Title

Building a custom high-throughput platform at the Joint Genome Institute for DNA construct design and assembly—present and future challenges

Journal

Synthetic Biology, 5(1)

ISSN

2397-7000

Authors

Blaby, Ian K
Cheng, Jan-Fang

Publication Date

2020

DOI

10.1093/synbio/ysaa023

Peer reviewed

BUILDING A CUSTOM HIGH THROUGHPUT PLATFORM AT THE JOINT GENOME INSTITUTE FOR DNA CONSTRUCT DESIGN AND ASSEMBLY – PRESENT AND FUTURE CHALLENGES

Ian K. Blaby[1,2,¶] and Jan-Fang Cheng[1,2¶]

[1]US Department of Energy, Joint Genome Institute, Lawrence Berkeley National Laboratory, 1 Cyclotron Road, Berkeley, CA 94720, USA [2]Environmental Genomics and Systems Biology Division, Lawrence Berkeley National Laboratory, Berkeley, CA, USA

Running title:

High throughput DNA construct design and assembly

[¶] To whom correspondence should be addressed

Keywords:

bio-CAD; DNA assembly; DNA synthesis; Platform development; Synthetic biology

ABSTRACT

The rapid design and assembly of synthetic DNA constructs has become a crucial component of biological engineering projects via iterative design build test learn (DBTL) cycles. In this perspective we provide an overview of the workflows used to generate thousands of constructs and libraries produced each year at the U.S. Department of Energy (DOE) Joint Genome Institute (JGI). Particular attention is paid to describing pipelines, tools used, types of scientific projects enabled by the platform and challenges faced in further scaling output.

INTRODUCTION

The redesign of biological systems has historically been impeded by an incomplete understanding of the system and the difficulty in unhinging molecular components from the constraints, such as regulation and enzyme metabolite preference, to which they have evolved. With the broad objective of overcoming the latter and providing novel and accelerated approaches towards the former, synthetic biology lies at the interface of a number of biological disciplines and engineering. To help overcome incomplete knowledge of biological complexity, design-build-test-learn cycles are commonly employed as a means to iterating towards exploiting biological machinery with applications ranging from energy to agriculture to health. This approach has been enabled by technological achievements in the last decade: just as sequencing costs have decreased over the last 15 years, chemical DNA synthesis has declined to <$0.1/bp, while synthesized sequence accuracy and fragment

length have both increased. In recognition of this potential the Joint Genome Institute (JGI), a user facility that provides capabilities and scientific expertise on a competitive peer reviewed basis for DOE-relevant research, added to the suite of genomics capabilities by initiating the DNA synthesis program for synthetic biology in 2012. With similar goals to the JGI platform, a growing number of bio-foundries have been built that serve the regional needs for synthetic biology and bio-manufacturing process engineering (*1*).

For several decades, core facilities at academic institutes have been providing simple access to DNA sequencing services with rapid and cost-effective turnaround. Given cost reductions, development of biological computer aided design (bio-CAD) tools and laboratory automation, it is conceivable that DNA synthesis and construct assembly could become another major service of core facilities. This perspective aims to provide an overview of our experience in building a high throughput platform incorporating construct design, assembly, cloning, and sequence verification processes at scale for synthetic biology products.

CAPABILITIES OFFERED BY THE PLATFORM AND USER PROJECTS

The DNA synthesis platform at JGI comprises an end-to-end pipeline from design to assembled constructs and presently generates ~7Mbp/annum of custom DNA synthesis and assemblies. The platform team is divided between bioinformatics, production and research, and supports the workflow by developing new tools for design and optimization, assembly of constructs for users and development of new capabilities respectively. Presently, the platform focuses on four classes of project (Figure 1). These include: a) small inserts, such as single or a few small genes (typically <5kb/insert total); b) large inserts including complete pathways or

multiple operons (up to ~50 kb); c) combinatorial constructs or libraries and d) small size and high degree of variants libraries, such as gRNA or promoter libraries (Figure 1). DNA is not synthesized internally at the JGI, but ordered as linear or clonal fragments by commercial vendors, and is used to generate the building blocks of each construct (or oligonucleotide pools in the case of small size libraries; see Supplemental File for details). These are then seamlessly assembled via Gibson (*2*), Golden Gate or MoClo assembly (*3, 4*), or yeast recombinase mediated cloning (*5, 6*) as appropriate for each project. For large constructs exceeding ~25kb sequential rounds of these methods are employed.

As a DOE BER user-facility all granted proposals at JGI contribute to BER priorities (*7*), and more broadly, DOE mission. Consequently, many projects aim to address or query gene function, for example generating large numbers of enzymes for biochemical or biophysical characterization, or for performing mutant library screens to identify all genes responding to a given condition. Other projects focus on optimizing a previously characterized pathway by combinatorially arranging components (*eg* promoters, coding regions and terminators) from different sources to achieve elevated product level.

Projects can be further enhanced in scope by additional JGI capabilities that intersect with DNA synthesis. Multiple projects have benefited from data mining sequence based databases developed and maintained by JGI (such as the Integrated Microbial Genomes and Microbiomes (*8*), Phytozome (*9*) and Mycocosm (*10*)) to maximize the breadth of phylogenetic diversity surveyed, contributing to the design of DBTL the cycle. Other opportunities exist with other technologies, for example, synthesizing genes encoding transcription factors for regulon

interrogation by DapSeq (*11*), or for metabolomics to investigate the metabolic consequences of introduced pathways (supporting test and learn phases of DBTL). Frameworks are also in place allowing DNA synthesis (and other JGI capabilities) to be coupled with other complementary DOE user facilities.

Many projects aim to obtain a greater understanding of protein function and focus on characterizing proteins expressed from synthetic DNA constructs either *in vitro* in cell free extracts or *in vivo* using model organisms. An affinity purification tag is usually used for expressing genes heterogeneously in model organisms to yield sufficient protein quantity and purity for downstream enzymology and structural analysis. To mitigate potential protein insolubility or toxicity there is growing interest in expressing the same set of genes in multiple organisms, thus elevating the likelihood of obtaining the desired protein. For both identifying specific sequences to work with and ensuring taxonomic diversity ensuring the breadth of a protein family is captured, sequence data mining often constitutes part of the design of gene function discovery-based projects. Some examples include characterization of terpene biosynthesis (*12*-*15*) and the glycoside hydrolase protein families (*16*-*19*). Such projects often require single gene constructs, but increasingly interest is growing in much larger fragments, for example functional interrogation of biosynthetic gene clusters (BGCs).

Another growing area of interest is CRISPR-related projects for tool development and libraries of engineered strains exploiting CRISPR nuclease, interference or activation-based screens (CRISPRi (*20*) and CRISPRa (*21*) respectively). Multiple libraries have been generated and are presently being worked on ranging from prokaryotes to yeasts (*22*) and algae.

The high capacity of the platform facilitates the synthesis of large construct numbers allowing for pathway screening. Examples of these approaches include the use of combinatorial screens to uncover novel and optimal activities (*23-25*) and metabolic engineering for the development of new pathways (*26*).

PROJECT MANAGEMENT AND WORKFLOWS

The general workflow of the JGI platform has been modelled on DBTL engineering cycles, with a typical project beginning with an initiating conference call between JGI staff scientists and the proposers to discuss the project goals, experimental design, and DNA assembly strategies suggested by JGI staff to aid users with their research (Supplemental Figure 1). In this regard, work by JGI scientists provides the Design and Build capabilities, and, depending on the nature of the project, can contribute to limited capabilities of Test and Learn in the researcher's own laboratory (Supplement Figure 1).

Subsequent to receiving sequence information, the data are processed through a suite of custom software pipelines, as appropriate for the individual project, and the final designs sent back to the user for final confirmation prior to ordering individual fragments (termed "building blocks"). These computational tools have recently been comprehensively reviewed (*27*), but are summarized in the Supplemental File. Once all of the synthetic DNA fragments and oligonucleotides required for a project have been received, their delivery is recorded into our LIMS system for tracking and the molecular workflows are initiated (Supplemental File; Supplemental Figure 2)

CHALLENGES & LIMITATIONS OF SCALING OUTPUT

Total output of the platform has grown steadily each year (Supplemental Figure 3). In 2019, 41 user projects were initiated. 2019 saw the construction of 4182 and 338 cloned fragments of < and > 5kb respectively, which totals 7.44Mb of constructs delivered (Supplemental Figure 3). Of these 4520 requested constructs 4030 were delivered (89%). 7.44Mb delivered compares to 6.59Mb synthesized; this discrepancy is accounted for by several projects requiring only PCR amplification from genomic or plasmid DNA templates. In addition, 9 libraries of high degree variants were constructed and delivered. Users are predominantly based in the USA, but the platform is globally accessible and projects also originate from institutes from other countries. Nevertheless, the platform is not presently approaching maximum capacity; the potential for scaling beyond these numbers is dependent upon multiple factors, as discussed below. Detailed information regarding our molecular workflows, apparatus and protocols are provided in Supplemental File and summarized in Supplemental Figure 2.

A limitation in achieving the theoretical maximum capacity of our pipeline infrastructure is that user projects are often split into smaller batches. This enables the collaborator to pilot test their vector or downstream assays before committing with their full request, but frequently results in non-filled plates during the assembly process, leading to reduced machine and staff time efficiencies. One possible option that we are currently exploring to mitigate this is to combine projects on plates to fill every well, and deconvolute constructs post completion.

Standardization of project type has generally been resisted in order to maximize project flexibility, and accordingly project goals and the scientific questions posed by users vary significantly. Whether fragments originate from PCR amplicons or

synthesized DNA, fragments are assembled into the user's vector of choice precisely as agreed upon in discussions. Often this necessitates vectors being modified or built prior to assembly of the final constructs. A further complication that occasionally derives from custom vectors is incorrect sequence data. By default, platform staff sequence validate all incoming vectors to ensure constructs are assembled as intended. On rare occasions point mutations or larger sequence deviations identified in the provided plasmid DNA must be repaired before a project can proceed. While this vector flexibility, which most companies do not offer (or require on-boarding fees for new vectors and/or additional cloning costs), represents a limitation to the potential platform output, this approach ensures the final constructs are of maximal utility to collaborators for addressing their scientific questions.

Another restriction to throughput is that for some projects DNA synthesis is inappropriate or not applicable. For instance, if the DNA sequence of the final construct cannot be refactored (such as for gene-flanking regions for homologous recombination mediated gene deletions where the cloned regions must be identical at the nucleotide level) and/or synthesis constraints cannot be overcome, PCR amplification from the source DNA or direct cloning are the only affordable options. Conversely, DNA synthesis provides an opportunity to study coding capacity if the naturally occurring DNA is not available, such as the characterization of genes encoded on a sequenced environmental sample or unculturable microbes. Since projects that depend upon significant PCR amplification from genomic DNA are more prone to failure to achieve all required building block fragments (due to PCR limitations of high or low GC skew, repetitive sequence, secondary structure, and the possibility of point mutations being introduced by the polymerase into the

amplified DNA), projects utilizing synthetic DNA generally yield higher delivery rates and are more amenable to both automation and high-throughput approaches. Operonic structures and biosynthetic gene clusters present a different challenge since (depending on the use case) coding regions may be refactored, but alteration of regulatory regions such as promoters or terminators from the native sequence is generally undesirable. As well as surmounting synthesis of problematic sequence, refactoring provides opportunities to modify each codon, and potentially impacting gene expression levels by mimicking codon usage of the host organism's genome.

In our present workflows, some steps may be performed manually using multi-channel pipettes instead of automation for small numbers of assemblies or if a machine is in use. While affording maximal flexibility with methods for assembly and general workflows, this approach both prevents maximal capacity from being achieved and necessitates constant human oversight. To help overcome this, one avenue that is presently being explored is a complete end-to-end integrated system for the automated assembly of constructs via Gibson assembly, for which the inputs would be DNA building blocks, oligonucleotides and reagents, and the output would be complete final constructs arrayed in plates.

Finally, an ongoing challenge faced is balancing the testing of emergent technologies to identify novel approaches, potentially leading to efficiency gains or new product types, versus investing in scaling up methods already in place for which robust protocols have been developed. New products are frequently assessed for their possible benefits to the platform as they become available, and subsequent to validating for compatibility with existing workflows and protocol robustness, are incorporated into the platform.

FUTURE DIRECTIONS

Catalyzed by technological advances and the resulting decreases in price and turnaround times, the applications of DNA synthesis remain fast-moving technologies. Accordingly, a series of computational and biological applications are presently in development with potential future availability to users, briefly summarized here.

Extensive genome sequencing has revealed new coding capacity whose function has yet to be discovered. But functional characterization of these sequences has been impeded by difficulty in expressing from traditional model organisms, possibly due to a combination of misfolding of proteins, absence of required precursor metabolites and/or low tolerance to gene products. To help overcome this, an area of active development is the engineering of diverse microorganisms as modular chassis strains for heterogeneous expression of synthetic genes and pathways. Presently this includes around two dozen new strains in δ-Proteobacteria (*28, 29*), with work progressing on additional prokaryotic lineages and unicellular eukaryotes.

Finally, if trends from the past decade continue, notwithstanding progress in new technologies, the annual capacity of the platform might be expected to continue to increase, leading to possible increases in the scope and/or number of projects worked on.

ACKNOWLEDGMENTS

This work has been supported by the DOE Joint Genome Institute (http://jgi.doe.gov) by the U.S. Department of Energy, Office of Science, Office of Biological and Environmental Research, through Contract DE-AC02-05CH11231 between Lawrence Berkeley National Laboratory and the U.S. Department of Energy. We are extremely grateful to Sangeeta Nath for input on drafting figures.

FIGURE LEGENDS

Figure 1 Overview of construct types. (a) Single or multiple small genes are typically constructed by Gibson assembly. gRNA libraries can be similarly constructed using oligonucleotide pools flanked with vector homology. (b) Pathways and multiple operons can be compiled by digestion with a type IIS restriction enzyme (whose recognition site either does not occur in composite sequences or has been removed by sequence refactoring) followed by Golden Gate assembly or inclusion of overlapping sequences and yeast-mediated recombination. (c) Combinatorial libraries are constructed by Golden Gate assembly or for libraries with increased complexity using modular cloning (MoClo) approaches. (d) Libraries containing higher degrees of variants are generated using multiple compatible inserts and assembled into vectors via Gibson or Golden Gate assembly. Regions of overlapping homology for Gibson assembly or yeast recombination are signified by matching colors. Promoters, terminators and enzyme cut sites are designated by green arrows, red Ts and scissor cartoons respectively.

REFERENCES

1. N. Hillson *et al.*, Building a global alliance of biofoundries. *Nat Commun* **10**, 2040 (2019).
2. D. G. Gibson *et al.*, Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nat Methods* **6**, 343-345 (2009).
3. E. Weber, C. Engler, R. Gruetzner, S. Werner, S. Marillonnet, A modular cloning system for standardized assembly of multigene constructs. *PLoS One* **6**, e16765 (2011).
4. C. Engler, R. Kandzia, S. Marillonnet, A one pot, one step, precision cloning method with high throughput capability. *PLoS One* **3**, e3647 (2008).
5. T. M. Joska, A. Mashruwala, J. M. Boyd, W. J. Belden, A universal cloning method based on yeast homologous recombination that is simple, efficient, and versatile. *J Microbiol Methods* **100**, 46-51 (2014).
6. N. Kouprina, V. Larionov, Transformation-associated recombination (TAR) cloning for genomics studies and synthetic biology. *Chromosoma* **125**, 621-632 (2016).
7. BERAC, Grand Challenges for Biological and Environmental Research: Progress and Future Vision. *A Report from the Biological and Environmental Research Advisory Committee* DOE/SC–0190, science.energy.gov/~/media/ber/berac/pdf/Reports/BERAC-2017-Grand-Challenges-Report.pdf, (2017).
8. I. A. Chen *et al.*, IMG/M v.5.0: an integrated data management and comparative analysis system for microbial genomes and microbiomes. *Nucleic Acids Res* **47**, D666-D677 (2019).
9. D. M. Goodstein *et al.*, Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Res* **40**, D1178-1186 (2012).
10. I. V. Grigoriev *et al.*, MycoCosm portal: gearing up for 1000 fungal genomes. *Nucleic Acids Res* **42**, D699-704 (2014).
11. A. Bartlett *et al.*, Mapping genome-wide transcription-factor binding sites using DAP-seq. *Nat Protoc* **12**, 1659-1672 (2017).
12. K. M. Murphy, L. T. Ma, Y. Ding, E. A. Schmelz, P. Zerbe, Functional Characterization of Two Class II Diterpene Synthases Indicates Additional Specialized Diterpenoid Pathways in Maize (Zea mays). *Front Plant Sci* **9**, 1542 (2018).
13. K. A. Pelot *et al.*, Functional Diversity of Diterpene Synthases in the Biofuel Crop Switchgrass. *Plant Physiol* **178**, 54-71 (2018).
14. Y. Ding *et al.*, Multiple genes recruited from hormone pathways partition maize diterpenoid defences. *Nat Plants* **5**, 1043-1056 (2019).
15. K. A. Pelot, D. M. Hagelthorn, Y. J. Hong, D. J. Tantillo, P. Zerbe, Diterpene Synthase-Catalyzed Biosynthesis of Distinct Clerodane Stereoisomers. *Chembiochem* **20**, 111-117 (2019).

16. S. S. Macdonald *et al.*, Development and Application of a High-Throughput Functional Metagenomic Screen for Glycoside Phosphorylases. *Cell Chem Biol* **26**, 1001-1012 e1005 (2019).
17. E. M. Glasgow *et al.*, Extent and Origins of Functional Diversity in a Subfamily of Glycoside Hydrolases. *J Mol Biol* **431**, 1217-1233 (2019).
18. K. Deng *et al.*, Development of a High Throughput Platform for Screening Glycoside Hydrolases Based on Oxime-NIMS. *Front Bioeng Biotechnol* **3**, 153 (2015).
19. R. A. Heins *et al.*, Phylogenomically guided identification of industrially relevant GH1 beta-glucosidases through DNA synthesis and nanostructure-initiator mass spectrometry. *ACS Chem Biol* **9**, 2082-2091 (2014).
20. M. H. Larson *et al.*, CRISPR interference (CRISPRi) for sequence-specific control of gene expression. *Nat Protoc* **8**, 2180-2196 (2013).
21. P. Perez-Pinera *et al.*, RNA-guided gene activation by CRISPR-Cas9-based transcription factors. *Nat Methods* **10**, 973-976 (2013).
22. C. Schwartz *et al.*, Validating genome-wide CRISPR-Cas9 function improves screening in the oleaginous yeast Yarrowia lipolytica. *Metab Eng* **55**, 102-110 (2019).
23. L. Xie *et al.*, Methylglucosylation of aromatic amino and phenolic moieties of drug-like biosynthons by combinatorial biosynthesis. *Proc Natl Acad Sci U S A* **115**, E4980-E4989 (2018).
24. Z. Y. Dossani *et al.*, A combinatorial approach to synthetic transcription factor-promoter combinations for yeast strain engineering. *Yeast* **35**, 273-280 (2018).
25. X. Wang *et al.*, Rational Reprogramming of O-Methylation Regioselectivity for Combinatorial Biosynthetic Tailoring of Benzenediol Lactone Scaffolds. *J Am Chem Soc* **141**, 4355-4364 (2019).
26. T. Schwander, L. Schada von Borzyskowski, S. Burgener, N. S. Cortina, T. J. Erb, A synthetic pathway for the fixation of carbon dioxide in vitro. *Science* **354**, 900-904 (2016).
27. E. Oberortner *et al.*, An integrated computer-aided design and manufacturing workflow for synthetic biology. In *DNA Cloning and Assembly Methods and Protocols* 3-18 Humana Press, New York, NY (2020)
28. J. Ke, Y. Yoshikuni, Multi-chassis engineering for heterologous production of microbial natural products. *Curr Opin Biotechnol* **62**, 88-97 (2019).
29. G. Wang *et al.*, CRAGE enables rapid activation of biosynthetic gene clusters in undomesticated bacteria. *Nat Microbiol* **4**, 2498-2510 (2019).