Accelerated discovery via a whole-cell model

Jayodita C Sanghvi1,3, Sergi Regot1, Silvia Carrasco1, Jonathan R Karr2, Miriam V Gutschow1, Benjamin Bolival Jr1 & Markus W Covert1

To test the promise of whole-cell modeling to facilitate scientific inquiry, we compared growth rates simulated in a whole-cell model with experimental measurements for all viable single-gene disruption Mycoplasma genitalium strains. Discrepancies between simulations and experiments led to predictions about kinetic parameters of specific enzymes that we subsequently validated. These findings represent, to our knowledge, the first application of whole-cell modeling to accelerate biological discovery.

As high-throughput experiments yield large and complex data sets, there is an increasing need for corresponding computational methods to derive knowledge from those data. Whole-cell models that incorporate every known gene function to make predictions about integrated and complex biological phenomena may address this need. In the 1980s, bacteriologists realized that the simple Mycoplasma spp. might be tractable for the first comprehensive description of a cell, and since that time several groups have worked toward building models of various aspects of M. genitalium, the simplest self-replicating organism2–4. Most recently, members of our laboratory constructed a whole-cell model of M. genitalium that incorporates every known gene function to simulate a complete life cycle. The simulations generated by this model compared favorably with published data sets, which included metabolic fluxes, metabolite concentrations, growth rates, and gene and protein expression5.

Although these comparisons represented perhaps the broadest validation of any biological model, the most exciting promise of whole-cell modeling, prediction of previously undiscovered biology, remains to be fulfilled. In the process of model-driven discovery, for any given complex experiment, the whole-cell model can be used to produce a corresponding simulation, resulting in two sets of data, one computational and one experimental, which can be directly compared to determine how well the model describes observations. Any discrepancy between predictions and observations represents a high-probability opportunity for a discovery as the discrepancy is resolved, and new discoveries are incorporated back into the model (Fig. 1a).

We implemented this model-driven discovery approach using our whole-cell model. We compared model predictions and experimental measurements of the specific growth rates of a nonessential single-gene disruption library of M. genitalium, for all 86 strains for which an experimental rate could be determined5–6 (Fig. 1b). We also compared our predictions to a metabolic model based on flux-balance analysis7 and found that the whole-cell model made more quantitative predictions (Supplementary Fig. 1), underlining the need for the whole-cell model.

For 84% of the strains, the specific growth rates determined by experiment and simulation were statistically indistinguishable. Although our previous analysis showed that the model could predict the phenotypes of all 525 gene disruption strains with high accuracy (P < 10−5)5, a null test applied only to the set of 86 viable strains (wherein the null hypothesis is that all viable strains grow at the same rate as the wild type) would yield a success rate of 94%. In other words, most of the viable strains grow at rates that are essentially the same as those of the wild type, and some of these are not captured by the model.

Therefore, the greatest value of this data set is found by considering the discrepancies between the model and experimental data. By combining our quantitative growth rate measurements and predictions with the qualitative information from our previous work, we produced a detailed map of model-experiment comparisons for all 525 genes in the chromosome (Fig. 1c). To our knowledge, this is the most comprehensive and quantitative comparison of any large-scale cellular model’s predictions to growth-phenotype data, as other studies (including our own work) either considered only a small fraction of the total nonessential genes or else made strictly qualitative (growth or no growth) predictions5,8,9.

Scrutiny of this comparison map highlighted a small group of discrepancies, the resolution of which we hypothesized would be most likely to lead to new discoveries. The model–experimental data comparisons fell into seven categories, depending on the nature of the model prediction and whether a gene’s function was well-enough annotated for functional inclusion in the model (Supplementary Table 1). Two categories had the richest information content. The first group included 13 strains (Fig. 1c) for which the model could predict the qualitative essentiality but not the quantitative growth rate (P ≤ 0.01). The second group consisted of five strains for which the model failed qualitatively, predicting a growth rate that was insufficient to sustain life (‘lethal zone’; Fig. 1b); we therefore labeled the corresponding genes as ‘false essential’.

For four of the 18 strains in these two categories, the difference in growth rate between model and experiment was

1Department of Bioengineering, Stanford University, Stanford, California, USA. 2Graduate Program in Biophysics, Stanford University, Stanford, California, USA. 3Present address: Institute for Quantitative Biosciences, University of California Berkeley, Berkeley, California, USA. Correspondence should be addressed to M.W.C. (mcovert@stanford.edu).

RECEIVED 24 JULY; ACCEPTED 10 SEPTEMBER; PUBLISHED ONLINE 3 NOVEMBER 2013; DOI:10.1038/NMETH.2724
small (<20%). For the remaining 14 strains, five of the corresponding genes are associated with metabolism, two are linked to gene expression, three are involved in chromosome condensation and the remaining four genes had little or no functional annotation (Supplementary Table 2). For each of these genes, we explored the model’s inability to predict the experimental data, looking for a possible mechanism that could explain the discrepancy. We used the whole-cell model and the literature to examine the ‘molecular pathology’ of each single-gene disruption, as described in our earlier work. Using this analysis, we hypothesized a previously misrepresented or missing function for each of the hits for which there was a well-characterized gene annotation (Supplementary Results and Supplementary Fig. 2).

Three ‘hits’ were of particular interest because we could test model predictions for them using established methods. These hits were all metabolic genes: thyA (thymidylate synthase) and deoD (deoxyribose-phosphate aldolase), as previously reported, and MG_039 (glycerol phosphate). Prediction failures of the model were quantitative, not qualitative, for all three genes. (Supplementary Table 1). For each gene, we identified an alternate metabolic route that could compensate for its disruption. We used a strategy based on reduced costs, which are calculated as part of the metabolic submodel’s linear optimization method, to determine the metabolic fluxes that according to the model were limiting cell growth. Changing the parameter values for these fluxes would have the greatest impact on the specific growth rate calculated by the model, making them the most likely candidates for parameter estimation. For the thyA strain, we found only two metabolic reactions by reduced-cost analysis that did not also appear in the reduced-cost analysis of the wild-type strain (Fig. 2a).

Both reactions were catalyzed by thymidine kinase (Tdk). ThyA and one of the Tdk reactions have a common metabolic product, dTMP, which is required for DNA replication (Fig. 2b). Reduced-cost analysis of the deoD strain led to three candidate reactions, catalyzed by ThyA, Tdk and pyrimidine-nucleoside phosphorylase (Pdp) (Fig. 2c). Of these, only Pdp could compensate for the production of uracil by DeoD (Fig. 2d). For comparison, we performed the same reduced-cost analysis using a simple FBA framework without the additional constraints imposed by the whole-cell model. The stand-alone FBA model was unable to identify notable differences in reduced cost between the wild-type and disruption strains (Supplementary Table 3).

The reduced-cost analysis for MG_039, whose product converts dihydroxyacetone phosphate to glycerol-3-phosphate, did not highlight compensating reactions to MG_039. We therefore adopted a different approach, reasoning that growth of the disruption strain would be more susceptible than that of the wild-type strain to the inhibition of any candidate enzyme that could compensate for MG_039. We searched for metabolic reactions that, if constrained in the single-gene disruption, had a stronger effect on the calculated specific growth rate than in the wild type. Constraining glycerol kinase (GlpK) had the most pronounced difference in effect between the gene disruption strain and the wild type, sharply reducing the specific growth rate at a flux constraint that was over tenfold higher (Fig. 2e). GlpK catalyzes the production of glycerol-3-phosphate from glycerol as part of a complex fatty acid synthesis network with at least three interlocking...
cycles (Fig. 2f). The complexity of this subnetwork is the most likely reason for the failure of the reduced-cost analysis.

In other words, a compensating reaction existed, and was encoded in the model, for each of the three cases, yet the model failed to correctly predict the specific growth rates of the corresponding strains. We hypothesized that the failure was in the kinetic rates of the compensating enzymes. The metabolic component of the whole-cell model uses the rate of catalysis ($k_{\text{cat}}$), multiplied by the enzyme abundance (a dynamic variable in the model) as the upper bound of the metabolic flux for a given reaction. Many of the kinetic parameters required to build the original whole-cell model had not previously been measured in *M. genitalium* and had to be approximated from measurements in different organisms, including the parameters for Tdk, Pdp and GlpK. We therefore used the whole-cell model to determine a range for $k_{\text{cat}}$ values for Tdk, Pdp and GlpK that would reconcile the $\Delta$thylA, $\Delta$deoD and $\Delta$MG_039 growth phenotypes while minimizing the effects on growth of the wild type. For each strain, we plotted the absolute value of the difference between the model-predicted and experimentally observed specific growth rates as a function of $k_{\text{cat}}$ (Fig. 2g–i). We then examined all three plots together to determine a common difference cutoff between experimental and simulated specific growth rates. The cutoff that produced minimum discrepancy for all of the strains was 0.015 $h^{-1}$, which corresponds to roughly 20% of the specific growth rate of the wild type.

Implementing this cutoff, we observed a single range of $k_{\text{cat}}$ values that minimized the discrepancy between the model and experimental measurements for each enzyme. The range for Pdp differed in that the upper limit was unbounded. We confirmed that the predicted range of $k_{\text{cat}}$ values mapped well to the distribution of experimental measurements of growth rates of single-gene disruption strains (Fig. 2j–l).

The $k_{\text{cat}}$ ranges, calculated directly from model predictions, were experimentally testable. We expressed the *M. genitalium* genes in *Escherichia coli*, purified the proteins and performed kinetic assays (Fig. 3a–c and Supplementary Fig. 3). The measured values corresponded well with the model’s predictions and differed by at least one and up to four orders of magnitude from the values originally used to train the model (Fig. 3d). These results indicate that our model can make accurate quantitative predictions about previously unmeasured cellular properties.

As a final step in the discovery process, we incorporated all of the newly determined experimental parameters into the whole-cell model to test whether interactions between these parameters existed that might make the resulting model less predictive than expected. The new parameters led to better predictions for all strains of interest without compromising predictions of the...
Supplementary Fig. 5. Another possibility is that the *v* value, *k*_cat, of Tdk (a), Pdp (b) and GlpK (c). Error bars, s.d. of three technical replicates per substrate concentration. Linear regression was used to obtain a *v* value, *k*_cat, (99% confidence interval indicated on plot) was calculated from the enzyme concentration and *v* / *k*_cat, t, time. (d) Comparison of *k*_cat values used to train the model (‘original’), estimated from organisms other than *M. genitalium*, with model-based predictions and experimental measurements. (e) Predicted and measured *k*_cat values were input into the whole-cell model (*n* = 6) and compared to the experimentally measured (*n* = 5) specific growth rates and model predictions with original *k*_cat values. P values were determined by two-tailed t-test, *P* ≤ 0.01. Dashed gray lines, experimental mean. Error bars, s.d.

**METHODS**

Methods and any associated references are available in the online version of the paper.

**ACKNOWLEDGMENTS**

We acknowledge support from: an US National Institutes of Health Pioneer Award (5DP1LM01150-05) and an Allen Distinguished Investigator Award to M.W.C.; Siebel Scholars, US National Science Foundation and Stanford University Bio-X Fellowships to J.C.S.; an European Molecular Biology Organization Fellowship (ALTF1371-2011) to S.R.; US National Defense Science and Engineering Graduate Fellowship, US National Science Foundation Fellowship, and Stanford Graduate Fellowship to J.R.K.; and Weill and Rensselaer Engineering Fellowships to M.V.G.

**AUTHOR CONTRIBUTIONS**

J.C.S., J.R.K. and M.W.C. performed the growth rate measurements; J.C.S., S.R. and S.C. performed the simulations and computational analysis; J.C.S., J.R.K., M.V.G. and B.B. performed the growth rate measurements; J.C.S., S.R. and M.W.C. wrote the paper; and M.W.C. supervised the project.

**COMPETING FINANCIAL INTERESTS**

The authors declare competing financial interests: details are available in the online version of the paper.

Reprints and permissions information is available online at [http://www.nature.com/reprints/index.html](http://www.nature.com/reprints/index.html).

1. Morowitz, H.J. Isr. J. Med. Sci. 20, 750–753 (1984).
2. Suthers, P.F. et al. PLOS Comput. Biol. 5, e1000285 (2009).
3. Browning, S.T. & Shuler, M.L. Biotechnol. Bioeng. 76, 187–192 (2001).
4. Tomita, M. et al. Bioinformatics 15, 72–84 (1999).
5. Kari, J.R. et al. Cell 150, 389–401 (2012).
6. Glass, J.I. et al. Proc. Natl. Acad. Sci. USA 103, 425–430 (2006).
7. Orth, J.D., Thiele, I. & Palsson, B.O. Nat. Biotechnol. 28, 245–248 (2010).
8. Covert, M.W., Knight, E.M., Reed, J.L., Herrgard, M.L. & Palsson, B.O. Nature 429, 92–96 (2004).
9. Edwards, J.S. & Palsson, B.O. BMC Bioinformatics 1, 1 (2000).
10. Savinell, J.M. & Palsson, B.O. J. Theor. Biol. 155, 215–242 (1992).
11. Maier, T. et al. Mol. Syst. Biol. 7, 511 (2011).

**Figure 3 | Model-driven discovery.** (a–c) Hanes-Woolf plots of kinetic assays to measure the maximum reaction velocity (*v*_max) of Tdk (a), Pdp (b) and GlpK (c). Error bars, s.d. of three technical replicates per substrate concentration. Linear regression was used to obtain a *v*_max, *k*_cat value (99% confidence interval indicated on plot) was calculated from the enzyme concentration and *v*_max, t, time. (d) Comparison of *k*_cat values used to train the model (‘original’), estimated from organisms other than *M. genitalium*, with model-based predictions and experimental measurements. (e) Predicted and measured *k*_cat values were input into the whole-cell model (*n* = 6) and compared to the experimentally measured (*n* = 5) specific growth rates and model predictions with original *k*_cat values. P values were determined by two-tailed t-test, *P* ≤ 0.01. Dashed gray lines, experimental mean. Error bars, s.d.
**ONLINE METHODS**

**Single-gene disruption strain simulations and growth assays.** *M. genitalium* single-gene disruption and wild-type strains were obtained from J. Craig Venter Institute. Specific growth rates were determined according to the colorimetric protocol described in ref. 5. Simulations were performed as described. Simulation results can be found at SimTK (https://simtk.org/home/wholecell/). Simulated cells were modeled in an environment based on Spiroplasma 4 medium. At least six experimental replicates and five simulations were run for each disruption strain.

**Protein expression.** *M. genitalium* tdk, pdp and glpK genes were synthesized with GenScript with *E. coli* codon optimization. Genes were inserted into pGEX-6P-1 glutathione S-transferase (GST) expression vectors (GE Healthcare) by Gibson assembly, and sequences were verified. Resulting vectors were expressed in DH5α *E. coli* cells. Cells were lysed by sonication, and GST-tagged proteins were pulled down using glutathione sepharose beads (GE Healthcare). Proteins were cleaved from the beads by PreScission Protease (GE Healthcare), and quantified against a BSA standard by SDS electrophoresis (Supplementary Fig. 3a–c). Final concentrations were: Tdk, 1.3 mg ml⁻¹; Pdp, 0.2 mg ml⁻¹; and GlpK, 0.14 mg ml⁻¹.

**Kinetic assays.** Tdk activity was measured using a spectrophotometric assay similar to that described previously. Reaction schema: thymidine + ATP ↔ dTMP + ADP + H⁺ (catalyzed by Tdk); ADP + phosphoenolpyruvate + H⁺ ↔ pyruvate + ATP (catalyzed by pyruvate kinase); and pyruvate + NADH + H⁺ ↔ lactate + NAD⁺ (catalyzed by lactate dehydrogenase).

In a 75-µl reaction, we added 0.05 M Tris pH 7.2, 1 mM dithiothreitol, 2.5 mM MgCl₂, 5 mM ATP, 0.18 mM NADH, 0.21 mM phosphoenolpyruvate, 2.4 µg pyruvate kinase and 1.5 µg lactate dehydrogenase in H₂O (Sigma). Pyruvate kinase and lactate dehydrogenase were added in excess such that they were not rate-limiting (Supplementary Fig. 3e). Thymidine concentrations varied between 0.1 mM and 1.5 mM, and each reaction contained 13 µg of Tdk. Reactions were performed in triplicate. We measured the loss of NADH by measuring absorbance at 340 nm at 37 °C for 30 s intervals for 10–30 min.

GlpK activity was measured using a spectrophotometric assay similar to that described previously. Reaction schema: glycerol + ATP ↔ ADP + glycerol-3-phosphate (catalyzed by GlpK); ADP + phosphoenolpyruvate + H⁺ ↔ pyruvate + ATP (catalyzed by pyruvate kinase); and pyruvate + NADH + H⁺ ↔ lactate + NAD⁺ (catalyzed by lactate dehydrogenase).

In a 75-µl reaction, we added 0.05 M Tris pH 7.2, 1 mM dithiothreitol, 0.1 M KCl, 5.25 mM phosphoenolpyruvate, 2.5 mM MgSO₄, 0.2 mM NADH, 2.25 µg pyruvate kinase and 1.125 µg lactate dehydrogenase in H₂O (Sigma). Pyruvate kinase and lactate dehydrogenase were added in excess such that they were not rate-limiting (Supplementary Fig. 3f). ATP concentrations varied between 0.033 mM and 1 mM, and each reaction contained 1.8 µg of GlpK. Reactions were performed in triplicate. We measured the loss of NADH by measuring absorbance at 340 nm at 37 °C for 30 s intervals for 10–30 min.

Pdp activity was measured using a spectrophotometric assay similar to that described in ref. 15. Reaction schema: deoxyuridine + phosphate ↔ deoxyribose-1-phosphate + uracil (catalyzed by Pdp).

In a 400-µl reaction, we added 10 mM Tris pH 7.3, 10 mM phosphate pH 7.3 and 1 mM EDTA in H₂O. Deoxyuridine (Sigma) concentrations varied between 1.25 mM and 12.5 mM. Each reaction contained of 1 µg of Pdp and was conducted at 37 °C for 9 min. At 45-s intervals, we added 70 µl of 0.5 M NaOH to 30 µl of reaction mix to stop the reaction. We determined uracil production by measuring absorbance at 290 nm.

Blank controls were performed for all reactions with H₂O instead of enzyme. NADH and uracil concentrations were determined from measurements of absorbance at 340 nm and 290 nm, respectively, using NADH and uracil standard curves. The maximal slope of each reaction curve was used to determine its velocity, and Hanes–Woolf plots were used to determine the vₘₐₓ and kₖₐt (Fig. 3a–c). Hanes–Woolf results were compared to Michaelis–Menten nonlinear regression results (Supplementary Fig. 3g,h). Both methods yielded comparable results.

**Comparison of experimental data and model predictions.** The following statistical measures were taken in analyzing the results shown in Figures 1b and 3e: at least six replicates were performed for each gene disruption–specific growth rate measurement, and 15 replicates were performed for the wild type. At least five simulations were run for each model prediction, and 128 simulations were run for the wild type. A heteroskedastic two-tailed t-test was performed between each set of experimental measurements and model predictions. We considered experimental and model results to be significantly different if P values were ≤0.01. Owing to the small sample size, we wanted to be sure that our results were not biased by the distributions of the data or by extreme outliers. We performed a nonparametric Wilcoxon rank-sum test (also with cutoff of P ≤ 0.01), which identified the same set of gene ‘hits’. Finally, from the list of genes with P ≤ 0.01 (18 genes), we only considered those that were over or underpredicted by at least 20% (14 genes).

**Linear regression of kinetic assay data.** The following statistical measures were taken in computing the results of Figure 3a–c: kinetic reactions were performed in triplicate for seven substrate concentrations for each enzyme. A linear regression was fit to the data, and 99% confidence intervals were determined using the standard error of the slope and a t distribution obtained using α = 0.01 and degrees of freedom = n – 2 = 19.

**Enzyme quantification and kₖₐt range calculations.** Isolated enzymes were quantified by SDS electrophoresis against known quantities of BSA. Band volumes were quantified using Quantity One v.4.6.9 software (intensity × area). Linear regressions of the BSA standard curve were used to determine the isolated enzyme concentration. 95% confidence intervals of the linear regressions were determined using the standard errors of the slope and intercept and t distributions obtained using α = 0.05 and degrees of freedom = n – 2 (Supplementary Fig. 3a–c). The lower enzyme concentration bound and upper vₘₐₓ bound for each enzyme was used to calculate the upper bound of the kₖₐt, and the upper enzyme concentration bound, and lower vₘₐₓ bound for each enzyme was used to calculate the lower bound of the kₖₐt (Supplementary Fig. 3d).

The final predicted values of kₖₐt used to constrain the whole-cell model were 0.215 s⁻¹ for Tdk, 0.5 s⁻¹ for Pdp and GlpK = 0.46 s⁻¹.
for GlpK. The measured $k_{\text{cat}}$ values used were 0.215 s$^{-1}$ for Tdk, 78.1 s$^{-1}$ for Pdp and 0.80 s$^{-1}$ for GlpK (Fig. 3e). The $k_{\text{cat}}$ values were input into the whole-cell model and compared to the experimentally measured specific growth rates and model predictions with original $k_{\text{cat}}$ values. $P$ values were determined by two-tailed t-test, $P \leq 0.01$.

**Flux balance analysis.** Flux balance analysis (FBA) has been used to make predictions of cellular growth rates based on a given metabolic network and environmental conditions. However, other studies have noted the difficulty in quantitatively predicting the short-term effects of gene deletion on cell growth using FBA. To test the capacity of FBA to quantitatively predict phenotypes of single-gene disruption strains for the *M. genitalium* metabolic network, we analyzed phenotypes of single-gene disruption strains of the metabolic genes in the same data set using FBA. We found that these predictions fell into only two categories: either zero growth rate or growth rate essentially similar to that of the wild type (Supplementary Fig. 1a). In contrast, the whole-cell model predictions included specific growth rates across the range of 0–115% of the growth rate of the wild type (Supplementary Fig. 1b).

The distribution was more descriptive than FBA, even in the case where only metabolic genes were considered. This most likely stemmed from two causes: first, the variation near to the growth rate similar to that of the wild type arose predominantly from the stochastic aspects of the whole-cell model; and second, predicted specific growth rates more distant from that of the wild type were due to the substantial constraints on the whole-cell model’s metabolic network. Specifically, the metabolic module of the whole-cell model was solved using a similar linear optimization strategy to FBA but with 63% of the catalysis reactions constrained by rate parameters as opposed to none of the catalysis reactions typically being constrained in FBA studies. The tight, detailed constraints on almost every metabolic reaction in the whole-cell model arise from a combination of many cellular processes in the whole-cell framework including transcription, tRNA aminoacylation, translation, protein processing and modification, protein translocation and folding, and macromolecular complexation. Previous studies have shown that the addition of new constraints to FBA improves predictive ability, and our comparison underscores that such constraints may be essential to make accurate quantitative predictions about cellular growth rates. We conclude that the quantitative specific growth rates and hypotheses presented here required the use of the whole-cell model and would not have been achieved by FBA alone.

12. Gibson, D.G. et al. Nat. Methods **6**, 343–345 (2009).
13. Schelling, P., Folkers, G. & Scapozza, L. Anal. Biochem. **295**, 82–87 (2001).
14. Lester, L.M., Rusch, L.A., Robinson, G.J. & Speckhard, D.C. Biochemistry **37**, 5349–5355 (1998).
15. Leer, J.C., Hammer-Jespersen, K. & Schwartz, M. Eur. J. Biochem. **75**, 217–224 (1977).
16. Fong, S.S. & Palsson, B.O. Nat. Genet. **36**, 1056–1058 (2004).