Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in
Systems Biology

Furthering genome design using models and algorithms

Joshua Rees-Garbutt[1,2], Jake Rightmyer[1], Jonathan R. Karr[3], Claire Grierson[1,2] and Lucia Marucci[2,4,5]

Abstract

Large-scale *in silico* genome designs are on the brink of being engineered *in vivo*, offering a potential paradigm shift for cellular research (previous designs relied on fractured available knowledge and *in vivo* engineering iteration) by integrating computational design, *in silico* models and algorithms, with laboratory construction. However, several challenges remain. If *in vivo* engineering is successful, designing genomes can be used to gain new understanding of cellular life, improve the metabolite production process and reduce the risk of unintended genetic modification and release. Here, we review the progress so far. We suggest improvements on recent models and algorithms, illustrate the next steps for integrating computational and laboratory engineering and offer our opinions on the future of the field.

Addresses

[1] School of Biological Sciences, University of Bristol, Bristol Life Sciences Building, 24 Tyndall Avenue, Bristol, BS8 1TQ, UK
[2] BrisSynBio, University of Bristol, Bristol, BS8 1TQ, UK
[3] Icahn Institute for Data Science and Genomic Technology and Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY, 10029, USA
[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

Corresponding authors: Grierson, Claire (lacsg@bristol.ac.uk); Rees-Garbutt, Joshua (joshua.rees@bristol.ac.uk); Marucci, Lucia (lucia.marucci@bristol.ac.uk); Karr, Jonathan R (karr@mssm.edu)

Current Opinion in Systems Biology 2020, **24:**120–126

This review comes from a themed issue on **Synthetic Biology (2020)**

Edited by **Matteo Barberis** and **Tom Ellis**

For complete overview of the section, please refer the article collection - Synthetic Biology (2020)

Available online 16 October 2020

https://doi.org/10.1016/j.coisb.2020.10.007

Keywords

Genome-scale design, Genome-scale engineering, Algorithms, Whole-cell models, *Insilico*, Minimal genomes.

Introduction

Although we now have substantial knowledge to design individual cellular components, our ability to design entire cells is limited by the fractured nature of our knowledge. Novel biological design has been constrained to small-scale cellular research, resulting in the engineering of genetic circuits [1] and protocells [2], improving metabolite production [3] and designing modular production strains *in silico* [4] (prototyped *in vivo* [5]). Large-scale (hereafter genome-scale) cellular research has designed and engineered recoded genomes [6–8], chassis [9,10] and near-minimal genomes [11,12] (reviewed recently by Landon et al. [13]), but the engineering of novel cells is yet to be achieved [14]. This review introduces models and algorithms and their integration into genome-scale design since 2012, with context from the last 25 years. We also outline the biggest open challenges and provide a roadmap for overcoming them.

Test beds for designing genomes: minimal cells

For genome-scale design, minimal genomes are currently the best proof of concept [15]; they are briefly introduced here as they are used as illustration throughout this paper. Minimal genome research [16] attempts to understand the minimum requirements and origins of life [17]. In this article, we consider 'minimal' to be where no one single-protein–coding gene can be removed without preventing successful reproduction, given an appropriately rich medium and no external stresses [12]. Minimal genome research has focused on natural species, as we lack the knowledge to design genomes from scratch. *Mycoplasma genitalium* (*M. genitalium*) is one such species, because of its small genome size (0.58 mb, 525 genes) and early sequencing [18]. Recent genome reductions, relying on *in vivo* iteration, resulted in 15%–50% reductions in *Mycoplasma mycoides* [11], *Escherichia coli* (*E. coli*) [9,10,19–22] and *Bacillus subtilis* [23–25] and a recoded and synthetic *E. coli* genome being successfully engineered and transplanted [6]. This iterative development was caused by needing to identify contextual essentiality as part of research efforts. Gene essentiality depends both on the environmental context (i.e. how cells are grown) [26] and on the genomic context (i.e. what other genes are present)

[15], making it contextual, but we only have access to single or double knockout gene essentiality data [27,28].

There are other ways of defining 'minimal' in the literature, for example, reducing total genome size (by removing noncoding elements [29]), reducing the number of codons (i.e. genome recoding [6]), removing nonessential enzymes [30] and streamlining the metabolism through protein colocalisation [31,32].

Genome-scale design is complicated by scale; if you consider the removal of protein-coding genes for minimal genome designs using brute force and no assumptions, the number of possible genome-scale designs with *M. genitalium's* 525 genes is 2^{525}. This is infeasible *in vivo*; laboratories can only follow a small number of research avenues [15]. High-quality computational models can investigate many more research avenues and model contextual essentiality, assuming that cellular interactions are modelled correctly and emergent behaviour can occur (compensating for lack of knowledge). They also cost less, if the computational infrastructure is available.

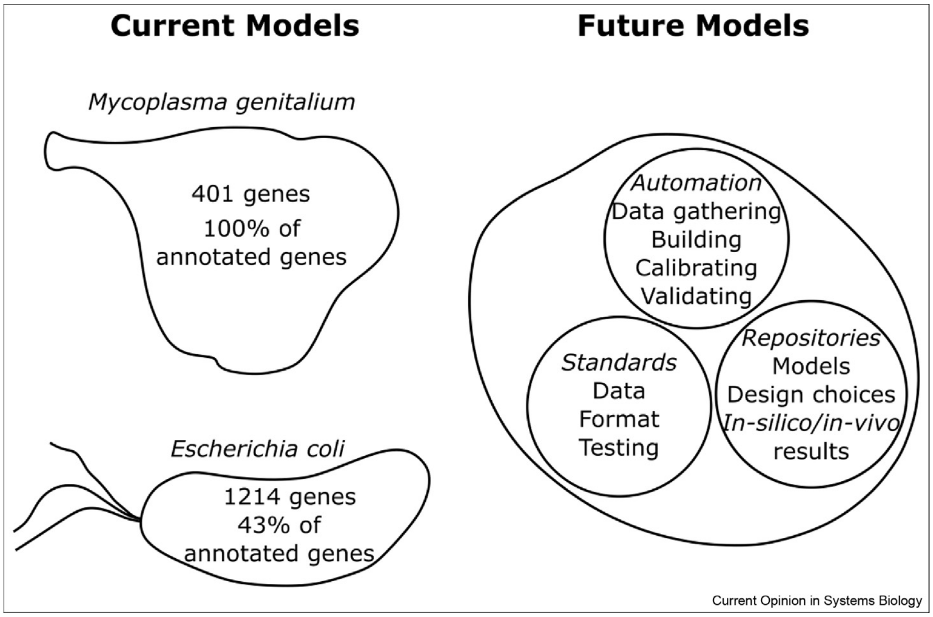
Whole-cell computational models

Whole-cell models (WCMs) are computational models that simulate the dynamics of an entire cell (Figure 1) [33,34]. They are capable of modelling individual molecules and interactions, integrating the fractured available knowledge and making it operational for designing genomes, and aim to include the function of all known gene products. Three WCMs have been published: *M. genitalium* [33], *E. coli* [34] and *Saccharomyces cerevisiae* [35], with WCMs for *Mycoplasma pneumoniae*, H1 human embryonic stem cell and an archetypal bacterium in development (Whole-Cell Modeling; URL: https://www.wholecell.org/models/). Of the bacterial WCMs, *M. genitalium* has 100% of well-annotated genes modelled, and while *E. coli* has only 43% of well-annotated genes modelled it can simulate multiple generations and 50 times the number of molecules [34]. The *M. genitalium* WCM has been used to compare model and real-world measurements [36], design genetic circuits in cellular context [37], analyse existing antibiotics against new targets [38], discover novel *in silico* minimal genomes [39] and test theoretical minimal genomes *in silico* [40]. The *E. coli* WCM was developed to conduct extensive comparison of *E. coli* data in the literature [34] and is the basis for colony simulations using WCMs (Vivarium; URL: https://wc-vivarium.readthedocs.io/). Box 1.

Box 1. Proposed strategies for enabling more predictive whole-cell models.

- Further studies and tool development to collect greater quality and quantity of cellular data for model inclusion and modification.
- Further development of data extraction and formatting tools and, eventually, development of automated model construction.
- Adoption of standards for models, data and testing.
- Use of standardised repositories for sharing models, data, i*n silico* and *in vivo* results, and *in silico* calibration efforts.

Figure 1



Current and future bacterial whole-cell models. Features of current bacterial whole-cell models and improvements that should be incorporated into future bacterial whole-cell models.

Despite progress, the current models are still not perfect representations of reality. Through necessity, the *M. genitalium* WCM is based on data from other species, and the single-generation-only simulations resulted in genome-scale designs that were highly unlikely to produce a second generation [39]. In addition, the *E. coli* WCM is based on data from multiple *E.coli* strains and needs improvements to its growth rate implementation to fully test gene edits *in silico*.

In silico genome-scale designs are currently untested and *in vivo* engineering may fail if gaps in the available knowledge are large enough (developers cannot validate models against missing laboratory data or model genes with unknown functions). The accuracy of *in silico* genome-scale designs would improve with greater quality and quantity of data; however, we currently lack the ability to measure (at a fine-grained level) the state of individual cells over time and over changing environmental conditions, instead relying on a combination of experimental estimation and inference [41]. As *in vivo* testing occurs, specific cellular processes, conditions and phenotypes will be highlighted for inclusion or modification in WCM models to enable more accurate and comprehensive predictions.

Because of the extensive effort required to build WCMs, new tools are needed to help researchers (detailed previously [41—43]). One of the biggest barriers to WCMs has been gathering and organising the data needed for building and calibrating models. The Karr lab recently developed Datanator [44], a database which integrates many types of WCM-required data. Further development of data extraction and reformatting tools will reduce development times of WCMs and will, eventually, enable automated model construction tools similar to ModelSEED [45] and CarveME [46] for metabolic models.

To foster WCM research, new standardised formats are needed to represent models and their semantic meaning, history and validation [47]. The adoption of data and model standards would enable modellers to communicate their design choices in greater detail (granularity level, format, units, assumptions [42]). To start, we encourage adoption of formats such as the Systems Biology Markup Language for representing models [48] (for further discussion see Waltemath et al. [49]).

To enhance WCM reproducibility, we encourage the sharing of model calibration and validation results more transparently and the unification of *in silico* predictions in a single platform for comparison with *in vivo* results. This pooling of development data would surface shared issues, highlighting requirements for new and improved WCM tools.

We also encourage adopting unit testing (e.g. SciUnit) for models, with the eventual goal of creating a testing framework explicitly for WCMs.

Producing these WCMs of higher quality and greater fidelity, using automated production, testing and validation, could create WCMs which are reusable, easily assessed and easily shared. This streamlined construction is required to increase the number and complexity of WCMs, especially for the successful development of a human WCM [43].

However, models by themselves are not enough. Currently, the largest academic supercomputer in the world is Frontera (the University of Texas) with 448,448 cores (Top 500 Computer Systems; URL: https://www.top500.org/lists/top500/2020/06/). Although this quantity of computational power can investigate many thousands more research avenues than *in vivo* research, brute forcing genome-scale design *in silico* is still infeasible. Improving simulation programs to simulate large hybrid models (i.e. WCMs) more quickly would help, but we would still require a method of selecting genome-scale designs to cover the largest range of possible designs (within CPU, time and data storage limitations).

Algorithms for designing genomes using WCMs

Algorithms (a series of steps that attempt to solve a problem) can search intelligently across all possible genome-scale designs. Minimal genome design algorithms need to recommend designs with the fewest number of genes, in the form of concrete genome sequences. Selecting genes for a genome design, mathematically, is a combinatorial optimisation problem, that is, there are multiple solutions, which require searching across all solutions to determine the optimum. This 'nonconvex' optimisation is a nondeterministic polynomial-time hard (NP-hard) problem, which very roughly equates to there is no known algorithm that can efficiently find the optimal solution [50,51].

Nevertheless, heuristics (approximate solution techniques) can be used to quickly find a solution with no guarantee that it will be the best one [50,52,53]. Metaheuristics are frameworks for developing these heuristic algorithms, although, confusingly, the literature does also refer to individual algorithms as metaheuristics [51]. For a pointed review of (meta)heuristic algorithms, refer the study by Sörensen (2015) [51]. There are two classes of (meta)heuristics: trajectory-based methods [50,53] and population-based methods [54,55]. For a review that compares (meta)heuristic algorithms conceptually, see Lones 2019 [55]. Box 2.

Box 2. Proposed future algorithms to apply to whole-cell models.

- Apply off-the-shelf trajectory-based and population-based (meta)heuristics to genome design problems.
- Develop *ad hoc* or hybridised algorithms and implement intermediary algorithms, for genome-scale design through interdisciplinary collaborations.
- Develop automatic software to assist in the selection of algorithm-produced designs.

There are three algorithms in the literature for designing minimal genomes: Minesweeper, Guess/Add/Mate Algorithm (GAMA) [39] and MinGenome [56] (for algorithms that predict individual gene essentiality but do not propose deletions, refer the study by Lin and Zhang [57] and Liu et al. [58]). MinGenome [56], by combining a genome-scale metabolic model with biological knowledge (i.e. gene location and essentiality, operons and promoter site information, and transcription factors), identified all unbroken stretches of nonessential genes in the *E. coli* MG1655 genome. These were compared with existing *E. coli* genome reductions, highlighting new deletions to test *in vivo*. Minesweeper is an *ad hoc*, nondeterministic, concurrent algorithm. It is inspired by laboratory fragment-cassette-fraction genome engineering [11], similar to divide and conquer, and reduces the genes under consideration for removal from 401 (all modelled genes [33]) to 152 (nonessential modelled genes [39]), making 2^{152} the number of possible minimal genome-scale designs within the *M. genitalium* WCM. It is nondeterministic because of the simulations of each stage running concurrently; as the order of gene deletions matters (because of contextual essentiality and redundant pathway selection [39]), different outcomes can result. GAMA is a stochastic, 'biased' genetic algorithm (see Table 1), which reduces the number of possible minimal genome-scale designs to 2^{152} by only considering nonessential genes initially, before later restoring to 2^{401}

Table 1

Existing optimisation algorithms with potential for genome design.

Trajectory-based (meta)heuristics	
Iterated local search/variable neighbourhood search	Improves on the current solutions by investigating close variants, identifying local optima quickly. Even the most basic trajectory-based (meta)heuristics can be competitive starting points [52].
Simulated annealing	Analogous to heat treatment of metal, the occasional selection of inferior solutions allows escape from local optima, in hope of finding global optima.
Taboo search	Uses short-term memory to prevent repetition of searches already conducted (i.e. prior searches are considered taboo).
Population-based (meta)heuristics	
Genetic algorithms	Use evolutionary concepts of random mutation, selection, recombination to produce a "population" of solutions, where selected "parent" solutions produce new "offspring".
Ant colony optimisation	Simulates ants in a colony using pheromone trails to guide future expeditions (i.e. future searches are informed by past searches).
Particle swarm optimisation	Computes a swarm/flock/school of individuals, with movement dictated by an individual's velocity and the movement of the rest of the swarm/flock/school.
Ad hoc (meta)heuristics	
Some researchers argue that the best results are obtained from *ad hoc* (meta)heuristics designed and tuned for each particular problem [50,52].	
Intermediary algorithms	
Approximation algorithms	Intermediaries ('heuristics with a guarantee') can guarantee to get within a certain percentage of the optimal solution [51]; they are commonly run alongside other (meta)heuristics to both guarantee a solution and have a chance to better it [50].
Hybridised algorithms	
Matheuristics [51]	Approximation algorithms hybridised with heuristic algorithms. Branch and bound (an approximation algorithm that divides the solution space) has been merged with evolutionary [65] and genetic [66,67] algorithms.
Memetic algorithms [68]	Merge population-based (meta)heuristics with iterated local search to exploit available knowledge about the problem.
Hybridising named algorithms	Examples include FOCuS [69], which merges a nature-inspired flower pollination algorithm and an immune system-inspired clonal selection algorithm, and GACOFBA [70] which merges three algorithms (genetic algorithm, ant colony optimisation and flux balance analysis). These combined algorithms have already been used for genome design based on genome-scale metabolic models [13].

by considering all modelled genes. The genome design suite tool (O Chalkley et al., bioRxiv https://doi.org/10.1101/681270) was developed to run the GAMA algorithm designed simulations in parallel across multiple supercomputers.

As Minesweeper is an *ad hoc* algorithm, it is not generalisable to other genome design problems; GAMA, even with solution space reductions and use of the genome design suite, is a genetic algorithm, which have prohibitive run times [50], meaning other algorithms could be better for genome-scale design (Table 1). We encourage interested researchers to apply off-the-shelf trajectory-based and population-based (meta)heuristics and fully endorse interdisciplinary collaborations with maths/computational researchers to aid in the development of *ad hoc* or hybridised algorithms and the implementation of intermediary algorithms.

Because the problem is very hard, algorithms alone might not be sufficient. Interactive graphical user interface (GUI) software, akin to Genome Dashboards [59] or BioCAD [60,61], could be used to assist researchers by automatically identifying good genome-scale designs and providing visual representations, enabling quick manual reviews and increasing the speed of the research.

Conclusions

The integration of WCMs with *ad hoc* and heuristic algorithms has advanced genome-scale design and can advance genome-scale engineering. Next steps for the field involve testing genome-scale designs *in vivo* (likely using CRISPR-cas9 homologous recombination techniques [62—64]). Constructing a (likely minimal) genome this way would integrate the *in silico* design and *in vivo* editing at a greater scale than seen before. This could change the methodology for future large-scale cellular research, move the field onto new genome design goals, increase momentum for improving the WCM ecosystem and integrate with laboratory automation. Coupling established CRISPR-cas9 gene editing with biological programming languages (such as Antha [URL: https://docs.antha.com/]) and laboratory automation tools, directed by *in silico* predictions, could revolutionise design——build——test cycles [41].

Future WCMs, especially the H1 human embryonic stem cell 1, have the potential to transform both science and medicine [42]. Knowledge gained from constructing genome-scale designs *in vivo* could be parlayed into constructing an optimal chassis for industrial metabolite production [4,13] and, later, the development of a novel cell, constructed from the 'best' components of individual bacterial species [14]. Further development of WCMs and genome-scale design algorithms will push us toward this future and, eventually, enable engineers to design and construct entire cells.

Funding

J.R-G. was supported by BrisSynBio, a BBSRC/Engineering and Physical Sciences Research Council Synthetic Biology Research Centre (BB/L01386X/1), with a funded PDRA position and research grant. J.R.K. was supported by NSF award 1935372. L.M. was supported by the Engineering and Physical Sciences Research Council (grant 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 a BrisSynBio (BB/L01386X/1) flexi-fund grant.

Conflict of interest statement

Nothing declared.

References

Papers of particular interest, published within the period of review, have been highlighted as:

* of special interest
** of outstanding interest

1. Brophy JAN, Voigt CA: **Principles of genetic circuit design**. *Nat Methods* 2014, **11**:508–520.

2. Dzieciol AJ, Mann S: **Designs for life: protocell models in the laboratory**. *Chem Soc Rev* 2012, **41**:79–85.

3. Calero P, Nikel PI: **Chasing bacterial chassis for metabolic engineering: a perspective review from classical to non-traditional microorganisms**. *Microb Biotechnol* 2019, **12**:98–124.

4. Garcia S, Trinh CT: **Multiobjective strain design: a framework for modular cell engineering**. *Metab Eng* 2019, **51**:110–120.

5. Garcia S, Trinh CT: **Modular design: implementing proven engineering principles in biotechnology**. *Biotechnol Adv* 2019, **37**:107403.

6. ** Fredens J, Wang K, de la Torre D, Funke LFH, Robertson WE, Christova Y, Chia T, Schmied WH, Dunkelmann DL, Beránek V, *et al.*: **Total synthesis of Escherichia coli with a recoded genome**. *Nature* 2019, **569**:514–518.
This study demonstrates the first insertion and assembly of a completely synthetic *E. coli* genome, with genome recoding conducted as part of the synthetic redesign, creating *E. coli* Syn61.

7. Richardson SM, Mitchell LA, Stracquadanio G, Yang K, Dymond JS, DiCarlo JE, Lee D, Huang CLV, Chandrasegaran S, Cai Y, *et al.*: **Design of a synthetic yeast genome**. *Science* 2017, **355**:1040–1044.

8. Ostrov N, Landon M, Guell M, Kuznetsov G, Teramoto J, Cervantes N, Zhou M, Singh K, Napolitano MG, Moosburner M, *et al.*: **Design, synthesis, and testing toward a 57-codon genome**. *Science* 2016, **353**:819–822.

9. Hirokawa Y, Kawano H, Tanaka-Masuda K, Nakamura N, Nakagawa A, Ito M, Mori H, Oshima T, Ogasawara N: **Genetic manipulations restored the growth fitness of reduced-genome Escherichia coli**. *J Biosci Bioeng* 2013, **116**:52–58.

10. Park MK, Lee SH, Yang KS, Jung S-C, Lee JH, Kim SC: **Enhancing recombinant protein production with an Escherichia coli host strain lacking insertion sequences**. *Appl Microbiol Biotechnol* 2014, **98**:6701–6713.

11. ** Hutchison CA, Chuang RY, Noskov VN, Assad-Garcia N, Deerinck TJ, Ellisman MH, Gill J, Kannan K, Karas BJ, Ma L,

et al.: **Design and synthesis of a minimal bacterial genome**. *Science* 2016, **351**. 1414–U73.
This study produced *JCVI-syn3.0*, a synthetic 50% reduction of the *Mycoplasma mycoides* genome, currently the smallest, free-living, dividing cell.

12. Glass JI, Merryman C, Wise KS, Hutchison 3rd CA, Smith HO: **Minimal cells-real and imagined**. *Cold Spring Harb Perspect Biol* 2017, **9**.

13. Landon S, Rees-Garbutt J, Marucci L, Grierson C: **Genome-driven cell engineering review: in vivo and in silico metabolic and genome engineering**. *Essays Biochem* 2019, **63**:267–284.

14. Vickers CE, Blank LM, Krömer JO: **Grand challenge commentary: chassis cells for industrial biochemical production**. *Nat Chem Biol* 2010, **6**:875–877.

15. ** Rancati G, Moffat J, Typas A, Pavelka N: **Emerging and evolving concepts in gene essentiality**. *Nat Rev Genet* 2018, **19**:34–49.
This review introduced the quantitative gene essentiality definitions of no essentiality, low essentiality, high essentiality, and complete essentiality, incorporating the environmental and genetic context for gene essentiality.

16. * O'Malley MA, Powell A, Davies JF, Calvert J: **Knowledge-making distinctions in synthetic biology**. *Bioessays* 2008, **30**: 57–65.
This essay investigates the fundamental themes and distinctions of synthetic biology and discusses the different schools/research themes it contains.

17. Mushegian A: **The minimal genome concept**. *Curr Opin Genet Dev* 1999, **9**:709–714.

18. Fraser CM, Gocayne JD, White O, Adams MD, Clayton RA, Fleischmann RD, Bult CJ, Kerlavage AR, Sutton G, Kelley JM, *et al.*: **The minimal gene complement of mycoplasma-genitalium**. *Science* 1995, **270**:397–403.

19. Iwadate Y, Honda H, Sato H, Hashimoto M, Kato J-I: **Oxidative stress sensitivity of engineered Escherichia coli cells with a reduced genome**. *FEMS Microbiol Lett* 2011, **322**:25–33.

20. Hashimoto M, Ichimura T, Mizoguchi H, Tanaka K, Fujimitsu K, Keyamura K, Ote T, Yamakawa T, Yamazaki Y, Mori H, *et al.*: **Cell size and nucleoid organization of engineered Escherichia coli cells with a reduced genome**. *Mol Microbiol* 2005, **55**: 137–149.

21. Posfai G, Plunkett G, Feher T, Frisch D, Keil GM, Umenhoffer K, Kolisnychenko V, Stahl B, Sharma SS, de Arruda M, *et al.*: **Emergent properties of reduced-genome Escherichia coli**. *Science* 2006, **312**:1044–1046.

22. Mizoguchi H, Sawano Y, Kato J, Mori H: **Superpositioning of deletions promotes growth of Escherichia coli with a reduced genome**. *DNA Res* 2008, **15**:277–284.

23. Ara K, Ozaki K, Nakamura K, Yamane K, Sekiguchi J, Ogasawara N: **Bacillus minimum genome factory: effective utilization of microbial genome information**. *Biotechnol Appl Biochem* 2007, **46**:169–178.

24. Morimoto T, Kadoya R, Endo K, Tohata M, Sawada K, Liu S, Ozawa T, Kodama T, Kakeshita H, Kageyama Y, *et al.*: **Enhanced recombinant protein productivity by genome reduction in Bacillus subtilis**. *DNA Res* 2008, **15**:73–81.

25. ** Reuß DR, Altenbuchner J, Mäder U, Rath H, Ischebeck T, Sappa PK, Thürmer A, Guérin C, Nicolas P, Steil L, *et al.*: **Large-scale reduction of the Bacillus subtilis genome: consequences for the transcriptional network, resource allocation, and metabolism**. *Genome Res* 2017, **27**:289–299.
This study details the planning and execution of producing reduced *B.subtilis* strains (~36%), followed by a multi-omics analysis of the resulting strains, demonstrating remarkably thorough research.

26. D'Elia MA, Pereira MP, Brown ED: **Are essential genes really essential?** *Trends Microbiol* 2009, **17**:433–438.

27. Baba T, Ara T, Hasegawa M, Takai Y, Okumura Y, Baba M, Datsenko KA, Tomita M, Wanner BL, Mori H: **Construction of Escherichia coli K-12 in-frame, single-gene knockout mutants: the Keio collection**. *Mol Syst Biol* 2006, **2**.

28. Butland G, Babu M, Diaz-Mejia JJ, Bohdana F, Phanse S, Gold B, Yang WH, Li J, Gagarinova AG, Pogoutse O, *et al.*: **eSGA: E. coli synthetic genetic array analysis**. *Nat Methods* 2008, **5**: 789–795.

29. Roberts TC, Morris KV: **Not so pseudo anymore: pseudogenes as therapeutic targets**. *Pharmacogenomics* 2013, **14**: 2023–2034.

30. de Crécy-Lagard V, Marck C, Brochier-Armanet C, Grosjean H: **Comparative RNomics and modomics in Mollicutes: prediction of gene function and evolutionary implications**. *IUBMB Life* 2007, **59**:634–658.

31. Li C, Zhang R, Wang J, Wilson LM, Yan Y: **Protein engineering for improving and diversifying natural product biosynthesis**. *Trends Biotechnol* 2020, **38**:729–744.

32. Xavier JC, Patil KR, Rocha I: **Systems biology perspectives on minimal and simpler cells**. *Microbiol Mol Biol Rev* 2014, **78**: 487–509.

33. ** Karr JR, Sanghvi JC, Macklin DN, Gutschow MV, Jacobs JM, Bolival Jr B, Assad-Garcia N, Glass JI, Covert MW: **A whole-cell computational model predicts phenotype from genotype**. *Cell* 2012, **150**:389–401.
This paper details the production of the first whole-cell model, the *M. genitalium* whole-cell model. This author (J.R-G) is of the opinion, depending on the future trajectory of whole-cell models (including the human whole-cell model), that this publication could eventually have the same level of impact as another paper published that year (Jinek et al. 2012 "A Programmable Dual-RNA–GuidedDNA Endonuclease in AdaptiveBacterial Immunity"), which resulted in CRISPR-cas9.

34. ** Macklin DN, Ahn-Horst TA, Choi H, Ruggero NA, Carrera J, Mason JC, Sun G, Agmon E, DeFelice MM, Maayan I, *et al.*: **Simultaneous cross-evaluation of heterogeneous E. coli datasets via mechanistic simulation**. *Science* 2020:369.
This paper details the production of the second whole-cell model produced, the *E. coli* whole-cell model. Although currently only accounting for 43% of the well annotated genes, it introduces multi-generational simulations and greater integration of species-specific data.

35. Münzner U, Klipp E, Krantz M: **A comprehensive, mechanistically detailed, and executable model of the cell division cycle in Saccharomyces cerevisiae**. *Nat Commun* 2019, **10**:1308.

36. Sanghvi JC, Regot S, Carrasco S, Karr JR, Gutschow MV, Bolival Jr B, Covert MW: **Accelerated discovery via a whole-cell model**. *Nat Methods* 2013, **10**:1192–1195.

37. Purcell O, Jain B, Karr JR, Covert MW, Lu TK: **Towards a whole-cell modeling approach for synthetic biology**. *Chaos* 2013, **23**, 025112.

38. Kazakiewicz D, Karr JR, Langner KM, Plewczynski D: **A combined systems and structural modeling approach repositions antibiotics for Mycoplasma genitalium**. *Comput Biol Chem* 2015:91–97. 59 Pt B.

39. * Rees-Garbutt J, Chalkley O, Landon S, Purcell O, Marucci L, Grierson C: **Designing minimal genomes using whole-cell models**. *Nat Commun* 2020, **11**:836.
This paper is the first application of algorithms to WCMs. Two minimal genome design algorithms, Minesweeper and GAMA, were used to find novel *in-silico* minimal genomes and identified contextual essential genes, simulating 4620 and 53,451 genome-scale designs, respectively.

40. Rees-Garbutt J, Rightmyer J, Chalkley O, Marucci L, Grierson C: *Testing theoretical minimal genomes using whole-cell models.* 2020, https://doi.org/10.1101/2020.03.26.010363.

41. Marucci L, Barberis M, Karr J, Ray O, Race PR, de Souza Andrade M, Grierson C, Hoffmann SA, Landon S, Rech E, *et al.*: *Computer-aided whole-cell design: taking a holistic approach by integrating synthetic with systems biology*. 2020. arXiv [q-bioQM].

42. Szigeti B, Roth YD, Sekar JAP, Goldberg AP, Pochiraju SC, Karr JR: **A blueprint for human whole-cell modeling**. *Curr Opin Struct Biol* 2018, **7**:8–15.

43. Goldberg AP, Szigeti B, Chew YH, Sekar JA, Roth YD, Karr JR: **Emerging whole-cell modeling principles and methods**. *Curr Opin Biotechnol* 2018, **51**:97–102.

44. Roth YD, Lian Z, Pochiraju S, Shaikh B, Karr JR: *Datanator: an integrated database of molecular data for quantitatively modeling cellular behavior*. 2020, https://doi.org/10.1101/2020.08.06.240051.

45. Devoid S, Overbeek R, DeJongh M, Vonstein V, Best AA, Henry C: **Automated genome annotation and metabolic model reconstruction in the SEED and Model SEED**. *Methods Mol Biol* 2013, **985**:17–45.

46. Machado D, Andrejev S, Tramontano M, Patil KR: **Fast automated reconstruction of genome-scale metabolic models for microbial species and communities**. *Nucleic Acids Res* 2018, **46**:7542–7553.

47. Medley JK, Goldberg AP, Karr JR: **Guidelines for reproducibly building and simulating systems biology models**. *IEEE Trans Biomed Eng* 2016, **63**:2015–2020.

48. Keating SM, Waltemath D, König M, Zhang F, Dräger A, Chaouiya C, Bergmann FT, Finney A, Gillespie CS, Helikar T, *et al.*: **SBML Level 3: an extensible format for the exchange and reuse of biological models**. *Mol Syst Biol* 2020, **16**, e9110.

49. Waltemath D, Karr JR, Bergmann FT, Chelliah V, Hucka M, Krantz M, Liebermeister W, Mendes P, Myers CJ, Pir P, *et al.*: **Toward community standards and software for whole-cell modeling**. *IEEE Trans Biomed Eng* 2016, **63**:2007–2014.

50. Skiena SS: *The algorithm design manual*. Springer Publishing Company; 2008.

51. Sörensen K: **Metaheuristics-the metaphor exposed**. *Intl Trans*
* *in Op Res* 2015, **22**:3–18.
This review is a very good introduction to a field, while also providing a laudably pointed criticism of its current state, making its point through very entertaining scientific writing (e.g. "the mid-2000s turned out to be a time so plentiful of "novel" metaphors as to make the Cambrian explosion pale in comparison").

52. Žerovnik J: **Heuristics for NP-hard optimization problems - simpler is better!?** *Logistics & Sustainable Transport* 2015, **6**: 1–10.

53. Gendreau M, Potvin JY: *Handbook of metaheuristics*. Boston, MA: Springer; 2010.

54. Phogat M, Kumar D: **A survey of Meta-heuristics Approaches for application in Genomic data**. *Int J Comput Sci Eng* 2017, **5**: 51–55.

55. Lones MA: **Mitigating metaphors: a comprehensible guide to**
* **recent nature-inspired algorithms**. *SN Computer Science* 2019, **1**:49.
This review, which is also a very good introduction to the field, argues the same criticisms as Sörensen (2015), but takes an opposing tack, outlining the uniformity of a variety of algorithms by translating their unique descriptions into plainer text and cross examining them thoroughly.

56. Wang L, Maranas CD: **MinGenome: an in silico top-down approach for the synthesis of minimized genomes**. *ACS Synth Biol* 2018, **7**:462–473.

57. Lin Y, Zhang RR: **Putative essential and core-essential genes in Mycoplasma genomes**. *Sci Rep* 2011, **1**:53.

58. Liu W, Fang L, Li M, Li S, Guo S, Luo R, Feng Z, Li B, Zhou Z, Shao G, *et al.*: **Comparative genomics of Mycoplasma: analysis of conserved essential genes and diversity of the pangenome**. *PloS One* 2012, **7**, e35698.

59. Li Z, Sun R, Bishop TC: **Genome Dashboards: framework and examples**. *Biophys J* 2020, **118**:2077–2085.

60. Patané A, Conca P, Carapezza G, Santoro A, Costanza J, Nicosia G: **Metabolic circuit design automation by multi-objective BioCAD**. In *Machine learning, optimization, and big data*. Springer International Publishing; 2016:30–44.

61. Oberortner E, Evans R, Meng X, Nath S, Plahar H, Simirenko L, Tarver A, Deutsch S, Hillson NJ, Cheng J-F: **An integrated computer-aided design and manufacturing workflow for synthetic biology**. In *DNA cloning and assembly: methods and protocols*. Edited by Chandran S, George KW, Springer US; 2020:3–18.

62. Jiang WY, Bikard D, Cox D, Zhang F, Marraffini LA: **RNA-guided editing of bacterial genomes using CRISPR-Cas systems**. *Nat Biotechnol* 2013, **31**:233–239.

63. Zerbini F, Zanella I, Fraccascia D, König E, Irene C, Frattini LF, Tomasi M, Fantappiè L, Ganfini L, Caproni E, *et al.*: **Large scale validation of an efficient CRISPR/Cas-based multi gene editing protocol in Escherichia coli**. *Microb Cell Factories* 2017, **16**:68.

64. Reisch CR, Prather KLJ: **The no-SCAR (Scarless Cas9 Assisted Recombineering) system for genome editing in Escherichia coli**. *Sci Rep* 2015, **5**:15096.

65. Cotta C, Troya JM: **Embedding branch and bound within evolutionary algorithms**. *Appl Intell* 2003, **18**:137–153.

66. French AP, Robinson AC, Wilson JM: **Using a hybrid genetic-algorithm/branch and bound approach to solve feasibility and optimization integer programming problems**. *J Heuristics* 2001, **7**:551–564.

67. Cotta C, Aldana JF, Nebro AJ, Troya JM: **Hybridizing genetic algorithms with branch and bound techniques for the resolution of the TSP**. In *Artificial neural nets and genetic algorithms*. Edited by Pearson DW, Steele NC, Albrecht RF, Vienna: Springer; 1995:277–280.

68. Norman MG, Moscato P: **A competitive and cooperative approach to complex combinatorial search**. In *Proceedings of the 20th informatics and operations research meeting*. Citeseer; 1991:3–15.

69. Mutturi S: **FOCuS: a metaheuristic algorithm for computing knockouts from genome-scale models for strain optimization**. *Mol Biosyst* 2017, https://doi.org/10.1039/c7mb00204a.

70. Salleh AHM, Mohamad MS, Deris S, Omatu S, Fdez-Riverola F, Corchado JM: **Gene knockout identification for metabolite production improvement using a hybrid of genetic ant colony optimization and flux balance analysis**. *Biotechnol Bioproc Eng* 2015, **20**:685–693.