ribose-1-phosphate (Rib-1-P). Rib-1-P is converted to ribose-5-phosphate (Rib-5-P) by phosphopentomutase (PPm). Phosphoribosylpyrophosphate synthetase (PRPPs) phosphorylates Rib-5-P to produce 5-phosphoribosol-1-pyrophosphate (PRPP). Guanine phosphoribosyltransferase (GPT) produces GMP from the reaction of PRPP and guanine, which is a by-product of tRnA metabolism. (**b**) CmP produced during mRnA degradation is recycled to CTP using cytidylate kinase (CMPK) and nucleoside-diphosphate kinase (NDK-CDP). (**c**) The ME-Model uses the canonical glycolytic pathway, whereas with the M-Model can circumvent portions during optimal growth simulations. The canonical pathway involves phosphorylation of D-glucose (D-Glc) to glucose-6-phospate (G6P) by hexokinase (HK1). G6P is isomerized to fructose-6-phosphate (F6P) by phosphoglucose isomerase (PGI). F6P is phosphorylated to fructose-1,6-bisphosphate (FBP) by phosphofructokinase (PFK). FBP is metabolized to glyceraldehyde-3-phosphate (G3P) and dihydroxyacetone phosphate (DHAP) by FBP aldolase (FBA). The M-Model can circumvent the HK1/PGI portion with glucose/xylose isomerase (GXI) and fructokinase (FRK); however, HK1 or PGI must also be expressed because G6P is an essential metabolite. PFK can be circumvented by diphosphate-fructose-6-phosphate 1-phosphotransferase (PPi-PFK). FBA can be circumvented by a pathway using 1-phosphofructokinase (FRUK), fructose-1-phosphate aldolase (FPA), alcohol dehydrogenase (ADH(glycerol)), glycerol kinase (GLYK), glycerol-3-phosphate dehydrogenase (GPDH) and triose phosphate isomerase (TPI). Enzyme commission numbers are provided for each reaction. mRNA and protein expression (and quantile) values are provided. Flux variability analysis was performed for simulated growth on maltose minimal medium. Blue arrows: reactions required for optimally efficient growth with the ME-Model, but not the M-Model. Green arrows: active reactions in a single maltose minimal medium simulation shown to put results into pathway context. Grey arrows: alternate optimal pathways in the M-Model.
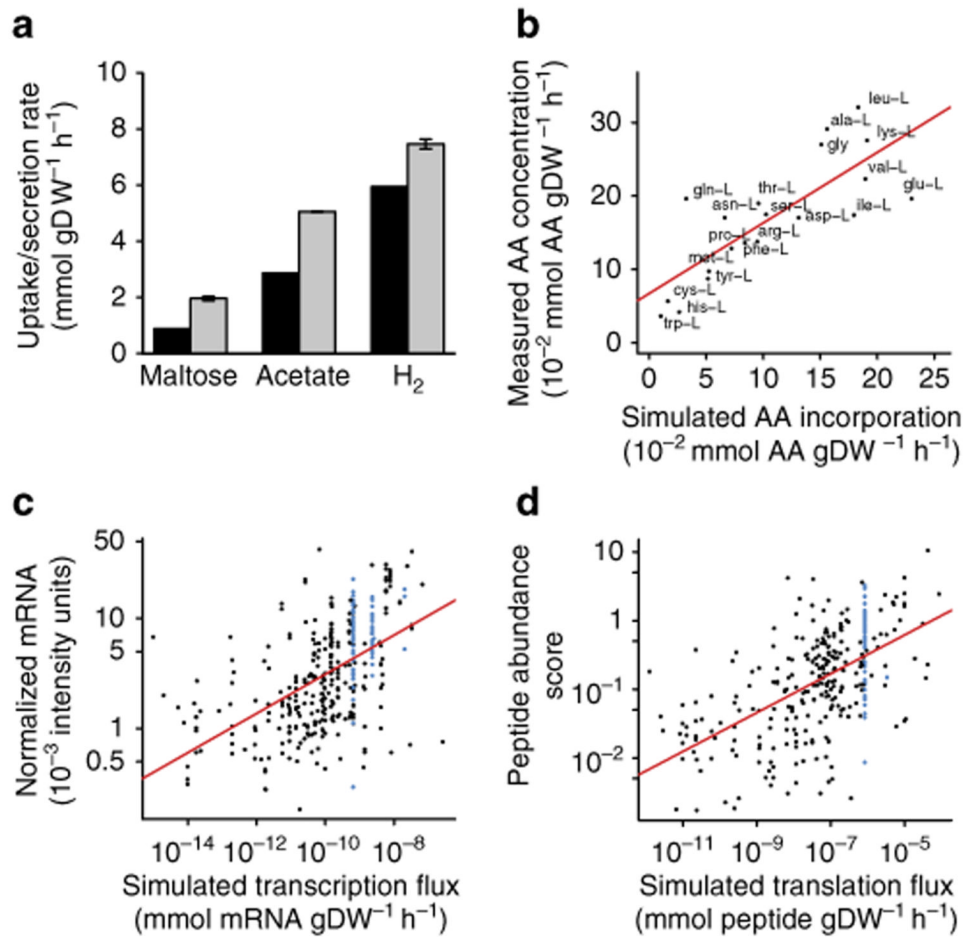
**Figure 5. The ME-Model accurately simulates molecular phenotypes during log-phase growth**
(**a**) The ME-Model accurately simulates $H_2$ and acetate secretion with maltose uptake when constrained with a measured growth rate ($n = 2$). Experiment: grey bars, simulation: black bars. (**b**) The *in silico* ribosome incorporates the 20 amino acids at rates proportional (Pearson correlation coefficient = 0.79; $P < 4.1 \times 10^{-5}$ *t*-test) to the bulk amino-acid composition of a *T. maritima* cell as measured by high-performance liquid chromatography ($n = 1$). (**c**) simulated transcriptome fluxes are significantly ($P < 2.2 \times 10^{-16}$ *t*-test) and positively correlated (Pearson correlation coefficient = 0.54) with semiquantitative *in vivo* transcriptome measurements ($n = 4$). RnAs containing ribosomal proteins (blue) were expressed stoichiometrically in simulations but exhibited variability in measurements. (**d**) simulated translation fluxes are significantly ($P < 2.2 \times 10^{-16}$ *t*-test) and positively correlated (Pearson correlation coefficient = 0.57) with semiquantitative *in vivo* proteomic measurements ($n = 3$). Ribosomal proteins (blue) were expressed stoichiometrically in simulations but exhibited variability in measurements.
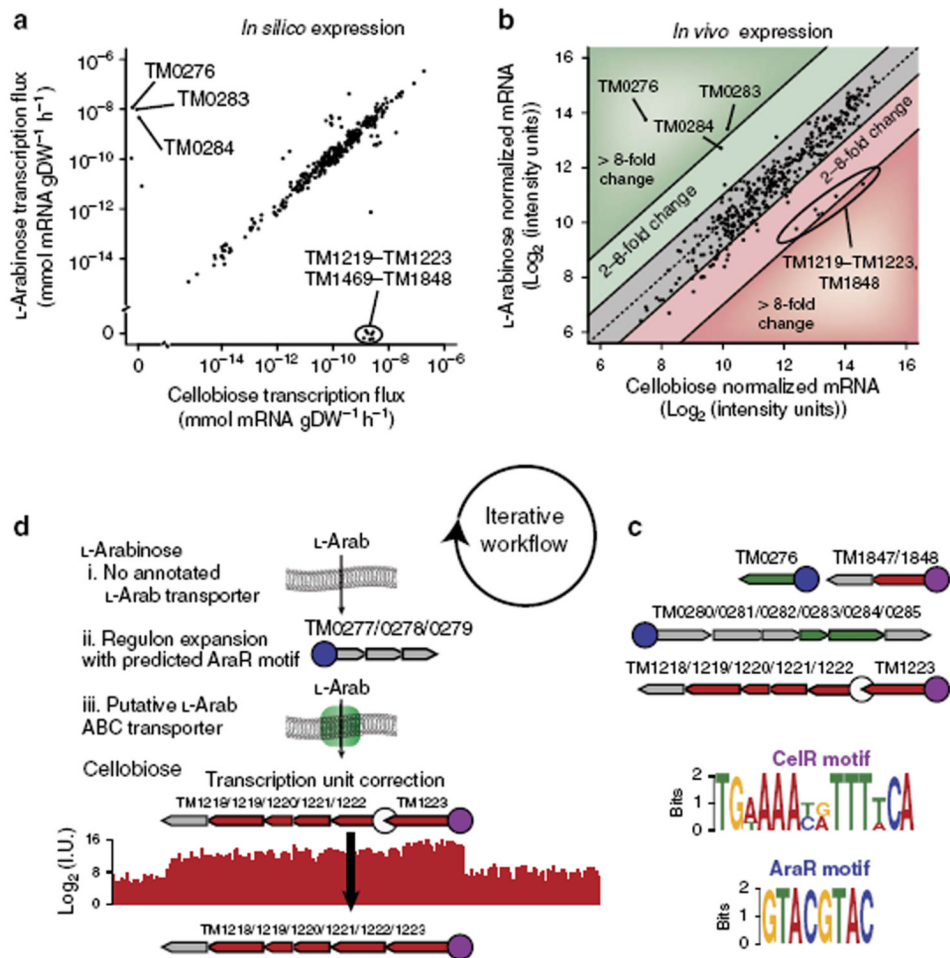
**Figure 6. *In silico* transcriptome profiling drives biological discovery**
(**a**)*In silico* comparative transcriptomics identifies sets of genes that are differentially regulated for growth in L-arabinose (L-Arab) versus growth in cellobiose minimal media. Tm0276, Tm0283 and Tm0284 are essential for metabolizing l-Arab, whereas TM1219–TM1223, TM1469 and TM1848 are essential for metabolizing cellobiose. (**b**)*In vivo* transcriptome measurements ($n = 2$) confirm the *in silico* transcriptomics predictions for differential expression of genes when metabolizing l-Arab or cellobiose. (**c**) Two distinct putative TF-binding motifs are present upstream of the TUs containing genes differentially expressed *in silico* when simulating growth in l-Arab versus cellobiose minimal media. The motif upstream of the genes upregulated during growth in l-Arab medium is termed AraR, whereas the motif of the genes upregulated during growth in cellobiose medium is termed CelR. Genes (grey: not in the model, green: upregulated by l-arabinose, red: upregulated by cellobiose) organized into TUs involved in the shift are shown. Each TU contains a promoter region (circle) arbitrarily taken to be 75 base pairs upstream of the first gene in the TU. Promoters found to contain the AraR or CelR motifs are coloured blue and purple, respectively. (**d**) searching *T. maritima*'s genome for additional AraR and CelR motifs results in new biological knowledge. Although *T. maritima* can metabolize l-Arab, there is no annotated transporter in the current genome. We identified a putative AraR motif in a single TU (Tm0277/0278/0279) not contained in the ME-Model. Analysis of the Tm0277/0278/0279 TU with the SEED RAST server indicated that the genes are likely components of an ABC transporter that may be associated with l-Arab transport. The CelR

motif was not present in the promoter region upstream of the cellobiose transporter operon (Tm1218/1219/1220/1221/1222); however, the CelR motif was present in the promoter of the TU (TM1223) directly upstream of the cellobiose transport operon. Examination of the *in vivo* transcriptome measurement indicates that the cellobiose transporter operon belongs to the same TU as that of TM1223.