SCIENCE GALLEY

SYNTHETIC BIOLOGY

Technological Challenges and Opportunities for Writing Genomes

Insert Deck Here

By **Nili Ostrov [1], Jacob Beal [2], Tom Ellis [3], D. Benjamin Gordon [4], Bogumil J. Karas [5], Henry H. Lee [1], Scott C. Lenaghan [6], Jeffery A. Schloss [7*], Giovanni Stracquadanio [8], Axel Trefzer [9], Joel S. Bader [10], George M. Church [1,11], Cintia M. Coelho [12], J. William Efcavitch [13], Marc Güell [14], Leslie A. Mitchell [15], Alec A.K. Nielsen [16], Bill Peck [17], Alexander C. Smith [18], C. Neal Stewart, Jr. [19], Hille Tekotte [20]**

See supplementary material for author affiliations
Correspondence to Jeffery A. Schloss, schlossjeff500@gmail.com

Engineering biology with recombinant DNA, broadly called synthetic biology, has progressed tremendously in the last decade, owing to continued industrialization of DNA synthesis, discovery and development of molecular tools and organisms, and increasingly sophisticated modeling and analytic tools. However, we have yet to understand the full potential of engineering biology due to our inability to write and test whole genomes, which we call synthetic genomics. Dramatic improvements are needed to reduce the cost and increase the speed and reliability of genetic tools.

Here, we identify emerging technologies and improvements to existing methods that will be needed in four major areas to advance synthetic genomes within the next 10 years: Genome design, DNA synthesis, Genome editing, and Chromosome construction (Table 1). Similar to other large-scale projects for the responsible advancement of innovative technologies, such as the Human Genome Project, an international, cross-disciplinary effort consisting of public and private entities will likely yield maximal return-on-investment and open new avenues of research and biotechnology.

The ability to readily design and write the genomes of living cells provides unique opportunities to tackle problems that are intractable with current technologies. These transformative technologies will undoubtedly have widespread scientific, social and economic impacts, thus their development and adoption requires proactive identification of potential pitfalls through ongoing public discussion. Accordingly, the outlook presented here is part of a broader effort by the international Genome Project-write (GP-write) consortium *(1)* to encourage inclusive conversation among scientists, lawyers, ethicists, educators, environmentalists, other experts and stakeholders as well as the general public and ensure responsible, safe, and coordinated implementation of these exciting new technologies.

Synthetic genomics is a relatively new field and the majority of writing technologies (with the exception of commercial DNA synthesis) are still developed by academic research laboratories. Thus, unlike the reasonably predictable progress of engineering in more established industries, such as semiconductors, prediction of timelines and costs in these nascent fields remains highly speculative (Table 1).

GENOME DESIGN

Genome design aims to encode higher-level design criteria into precise DNA sequences at the chromosome scale. This will require computer-aided design (CAD) technologies to (i) reliably produce desired phenotypes, (ii) maximize the impact of the design, both in terms of experimental feedback and technical feasibility, and (iii) facilitate collaboration by employing standards for handling and exchanging design information.

Current synthetic genomic CAD software, such as used for Sc2.0 *(2)*, employs automation and collaboration tools for scaling DNA design from plasmids to entire chromosomes. Currently, estimation of the functional effects of edits (e.g., gene deletions) are left to humans experts. Looking forward, these tools will need to include capabilities to accurately predict the viability and phenotype of a cell from its genome design.

Although simple models are sufficient to handle silent edits (e.g., synonymous editing of coding sequences for watermarking), models of increasing complexity and precision will be necessary to predict how changes to genome sequence impact gene regulation and protein function. While comprehensive mechanistic models of higher eukaryotes are likely decades away, machine learning approaches could improve phenotype prediction by leveraging the wealth of high-dimensional, high-throughput systems biology data in public databases and generated from genome writing projects, similar to how AI techniques have been recently used to predict protein structures.

Leveraging these models, there is also a need for new tools for experimental design to minimize the number of expensive iterations required to obtain successful genome designs. For instance, successful redesign of the relatively small *Mycoplasma* genome required four iterations, as well as genomic screens to identify essential genes *(3)*. Larger projects will require many more intermediate stages of construction and supporting experiments. To provide the most valuable feedback and leverage the resulting data in subsequent designs, new algorithms are required to automate

design of experiments and selection of appropriate engineering technologies for their implementation (e.g., “write vs. edit”).

A related, unaddressed pragmatic need is to ensure compatibility of designed DNA with constraints from downstream synthesis, assembly, delivery, and analytical stages. For example, software-guided parsing of chromosome sequence into synthesis-compatible fragments, or introduction of designated sequences to facilitate assembly and delivery. Tools will also need to be adaptable enough to anticipate continuing advances in writing technologies.

All of these efforts rely on and generate large datasets, which require proper stewardship to streamline model definition and facilitate sharing experimental results. Two major barriers for integrating biological information are data incompatibility and a lack of sufficiently descriptive metadata. We encourage researchers to use widely-adopted data-exchange formats, e.g., genomics formats such as GenBank and GFF, and to continue to establish and adhere to standards for experimental metadata such as the Synthetic Biology Open Language (SBOL)..

Finally, the emerging genome engineering community will involve large teams with diverse expertise, working on a multitude of different projects. Funding agencies and industrial stakeholders should recognize and prioritize support for long-term development of specialized software and standard data formats, including collaboration, visualization, and quality control capabilities, similar to initiatives already in place for genomics, such as the Wellcome Trust Open Research fund and the Chan-Zuckerberg BioHub. We anticipate that synthetic genomics software will not only enhance our ability to plan and execute large-scale genetic projects, but will also lead to fundamental methodological advances to move from correlation to causation of genotype-phenotype relationships.

DNA SYNTHESIS

Genome writing projects depend heavily on the availability of large numbers of long (>5,000 bp) and highly-precise synthetic DNA constructs *(4, 5)*. However, chemical synthesis of DNA remains limited to production of short oligonucleotides (oligos), commonly 200 bp long. While oligos have driven substantial advances in recombinant DNA technologies to date, larger DNA constructs require assembly of multiple oligonucleotides, a process which is laborious and lossy. Therefore, routine production of long, precise fragments of synthetic DNA would be desirable for chromosome-scale engineering.

While DNA has become more available through commercial vendors in recent years due to industrialization by parallelization and miniaturization, there has been little improvement to the underlying phosphoramidite chemistry which limits DNA length, production speed, and cost. Accordingly, construction of whole chromosomes remains prohibitively expensive and time-consuming. For example, array synthesis of oligonucleotides costs approximately $5E-4 per nucleotide, yielding an estimate of $1.5M just for synthesizing 3 gigabases of DNA - the size of a human genome. Radical new approaches to DNA assembly, purification and synthesis processes are thus required to achieve substantial advancement on cost and ease.

Innovations to minimize or eliminate the need for assembly, error correction, and cloning of DNA fragments assembled from oligos could boost productivity of current DNA synthesis infrastructure. To increase the yield of perfect sequences, currently ranging from 5 to 60% *(6)*, hybridization and error-correction can benefit from the engineering of high-fidelity polymerases and ligases. These advancements, largely driven by industry, will decrease operating costs and production time. Cloning efficiency can also be enhanced by harnessing hosts with rapid division and/or high recombination rates, or altogether obviated using cell-free cloning techniques and artificial cells. These emerging technologies will require further fundamental research before they reach commercial readiness.

New technologies capable of synthesizing high-quality long DNA fragments would fundamentally alter the process of chromosome-scale engineering. Recently, methods for rapid production of short sequence-defined single-stranded DNA (ssDNA) have been reported using the template-independent DNA polymerase TdT (terminal deoxynucleotidyl transferase) *(7, 8)*. TdT offers potential for directly synthesizing multi-kilobase sequences with increased polymerization rate and higher coupling efficiencies. To compete with existing phosphoramidite chemistry, enzymatic synthesis approaches should be further developed to address complex sequences and achieve precise, high-quality DNA in an automated and cost-effective fashion. These efforts, initiated by startup companies, will benefit greatly from continued investment in fundamental research to elucidate the molecular mechanisms of enzymatic terminal transferase reactions.

Overall, genome-scale projects will substantially increase the demand for DNA synthesis, requiring a larger and more robust infrastructure. To support the scale and quality of required DNA, enhancement of electro-mechanical systems as well as novel biological tools are needed. Continued increase in throughput may be achieved by further parallelization and miniaturization, for example by semiconductor fabrication or droplet-based techniques. Increases in DNA quality and production speed will be influenced by utilization of biological tools such as enzymes and organisms.

GENOME EDITING

In recent years, powerful new DNA editing tools have lowered the technical barriers for performing highly precise genetic and epigenetic modifications. Multiplexed editing of an intact genome could dramatically decrease the time and labor required to generate a large number of modifications and, in some instances, circumvent the need for *de novo* synthesis and chromosome assembly.

Yet despite remarkable success using programmable nucleases such as Cas9, TALEN, and ZFN in multiple cell types with exquisite temporal and tissue-specific control, genome-scale editing remains limited. Large scale editing requires simultaneous DNA alterations at multiple genomic loci in the same cell. A localized, nuclease-induced double strand break can be used to increase editing efficiency at each locus, but multiple simultaneous breaks often cause cellular toxicity. To avert toxicity, ‘Base

editor' enzymes were engineered in which the nuclease is replaced by enzymatic base modification *(9)*, achieving simultaneous editing of over 13,000 ALU repetitive elements in human cells using a small number of guides *(10)*. Other engineered Cas9 tools are used for repression, activation, or targeted insertion of DNA. However, a major bottleneck for multiplex genome-wide editing remains the delivery of gRNAs, as multiple unique genome changes necessitate the presence of multiple unique guide RNAs in the same cell.

In light of the rapid pace of innovations and significant investment in editing technologies in recent years, we anticipate this barrier, as well as off-target mutagenesis and constraints paused by sequence-specificity of editing enzymes (e.g. PAM sequence requirement) will be overcome to enable routine multiplex editing.

Genome-scale editing can also be accomplished by oligo recombineering *(11)*, a technique that relies on homologous recombination (HR) and has reduced *in vivo* toxicity. However, this technique is currently limited to a handful of organisms where high-efficiency HR can be catalyzed by a recombinase that uses a donor ssDNA for targeting. To enable recombineering to edit plants and mammalian cells, new recombinase enzymes must be discovered or designed. It may also be necessary to map and modulate an organism's repair pathways to improve HR.

Finally, comprehensive suites of molecular tools should be generated to accelerate testing and optimization of genome editing. For example, a complete set of programmable TALEN or ZFN nucleases can be generated for targeting all UAG stop codons in human cells. Similarly, CRISPR-Cas9 guide libraries targeting all PAMs can be used to explore multiplexed, allele-specific targeting in plant, human, or fungal cells. Efforts to generate these genome-wide resources will provide experimental evidence of 'accessibility maps' that reflect editing efficiency variability across genomic targets. Such data will optimize the choice of target sequences, inform predictive computational models, and deepen our knowledge of underlying chromosome structure, folding, and repair.

Chromosome Construction

The most critical hurdle facing synthetic genomics is the assembly and introduction of synthetic chromosomes into host cells. How does one stitch together all the DNA pieces required to construct a fully functional chromosome? Once constructed, how can we control chromosome localization and architecture to ensure cell viability? How do we replace all chromosome copies in polyploid organisms? As the genomes of most free living organisms are larger than 2 Mb, methods for routine manipulation of large DNA fragments are critically needed.

Despite recent improvements in DNA synthesis and *in vitro* cloning techniques, such methods are not efficient for construction of entire chromosomes. Higher-order assembly of chromosomes at least 1 Mb in length can be performed by *in vivo* homologous recombination (HR) in the yeast *Saccharomyces cerevisiae*, a robust technique used in all synthetic chromosomes reported to date, including viral, bacterial, yeast and algal chromosomes as well as fragments of mice and human genomes *(12, 13)*.

The efficiency of DNA assembly in *S. cerevisiae* has not been found in other genetically tractable organisms. To expand the toolkit for writing specialized synthetic chromosomes that are difficult to assemble in *S. cerevisiae,* new HR-proficient cloning organisms should be developed that tolerate high GC, direct repeats, and desired post-transcriptional modifications. In addition, organisms compatible with extreme environments such as desert, deep ocean, or space travel may provide new routes for DNA assembly, such as the one found in the polyextremophile *Deinococcus radiodurans*.

Once constructed, chromosome delivery and manipulation becomes the primary engineering bottleneck in the majority of desired hosts. To deliver megabase-scale constructs, robust, high-throughput DNA transformation methods must be developed that work in a variety of organisms spanning genera. For example, breakthroughs in DNA delivery can revolutionize plant engineering, currently hindered by species-specific, labor intensive transformation methods and limited by traditionally conservative funding. High-risk, high-reward funding to support modernization of plant research, such as development of tissue-culture-independent DNA delivery methods, are pertinent for synthetic biology-enabled improvements of agricultural organisms.

Automation of highly specialized methods for chromosome transfer between yeast, bacteria, plants, and mammalian cells (such as cell fusion, genome transplantation, or microinjection) requires multidisciplinary funding opportunities aimed at bridging microfluidics with traditional cell and molecular biology work. To explore innovative solutions, it is essential that early proof-of-concept efforts be supported at government and foundation levels.

Finally, many of the cellular forces that shape genome structure and function remain largely unknown. Fundamental studies are needed to elucidate the mechanisms by which sequence and epigenetic regulation guide inter- and intra-chromosomal interactions and determine genome architecture. Emerging technologies for programmable modifications of chromatin, such as insulators guiding chromatin remodelling, safe harbor sites for DNA insertion, and orthogonal recombinase enzymes, will be necessary for developing gene therapies *(14)*. Better understanding of organelle genomes (plastid, mitochondrion), which remain extremely difficult to engineer, would offer new routes for stable maintenance and incorporation of artificial chromosomes.

A final challenge when introducing synthetic constructs is whether they perform as desired in the destination cell. A key factor for all large-scale genome engineering, particularly of mammalian and plant systems, is how quickly success or failure can be determined. Whole genome DNA and RNA sequencing will no doubt serve as a first-pass verification of chromosome integrity. In addition, specifically tailored cell lines with appropriate phenotypic reporters may be developed for assessing the performance of large synthetic constructs. Finally, reliable organoid models and a clear understanding of the regulation and expression changes that drive organismal development will be key to

extrapolating results from single cells to the design of functional chromosomes for multicellular organisms.

Innovation and funding sources for chromosome construction technologies are thus broadly applicable across bioscience institutions. Innovation will be driven by government grants as well as powerhouse genomic and cancer institutes, with a growing role for BioFoundries, emerging hubs for automation of bioengineering. By curating collaborative efforts across institutions and disciplines, GP-write consortium can accelerate the development of technologies related to synthetic genomics.

A GLOBAL, MULTIDISCIPLINARY EFFORT

New technologies may come from efforts in synthetic biochemistry, such as programmable synthetic protocells, and from progress at the interface of hardware and wetware such as solid-phase DNA assembly platforms. Or as is often the case, they may emerge from findings in basic bioscience research, for example, by uncovering valuable new enzymes or delivery systems. It is thus important to direct innovation to broadly enable genome writing across all domains of life.

Synthetic genomes are a maturing technology that has the potential for far reaching impact on society. A coordinated global endeavour would undoubtedly drive innovation in genome writing. A highly interdisciplinary, multinational effort from government and private sectors will help achieve and disseminate these advances to make an impact in biomedical, pharmaceutical, agricultural, and chemical industries.

References and Notes

1. J. D. Boeke, G. Church, A. Hessel, N. J. Kelley, A. Arkin, Y. Cai, R. Carlson, A. Chakravarti, V. W. Cornish, L. Holt, F. J. Isaacs, T. Kuiken, M. Lajoie, T. Lessor, J. Lunshof, M. T. Maurano, L. A. Mitchell, J. Rine, S. Rosser, N. E. Sanjana, P. A. Silver, D. Valle, H. Wang, J. C. Way, L. Yang, GENOME ENGINEERING. The Genome Project-Write, *Science* **353**, 126–127 (2016).

2. S. M. Richardson, L. A. Mitchell, G. Stracquadanio, K. Yang, J. S. Dymond, J. E. DiCarlo, D. Lee, C. L. V. Huang, S. Chandrasegaran, Y. Cai, J. D. Boeke, J. S. Bader, Design of a synthetic yeast genome, *Science* **355**, 1040–1044 (2017).

3. C. A. Hutchison 3rd, R.-Y. Chuang, V. N. Noskov, N. Assad-Garcia, T. J. Deerinck, M. H. Ellisman, J. Gill, K. Kannan, B. J. Karas, L. Ma, J. F. Pelletier, Z.-Q. Qi, R. A. Richter, E. A. Strychalski, L. Sun, Y. Suzuki, B. Tsvetanova, K. S. Wise, H. O. Smith, J. I. Glass, C. Merryman, D. G. Gibson, J. C. Venter, Design and synthesis of a minimal bacterial genome, *Science* **351**, aad6253 (2016).

4. J. Fredens, K. Wang, D. de la Torre, L. F. H. Funke, W. E. Robertson, Y. Christova, T. Chia, W. H. Schmied, D. L. Dunkelmann, V. Beránek, C. Uttamapinant, A. G. Llamazares, T. S. Elliott, J. W. Chin, Total synthesis of Escherichia coli with a recoded genome, *Nature* **569**, 514–518 (2019).

5. N. Ostrov, M. Landon, M. Guell, G. Kuznetsov, J. Teramoto, N. Cervantes, M. Zhou, K. Singh, M. G. Napolitano, M. Moosburner, E. Shrock, B. W. Pruitt, N. Conway, D. B. Goodman, C. L. Gardner, G. Tyree, A. Gonzales, B. L. Wanner, J. E. Norville, M. J. Lajoie, G. M. Church, Design, synthesis, and testing toward a 57-codon genome, *Science* **353**, 819–822 (2016).

6. N. B. Lubock, D. Zhang, A. M. Sidore, G. M. Church, S. Kosuri, A systematic comparison of error correction enzymes by next-generation sequencing, *Nucleic Acids Res.* **45**, 9206–9217 (2017).

7. S. Palluk, D. H. Arlow, T. de Rond, S. Barthel, J. S. Kang, R. Bector, H. M. Baghdassarian, A. N. Truong, P. W. Kim, A. K. Singh, N. J. Hillson, J. D. Keasling, De novo DNA synthesis using polymerase-nucleotide conjugates, *Nat. Biotechnol.* (2018), doi:10.1038/nbt.4173.

8. H. H. Lee, R. Kalhor, N. Goela, J. Bolot, G. M. Church, Terminator-free template-independent enzymatic DNA synthesis for digital information storage, *Nat. Commun.* **10**, 2383 (2019).

9. H. A. Rees, D. R. Liu, Base editing: precision chemistry on the genome and transcriptome of living cells*Nature Reviews Genetics* **19**, 770–788 (2018).

10. C. J. Smith, O. Castanon, K. Said, V. Volf, P. Khoshakhlagh, A. Hornick, R. Ferreira, C.-T. Wu, M. Güell, S. Garg, H. Myllykallio, G. M. Church, Enabling large-scale genome editing by reducing DNA nicking*bioRxiv* , 574020 (2019).

11. M. J. Lajoie, A. J. Rovner, D. B. Goodman, H.-R. Aerni, A. D. Haimovich, G. Kuznetsov, J. A. Mercer, H. H. Wang, P. A. Carr, J. A. Mosberg, N. Rohland, P. G. Schultz, J. M. Jacobson, J. Rinehart, G. M. Church, F. J. Isaacs, Genomically recoded organisms expand biological functions, *Science* **342**, 357–360 (2013).

12. B. J. Karas, Y. Suzuki, P. D. Weyman, Strategies for cloning and manipulating natural and synthetic chromosomes, *Chromosome Res.* **23**, 57–68 (2015).

13. L. A. Mitchell, A. Wang, G. Stracquadanio, Z. Kuang, X. Wang, K. Yang, S. Richardson, J. A. Martin, Y. Zhao, R. Walker, Y. Luo, H. Dai, K. Dong, Z. Tang, Y. Yang, Y. Cai, A. Heguy, B. Ueberheide, D. Fenyö, J. Dai, J. S. Bader, J. D. Boeke, Synthesis, debugging, and effects of synthetic chromosome consolidation: synVI and beyond, *Science* **355** (2017), doi:10.1126/science.aaf4831.

14. F. Ceroni, T. Ellis, The challenges facing synthetic biology in eukaryotes, *Nat. Rev. Mol. Cell Biol.* **19**, 481–482 (2018).

Acknowledgments

We thank the following members of the GP-write technology and infrastructure working group for helpful discussions: O. Amirav-Drory, B. Bishop, A. Hessel, J. Hollenhorst, A. Khan, V. Martin, D. McClymont, N. McCorkle, J. Medford, N. Menon, M. Oshimura, B. Panda, E. Rech, N. Roehner and V. Zhirnov. We also thank N. J Kelley and A. Schwartz for administration and coordination of group discussions. **Competing interests:** Bogumil J. Karas is a co-founder, Chief Executive Officer, and shareholder of Designer Microbes Inc. Henry H. Lee is a co-founder of Kern Systems. Joel S. Bader is a founder and director of Neochromosome, Inc. George M. Church's financial interests are listed at http://arep.med.harvard.edu/gmc/tech.html. J. William Efcavitch is a co-founder and Board member of Molecular Assemblies Inc. Leslie A. Mitchell is a co-founder of Neochromosome, Inc. Alec A.K. Nielsen is a co-founder, Chief Executive Officer, and shareholder of Asimov Inc. These author declares no conflict of interest: Nili Ostrov, Jacob Beal, Tom Ellis, D. Benjamin Gordon, Scott C. Lenaghan, Jeffery A. Schloss, Giovanni Stracquadanio, Axel Trefzer, Cintia M. Coelho, Marc Güell, Bill Peck, Alexander C. Smith, C. Neal Stewart Jr, Hille Tekotte.

10.1126/science.aay0339