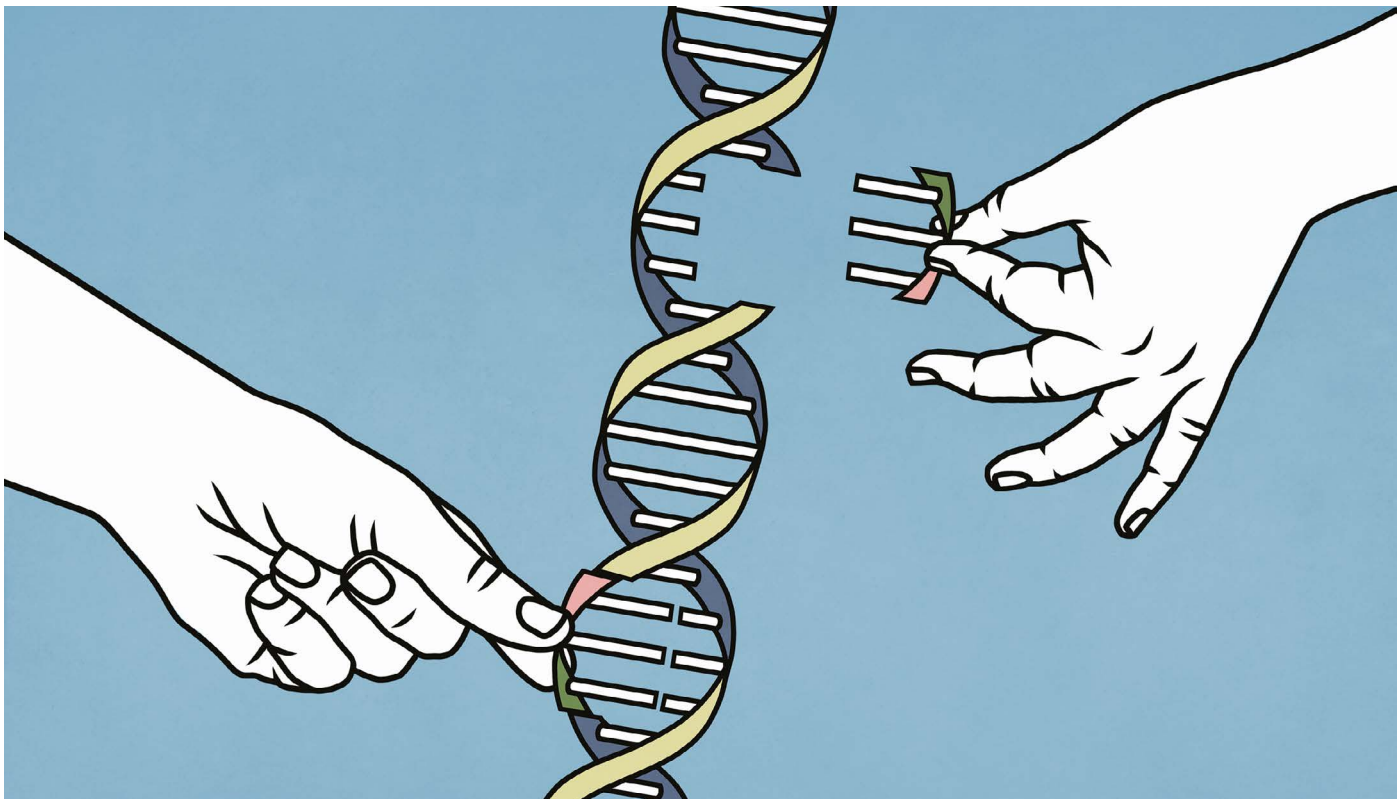


MALTE MUELLER/GETTY

WHY IS IT SO HARD TO REWRITE A GENOME?

Synthetic biologists can retool whole genomes, but the complexity of biological systems continues to surprise them. **By Michael Eisenstein**

When Patrick Yizhi Cai reflects on the state of synthetic genomics, he recalls the Big DNA Contest. Launched in 2004, the competition challenged synthetic biologists to design a novel, functional 40,000-base-pair DNA sequence that the contest sponsor, US DNA-synthesis firm Blue Heron Biotech (now Eurofins Genomics Blue Heron) would manufacture for free.

It was no small prize: at the time, producing this modest slab of DNA — less than 1% the length of the *Escherichia coli* genome — would have cost roughly US$250,000. The company's aim was to energize the then-nascent field of synthetic biology. "In the end, zero applications were received," says Cai, a synthetic biologist at the University of Manchester, UK. "That just tells you that even if you could make synthetic DNA for free, nobody really had enough imagination 20 years ago."

Today, steady progress in genomics and computational biology — not to mention DNA synthesis and assembly — have yielded multiple examples of what an ambitious, imaginative genome-writing effort can achieve. The synthetic bacterial strain JCVI-syn3A, developed at the J. Craig Venter Institute (JCVI) in La Jolla, California, is a streamlined version of *Mycoplasma mycoides* that survives and replicates despite having had several hundred non-essential genes stripped away[1]. Various groups are engineering *E. coli* strains in which the genetic code has been altered to enable production of proteins containing amino acids beyond the 20 typically observed in nature. And last year, the multinational Synthetic Yeast Genome Project (Sc2.0) completed construction of heavily engineered versions of every chromosome in the eukaryotic budding yeast, *Saccharomyces cerevisiae* — comprising some 12 million base pairs in all.

These efforts have been invaluable learning experiences, says Akos Nyerges, a synthetic-genomics researcher involved with the *E. coli* rewriting effort in George Church's lab at Harvard Medical School in Boston, Massachusetts. "You can mimic and test evolutionary steps which otherwise would have taken billions of years to evolve — or wouldn't have evolved ever," he says. But they have also laid bare how much we still don't understand about the fundamental language of the genome. Every genome-rewriting program so far has grappled with substantial and unexpected challenges, and the era of made-to-order genomes remains out of reach. When it comes to heavily modified genomes, says Nyerges, "we underestimated how complex biology is".

Back to basics

Most synthetic-genome projects are 'top-down' efforts that take a naturally occurring organism and pare away or redesign its DNA. That provides a valuable initial framework relative to 'bottom-up' approaches, in which the goal is to build a working genome from scratch. After all, explains Farren Isaacs, a genome engineer at Yale University in New Haven, Connecticut, when it comes to tinkering with genomes, the margin for error is surprisingly slim. "If you create an error in an essential gene, you're going to wipe the organism out."

A key goal of the JCVI and Sc2.0 projects was to determine which genes are truly

essential — a characteristic that is surprisingly hard to predict. John Glass, leader of the JCVI's synthetic-biology programme, says that when he and his team published[2] their 2016 report on their first minimal cell, nearly one-third of the cell's remaining genes (149 of 473) had no known function. "I'd say it's 78 now," he adds.

To determine which genes were necessary, both projects used random mutagenesis — basically, introducing untargeted perturbations throughout the genome and asking which ones the cells could tolerate and which ones severely undermined cellular viability.

But essentiality is a slippery concept, particularly given that most genomes contain redundancies and 'fail-safe' mechanisms to minimize the impact of individual mutations. Glass and his colleagues encountered dozens of instances in which mutagenesis revealed pairs of seemingly dissimilar genes that unexpectedly performed overlapping functions. As a result, there is no single minimal genome, he explains. "You take away one [gene], and with each choice, you're going down a different road to a slightly different minimal cell." Furthermore, many bacterial genes have multiple jobs, making it difficult to recognize which is the essential function. Glass cites the example of enolase, an enzyme with a well-known role in carbohydrate metabolism that also, it turns out, helps to degrade unwanted RNA.

Increasingly sophisticated computational 'whole-cell models' could help to remove some of the guesswork from future genome-trimming efforts. In 2020, mathematician Lucia Marucci and synthetic biologist Claire Grierson, both at the University of Bristol, UK, led an effort to simulate genome-reduction strategies in a whole-cell model of *Mycoplasma genitalium* — a close relative of the microorganism edited by the JCVI[3]. Their analysis, which used elaborate models of cellular processes and their interactions, suggested two redesigns with distinct sets of genes deleted, each yielding genomes that were roughly 40% smaller than the natural *M. genitalium* genome.

More recently, Marucci and Grierson have begun working with sophisticated whole-cell models of *E. coli*. As described in a 2024 preprint[4], their current efforts combine mechanistic models with machine learning to predict the consequences of genome manipulation across a broad range of biological functions. These are described by thousands of interlinked equations, yielding blueprints for bacteria that have 40% fewer genes than wild-type *E. coli*. "We now have a bunch of minimized reduced genomes that we want to test in the laboratory," Marucci says.

Find and replace

Rather than making abridged editions of the genome, other groups have set out to subtly reword the genetic text — encountering an entirely different set of challenges.

Protein-coding sequences are built of nucleotide triplets known as codons. With 61 possible codons for the 20 naturally occurring amino acids as well as 3 'stop' codons that terminate protein synthesis, there is considerable redundancy in the resulting code. Various teams have shown that, by comprehensively converting each instance of a given codon to one of its 'synonyms', one can repurpose that codon. This month, for example, Isaacs and his colleagues described an *E. coli* strain called Ochre in which two stop codons were reassigned to direct the incorporation of the non-natural amino acids *para*-acetyl-L-phenylalanine and N^{ε}-Boc-L-lysine[5]. These amino acids have chemical properties and functions that don't exist in nature, but recoding can also serve as a 'firewall' that prevents the interaction and exchange of genetic material with other organisms in natural environments.

"With each choice, you're going down a different road to a slightly different minimal cell."

Such work might sound straightforward — simply substituting one codon for another — but genome recoding requires much planning and effort. After researchers have found all instances of the codon they wish to eliminate, they must then work out how to replace it without disrupting the affected genes or regulatory machinery. Bacterial genes often contain regulatory sequences in the protein-coding sequence, Nyerges points out, and a gene on one DNA strand can overlap with a gene on the opposite strand. Seemingly minor changes can thus have major, unexpected consequences.

Nyerges, Church and their colleagues are grappling with this challenge at an unprecedented scale as they finalize a heavily recoded variant of *E. coli* that uses only 57 of the 61 naturally occurring amino-acid codons[6]. This effort has entailed more than 73,000 changes to the strain's 4-megabase genome, which inevitably creates unintended effects. "Some things will happen readily with no impact on growth or fitness, while others have a striking impact," says Nyerges. Some changes disabled existing regulatory elements or unwittingly created new one; others established new protein-coding sequences. "And we're only learning about this as we go."

Sorting out these issues is a substantial undertaking in its own right. For example, throughout the recoding process for their Ochre strain, Isaacs and his team used extensive 'multi-omic' analyses to characterize the bacterium. "We collected metabolic profiling data under different [culture] conditions," he says. "We also collected proteomics data comparing the recoded cell to a few different progenitors, including wild-type cells." In this way, they systematically tweaked the genome until the cells were able to grow under standard culture conditions at roughly the same rate as unmodified bacteria — a non-trivial result, given that genome recoding often impairs growth. Nyerges and his colleagues likewise turned to multi-omics to troubleshoot their 57-codon genome. They also used an experimental strategy that spurs rapid evolution of bacteria in culture, to promote the selection of genome mutations that improve fitness.

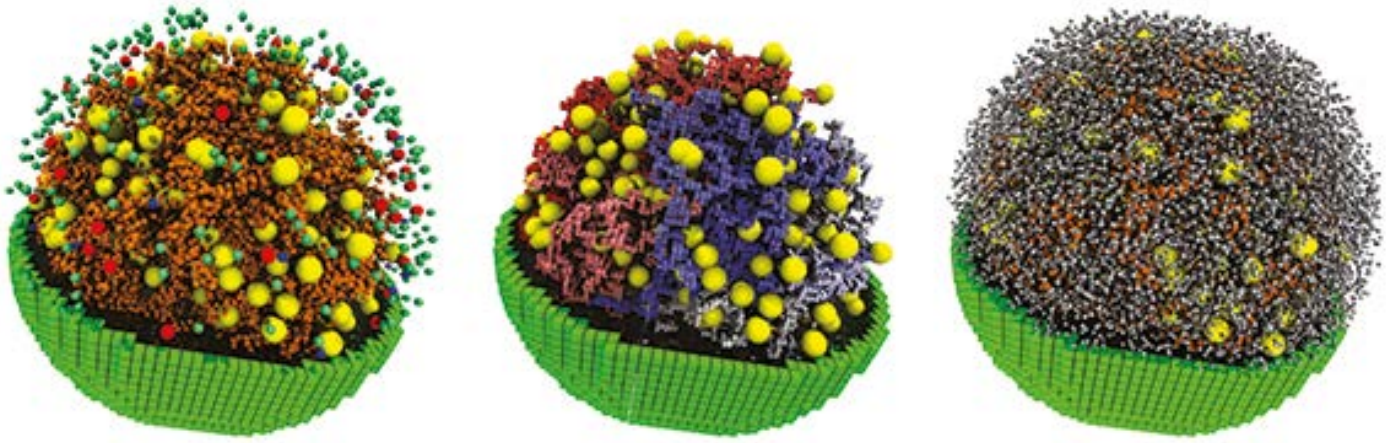
Algorithmic tools are also helping researchers to model and predict the outcomes of some genome-rewriting experiments in advance. For instance, synthetic biologist Howard Salis's team at Pennsylvania State University in University Park uses quantitative data from high-throughput screens of both genetically modified cells and strands of synthetic DNA to develop algorithms that can define, characterize and even design sequences that govern processes such as transcription and translation. "A typical paper for us nowadays is anywhere from 10,000 to 100,000 different defined, designed experiments," says Salis. The results are used to extract testable physical principles that allow the algorithms to predict, for example, how changes to a gene's promoter sequence alter downstream expression.

"You can ground-truth everything," says Salis. "And we can combine our existing models to design the next experiments to understand the remaining misunderstood stuff." Indeed, the Church lab has used several of Salis's tools to design its 57-codon microbe. Nyerges says such algorithms have been a substantial asset — albeit not enough to prevent considerable troubleshooting. "Even very tiny changes can cumulatively lead to significant fitness problems once you add up thousands of genes in a genome," he says.

Brewing progress in eukaryotes

Small and self-contained, bacterial genomes are ideal test beds for developing synthetic genomics tools. But the Sc2.0 team's remarkable progress shows that similar feats can be achieved in eukaryotes.

Unlike *E. coli*, which has a single circular chromosome of roughly 5 million base pairs, the *S. cerevisiae* genome encompasses more than 12 million bases across 16 linear chromosomes. Since 2011, the Sc2.0 team, led by geneticist Jef Boeke at New York University, has systematically redesigned, constructed and debugged synthetic versions of all those chromosomes. Its goals include recoding the genome to liberate one of three stop codons for alternative use; deleting transposons and other mobile elements; and moving all genes encoding transfer RNAs to a 17th 'neochromosome'.



LABORATORY OF ZAIDA LUTHEY-SCHULTEN

Computer-generated models of the synthetic minimal cell created by researchers at the J. Craig Venter Institute in La Jolla, California.

In addition, the Sc2.0 project used a system called SCRaMbLE, in which yeast genes deemed non-essential are flanked by DNA sequences that allow them to be snipped out and rearranged by enzymes. SCRaMbLE enabled the team to generate and test chromosome variants containing different gene deletions and structural rearrangements, providing a platform for fitness testing. "In engineering, it's very difficult to imagine building a billion airplanes and trying to fly them to see which one does not crash," says Cai. But the researchers were able to do just that with yeast, systematically probing how far the genome could be tweaked before it broke.

Some engineering tasks were simpler in yeast than in bacteria — for example, yeast genomes have less genetic crowding. "We haven't seen evidence for overlapping genes or promoters embedded within genes," says Boeke. Still, Cai estimates that two-thirds of the team's effort was focused on debugging rather than construction, and surprises were commonplace. Boeke says many challenges arose from poor annotation of the yeast genome that the team initially used for their design efforts. "There was at least one chromosome that had a lot of errors in it," he says. There were also several cases in which using SCRaMbLE to delete non-essential genes unwittingly disrupted the function or regulation of other, nearby genes with more essential roles in the cell.

The team fine-tuned its designs using an iterative process of design–build–test cycles. "We used recombination to allow us to replace the wild-type sequence step by step," says Yue Shen, chief scientist of synthetic biology at BGI Research in Shenzhen, China, whose lab group worked on three of the yeast chromosomes for Sc2.0. This allowed the researchers to assess the specific impact of each stretch of recoded chromosome. In parallel, Sc2.0 researchers used a multi-omics strategy similar to that used in bacteria to diagnose and correct issues of cell viability and health.

But Sc2.0 is also grappling with combinatorial effects that emerge only when multiple rewritten chromosomes are introduced into a cell at the same time. In 2023, Boeke and his colleagues described the diagnosis and repair of one such system bug, arising from unexpected incompatibility between synthetic chromosomes III and X — a modified gene on one chromosome impaired translation of an essential gene on the other[7]. "We hope that there aren't too many more like that," says Boeke. So far, the team has combined 7.5 synthetic chromosomes in a single cell — representing more than 50% of the yeast genome — and Boeke hopes to complete the assembly process for all 17 chromosomes in the next 6 months.

At a crossroads

Today, the synthetic-genomics field is at a crossroads. Whereas many researchers plan to dig deeper into their model organisms of choice, others are eyeing new terrain. Cai's group aspires to redesign human and potato chromosomes, for example, and some groups are contemplating opportunities for true bottom-up genome design. Salis sees this as an exciting opportunity to develop optimized organisms for biotechnology purposes, enabling much greater complexity than would be possible by just tinkering with existing genomes. "You can basically take the best of the best of what you want — and importantly, you know exactly what you put in," he says.

But progress will require solutions to some pressing challenges. For one, the cost of large-scale precision DNA synthesis remains high. Cai estimates that commercially synthesized DNA building blocks of up to 10 kilobases might cost roughly $0.10 per base to produce. But many eukaryotic genomes contain repetitive elements that are hard to synthesize, and Cai says that the assembly of all the components "can easily double the cost of the starting material". This is why one of Shen's initiatives at BGI Research aims to develop scalable solutions for efficient production of genome building blocks[8]. "We are hoping maybe that for the next synthetic yeast genome, we will be able to finish that in 2 or 3 years instead of 12," she says.

Sophisticated design algorithms with greater predictive power could cut these costs by generating more accurate genomic blueprints that streamline the testing and optimization process. For instance, several groups have demonstrated the potential of generative artificial intelligence (AI) for constructing functional DNA molecules that are based on patterns learnt from vast training sets of sequence data. But Salis is wary of leaning too heavily on AI: "It's not science any more — it's literally a black box." Instead, he hopes to see progress in building mechanistic machine-learning models that are trained on well-defined, carefully annotated experiments. But this is a slow and costly process, and Salis estimates that models for complex eukaryotic genomes are "probably about 25 years behind" equivalent microbial models.

Still, opportunity abounds. Cai compares the state-of-the-art in synthetic biology to early forays into computer coding. "In the early days, you would just try to write an app and compile and hope that there's no error," he says. "But I think once you get past that first stage of 'hello world', the next phase will be much more intention-driven design."

Michael Eisenstein is a science writer in Philadelphia, Pennsylvania.

1. Breuer, M. *et al. eLife* **8**, e36842 (2019).
2. Hutchison, C. A. *et al. Science* **351**, aad6253 (2016).
3. Rees-Garbutt, J. *et al. Nature Commun.* **11**, 836 (2020).
4. Gherman, I. M. *et al*. Preprint at bioRxiv https://doi.org/10.1101/2023.10.30.564402 (2024).
5. Grome, M. W. *et al. Nature* https://doi.org/10.1038/s41586-024-08501-x (2025).
6. Nyerges, A. *et al*. Preprint at bioRxiv https://doi.org/10.1101/2024.06.16.599206 (2024).
7. Zhao, Y. *et al. Cell* **186**, 5220–5236 (2023).
8. Zhang, X. *et al*. Preprint at bioRxiv https://doi.org/10.1101/2024.10.30.619547 (2024).