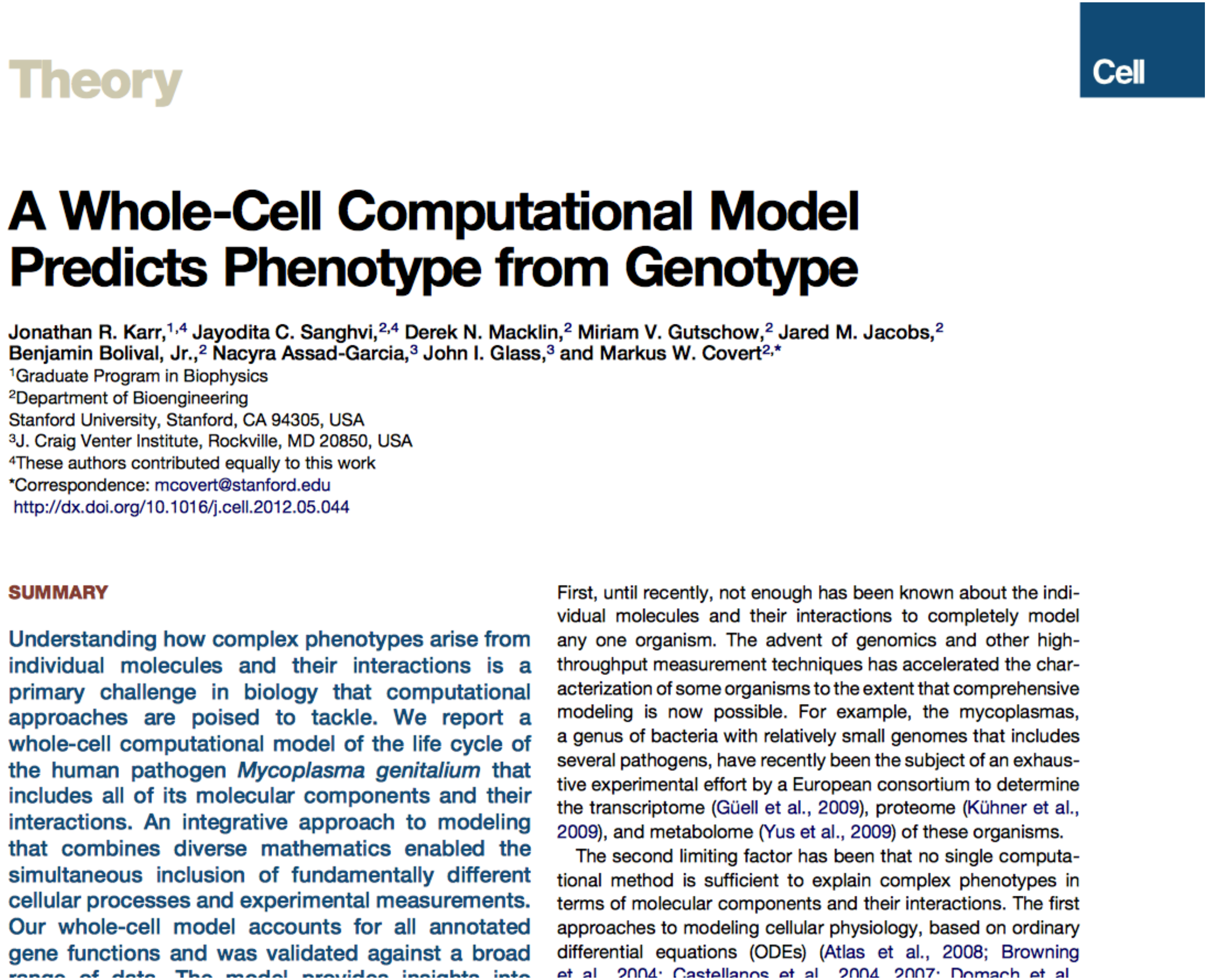
A Whole-Cell Computational Model Predicts Phenotype from Genotype



Theory

Cell

A Whole-Cell Computational Model Predicts Phenotype from Genotype

Jonathan R. Karr,[1,4] Jayodita C. Sanghvi,[2,4] Derek N. Macklin,[2] Miriam V. Gutschow,[2] Jared M. Jacobs,[2]
Benjamin Bolival, Jr.,[2] Nacyra Assad-Garcia,[3] John I. Glass,[3] and Markus W. Covert[2,*]
[1]Graduate Program in Biophysics
[2]Department of Bioengineering
Stanford University, Stanford, CA 94305, USA
[3]J. Craig Venter Institute, Rockville, MD 20850, USA
[4]These authors contributed equally to this work
*Correspondence: mcovert@stanford.edu

http://dx.doi.org/10.1016/j.cell.2012.05.044

SUMMARY

Understanding how complex phenotypes arise from individual molecules and their interactions is a primary challenge in biology that computational approaches are poised to tackle. We report a whole-cell computational model of the life cycle of the human pathogen *Mycoplasma genitalium* that includes all of its molecular components and their interactions. An integrative approach to modeling that combines diverse mathematics enabled the simultaneous inclusion of fundamentally different cellular processes and experimental measurements. Our whole-cell model accounts for all annotated gene functions and was validated against a broad range of data. The model provides insights into

First, until recently, not enough has been known about the individual molecules and their interactions to completely model any one organism. The advent of genomics and other high-throughput measurement techniques has accelerated the characterization of some organisms to the extent that comprehensive modeling is now possible. For example, the mycoplasmas, a genus of bacteria with relatively small genomes that includes several pathogens, have recently been the subject of an exhaustive experimental effort by a European consortium to determine the transcriptome (Güell et al., 2009), proteome (Kühner et al., 2009), and metabolome (Yus et al., 2009) of these organisms.

The second limiting factor has been that no single computational method is sufficient to explain complex phenotypes in terms of molecular components and their interactions. The first approaches to modeling cellular physiology, based on ordinary differential equations (ODEs) (Atlas et al., 2008; Browning et al., 2004; Castellanos et al., 2004, 2007; Domach et al.,

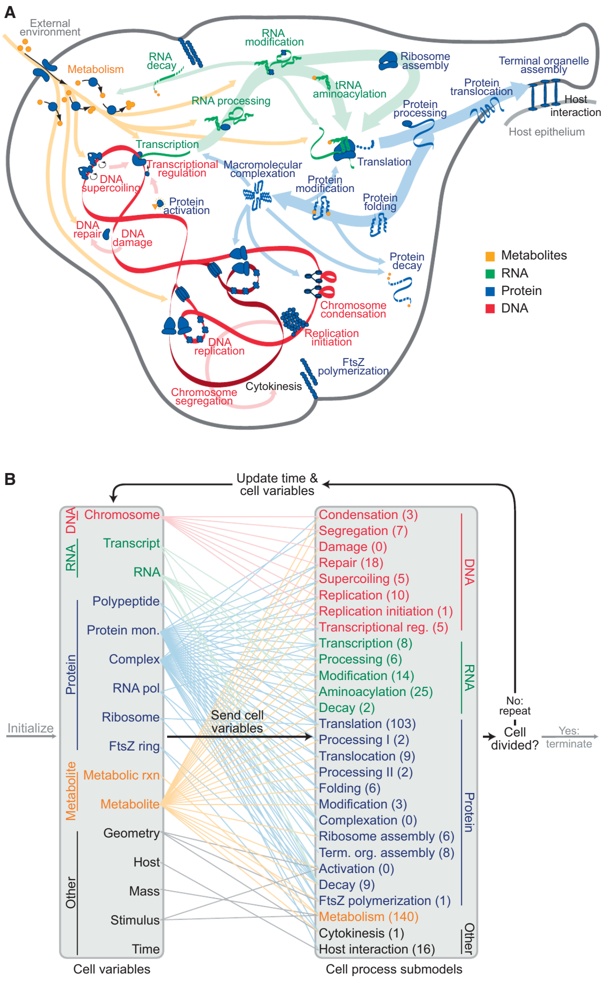


Figure 1. *M. genitalium* Whole-Cell Model Integrates 28 Submodels of Diverse Cellular Processes

(A) Diagram schematically depicts the 28 submodels as colored words—grouped by category as metabolic (orange), RNA (green), protein (blue), and DNA (red)—in the context of a single *M. genitalium* cell with its characteristic flask-like shape. Submodels are connected through common metabolites, RNA, protein, and the chromosome, which are depicted as orange, green, blue, and red arrows, respectively.

(B) The model integrates cellular function submodels through 16 cell variables. First, simulations are randomly initialized to the beginning of the cell cycle (left gray arrow). Next, for each 1 s time step (dark black arrows), the submodels retrieve the current values of the cellular variables, calculate their contributions to the temporal evolution of the cell variables, and update the values of the cellular variables. This is repeated thousands of times during the course of each simulation. For clarity, cell functions and variables are grouped into five physiologic categories: DNA (red), RNA (green), protein (blue), metabolite (orange), and other (black). Colored lines between the variables and submodels indicate the cell variables predicted by each submodel. The number of genes associated with each submodel is indicated in parentheses. Finally, simulations are terminated upon cell division when the septum diameter equals zero (right gray arrow).

submodels together. We defined 28 modules (Figure 1A) and independently built, parameterized, and tested a submodel of each. Some biological processes have previously been studied quantitatively and in depth, whereas other processes are less well characterized or are hardly understood. Consequently, each module was modeled using the most appropriate mathematical representation. For example, metabolism was modeled using flux-balance analysis (Suthers et al., 2009), whereas RNA and protein degradation were modeled as Poisson processes.

A key challenge of the project was to integrate the 28 submodels into a unified model. Although we and others had previously developed methods to integrate ODEs with Boolean, probabilistic, and constraint-based submodels (Covert et al., 2001, 2004, 2008; Chandrasekaran and Price, 2010), the current effort involved so many different cellular functions and mathematical representations that a more general approach was needed. We began with the assumption that the submodels are approximately independent on short timescales (less than 1 s). Simulations are then performed by running through a loop in which the submodels are run independently at each time step but depend on the values of variables determined by the other submodels at the previous time step. Figure 1B summarizes the simulation algorithm and the relationships between the submodels and the cell variables. Data S1 (available online) provides a detailed description of the complete modeling process, including reconstruction and computational implementation.

Model Training and Parameter Reconciliation

Our model is based on a synthesis of over 900 publications and includes more than 1,900 experimentally observed parameters. Most of these parameters were implemented as originally reported. However, several other parameters were carefully reconciled; for example, the experimentally measured DNA content per cell (Morowitz et al., 1962; Morowitz, 1992) represents less than one-third of the calculated mass of the mycoplasma chromosome. Data S1 details how we resolved this and several similar discrepancies among the experimentally observed parameters.

Once the model was implemented and all parameters were reconciled, we verified that the model recapitulates key features of our training data. We simulated 128 wild-type cells in a typical *Mycoplasma* culture environment, with each simulation predicting not only cellular properties such as the cell mass and growth rate but also molecular properties including the count, localization, and activity of each molecule (Movie S1 illustrates the life cycle of one in silico cell). We found that the model calculations were consistent with the observed doubling time (Figures 2A and 2B), cellular chemical composition (Figure 2C), replication of major cell mass fractions (Figure 2D), and gene expression (R^2 = 0.68; Figure S1A).

Model Validation against Independent Experimental Data

Next, we validated the model against a broad range of independent data sets that were not used to construct the model and which encompass multiple biological functions—metabolomics, transcriptomics, and proteomics—and scales from single cells to populations. In agreement with earlier reports (Yus et al., 2009), the model predicts that the flux through glycolysis is >100-fold more than that through the pentose phosphate and lipid biosynthesis pathways (Figure 2E). Furthermore, the predicted metabolite concentrations are within an order of magnitude of concentrations measured in *Escherichia coli* for 100% of the metabolites in one compilation of data (Sundararaj et al., 2004) and for 70% in a more recent high-throughput study (Bennett et al., 2009; Figure 2F). Our model also predicts ‘‘burst-like'' protein synthesis due to the local effect of intermittent messenger RNA (mRNA) expression and the global effect of stochastic protein degradation on the availability of free amino acids for translation, which is comparable to recent reports by Yu et al. (2006) and So et al. (2011) (Figure 2G). The mRNA and protein level distributions predicted by our model are also consistent with recently reported single-cell measurements (Figure 2H; compare to Taniguchi et al., 2010). Taking all of these specific tests of the model's predictions together, we concluded that our model recapitulates experimental data across multiple biological functions and scales.

Prediction of DNA-Binding Protein Interactions

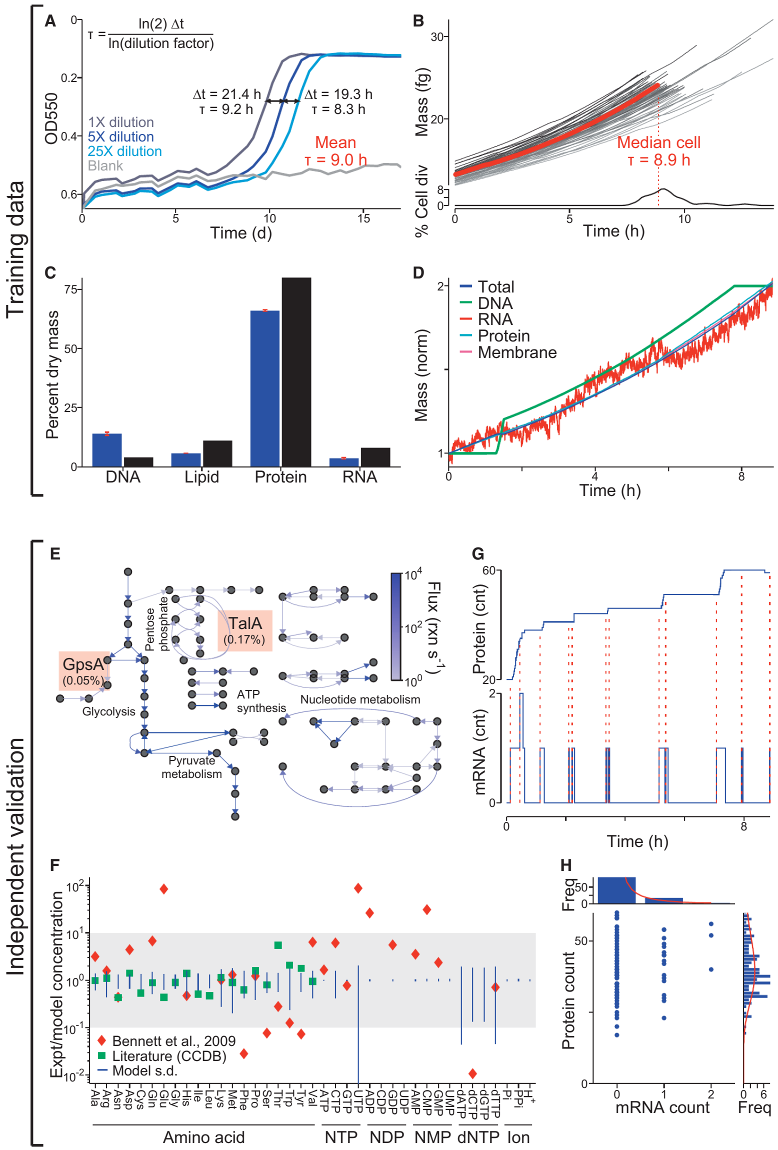
Models are often used to predict molecular interactions that are difficult or prohibitive to investigate experimentally, and our model offers the opportunity to make such predictions in the context of the entire cell. Whereas previous studies have either focused on the genomic distribution of DNA-binding proteins (Vora et al., 2009) or on the detailed diffusion dynamics of specific DNA-binding proteins (Bratton et al., 2011), the whole-cell model can predict both the instantaneous protein chromosomal occupancy as well as the temporal dynamics and interactions of every DNA-binding protein at the genomic scale at single-cell resolution. Figure 3A illustrates the average predicted chromosomal protein occupancy as well as the predicted chromosomal occupancies for DNA and RNA polymerase and the replication initiator DnaA, which are three of the 30 DNA-binding proteins represented by our model. Consistent with a recent experimental study by Vora et al. (2009), the predicted high-occupancy RNA polymerase regions correspond to highly transcribed ribosomal RNAs (rRNAs) and transfer RNAs (tRNAs). In contrast, the predicted DNA polymerase chromosomal occupancy is significantly lower and biased toward the terC (see below for further discussion).

The model further predicts that the chromosome is explored very rapidly, with 50% of the chromosome having been bound by at least one protein within the first 6 min of the cell cycle and 90% within the first 20 min (Figure 3B). RNA polymerase contributes the most to chromosomal exploration, binding 90% of the chromosome within the first 49 min of the cell cycle. On average, this results in expression of 90% of genes within the first 143 min (Figure 3C), with transcription lagging RNA polymerase exploration due to the significant contribution of nonspecific RNA polymerase-DNA interactions to RNA polymerase diffusion (Harada et al., 1999).

The model also predicts protein-protein collisions on the chromosome. Previous researchers have studied the collisions of pairs of specific proteins (Pomerantz and O'Donnell, 2010), but experimentally determining the collisions among all pairs of DNA-binding proteins at the genomic scale at single-cell resolution is currently infeasible. Our model predicts that over 30,000 collisions occur on average per cell cycle, leading to the displacement of 0.93 proteins per second. Figure 3D illustrates the binding dynamics of the same proteins depicted in Figure 3A over the course of the cell cycle for one representative simulation and highlights several protein-protein collisions. Further categorization of the predicted collisions by chromosomal location indicates that the frequency of protein-protein collisions correlates strongly with DNA-bound protein density across the genome (Figure 3F) and that the majority of collisions are caused by RNA polymerase (84%) and DNA polymerase (8%), most commonly resulting in the displacement of structural maintenance of chromosome (SMC) proteins (70%) or single-stranded binding proteins (6%) (Figure 3E and Table S2F).

Identification of Metabolism as an Emergent Cell-Cycle Regulator

The model can also highlight interesting aspects of cell behavior. In reviewing our model simulations, we noticed variability in the cell-cycle duration (Figure 2B) and wanted to determine the source of that variability. The model representation of the *M. genitalium* cell cycle consists of three stages: replication initiation, replication itself, and cytokinesis. We found that there was relatively more cell-to-cell variation in the durations of the replication initiation (64.3%) and replication (38.5%) stages than in cytokinesis (4.4%) or the overall cell cycle (9.4%; Figure 4A). This data raised two questions: (1) what is the source of duration


Training data
A
τ = ln(2) Δt / ln(dilution factor)
OD550
0
0.2
0.4
0.6
Δt = 21.4 h
τ = 9.2 h
Δt = 19.3 h
τ = 8.3 h
1X dilution
5X dilution
25X dilution
Blank
Mean
τ = 9.0 h
0
5
10
15
Time (d)
B
Mass (fg)
30
20
8
0
% Cell div
Median cell
τ = 8.9 h
0
5
10
Time (h)
C
Percent dry mass
75
50
25
0
DNA
Lipid
Protein
RNA
D
Mass (norm)
2
1
Total
DNA
RNA
Protein
Membrane
0
4
8
Time (h)
Independent validation
E
Pentose phosphate
TalA
(0.17%)
GpsA
(0.05%)
Glycolysis
ATP synthesis
Nucleotide metabolism
Pyruvate metabolism
Flux (rxn s^{-1})
10^4
10^2
10^0
G
Protein (cnt)
60
20
mRNA (cnt)
2
0
0
8
Time (h)
F
Expt/model concentration
10^2
10^1
10^0
10^{-1}
10^{-2}
Bennett et al., 2009
Literature (CCDB)
Model s.d.
Ala
Arg
Asn
Asp
Cys
Gln
Glu
Gly
His
Ile
Leu
Lys
Met
Phe
Pro
Ser
Thr
Trp
Tyr
Val
ATP
CTP
GTP
UTP
ADP
CDP
GDP
UDP
AMP
CMP
GMP
UMP
dATP
dCTP
dGTP
dTTP
Pi
PPi
H$^+$
Amino acid
NTP
NDP
NMP
dNTP
Ion
H
Freq
50
0
Protein count
50
0
0
1
2
mRNA count
0
6
Freq

variability in the initiation and replication phases; and (2) why is the overall cell-cycle duration less varied than either of these phases?

With respect to the first question, replication initiation occurs as DnaA protein monomers bind or unbind stochastically and cooperatively to form a multimeric complex at the replication origin (Figure 4B, top) (Browning et al., 2004). When the complex is complete, DNA polymerase gains access to the origin, and the complex is displaced. We found a correlation (R^2 = 0.49) between the predicted duration of replication initiation and the initial number of free DnaA monomers (Figure 4C); however, the low correlation indicated that the duration depends on more than the initial conditions. In particular, we observed that the stochastic aspect of the transcription and translation submodels creates variability in the number of new DnaA monomers produced over time, as well as the DnaA-binding and -unbinding events themselves. This indicates that the variability in replication initiation duration depends not only on variability in initial conditions but also in the simulation itself.

As to the second question, because the replication submodel is substantially more deterministic than the initiation submodel, we expected to find a straightforward relationship between the progress of replication and the cell cycle. Instead, the model predicts that DNA replication proceeds at two distinct rates during the cell cycle. This is reflected in the motion and DNA-binding density of DNA polymerase (Figures 3A and 3D) and in the dynamics of DNA synthesis as compared to the synthesis of other macromolecules (Figure 4B, middle). Initially, replication proceeds quickly due to the free deoxyribonucleotide triphosphate (dNTP) content in the cell (Figure 4B, bottom). When DNA polymerase initially binds to the replication origin, dNTPs are abundant, and replication proceeds unimpeded. When the dNTP pool is exhausted, however, the rate of replication slows to the rate of dNTP synthesis. Accordingly, the duration of the replication phase in individual cells is more closely related to the free dNTP content at the start of replication than to the dNTP content at the start of the cell cycle (Figure 4D).

This change in the availability of dNTPs imposes a control on the cell-cycle duration. Specifically, the duration of the initiation and replication phases is inversely related to each other in single cells (Figure 4E), such that longer initiation times led to shorter replication times. This occurs because cells that require extra time to initiate replication also build up a large dNTP surplus, leading to faster replication. This interplay buffers against the high variability in the duration of replication initiation, giving rise to substantially less variability in the length of the cell cycle. The whole-cell model therefore presents a hypothesis of an emergent control of cell-cycle duration that is independent of genetic regulation.

Global Distribution of Energy

The model also provided an opportunity to develop a quantitative assessment of cellular energetics, which represents one of the most connected aspects of our model. To begin, we investigated the synthesis dynamics of the high-energy intermediates ATP, GTP, FAD(H_2), NAD(H), and NADP(H) and found that ATP and guanosine triphosphate (GTP) are synthesized at rates greater than 1,000-fold higher than the others (Figure 5A). Notably, the overall usage of ATP and GTP did not vary considerably in all but the very slowest of our simulations (Figure 5B), underscoring the role of metabolism in controlling the cell-cycle length. We then considered the processes that use ATP and GTP and found that usage is dominated by production of mRNA and protein (Figure 5C). We also found a large (44%) discrepancy between total energy usage and production (Figure 5D). Others have noted an uncoupling between catabolism and anabolism, attributing the difference to factors such as varying maintenance costs or energy spilling via futile cycles (Russell and Cook, 1995), and the model's prediction estimates the total energy cost of such uncoupling.

Determining the Molecular Pathologies of Single-Gene Disruption Phenotypes

Having considered these above-described model predictions for the wild-type *M. genitalium* strain, we next performed in silico genome perturbations to gain insight into the genetic requirements of cellular life. We performed multiple simulations of each of the 525 possible single-gene disruption strains (over 3,000 total simulations) and found that 284 genes are essential

Figure 2. The Model Was Trained with Heterogeneous Data and Reproduces Independent Experimental Data across Multiple Cellular Functions and Scales

(A) Growth of three cultures (dilutions indicated by shade of blue) and a blank control measured by OD550 of the pH indicator phenol red. The doubling time, τ, was calculated using the equation at the top left from the additional time required by more dilute cultures to reach the same OD550 (black lines).

(B) Predicted growth dynamics of one life cycle of a population of 64 in silico cells (randomly chosen from the total simulation set). Median cell is highlighted in red. Distribution of cell-cycle lengths is shown at bottom.

(C) Comparison of the predicted and experimentally observed (Morowitz et al., 1962) cellular chemical compositions. Red bars indicate model SD; Morowitz et al. (1962) did not report SD.

(D) Temporal dynamics of the total cell mass and four cell mass fractions of a representative in silico cell. Mass fractions are normalized to their initial values.

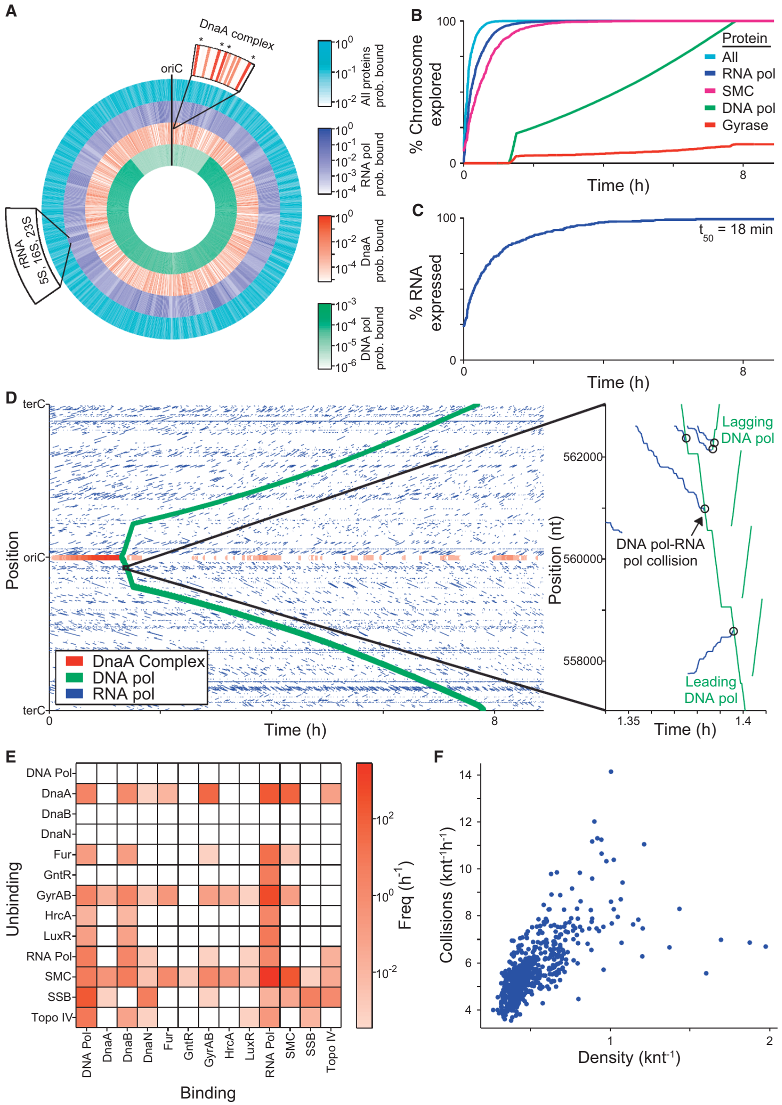
(E) Average predicted metabolic fluxes (see Figure S1B for metabolite and reaction labels). Arrow brightness indicates flux magnitude. The ratios of the GpsA and TalA fluxes to the Glk flux are indicated in orange boxes and are comparable to experimental data (Yus et al., 2009).

(F) Ratios of observed (Sundararaj et al., 2004; Bennett et al., 2009) and average predicted concentrations of 39 metabolites. Blue bars indicate model SD.

(G) Temporal dynamics of cytadherence high-molecular-weight protein 2 (HMW2, MG218) mRNA and protein expression of one in silico cell. Red dashed lines indicate the direct link between mRNA synthesis and subsequent bursts in protein synthesis.

(H) HMW2 mRNA and protein copy number distribution of an unsynchronized population of 128 in silico cells. Histograms indicate the marginal distributions of the copy numbers of mRNA (top) and protein (right). Red lines indicate log-normal regressions of these marginal distributions. The absence of correlation between the copy numbers of mRNA and protein and the shapes of the marginal distributions is consistent with recent single-cell measurements by Taniguchi et al. (2010).

See also Movie S1 and Tables S1 and S2.


A
DnaA complex
oriC
rRNA
5S, 16S, 23S
10^0
10^{-1}
10^{-2}
All proteins prob. bound
10^0
10^{-1}
10^{-2}
10^{-3}
10^{-4}
RNA pol prob. bound
10^0
10^{-2}
10^{-4}
DnaA prob. bound
10^{-3}
10^{-4}
10^{-5}
10^{-6}
DNA pol prob. bound
B
100
0
% Chromosome explored
0
8
Time (h)
Protein
All
RNA pol
SMC
DNA pol
Gyrase
C
100
0
% RNA expressed
t_{50} = 18 min
0
8
Time (h)
D
terC
oriC
terC
Position
0
8
Time (h)
DnaA Complex
DNA pol
RNA pol
562000
560000
558000
Position (nt)
Lagging DNA pol
DNA pol-RNA pol collision
Leading DNA pol
1.35
1.4
Time (h)
E
DNA Pol
DnaA
DnaB
DnaN
Fur
GntR
GyrAB
HrcA
LuxR
RNA Pol
SMC
SSB
Topo IV
Unbinding
DNA Pol
DnaA
DnaB
DnaN
Fur
GntR
GyrAB
HrcA
LuxR
RNA Pol
SMC
SSB
Topo IV
Binding
10^2
10^0
10^{-2}
Freq (h^{-1})
F
14
12
10
8
6
4
Collisions ($knt^{-1}h^{-1}$)
1
2
Density (knt^{-1})

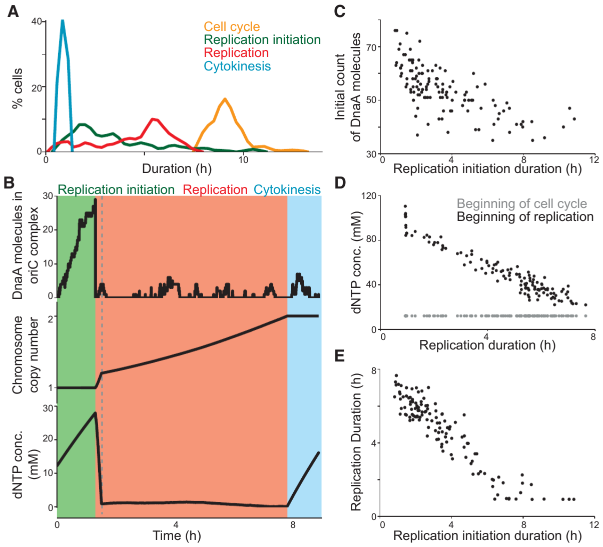


Figure 4. The Model Predictions Regarding Regulation of the Cell-Cycle Duration

(A) Distributions of the duration of three cell-cycle phases, as well as that of the total cell-cycle length, across 128 simulations.

(B) Dynamics of macromolecule abundance in a selected cell simulation. Top, the size of the DnaA complex assembling at the oriC (in monomers of DnaA); middle, the copy number of the chromosome; and bottom, the cytosolic dNTP concentration. The quantities of these macromolecules correlate strongly with the timing of key cell-cycle stages.

(C) Correlation between the initial cellular DnaA content and the duration of the replication initiation cell-cycle stage across the same 128 in silico cells depicted in (A).

(D) Correlation between the dNTP concentrations (both at the beginning of the cell cycle and at the beginning of replication) and the duration of replication across the same 128 in silico cells depicted in (A).

(E) Correlation between the duration of replication initiation and replication across the same 128 in silico cells depicted in (A).

to sustain *M. genitalium* growth and division and that 117 are nonessential. The model accounts for previously observed gene essentiality with 79% accuracy ($p < 10^{-7}$; Glass et al., 2006; Figure 6A).

In cases in which the model prediction agrees with the experimental outcome with respect to gene essentiality, we found that a deeper examination of the simulation can generate insight into why the gene product is required by the system. We examined the capacities of the 525 simulated gene disruption strains to produce major biomass components (RNA, DNA, protein, and lipid) and to divide. As shown in Figure 6B, the nonviable strains were unable to adequately perform one or more of these major functions. The most debilitating disruptions involved metabolic genes and resulted in the inability to produce any of the major cell mass components. The next most debilitating gene disruptions impacted the synthesis of a specific cell mass component, such as RNA or protein. Interestingly, in these cases, the model predicted an initial phase of near-normal growth followed by decreasing growth due to diminishing protein content. In some cases (Figure 6B, fifth column), the time required for the levels of specific proteins to fall to lethal levels was greater than one generation (Figures 6C and 6D). A third class of lethal gene disruptions impaired cell-cycle processes. For these, the model predicted normal growth rates and metabolism, but it also predicted incapacity to complete the cell cycle. The remaining lethal gene disruption strains grew so slowly compared to wild-type that they were considered nonviable (Figures 6B and S2). We conclude that the model can be used to classify cellular phenotypes by their underlying molecular interactions.

Figure 3. The Model Highlights the Central Physiological Role of DNA-Protein Interactions

(A) Average density of all DNA-bound proteins and of the replication initiation protein DnaA and DNA and RNA polymerase of a population of 128 in silico cells. Top magnification indicates the average density of DnaA at several sites near the oriC; DnaA forms a large multimeric complex at the sites indicated with asterisks, recruiting DNA polymerase to the oriC to initiate replication. Bottom left indicates the location of the highly expressed rRNA genes.

(B and C) Percentage of the chromosome that is predicted to have been bound (B) and the number of genes that are predicted to have been expressed (C) as functions of time. SMC is an abbreviation for the name of the chromosome partition protein (MG298).

(D) DNA-binding and dissociation dynamics of the oriC DnaA complex (red) and of RNA (blue) and DNA (green) polymerases for one in silico cell. The oriC DnaA complex recruits DNA polymerase to the oriC to initiate replication, which in turn dissolves the oriC DnaA complex. RNA polymerase traces (blue line segments) indicate individual transcription events. The height, length, and slope of each trace represent the transcript length, transcription duration, and transcript elongation rate, respectively. The inset highlights several predicted collisions between DNA and RNA polymerases that lead to the displacement of RNA polymerases and incomplete transcripts.

(E) Predicted collision and displacement frequencies for pairs of DNA-binding proteins.

(F) Correlation between DNA-binding protein density and frequency of collisions across the chromosome. Both (E) and (F) are based on 128 cell-cycle simulations.

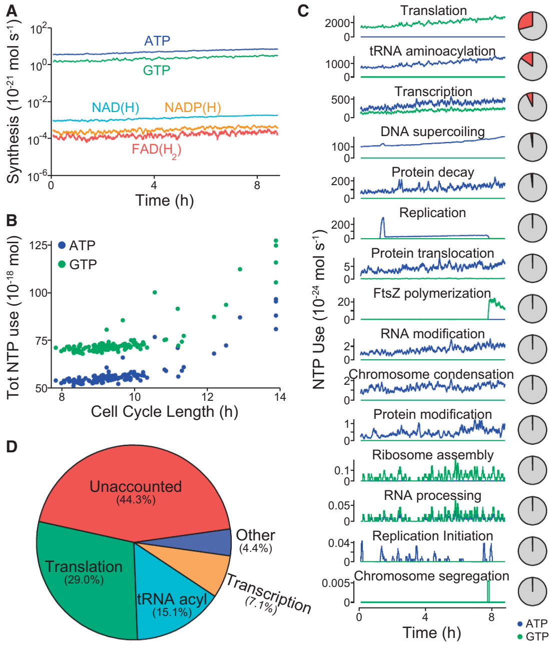


Figure 5. Model Provides a Global Analysis of the Use and Allocation of Energy

(A) Intracellular concentrations of the energy carriers ATP, GTP, FAD(H_2), NAD(H), and NADP(H) of one in silico cell.

(B) Comparison of the cell-cycle length and total ATP and GTP usage of 128 in silico cells.

(C) ATP (blue) and GTP (green) usage of 15 cellular processes throughout the life cycle of one in silico cell. The pie charts at right denote the percentage of ATP and GTP usage (red) as a fraction of total usage.

(D) Average distribution of ATP and GTP usage among all modeled cellular processes in a population of 128 in silico cells. In total, the modeled processes account for only 44.3% of the amount of energy that has been experimentally observed to be produced during cellular growth.

Model-Driven Biological Discovery

Using computational modeling as a complement to an experimental program has previously been shown to facilitate biological discovery (Di Ventura et al., 2006). This is often accomplished by reconciling model predictions that are initially inconsistent with observations (Covert et al., 2004). To test the utility of the whole-cell model in this context, we experimentally measured the growth rates of 12 single-gene disruption strains—ten of which were correctly predicted to be viable and two of which were incorrectly predicted to be nonviable—for comparison to our model's predictions (Figure 7A). We found that two-thirds of the predictions were consistent with the measured growth rates.

The most interesting of these comparisons concerned the *lpdA* disruption strain. The *lpdA* gene was originally determined to be nonessential (Glass et al., 2006). Consequently, we initially classified the model's prediction as false (Figure 6A). However, we did not detect growth using our colorimetric assay (Figure 7B), which was a discrepancy that warranted further investigation. An alternative method to determine the doubling time yielded a value that was 40% lower than the wild-type (Table S1). Taken together, the data suggested that disrupting the *lpdA* gene had a severe but noncritical impact on cell growth.

In an effort to resolve the discrepancy between our model and the experimental measurements, we determined the molecular pathology of the *lpdA* disruption strain. The *lpdA* gene product is part of the pyruvate dehydrogenase complex, which catalyzes the transfer of electrons to nicotinamide adenine dinucleotide (NAD) as a subset of the overall pyruvate dehydrogenase chemical reaction (de Kok et al., 1998). The viability of the *lpdA* disruption strain suggests that this reaction could be catalyzed by another enzyme with a lower catalytic efficiency.

Because previous studies have shown that many *M. genitalium* genes are multifunctional (Pollack et al., 2002; Cordwell et al., 1997), we searched the genome for candidates encoding an alternative NAD electron transfer pathway. We found that the Nox sequence was far more similar to the LpdA sequence than any other gene product in the genome, with 61% coverage, 25% identity, and an expectation value of less than 10^{-6} (Figure 7C). Furthermore, the *nox* gene product, NADH oxidase, has been shown to oxidize NAD (Schmidt et al., 1986). Moreover, the *nox* locus falls in a suboperon that contains two other pyruvate dehydrogenase genes and has been shown to be coexpressed with *pdhA* (Güell et al., 2009) (Figure 7D), strongly suggesting a functional relationship between the products of these two genes. Our model suggests that, to reproduce the observed growth rate in the absence of *lpdA*, the hypothetical Nox-dependent reaction would require a k_{cat} of ~50 s^{-1} (Figure 7E), which represents only ~5% of the maximum throughput of this enzyme. We therefore concluded that substrate promiscuity of Nox is likely to enable the *lpdA* disruption strain to survive.

Four gene disruption strains exhibited growth rates that were quantitatively different than those predicted by the model (Figure 7A); of these, we used the complete simulations for the *thyA* and *deoD* strains to determine the underlying pathology of the respective gene disruptions. The *thyA* gene product catalyzes thymidine monophosphate (dTMP) production and can be complemented by the *tdk* gene product. We therefore

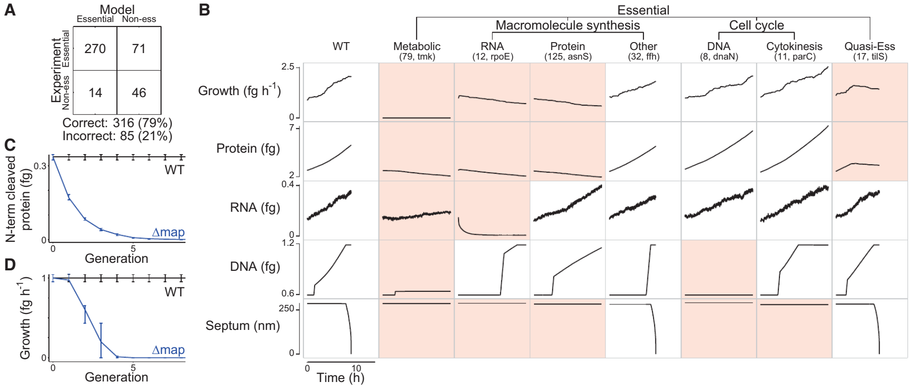


Figure 6. Model Identifies Common Molecular Pathologies Underlying Single-Gene Disruption Phenotypes

(A) Comparison of predicted and observed (Glass et al., 2006) gene essentiality. Model predictions are based on at least five simulations of each single-gene disruption strain; see Data S1 for details.

(B) Single-gene disruption strains were grouped into phenotypic classes (columns) according to their capacity to grow, synthesize protein, RNA, and DNA, and divide (indicated by septum length). Each column depicts the temporal dynamics of one representative in silico cell of each essential disruption strain class. Disruption strains of nonessential genes are not shown. Dynamics significantly different from wild-type are highlighted in red. The identity of the representative cell and the number of disruption strains in each category are indicated in parenthesis.

(C and D) Degradation and dilution of N-terminal protein content (C) of methionine aminopeptidase (*map*, MG172) disrupted cells causes reduced growth (D). Blue and black lines indicate the *map* disruption and wild-type strains, respectively. Bars indicate SD.

See also Figure S2 for the distribution of simulated growth rates.

hypothesized that, by reducing the k_{cat} value for Tdk in the model, we would see a reduction in the growth rate of the *tdk* disruption strain. Reducing the Tdk k_{cat} in the model did indeed reduce the predicted growth rate of the *thyA* strain, but it also affected the wild-type growth rate (Figure 7F). Only a small range of the k_{cat} values both reduced the *thyA* strain growth rate to the experimentally observed levels and was also consistent with the wild-type growth rate.

In a similar case, purine nucleoside phosphorylase (DeoD) catalyzes the conversion of deoxyadenosine to adenine and D-ribose-1phosphate; these products can also be produced by the *pdp* gene product from deoxyuridine. We identified a Pdp k_{cat} range for which the wild-type and *deoD* gene disruption strains produce the same growth rate (Figure 7G).

Significantly, these newly predicted k_{cat} values are consistent with previously reported values. In the original model reconstruction, to least constrain the metabolic model, we conservatively set each of these k_{cat}s to the least restrictive value found during the reconstruction process. For Tdk and Pdp, these values corresponded to distantly related organisms; however, the newly predicted k_{cat} values are consistent with reports from more closely related species (Figure 7H).

In each of these three cases (*lpdA*, *deoD*, and *thyA*), identifying a discrepancy between model predictions and experimental measurements led to further analysis, which resolved the discrepancy and also provided insight into *M. genitalium* biology (Figure 7I). These results support the assertion that large-scale modeling can be used to guide biological discovery (Kitano, 2002; Brenner, 2010).

DISCUSSION

We have developed a comprehensive whole-cell model that accounts for all of the annotated gene functions identified in *M. genitalium* and explains a variety of emergent behaviors in terms of molecular interactions. Our model accurately recapitulates a broad set of experimental data, provides insight into several biological processes for which experimental assessment is not readily feasible, and enables the rapid identification of gene functions as well as specific cellular parameters.

In contemplating these results, we make two observations based on comparing this work in whole-cell modeling with earlier work in whole-genome sequencing. First, similar to the first reports of the human genome sequence, the model presented here is a "first draft," and extensive effort is required before the model can be considered complete. Of course, much of this effort will be experimental (for example, further characterization of gene products), but the technical and modeling aspects of this study will also have to be expanded, updated, and improved as new knowledge comes to light.

Second, in whole-genome sequencing as well as in whole-cell modeling, *M. genitialium* was a focus of initial studies, primarily because of its small genome size. The goal of our modeling efforts, as well as that of early sequencing projects, was to

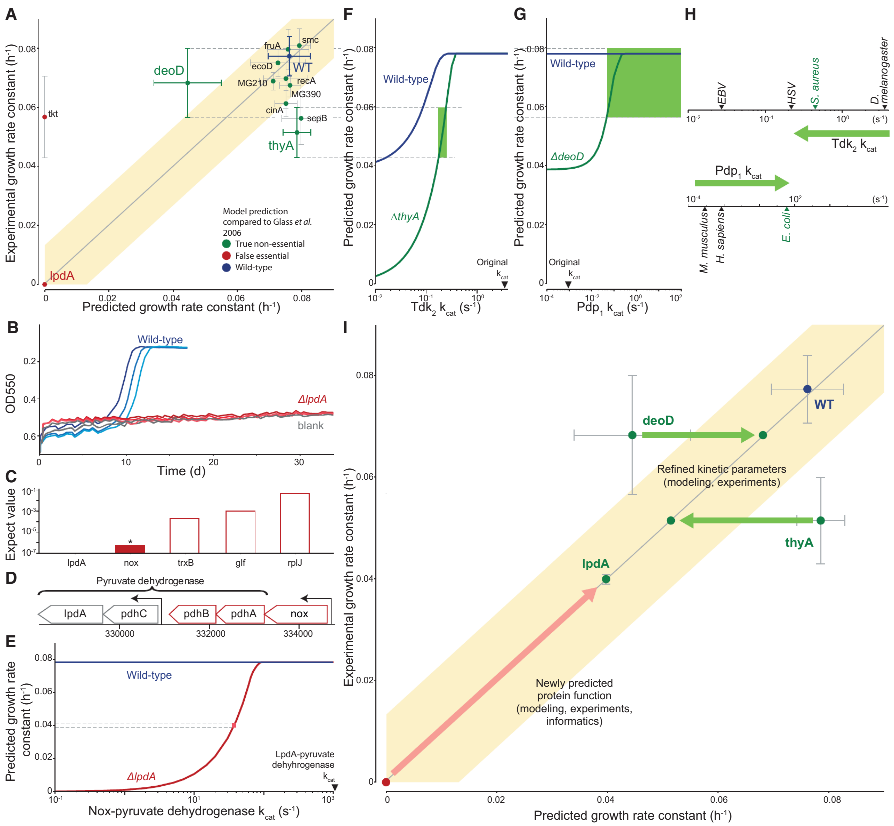


Figure 7. Quantitative Characterization of Selected Gene Disruption Strains Leads to Identification of Novel Gene Functions and Kinetic Parameters

(A) Comparison of measured and predicted growth rates for wild-type and 12 single-gene disrupted strains. Model predictions that fall within the shaded region were considered consistent with experimental observations; the region has a width of four times the SD of the wild-type strain growth measurement. Horizontal and vertical bars indicate predicted and observed SD.

(B) Growth curves for the wild-type and *lpdA* gene disruption strains and blank, similar to Figure 2A.

(C) Expectation values determined by performing a pBLAST search of the *M. genitalium* genome with the LpdA sequence as a query. The asterisk and colored bar indicate a significant match ($E < 10^{-6}$).

(D) Detail of the *M. genitalium* genome. The pyruvate dehydrogenase complex genes are indicated by the top bracket, and transcription units identified in *M. pneumoniae* (Güell et al., 2009) are indicated by arrows. The transcription unit including *nox* is highlighted in color.

(E) Allowing Nox to partially replace LpdA in pyruvate dehydrogenase reconciles model predictions and experimental observations. The blue and red lines represent the predicted wild-type and Δ*lpdA* strain growth rates as a function of the Nox-pyruvate dehydrogenase k_{cat}. The pink box indicates the k_{cat} at which the model predictions are consistent with both the wild-type and Δ*lpdA* strain experimentally measured growth rates.

(F and G) Diagnosing the discrepancy between predictions and experiment for the *thyA* (F) and *deoD* (G) gene disruption strains. Some of the functionalities of ThyA and DeoD can be replaced by the enzymes Tdk and Pdp, respectively. The predicted growth rates of the wild-type and gene disruption strains depend on the k_{cat} of these enzymes. The green region highlights the range of k_{cat} values that are consistent with the measured growth rates of both the wild-type and gene disruption strain.

(H) Newly predicted k_{cat} values are similar to values that were measured in closely related organisms. Measured values of k_{cat} for Tdk (top) and Pdp (bottom) are shown; green arrow indicates the initial and revised k_{cat} values. The nearest *M. genitalium* relative is highlighted in green.