Testing theoretical minimal genomes using whole-cell models

Joshua Rees-Garbutt [1,2*], Jake Rightmyer [2], Oliver Chalkley [1,3,4], Lucia Marucci [1,4,5 +*],

Claire Grierson [1,2 +*]

1. BrisSynBio, University of Bristol, Bristol BS8 1TQ, UK;
2. School of Biological Sciences, University of Bristol, Bristol Life Sciences Building, 24 Tyndall Avenue, Bristol, BS8 1TQ, UK;
3. Bristol Centre for Complexity Sciences, Department of Engineering Mathematics, University of Bristol, Bristol, BS8 1UB, UK
4. Department of Engineering Mathematics, University of Bristol, Bristol BS8 1UB, UK;
5. School of Cellular and Molecular Medicine, University of Bristol, Bristol BS8 1UB, UK;

+ Co-last authors * Corresponding authors

Corresponding authors: Joshua Rees-Garbutt (joshua.rees@bristol.ac.uk), Prof. Claire Grierson (claire.grierson@bristol.ac.uk), Dr. Lucia Marucci (lucia.marucci@bristol.ac.uk)

Summary

The minimal gene set for life has often been theorised, with at least ten produced for *Mycoplasma genitalium (M.genitalium)* [1–10]. Due to the difficulty of using *M.genitalium* in the lab [11], combined with its long replication time of 12 - 15 hours [12–14], none of these theoretical minimal genomes have been tested, even with modern techniques [12]. The publication of the *M.genitalium* whole-cell model [6] provided the first opportunity to test them, simulating the genome edits *in-silico*. We simulated eight minimal gene sets from the literature [1–8], finding that they produced *in-silico* cells that did not divide. Using knowledge from previous research [15], we reintroduced specific essential and low essential genes, producing dividing *in-silico* cells. This reinforces the need to identify species-specific low essential genes and their interactions [14,16].

Genome engineering builds on historical gene essentiality research. The sequencing of small bacterial genomes [17,18] led to comparative genomics, then, as genome sequencing capacity increased, minimal gene sets [1–10] were hypothesised. Minimal genomes are reduced genomes where no single gene can be removed without loss of viability [14], given an appropriately rich medium and no external stresses, and focusing solely on protein-coding genes. For a recent review of gene essentiality, see Rancati *et al*. [19].

M.genitalium is the focal point of minimal gene set creation due its naturally small genome size (0.58mb and 525 genes) and early sequenced genome [18]. Minimal gene sets are designed using three different approaches: protocells, comparative genomics, and single gene knockouts. Protocell designs [20] are not expected to function as true biological cells, instead functioning as a self-replicating, membrane-encapsulated collection of biomolecules [5]. Comparative genomics [21] compares multiple species to identify common genes. This is complicated [14,16] by non-orthologous gene displacements, i.e. independently evolved or diverged proteins that perform the same function but are not recognisably related [1,14], which can result in the removal of a large number of genes essential to one species. Design using single gene essentiality classifications should, in theory, not remove any essential genes; but if transposon mutagenesis is used, variance from different transposon variants, antibiotic resistance genes, and growth periods can result in differing essentiality classifications [22,23].

Ten minimal gene sets were found in the literature that were designed with *M.genitalium* genes [1–10], however two sets (Gil *et al.* (2004)[9]and Shuler *et al.*[10]) were excluded as they were considered derivative of the Gil (2014) set [8]; four genes differ in the Shuler *et al.* set (MG_056, MG_146, MG_388, MG_391) and four genes are absent in the Gil *et al.* (2004) set (MG_009, MG_091, MG_132, MG_460). Of the other eight sets, two (Tomita *et al.* [3] and Church *et al.* [5]) were designed as protocells, three (Mushegian and Koonin [1], Huang *et al.* [7], and Gil [8,9]) from comparative genomics, and three (Hutchison *et al.* [2], Glass *et al.* [4], and Karr *et al.* [6]) from single gene essentiality experiments. To prevent confusion, we named the sets after the main location where the set was constructed (Table 1). The Bethesda set directly compared *M.genitalium* and *Haemophilus influenzae* genomes (gram-positive and gram-negative bacteria respectively) [1]. The Rockville set applied global transposon mutagenesis to *M.genitalium in-vivo* to identify non-essential genes [2]. The Fujisawa set constructed an *in-silico* hypothetical cell from 127 *M.genitalium* genes using the E-Cell software [3]. The Rockville 2 set reapplied global transposon mutagenesis *in-vivo*, by isolating and characterising pure clonal populations [4]. The Nashville set listed 151 *E.coli* genes (compared to *M.genitalium* genes within the paper) theorised to produce a chemical system capable of replication and evolution [5]. The Stanford set was the result of *in-silico* single gene knockouts using the *M.genitalium* whole-cell model [6]. The Guelph set compared 186 bacterial genomes [7], whereas the Valencia set compared *M.genitalium* with genetic data of five insect endosymbionts [8].

We adapted these eight minimal gene sets for simulation within the *M.genitalium* whole-cell model. The Nashville, Fujisawa, and Stanford sets were unchanged, but the others had between 6 and 44 genes removed (Table 1) either because the genes were unmodelled (the gene's function is unknown [6]) or specific genes were listed twice. The protocell designs (Nashville, Fujisawa) predicted the smallest *in-silico* genome. Guelph contained substantially fewer genes than Valencia and Bethesda, due to comparing 186 species [7]. Stanford, Rockville, Rockville 2 had similar numbers of *in-vivo* deletions, but Rockville and Rockville 2 had the highest numbers of unmodelled genes (as genes can be disrupted *in-vivo* without knowing the gene's function). The *in-silico* genomes (and associated gene deletions) of the minimal gene sets are listed in Supplementary Data 1 and 2.

Prior to simulations, we analysed the comparative genomics and single gene essentiality designed minimal gene sets for design commonalities (the protocell designed minimal gene sets were excluded due to their much reduced size) and identified 96 genes in common

(Supplementary Data 1 (col L) and 3). 87 of these were classified as essential (genes that when removed stop the cell successfully dividing, classified *in-silico* previously [6,15]), eight were non-essential (removal did not prevent successful cell division), and one gene was unmodelled.

The 87 essential genes affect a range of cellular functions including: DNA (repair, supercoiling, chromosome replication, nucleotide synthesis/modification, sigma factors, ligation, transcription termination, and DNA polymerase); RNA (ribosome proteins, translation initiation factors, tRNA modification, ribonucleases, and RNA polymerase); and cellular processes (protein folding/modification, protein shuttling, protein membrane transport, metabolic substrates production/recycling, redox signalling, oxidation stress response, and the pyruvate dehydrogenase complex). Of the eight non-essential genes, four (MG_048, MG_072, MG_170, MG_297) are associated with the SecYEG complex [24] (protein transport across or into the cell membrane), while MG_172 removes protein synthesis targets from synthesised proteins, MG_305 and MG_392 assist in late protein folding, and MG_425 processes ribosomal RNA precursors. Although these eight genes are singly non-essential, by single gene deletion *in-silico* [6] and *in-vivo* [4], they all play a part in essential functions, hence their inclusion.

We also identified 14 genes deleted by all eight minimal gene sets (Supplementary Data 2 (col L) and 4). The functions of these genes include: fructose import, host immune response activation, chromosomal partition, amino acid transport, antibody binding, phosphonate transport, external DNA uptake, DNA repair, rRNA modification, membrane breakdown, toxin transport, quorum sensing, and a restriction enzyme. These had been previously classified as non-essential by single gene deletion *in-silico* [6] and *in-vivo* [4]. We placed these 14 common genes in an ‘Agreed set’ and a genome with these genes removed was also simulated.

We simulated each minimal gene set in the *M.genitalium* whole-cell model and found that every set, including the Agreed set, produced a non-dividing *in-silico* cell (30 repetitions, Supplementary Data 5).

Analysis found that every one of the sets deleted essential genes (classified *in-silico* previously [6,15]): Nashville deleted 121, Fujisawa deleted 112, Guelph deleted 107, Valencia deleted 69, Bethesda deleted 34, Rockville and Rockville 2 both deleted 9, and Stanford deleted 3 (Supplementary Data 6-14). This is especially surprising for the single gene essentiality minimal gene sets. For the Rockville sets, this is likely due to transposon mutagenesis issues. Rockville labelled six genes as non-essential in 1999, subsequently labelled essential in 2006. Additionally, Rockville grew cells in mixed pools with DNA isolated from these mixtures rather than from isolated pure colonies of cells [22]. For the Stanford set, the removal of MG_203, MG_250, and MG_470 is likely due to averaging multiple simulation's data together before computational assessment, genes instead found to be essential (Supplementary Table 3[15]).

In an attempt to restore *in-silico* division (Figure 1), we reintroduced essential genes to the minimal gene sets. Based on previous research [15], we also reintroduced low essential genes (i.e. genes dispensable in some contexts, such as redundant essential genes and gene complexes [19]). We did this by comparing the gene content of the individual minimal gene sets with a complete list of the *M.genitalium in-silico* genes and their essentiality classifications [15] (Supplementary Data 6-14). For example, the original Agreed set removed low essential genes MG_291 (phosphonate transport) and MG_412 (phosphate transport); by disrupting both these processes, the *in-silico* cell has no functioning source of phosphate, which has been established previously [15]. By reintroducing set-specific genes (Table 2, Supplementary Data 15), each modified set, including the Agreed set, was able to produce a dividing cell *in-silico* (30 repetitions, Supplementary Data 5).

In an attempt to gain further understanding, we investigated what processes the nine repaired minimal gene sets removed using gene ontology (GO) biological process term, and we compared the repaired minimal gene sets to the *in-silico M.genitalium* minimal genomes we produced previously [15]. The smallest repaired *in-silico* genomes (Nashville and Fujisawa, 260 genes) were larger than the prior *in-silico* minimal genomes (256 and 237 genes), though with the removal of the low essential phosphonate or phosphate transport genes, could match the larger of the two *in-silico* minimal genomes (Supplementary Table 6[15]). The other sets had different designs that, due to not systematically targeting non-essential genes, resulted in non-essential genes remaining in the genome, making them subsets of the repaired protocell sets and the prior *in-silico* minimal genomes (Supplementary Data 16). As such, the GO terms were also subsets and did not deviate from what we would expect to produce a dividing *in-silico* cell (Supplementary Data 17-25).

Analysis of the repaired sets found that 31 genes were reintroduced into five or more of the minimal gene sets (Supplementary Data 26). 26 were essential and 5 were low essential (Supplementary Table 5[15]). The corresponding cellular functions included: DNA (polymerase subunits, thymidine insertion, recycling of pyrimidine, chromosome segregation); RNA (polymerase subunit, tRNA modification, the 50S and 30S ribosomal subunits); transporters (cobalt, phosphonate, potassium); production (NAD, flavin, NADP, fatty acid/phospholipids); and dehydrogenation (glycerol and alpha-keto acids). Of the 26 reintroduced essential genes, 19 were already present in the single gene deletion minimal gene sets (Stanford, Rockville, Rockville 2). MG_137 (mycobacteria cell wall production) and MG_517 (plasma membrane stability) are genes specifically essential for *Mycoplasma* species, which were only identified as essential by the single gene deletion minimal gene sets. A further five genes were involved in cobalt transport, which increases the rates of DNA synthesis, fatty acid metabolism, and amino acid metabolism, and were also not identified as essential by the other design methodologies.

Of the five reintroduced low essential genes, Bethesda did not delete four, likely due to the direct comparison resulting in low essentiality genes being conserved to a greater degree.

We also looked at reintroductions to the protocell sets, as they could outline additional cellular requirements for the successful unification of protocell systems. The genes reintroduced to the Nashville set repaired functions that had been reduced (translation, glycolytic process, protein folding) and restored functions that had been removed including: cell (division, cycle, transport, redox homeostasis), DNA (topological change, transcription), rRNA processing (including pseudouridine synthesis), protein transport, and cellular processes (carbohydrate metabolic, glycerol metabolic, fatty acid biosynthesis, UMP salvage) (Supplementary Data 27). The glycolytic process had the most change, with 10 of 11 genes being reintroduced, and DNA repair had the least, with only one gene being reintroduced (MG_254), however, this did allow single strand DNA break repair. The genes reintroduced to the Fujisawa set additionally included tRNA processing and protein folding (Supplementary Data 28) with 8 out of 10 DNA replication genes reintroduced.

In conclusion, the repaired protocell minimal gene sets (Nashville and Fujisawa) produced the smallest genomes *in-silico* (Table 2), differing by 6 genes (Supplementary Data 29), but required the most gene reintroductions. The repaired comparative genomics minimal gene sets (Guelph, Valencia, Bethesda) required fewer genes reintroduced the fewer genomes compared in their original design. Interestingly, Stanford (a single gene essentiality set) produced a smaller *in-silico* genome than Bethesda (a comparative genomics set), as it did not target unmodelled gene deletions and only required eight genes to be reintroduced.

Without the ability to identify species-specific low essential genes, any minimal gene sets designed with the current incomplete gene essentiality information will require gene reintroductions to produce dividing cells.

This research has limitations associated with the use of the *M.genitalium* whole-cell model. Through necessity the *M.genitalium* whole-cell model bases some of its parameters on data from other bacteria [8] and is only capable of modelling a single generation, missing subgenerational gene expression and subsequent essentiality which affects >50% of the genes in some species [25]. Additional uncertainty exists around the unknown impact of the unmodelled genes on *in-vivo* experiments, as stated previously [15].

The computational predictions we have produced need to be tested in living cells, but with the advancement of gene synthesis and genome transplantation in other *Mycoplasma* species [12,26–30] this is becoming a more realistic proposition for *Mycoplasma* researchers.

Acknowledgements

We thank the Advanced Computing Research Centre (ACRC) and BrisSynBio, a BBSRC/EPSRC Synthetic Biology Research Centre, at the University of Bristol for access to the Bluegem supercomputer. L.M. was supported by the Engineering and Physical Sciences Research Council (grants EP/R041695/1 and EP/S01876X/1), and by the European Union's Horizon 2020 research and innovation programme (Grant agreement 766840). L.M. and C.G. were supported by BrisSynBio, a BBSRC/EPSRC Synthetic Biology Research Centre (BB/L01386X/1). J.R-G. was supported by a EPSRC Future Opportunity Scholarship. O.C. was supported by the Bristol Centre for Complexity Sciences (BCCS) Centre for Doctoral Training (CDT) EP/I013717/1.

Author Contributions

C.G., L.M., O.C., J.R-G were involved in ideation. O.C. conducted simulations comparing the Stanford and Rockville 2 sets, a prototype version of the research presented here. J.R-G. collated, simulated, analysed and repaired the nine sets, and wrote the paper and supplementary data 1 - 16, 26 - 29. J.R. analysed the nine sets and wrote supplementary data 17 - 25. C.G. and L.M supervised the project. C.G., L.M., J.R. edited the paper.

Competing Interests Statement

The authors declare no competing interests.

References

1. Mushegian, A. R. & Koonin, E. V. A minimal gene set for cellular life derived by comparison of complete bacterial genomes. *Proc. Natl. Acad. Sci. U. S. A.* **93**, 10268–10273 (1996).
2. Hutchison, C. A. *et al.* Global transposon mutagenesis and a minimal mycoplasma genome. *Science* **286**, 2165–2169 (1999).
3. Tomita, M. *et al.* E-CELL: software environment for whole-cell simulation. *Bioinformatics* **15**, 72–84 (1999).
4. Glass, J. I. *et al.* Essential genes of a minimal bacterium. *Proc. Natl. Acad. Sci. U. S. A.* **103**, 425–430 (2006).
5. Forster, A. C. & Church, G. M. Towards synthesis of a minimal cell. *Mol. Syst. Biol.* **2**, (2006).
6. Karr, J. R. *et al.* A whole-cell computational model predicts phenotype from genotype. *Cell* **150**, 389–401 (2012).
7. Huang, C. H., Hsiang, T. & Trevors, J. T. Comparative bacterial genomics: defining the minimal core genome. *Antonie Van Leeuwenhoek International Journal of General and Molecular Microbiology* **103**, 385–398 (2013).
8. Gil, R. The Minimal Gene-Set Machinery. *Reviews in Cell Biology and Molecular Medicine* (2014).
9. Gil, R., Silva, F. J., Pereto, J. & Moya, A. Determination of the core of a minimal bacterial gene set. *Microbiol. Mol. Biol. Rev.* **68**, 518–+ (2004).
10. Shuler, M. L., Foley, P. & Atlas, J. Modeling a Minimal Cell. in *Microbial Systems Biology* (ed. Navid, A.) vol. 881 573–610 (Humana Press, 2012).
11. Reich, K. A. The search for essential genes. *Res. Microbiol.* **151**, 319–324 (2000).
12. Benders, G. A. *et al.* Cloning whole bacterial genomes in yeast. *Nucleic Acids Res.* **38**, 2558–2569 (2010).
13. Hutchison, C. A. *et al.* Design and synthesis of a minimal bacterial genome. *Science* **351**,

1414–U73 (2016).

14. Glass, J. I., Merryman, C., Wise, K. S., Hutchison, C. A., 3rd & Smith, H. O. Minimal Cells-Real and Imagined. *Cold Spring Harb. Perspect. Biol.* (2017) doi:10.1101/cshperspect.a023861.
15. Rees-Garbutt, J. *et al.* Designing minimal genomes using whole-cell models. *Nat. Commun.* **11**, 836 (2020).
16. Lagesen, K., Ussery, D. W. & Wassenaar, T. M. Genome update: the 1000th genome - a cautionary tale. *Microbiology-Sgm* **156**, 603–608 (2010).
17. Fleischmann, R. D. *et al.* Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. *Science* **269**, 496–512 (1995).
18. Fraser, C. M. *et al.* THE MINIMAL GENE COMPLEMENT OF MYCOPLASMA-GENITALIUM. *Science* **270**, 397–403 (1995).
19. Rancati, G., Moffat, J., Typas, A. & Pavelka, N. Emerging and evolving concepts in gene essentiality. *Nat. Rev. Genet.* **19**, 34–49 (2018).
20. Dzieciol, A. J. & Mann, S. Designs for life: protocell models in the laboratory. *Chem. Soc. Rev.* **41**, 79–85 (2012).
21. Koonin, E. V. Comparative genomics, minimal gene-sets and the last universal common ancestor. *Nat. Rev. Microbiol.* **1**, 127–136 (2003).
22. Glass, J. I. *et al.* Essential genes of a minimal bacterium. *Proc. Natl. Acad. Sci. U. S. A.* **103**, 425–430 (2006).
23. Juhas, M., Eberl, L. & Glass, J. I. Essence of life: essential genes of minimal genomes. *Trends Cell Biol.* **21**, 562–568 (2011).
24. du Plessis, D. J. F., Nouwen, N. & Driessen, A. J. M. The Sec translocase. *Biochim. Biophys. Acta* **1808**, 851–865 (2011).
25. Macklin, D. N. *et al.* Simultaneous cross-evaluation of heterogeneous E. coli datasets via mechanistic simulation. *Science* **369**, (2020).

26. Tsarmpopoulos, I. *et al.* In-Yeast Engineering of a Bacterial Genome Using CRISPR/Cas9. *ACS Synth. Biol.* **5**, 104–109 (2016).

27. Karas, B. J. *et al.* Direct transfer of whole genomes from bacteria to yeast. *Nat. Methods* **10**, 410–+ (2013).

28. Gibson, D. G. *et al.* Creation of a Bacterial Cell Controlled by a Chemically Synthesized Genome. *Science* **329**, 52–56 (2010).

29. Gibson, D. G. *et al.* Complete chemical synthesis, assembly, and cloning of a Mycoplasma genitalium genome. *Science* **319**, 1215–1220 (2008).

30. Gibson, D. G. *et al.* One-step assembly in yeast of 25 overlapping DNA fragments to form a complete synthetic Mycoplasma genitalium genome. *Proc. Natl. Acad. Sci. U. S. A.* **105**, 20404–20409 (2008).

31. Kanehisa, M. & Goto, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **28**, 27–30 (2000).

32. Apweiler, R. *et al.* UniProt: the Universal Protein knowledgebase. *Nucleic Acids Res.* **32**, D115–9 (2004).

Tables / Figures

Minimal gene set	Code name	Design approach	*in-vivo* genome design size*	Unmodelled genes^ in genome design	Predicted *in-silico* genome size*	Predicted gene deletions *in-silico**
Forster and Church 2006 [5]	Nashville	Protocell	89	0	89	270
Tomita *et al.* 1999 [3]	Fujisawa	Protocell	98	0	98	261
Huang *et al.* 2013 [7]	Guelph	Comparative Genomics	123	5	118	241
Gil 2014 [8]	Valencia	Comparative Genomics	180	6	174	185
Mushegian and Koonin 1996 [1]	Bethesda	Comparative Genomics	253	12	241	118
Karr *et al.* 2012 [6]	Stanford	Single Gene Deletions	242	0	242	117
Glass *et al.* 2006 [4]	Rockville 2	Single Gene Deletions	258	44	302	57
Hutchison *et al.* 1999 [2]	Rockville	Single Gene Deletions	265	41	306	53
-	*M.genitalium* whole-cell Model*	-	359	124	359	-
-	*M.genitalium in-vivo* *	-	483	-	-	-

Table 1. Minimal Gene Sets from the literature, compared with *M.genitalium in-vivo* and the whole-cell model. *M.genitalium* has 42 RNA-coding genes that are not included in this table. * = protein-coding genes. ^ = due to unknown function.

Code name	*In-silico* gene deletions (cell did not divide)	*In-silico* gene deletions (cell divided)	Genes reintroduced	Size of *in-silico* genome
Nashville	270	141	129	260
Fujisawa	261	141	120	260
Guelph	241	129	112	272
Valencia	185	110	75	291
Stanford	117	109	8	292
Bethesda	118	82	36	319
Rockville 2	57	45	12	356
Rockville	53	43	10	358
Agreed	14	13	1	388

Table 2. Minimal gene sets that produce dividing *in-silico* cells after the reintroduction of essential and low essential genes. Size of *in-silico* genome minus the gene deletions that produced a dividing cell from the 359 protein-coding genes, but includes the 42 RNA-coding genes in the *M.genitalium* whole-cell model.

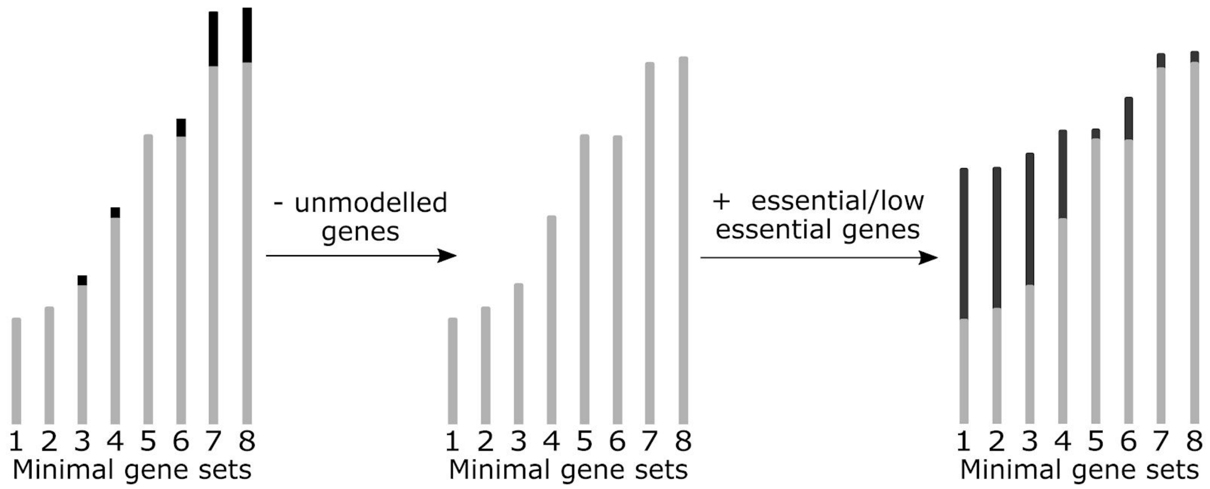

Figure 1. Testing and restoring the minimal gene sets. 124 genes were unmodeled in the *M.genitalium* whole-cell model due to unknown function. Essential genes are required by the cell to enable survival until successful division. Low essential genes are required by the cell in certain genomic and environmental contexts to enable survival to successful division. The restored minimal gene sets produced dividing *in-silico* cells.

Methods

Code Availability

All code created as part of this paper will be made available on Github (github.com/squishybinary, github.com/GriersonMarucciLab) under a GNU General Public License v3.0 (gpl-3.0). For more information see choosealicense.com/licenses/lgpl-3.0/.

Data Availability

The databases used to design the *in-silico* experiments, and compare the results to, includes Karr *et al.* [6] and Glass *et al.* [4] Supplementary Tables, and Fraser *et al. M.genitalium* G37 genome [18] interpreted by KEGG [31] and UniProt [32] as strain ATCC 33530/NCTC 10195. The output .fig files for all simulations referenced will be made available at the group's Research Data Repository (data-bris) at the University of Bristol.

Model Availability

The *M.genitalium* whole-cell model is freely available: github.com/CovertLab/WholeCell. The model requires a single CPU and can be run with 8GB of RAM. We run the *M.genitalium* whole-cell model on Bristol's supercomputers using MATLAB R2013b, with the model's standard settings. However, we use our own version of the SimulationRunner.m. MGGRunner.m (github.com/GriersonMarucciLab/Analysis_Code_for_Mycoplasma_genitalium_whole-cell_model) is designed for use with supercomputers that start hundreds of simulations simultaneously. It artificially increments the starting time-date value for each simulation, as this value is subsequently used to create the initial conditions of the simulation. Our research copy of the whole-cell model was downloaded 10th January 2017.

M.genitalium in-silico **Environmental Conditions**

M.genitalium is grown *in-vivo* on SP4 media. The *in-silico* media composition is based on the experimentally characterized composition, with additional essential molecules added (nucleobases, gases, polyamines, vitamins, and ions) in reported amounts to support *in-silico* cellular growth. Additionally, the *M.genitalium* whole-cell model represents 10 external stimuli including temperature, several types of radiation, and three stress conditions. For more information see Karr *et al.* Supplementary Tables S3F, S3H, S3R [6].

Equipment

For the *M.genitalium* whole-cell model we used the University of Bristol Advanced Computing Research Centres's BlueGem, a 900-core supercomputer, which uses the Slurm queuing system, to run whole-cell model simulations.

We used a standard office desktop computer, with 8GB of ram, to write new code, and interact with the supercomputer. We used the following GUI software on Windows 7: Notepad++ for code editing, Putty (ssh software) for terminal access to the supercomputer, FileZilla (ftp software) to move files in bulk to and from the supercomputer, and PyCharm (IDE software) as an inbuilt desktop terminal and for python debugging. The command line software used included: VIM for code editing, and SSH, Rsync, and Bash for communication and file transfer with the supercomputers.

Data Format

For the *M.genitalium* whole-cell model the majority of output files are state-NNN.mat files, which are logs of the simulation split into 100-second segments. The data within a state-NNN.mat file is organised into the 16 cellular variables. These are typically arranged as 3-dimensional matrices or time series, which are flattened to conduct analysis. The other file types contain summaries of data spanning the simulation. Each gene manipulated simulation can consist of up to 500 files requiring between 0.4GB and 0.9GB. Each simulation takes 5 to 12 hours to complete in real time, 7 - 13.89 hours in simulated time.

Data Analysis Process

For the *M.genitalium* whole-cell model, the raw data is automatically processed as the simulation ends. runGraphs.m carries out the initial analysis, while compareGraphs.m overlays the output on collated graphs of 200 unmodified *M.genitalium* simulations. Both outputs are saved as MATLAB .fig and .pdfs, though the .pdf files were the sole files analysed. The raw .mat files were stored in case of further investigation.

The GO biological process terms used for further analysis were downloaded from Uniprot [32] (strain ATCC 33530/NCTC 10195), processed by a created script (github.com/squishybinary/Gene_Ontology_Comparison_for_Mycoplasma_genitalium_whole-cell_model) in combination with lists of genes, organised manually into tables of GO terms that were unaffected, reduced, or removed entirely by gene deletions, and then analysed.

Modelling Scripts

There are six scripts used to run the *M.genitalium* whole-cell model. Three are the experimental files created with each new experiment (the bash script, gene list, experiment list), and three are stored within the whole-cell model and are updated only upon improvement (MGGrunner.m, runGraphs.m, and compareGraphs.m). The bash script is a list of commands for the supercomputer(s) to carry out. Each bash script determines how many simulations to run, where to store the output, and where to store the results of the analysis. The gene list is a text file containing rows of gene codes (in the format 'MG_XXX',). Each row corresponds to a single simulation and determines which genes that simulation should knockout. The experiment list is a text file containing rows of simulation names. Each row corresponds to a single simulation and determines the final location of the simulation output and analysis results.