

Organizing genome engineering for the gigabase scale

Bryan A. Bartley
Raytheon BBN Technologies
bryan.a.bartley@raytheon.com

Jacob Beal
Raytheon BBN Technologies
jakebeal@ieee.org

Jonathan R. Karr
Icahn School of Medicine at Mount Sinai
karr@mssm.edu

Elizabeth A. Strychalski
National Institute of Standards and Technology
elizabeth.strychalski@nist.gov

September 5, 2019

Abstract

Engineering the entire genome of an organism enables large-scale changes in organization, function, and external interactions, with significant implications for industry, medicine, and the environment. Improvements to DNA synthesis and organism engineering are already enabling substantial changes to organisms with megabase genomes, such as *Escherichia coli* and *Saccharomyces cerevisiae*. Simultaneously, recent advances in genome-scale modeling are increasingly informing the design of metabolic networks. However, major challenges remain for integrating these and other relevant technologies into workflows that can scale to the engineering of gigabase genomes.

In particular, we find that a major under-recognized challenge is coordinating the flow of models, designs, constructs, and measurements across the large teams and complex technological systems that will likely be required for gigabase genome engineering. We recommend that the community address these challenges by 1) adopting and extending existing standards and technologies for representing and exchanging information at the gigabase genomic scale, 2) developing new technologies to address major open questions around data curation and quality control, 3) conducting fundamental research on the integration of modeling and design at the genomic scale, and 4) developing new legal and contractual infrastructure to better enable collaboration across multiple institutions.

1 From Engineering Genes to Engineering Genomes

Engineering the entire genome of an organism will allow large-scale changes in organization, function, and environmental interactions, with major implications for applications of biotechnology broadly [1]. The past several decades have seen remarkable progress in our capability to create DNA and modify genomes [2–4]. Since Khorana created the first synthetic gene forty years ago [5], our capability to construct DNA sequences has doubled approximately every three years (Figure 1A), progressing from plasmids in the early 1990’s [6, 7], viruses in the early 2000’s [8], and gene clusters in the mid-2000’s [9, 10], to the first bacterial chromosome in 2008 [11, 12]. More recently, several groups have demonstrated the feasibility of total synthesis of the 4Mb genomes of *Escherichia coli* [13, 14] and *Salmonella typhimurium* [15], and the Sc 2.0 project [16, 17] has nearly completed re-engineering the 11.4 Mb genome of *Saccharomyces cerevisiae* [18]. Projecting

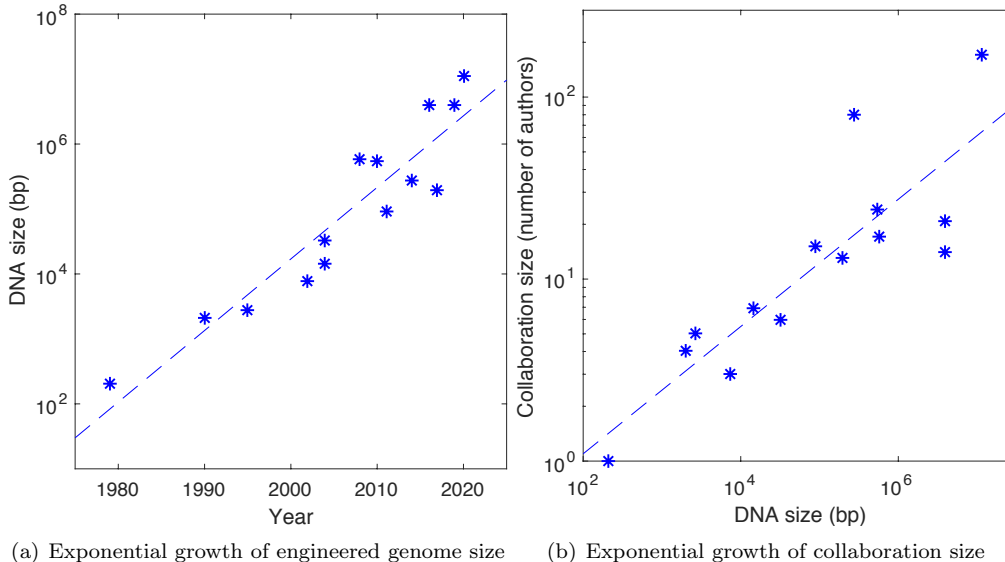


Figure 1. As capabilities for genome engineering advance rapidly, the size of teams involved in each genome engineering project also increase. **(a)** From 1980 to present, the size of the largest engineered genomes has grown exponentially, doubling approximately every three years. Extrapolating this trend projects gigabase engineering becoming feasible by 2050. **(b)** The sizes of the teams needed to produce these genomes has also grown exponentially, scaling with the cube root of genome size and suggesting that teams on the order of 500 investigators will be needed to engineer gigabase genomes. Data for this figure is provided in Supplementary Data 1.

further to the gigabase scale, in 2016 leaders from academia and industry formed the Genome Project-Write [1] consortium to roadmap new technical approaches and ethical frameworks for engineering the genomes of higher-order organisms. Since then, members of the consortium have proposed projects to develop human cell lines with engineered genomes, including a virus-resistant, ultra-safe cell line for pharmaceutical production [19].

Moving to the gigabase scale poses major technological and scientific challenges, with scaling up DNA synthesis and/or editing arguably foremost among these. These challenges related to synthesis and editing have been discussed extensively, including a recent summary in [20]. However, the challenges of managing the complex workflows and large teams needed for genome engineering not previously been analyzed. Figure 1B shows that the number of investigators needed to engineer a genome has also risen markedly with the size of the genome. If both of these trends continue, then the capability to engineer gigabase eukaryotic genomes would be projected to be realized in approximately 2050, with each such genome required a team of around 500 investigators. Thus, engineering gigabase genomes will likely require new approaches to coordinate complex workflows and large, interdisciplinary teams.

Accordingly, we have examined the emerging design-build-test-learn workflow for genome engineering and identified potential bottlenecks and associated solutions, as well as areas where additional research will likely be needed. In Section 2, we discuss the emerging, design-build-test-learn workflow for organism engineering and identify key points to integrate workflow steps across processes and/or organizations. We then discuss each integration point in detail in Section 3, identifying gaps that must be addressed to support gigabase engineering. Finally, in Section 4, we summarize and recommend priorities to advance genome engineering.

2 Toward Workflows for Gigabase Engineering

Gigabase engineering will likely require large-scale integration of efforts for the division, distribution, and coordination of labor and materials. Both genome engineering and smaller-scale organism engineering workflows have often been abstracted and organized in terms of design-build-test-learn cycles [3, 18, 21–26]. This iterative approach is helpful given the complexity and uncertainties in engineering biology, progressing in incremental steps, implementing genetic modifications in stages, and adjusting designs based on information learned from testing prototypes and partial implementations.

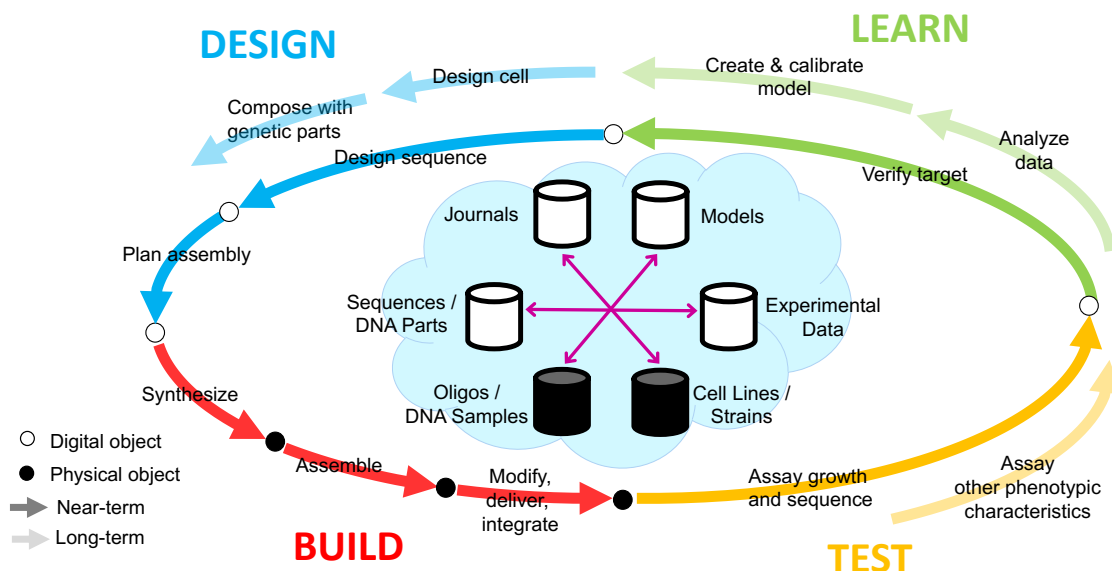


Figure 2. The emerging design-build-test-learn workflow for genome engineering is shown schematically with current (solid arrows) and predicted (transparent arrows) tasks, interfaces (circles), and digital (white cylinders) and physical (black cylinders) repositories.

Figure 2 shows the emerging design-build-test-learn workflow for genome engineering. The first step in this workflow is to *design* a target sequence using models and design heuristics. The second step is to *build* these sequences, producing and expressing DNA constructs. In the third step, the molecular and phenotypic characteristics and consequences of these sequences are *tested*. Finally, researchers and engineers improve models, design heuristics, and processes by *learning* from the data generated by testing, focusing on discrepancies between the predicted and observed results.

The primary loop in Figure 2 indicates the workflow necessary for current genome engineering projects, which have largely focused on “top-down” approaches to recode and refactor genomes, such as reducing genomes to essential sequences. As genome writing technology matures, the focus of genome engineering workflows may shift to more fully include the tasks in the outer loop of Figure 2. For example, one ultimate aim of synthetic biology is to design novel genomes that encode new phenotypes from the “bottom-up” by assembling organisms from modular parts [27]. Parts-based synthetic biology is already being used to engineer novel metabolic pathways for commercial production of high-value chemical products, but engineering on the genomic scale will be staggeringly more complex. In this case, sharing and reusing intermediate, mid-scale DNA constructs that constitute biological parts and devices, as well as large-scale DNA constructs more akin to synthetic chromosomes, will likely become increasingly important.

At each stage of the workflow, reagents, biological materials, and/or information must be transferred through an interface from one set of processes to another, often run by investigators with complementary areas of expertise and located in different groups or institutions. These products are both physical and digital in nature. For example, design steps should produce sequences described in a digital format, such as the Synthetic Biology Open Language (SBOL) [28, 29], which then serves as an input for experimentalists to build those sequences. Build steps produce DNA constructs and cell lines to be tested. Test steps produce data to be learned from. Learn steps produce data-driven and mechanistic models, ideally expressed in a digital format. Materials and data at all stages may also reside in sample and data repositories. Genome engineers will need interfaces to these repositories to efficiently browse, access, alter, and transfer materials, as well as to cross-reference information across the stages of a given workflow.

Beyond the technical challenges for each stage of the workflow, genome engineering must also contend with a number of cross-cutting issues to facilitate close coordination across many organizations. When technical information or materials are transferred, their recipients will need to know the associated contractual and legal obligations, such as information about intellectual property and licensing, safety, and any privacy or personally identifiable information (PII) concerns. Additionally, issues of cybersecurity, biosecurity, and/or biosafety may pertain at each interface.

For gigabase engineering, every one of the interfaces between tasks, repositories, and organizations will need to be optimized to mitigate bottlenecks to progress. For projects of this scale, ad hoc, human-centric, and bespoke interfaces will be impractical. Instead, every interface will require representations conducive to machine reasoning and automation, and appropriate accompanying cyber-infrastructure and tooling. Fortunately, most of the design-build-test-learn workflow is not unique to gigabase engineering, such that prior work in smaller-scale organism engineering provides a solid basis for paths forward from the state of the art towards realizing effective and routine engineering of large genomes.

3 Identifying And Closing Gaps At The State Of The Art

In this section, we discuss the integration challenges identified in the previous section, reviewing the state of the art in technologies and standards with respect to the emerging needs of gigabase genome engineering. Instead of focusing on specific evolving protocols and methods, which are likely to advance rapidly, we consider the information that must be communicated to enable protocols or methods to be composed into a comprehensive workflow. Through this analysis, we identify critical gaps and opportunities, where additional technologies and standards would facilitate workflows that are able to effectively deliver gigabase engineered genomes. Table 1 summarizes these recommendations.

Stage	Interface between workflow stages	Recommendation
Design	Genomic Design → Plan Assembly	Extend: GFF3 and/or SBOL with chromosomal coordinates
Design	Import materials from public repositories (e.g., NCBI)	Extend: existing FASTA/GenBank importers/API
Design	Sharing of genome-scale designs	Adopt: SynBioHub / ICE
Design	Models → Design Cell	Research: integrated CAD across multi-model repositories
Design	Cell Design → Composition of Genetic Parts	Research: comprehensive cell systems simulation for design evaluation
Build	Assembly Plan → Synthesize	Adopt: FASTA/GenBank, migrating to SBOL (function & plan for manufacturability flexibility)
Build	Short ssDNA/dsDNA Samples → Assemble	Adopt: SBOL Extend: FASTQ, GVF, and/or SBOL with quality metric requirements
Build	Long dsDNA samples → Modify/Deliver/Integrate	Extend: SBOL and/or GFF3 with integration context, modifications (e.g., BpForms)
Build	Sharing inventory of ssNA/dsNA samples, cell strains	Adopt: SBOL 2.2 Design/Build/Test, ICE
Build/Test	Sharing of biological samples (NAs & cells)	Develop: inventory-compatible legal/contractual framework
Test	Strains → Sequence Validation	Extend: FASTQ, GVF, and/or SBOL with quality metric requirements Develop: fitness metrics and associated specification language
Test	Strains → Phenotype assays	Research: integrate omics and other measurements with specifications of desired phenotype
Test	Cross-laboratory experimental data comparison	Extend: process control and calibration standards
Learn	Sharing of experimental data	Adopt: FAIRDOMHub, EDD or other open cloud solution
Learn	Experimental Data → Analysis Pipelines	Extend: SBOL Design/Build/Test + OBO & EFO ontologies, new metadata and knowledge curation tools
Learn	Analyzed Data → Create Model	Research: automated and scalable model generation and verification
Learn	Model sharing and composition	Extend: SBML, CellML, BioModels, and related COMBINE standards Research: standards for describing multi-scale modeling
Cross-Cutting Issues		Recommendation
Workflow Management		Adopt: CWL, container tools, PROV-O Develop: Interoperable laboratory automation languages
Database federation		Adopt: existing open DBMS solutions
Link designs → samples → strains → data → models		Extend: PROV-O, SBOL 2.2 Design/Build/Test
IP tracking and composition		Develop: based on OpenMTA, OSI / CC / ScienceCommons, PROV-O
Privacy management, public release timing		Develop: based on cross-domain information sharing protocols, PROV-O

Table 1. We assess the state of the art of gigabase-scale genome engineering. Recommendations are categorized into adopting or extending existing technologies for near-term solutions, developing new technologies for mid-term solutions, and the need for additional fundamental research for longer-term solutions. Color indicates technology readiness level: green indicates needs that can be fulfilled by adopting or extending relatively mature existing methods, yellow indicates potential solutions from extensions of the state of the art, and red indicates areas where more fundamental research is required.

3.1 Rationally refactoring and designing gigabase genomes

The first step of the design-build-test-learn workflow is to design a synthetic genome. Currently, most synthetic genomes are designed ad hoc using a combination of biological knowledge, engineering intuition, and informal design heuristics. This approach has enabled projects, such as reducing *Mycoplasma* genomes by eliminating non-essential elements [26], reordering genes into functional groups [17], and inserting small metabolic pathways [30]. However, this approach can be inefficient, slow, and expensive. In contrast, most other engineering fields, such as aerospace engineering, use computer-aided design (CAD) powered by mechanistic models to reliably design complex systems, such as commercial jets and spacecraft. Ultimately, CAD powered by predictive models will likely be needed to reliably design novel genomes that encode novel functions [31]. In turn, this will require comprehensive and detailed models that can help design of entire genomes. Substantial fundamental research is required to enable such model-driven design at the genome scale.

As genomes cannot currently be designed *de novo*, for the foreseeable future, genome design will likely involve modifying the sequences of existing organisms. Annotated genome sequences are available for diverse organisms. Public archives, such as the National Center for Biotechnology Information, the EMBL European Bioinformatics Institute, and the DNA Data Bank of Japan, which collectively constitute the International Nucleotide Sequence Database Collaboration (INSDC) [32], presently contain on the order of 10^5 bacterial genomes and hundreds of eukaryotic genomes [33–37].

Genome design and debugging can be further facilitated by supplementing genome sequences with additional functional annotations. For example, INSDC reference genomes primarily annotate all known transcriptionally active parts of the genome, but the engineer will also need to consider tissue-specific expression patterns. Other important sequence features include intergenic elements, including regulatory elements, such as enhancers and silencers, structural elements, replication origins, and clinically-significant sites of DNA recombination and instability. Much of this knowledge that can help the engineer is unfortunately distributed among many independent resources. Services such as NCBI Genome Viewer [38], WebGestalt [39] and DAVID [40] provide programmatic or web interfaces for integrating annotations. An ideal platform for genome design would integrate such resources with sequence editing and design tools.

The quality of annotations on a genome is another key consideration in the design process. Genomes can have significant differences in annotations depending on the toolchains that generate them, and these differences will likely result in trade-offs impacting engineering decisions. For example, the human reference genomes generated by the RefSeq and GENCODE projects have notable differences [41, 42]. GENCODE annotations have been shown to cover more coding regions and alternative splice regions and consequently are more likely to flag a loss-of-function when a sequence variant occurs [41]. The sensitivity and specificity of annotation tools are thus important factors in sequence design and debugging. A genome engineering effort would therefore benefit from adoption of a standard convention for labeling annotations with estimates of confidence and reliability, such as the RefSeq database does with the Evidence and Conclusion Ontology [43].

Collaborative design of genomes design at large scale will require an appropriate choice of data formats and schema. Many commonly used sequence formats, such as GenBank and EMBL, are formatted for human readability, but have significant limitations with respect to managing large genomic designs. Transferring genomic data over a network in these formats can be a significant bottleneck, due to their monolithic treatment of sequences. Moreover, these formats are not easy to merge or harmonize across multiple concurrent users. Fixed formats like GenBank and GFF do not facilitate integration of annotations from diverse resources, while open semantic formats, such as SBOL, more easily support such extensions.

For communicating a genomic design, however, there are already two good options to adopt and extend. One is the Generic Feature Format Version 3 (GFF), which allows hierarchical organization of sequence descriptions (e.g., genes may organized into clusters, and clusters into chromosomes) and makes use of the Sequence Ontology [44] for coherent sequence annotations. GFF has already been used in the Sc 2.0 genome engineering project [18]. The other is the Synthetic Biology Open Language (SBOL) [45]. Like GFF, SBOL supports hierarchical description with standardized vocabularies. SBOL, however, also allows representation of abstract architectures in which sequences are not yet fully specified (e.g., asserting that a certain set of genes will be used, but not yet the particular variants or their arrangement), which is useful both for interchange during the design process and also for representing sequence variants and combinatorial libraries. SBOL can also describe other molecular components and interactions in a cell (e.g., proteins, metabolites,

regulatory interactions), allowing specification of phenotype and its linkage to genotype. This also allows designs to be organized in terms of functional relations (e.g., metabolic pathways, cellular systems), rather than simply genomic proximity. Finally, the SBOL standard is interoperable with SBML, allowing bidirectional conversion between designs and simulation-ready models. For both GFF and SBOL, however, it would be useful to have a richer language for specifying the position of sequences in a chromosome: the current practice of using sequence index is fragile to unrelated distant changes impacting the alignment of design fragments.

To date, genome engineering projects have minimally leveraged tooling that integrates sequence-design and network-scale modeling. Models will become increasingly important as the target genomes increase in scale or diverge from native architecture. For example, models could help predict growth phenotypes, including non-viable states, that may result from engineering sequence elements. Network-scale models, such as genome-scale metabolic models [46, 47] and whole-cell models [48], can be constructed by cross-referencing annotations across multiple databases to identify gene-protein-reaction (GPR) associations. Resources, such as the Integrated Microbial Genomes database [49] and the SEED [50], maintain annotated genomes that enable automated generation of such models. Ultimately, more ambitious genome engineering projects, which move beyond refactoring and recoding into more complex changes of organism function, will likely need to make modeling a critical portion of the design process and will need to implement function not just at the sequence level but by composing separately characterized genetic parts and devices. Much fundamental research still needs to be conducted, however, before such approaches will become practical at the gigabase genomic scale.

Additional constraints on design come from the methods used to modify genomes. This necessitates adoption of consistent policies to resolve design conflicts, as well as to achieve the desired design objectives. Examples of design policies include the removal of elements, such as restriction sites, separation of overlapping features, replacement of codons, or optimization of nucleotide content for synthesis. In the Sc 2.0 project, design policies were semi-formalized through group consensus, but in practice were implemented by a combination of customized software tools or manual sequence editing. Such policies could be enhanced by encoding synthesis expertise in the form of rules-based ontologies [51, 52]. The BOOST tool developed by JGI provides a prototype for such a language [53]. Synthesis providers could also use such methods to encode their synthesis constraints, enabling automatic pre-qualification of designs as readily compatible with a providers' synthesis constraints.

The business interfaces of DNA synthesis providers could also improve. Current tools typically operate at the level of individual fragments, often ordered by a human-centric web interface or, less frequently, APIs, with *ad hoc* sequence redesign based on feedback from the tool on synthesis constraints. This makes it difficult for the designer to assess the overall impact of synthesis provider interactions in the context of the larger design, meaning that changes made to optimizing synthesis can inadvertently disrupt vital regulatory regions or sequences required for subsequent assembly, cloning, or sequence verification. Particular sequence features, such as repeats, secondary structure, and high GC content, can cause synthesis failures or problems assembling a DNA construct, often resulting in need for substantial redesign, project delays, and escalating costs. DNA synthesis providers may also have constraints specific to their platforms. Large-scale construction would be better supported by more standardized interfaces that would allow designers and synthesis providers to share a complete construct design in a common data format, including genomic context, annotation with desired functional biological features, and assembly plans (e.g., Gibson junctions). This would then allow a more collaborative interaction, in which the synthesis service can optimize for synthesis, or at least recommend solutions to the user, while remaining within the constraints of the original design and tracking the relation between suggested changes and the original target sequence. This also has an additional benefit for biosecurity, as the intent behind synthesis orders becomes more transparent. The SBOL format is well-suited for supporting such annotations, though GenBank and GFF could also potentially be extended to encode this information.

3.2 Building Engineered Genomes

Technology and protocols to build engineered genomes specified in the design process are advancing rapidly. Still, the current approach of assembling DNA from smaller pieces is likely to hold for the foreseeable future up to the gigabase scale. Building a gigabase genome will likely require synthesizing shorter DNA sequences,

assembling these into larger constructs, and realizing the final genome that passes to the testing portion of the engineering workflow. Discussion of scalable protocols for building gigabase-scale genomes may be found in [1] and [20]; here, we focus on the integration challenges of workflows with such protocols.

Depending on the specific host and intended function of the engineered organism, there are typically numerous alternative approaches and protocols for DNA synthesis, assembly, and delivery. These must result not only in the designed DNA sequence; synthesis, assembly, and delivery must proceed in a context that is ultimately compatible with and results in the stable expression of that genome within the host organism with the desired phenotype. This is no small task. Currently, there is an unmet need for guidance on best-practices for measuring, tracking, and sharing information regarding the engineered genome, as it is built step-by-step.

Building an engineered genome, then, involves manipulating DNA many times, which offers ample opportunities for reduced yield, breakage, error, and other sources of uncertainty in achieving the designed DNA sequence. Most DNA synthesis begins with shorter DNA fragments ordered from a service provider. Biases and other sources of uncertainty in the delivered sequence associated with a synthesis platform and method are not typically communicated in a useful fashion along with the delivered product. Rather, the DNA is often accompanied by a certificate asserting sequence verification. Sequencing by the recipient at this early step to verify and validate the received DNA is typically not done, due to cost and time constraints. However, the synthesized DNA may occasionally behave in biomolecular reactions in ways that are unexpected based on the requested sequence, such as during polymerase chain reactions with specific primers during amplification. This can be due to the presence of a secondary population of DNA with a slightly different sequence, which may not be apparent from sequencing (e.g., appearing as a second band after amplification in a gel electrophoresis separation). Large-scale operations should develop more comprehensive and quantitative approaches to measuring, tracking, and communicating information pertaining to quality control.

Protocols and commercial kits to assemble shorter DNA fragments into larger constructs often involve amplification, handling, purification, transformation, or other storage and delivery steps that can increase uncertainty regarding the quality and quantity of the DNA. Cell-based workflows may then incorporate growth and selection of colonies of cells that express the target constructs, thereby rejecting non-functional or unassembled DNA sequences that may have passed through the workflow up to this point. Still, depending on the approach, the resulting assembled DNA may include added sequences that are not biologically active, as in the case for some methods using restriction enzymes, or scars, such as occur may occur with Golden Gate Assembly [54] or MoClo [55]. Gibson Assembly [56] is scarless, but the yield and specific results may depend on the secondary structure of the DNA fragments. Thus, information to track through the workflow should include the assembly method, sequences required for assembly, and their location along the DNA molecule, such as landing pads, sequences for compatibility with “helper” strains (typically *E. coli* or yeast) used during construction, and DNA secondary structure. Larger DNA constructs are likely to be assembled using cell-based methods such as homologous recombination in yeast [57], and cell-based platforms that support eventual delivery to a variety of final host organisms and cell types would be desirable. Regardless of the specific approach, however, assembly must proceed in a context that supports the packaging, delivery, and eventual expression of the DNA in the target host, which may include imbuing the DNA with the correct epigenetic modifications for expression to achieve the desired phenotype.

All this information must be expressed in the inputs provided from the design stage, and while FASTA or GenBank may be sufficient for a simple sequence, more sophisticated representations, such as GFF or SBOL [29], are needed for hierarchical construction. SBOL can also convey functional information, allowing more manufacturing flexibility in the precise sequence chosen for construction, as well as full details of assembly plans and records [45], such as JGI manages with its BOOST software [53]. Sequence modifications may be expressed in an enhanced sequence encoding language, such as BpForms [58]. Sample inventories of partially assemblies maintained as in vitro constructs or within strains may be exchanged with databases already designed and used for such purposes, such as ICE [59] and SynBioHub [60].

The sequence of the assembled DNA is not typically verified at intermediate steps in the build process. However, good practice dictates that the DNA should be further cloned into a cell of interest, and colonies picked, sometimes at random, and the DNA sequenced to ensure assembly according to the intended design. Rather than sequence the entire DNA assembly, often only the specific region of interest is verified. Throughout the DNA assembly process, an abstracted description of the process would allow for machine

reasoning over the combination of assembly methods and verifications to achieve scale. Sequence verification results are typically produced in FASTQ, which is generally sufficient for smaller constructs. To operate on large-scale genomes, however, it may be important to assemble the information into more comprehensive descriptions of a genome and its variations using representations such as GVF [61] or SBOL.

Suitable options for the delivery of large, assembled DNA constructs and whole genomes are generally lacking. The yield of existing processes, such as electrical and chemical transformation or genome transplantation, could be improved significantly to increase their utility, and a broader range of approaches should be developed for use with any organism and cell type. This may also require identifying new cell-free environments or cell-based chassis for assembling and manipulating DNA that also have compatibility with genome packaging and delivery systems into host organisms. Similarly to DNA assembly, delivery protocols and their associated information regarding number of biological and technical replicate experiments, methods, measurements, and so on, should be available in a machine-readable format. This should include information regarding the host cell, such as its genotype, which is often not fully verified. It is unclear the extent to which the background cellular context in which the assembled DNA is inserted affects its expression and the ultimate phenotype of that cell.

The adoption of best practices from industrial biomanufacturing settings could provide a path forward toward integrating appropriate measurements, process controls, and information handling for large collaborations involved in engineering entire genomes. Appropriate implementation of laboratory information management systems (LIMS) into research settings could go a long way toward organizing, archiving, tracking, and sharing information within and across the individual laboratories of a larger collaboration. Advancing the use of automation to support the build step of the genome engineering workflow requires evaluation of which steps may reduce costs and speed results, the availability of automated methods, ways to effectively share those methods and adapt them across platforms and manufacturers, and straightforward ways to integrate automated steps into a larger automated workflow, as well as integrate that workflow with machine learning. Issues concerning best-practices for cybersecurity, cyberbiosecurity, and biosecurity should also be considered and implemented. Much the same arguments apply for the testing discussed in the next section.

3.3 Testing the function of engineered genomes

Once the final DNA constructs have been integrated to produce an engineered genome in the desired biological context, both the sequence and expression of that genome should be verified to test for the intended phenotype. Collaborating organizations should agree on specific measurements, along with control and calibration measurements, to ensure that results can be compared and used across the participant laboratories towards achieving the final goal. The challenges of sequence verification of the engineered genome will be similar to those described above for intermediate DNA constructs during the build process. In addition, to assess strain fitness and other phenotypic information, any number of omics and phenotypic measurements could be made, and measurements that return meaningful information to inform and advance the overall workflow should be prioritized. It will be necessary to decide at which scales of biological organization, such as single cell or population, and time, from milliseconds to days or years, the expression of the final DNA construct should be assayed to ensure the desired phenotype for the intended purpose. Increasingly, time-dependent attributes of cell processes and phenotype, as related to both gene expression internal to the cell and variable environmental conditions external to the cell, are expected to impact engineering decisions, including at the gigabase scale.

Details of particular assays are beyond the scope of this discussion, but it will likely be useful to develop standards and measurement assurance targeted for testing engineered genomes, for example, to help identify predictive relationships between genotype and phenotype or determine contributions of biological stochasticity and measurement uncertainty to overall variability in a measured trait. Standard protocols, reference cell lines, and the use of experimental design are examples of tools available to increase the rigor and confidence in conclusions drawn from testing. The development of methods for absolute quantitation would complement common relative measurements and facilitate calibration for comparability across measurement techniques and measurands along the central dogma. Calibration of biological assays will also aid in comparing results within a single laboratory over time and across different laboratories. Recent studies, for example, for fluorescence [62, 63], absorbance [64], and RNAseq [65] measurements, demonstrate the possibility of realizing scalable and cost-effective comparability in biological measurements. Best practices

from biomanufacturing may be of use in informing all of these decisions.

DNA constructs are often evaluated for their associated growth phenotypes to determine the nature and extent of unexpected consequences for cell function and fitness due to the revised genome sequence. This is complicated by the need for accepted definitions of fitness, along with methods to measure and quantify fitness. Often, determining fitness effects also involve measurements of metabolic burden. Guidance on how to choose the best measure of fitness for a specific purpose or to speed the workflow would be helpful. Engineered cell lines should also be evaluated for robustness to changes in the environmental context the cells are likely to experience during typical use in the intended application, as well as stability over relevant timescales to evolution or adaptation.

For a large collaboration, it will also be critical to establish shared representations and practices for metadata, process controls, and calibration. Both the target fitness specifications and the protocols to evaluate them will need to be described in sufficient detail to enable automation-assisted integration and comparison of data, metadata, process controls, and calibration across laboratories. These practices will help to ensure that testing includes measurements compatible with learning through modeling and simulation. Although some measurements will serve both test and learn portions of the engineering workflow, such as high-throughput metabolomics, some are likely to be different for testing the assembled DNA function and representing that function in a model. In general, however, we expect that it will be valuable to have sufficient metadata to track the data, conditions, and samples all the way back to the designs, in a machine-readable manner to allow operations at scale. Appropriate LIMS tooling and curation assistance software (e.g., [66]) will be vital for enabling such metadata to be created consistently, correctly, and in a timely fashion, by limiting the required input from human investigators.

3.4 Learning systematically from test results

Given our current limited knowledge of biology and current limited capabilities to design biology, learning the functions that genomes encode and learning how to design genomes will be critical to gigabase-scale genome engineering. This includes learning to predict the behavior of synthetic genomes, such as emergent interactions between new metabolic byproducts and existing machinery; learning fundamental biology, such as the functions of small non-coding RNA; learning design heuristics, such as preferentially using parts that only have one function; and learning better methods and processes, such as how to share information among teams of engineers. Developing the capabilities to predict and design the expression of genomes will likely become increasingly essential as our capability to construct synthetic organisms matures.

Design-build-test-learn workflows are an effective way to learn fundamental biology, design principles, and improved processes. The ideal workflow starts by using a predictive computational model, such as a whole-cell model [48] or a whole-organism model [67], to design a genome; proceeds by constructing, booting up, and testing this genome; and concludes by systematically learning from behavioral failures of the genome by using the data generated from testing the genome to improve the model, our biological knowledge, and the methods used to design, construct, and boot up the genome. One way to systematically learn better models is to identify the minimum set of changes that must be made to the model to align its predictions with observed behaviors of the synthetic genome. To realize these models, we should develop new experimental methods for better characterizing the relationship between genotype and phenotype, new tools discovering and aggregating the data needed for modeling (building on foundations such as the workflow model introduced in SBOL 2.2 [45], and ontology resources such as OBO [68] and the Experimental Factor Ontology [69]), new formalisms for modeling and simulating the combinatorial complexity of biology and the multiple scales between genomes and organismal behavior, new methods for high-performance simulation of large models, new tools verifying models, new frameworks for collaboratively building models, and new representations for the semantic meaning and provenance of models [70, 71]. To automatically learn these models from test data, we should develop new repositories of models of individual biological parts that can be composed into models [72]; new methods for generating variants of models that explain new observations by incorporating models of additional parts, alternative kinetic laws, or alternative parameter values; and develop new model selection techniques for non-linear multiscale models [73].

Until we have comprehensive predictive models, engineers will likely rely on ad hoc combinations of predictive models of parts of organisms, data-driven models, and heuristic design rules. BioModels [74], the NeuroML database [75], Open Source Brain [76], and the Physiome Model Repository [77] already contain

hundreds of models of parts of organisms that can be used to help design organisms. For example, metabolic engineers often use constraint-based models to design microbial factories [78]. Numerous data-driven models of parts of organisms can also already be used to help design synthetic organisms. For example, PSORTb [79] can help bioengineers design signal sequences to traffic proteins to specific compartments. Several design heuristics are also already commonly used. For example, bioengineers often optimize the GC content of parts to minimize their metabolic load on their host [80].

To close the design-build-test-learn cycle, the community should describe these models and heuristics with standard formats and share these models through public repositories. CellML [81], NeuroML [82], and Systems Biology Markup Language [83] (SBML) formats can already be used to describe several types of biochemical and physiological models, and the BioModels, NeuroML, Open Source Brain, and Physiome repositories can already be used to share models. Similarly, Docker containers and the DockerHub repository [84] can already be used to share data-driven models. New formats for describing multiscale models should be developed to help bioengineers use these models to design genomes. In particular, these formats should support more comprehensive and more detailed models. In addition, these formats should facilitate collaboration by concretely capturing the semantic meaning of each model component and capturing the provenance of each model element, including the data sources, assumptions, and design decisions that motivated each element. Finally, to help bioengineers reliably design genomes, these formats should also capture the predictive performance of models, including the behaviors, biological mechanisms, and environments which they can and cannot capture.

3.5 Coordinating complex workflows

Engineering gigabase genomes will require coordinating numerous heterogeneous tasks into clear, cohesive, reproducible workflows [85, 86]. Several technologies have already been developed to coordinate experimental protocols, computational workflows, and other complex tasks. Laboratory automation systems enable investigators to design and execute complex experimental protocols that involve numerous individual tasks, reagents, equipment, and personnel [87]. To execute protocols more reproducibly than possible with manual approaches, these systems capture the details of each task, such as the amount of each reagent, the order of tasks, and the results of each protocol. These systems can also be integrated with LIMS [88] to help track workflows and reagent stocks. Individual tasks could be described with languages such as Autoprotocol [89]. Although lab automation has not yet been widely adopted, some researchers are already using lab automation systems, such as Aquarium [90] and Antha [91], to engineer yeast [92]. As these systems improve and as their interfaces are standardized and integrated, the community should adopt these tools for specifying build and test protocols. This would help large numbers of researchers execute standardized protocols collaboratively.

On the informational side, computational workflow engines aim to enable the clear specification and reproducible execution of complex computational workflows that involve multiple steps and require multiple software programs and computing environments. Like lab automation systems, workflow engines aim to capture the details of each task, the relationships among the tasks, and the dependencies of the tasks so that workflows can be shared among investigators and reproducibly executed. Several workflow engines such as Cromwell [93], Galaxy [94], NextFlow [95], and Toil [96] have been developed for genomics. In addition, the Common Workflow Language [97] (CWL) enables workflows to be exchanged among several of the most popular engines. Furthermore, the Dockstore [98] and MyExperiment [99] repositories enable researchers to publish workflows. These tools should be adapted for gigabase engineering workflows. To make genome engineering workflows easy to understand, we recommend that the community develop a graphical tool for designing genome engineering workflows and an ontology for annotating the semantic meaning of workflow tasks. To make genome engineering workflows easy to reproduce and reuse, we recommend that the community include CWL files in COMBINE archives [100], develop REST or other programmatic interfaces for all of the databases involved in genome engineering workflows, containerize [101] all of the computational tools needed for genome engineering, and deposit these containers to a registry such as DockerHub [84].

Issue tracking systems, such as GitHub issues [102], can help teams coordinate the execution of complex tasks which are difficult to describe and automate. For example, software development teams frequently use issue tracking systems to coordinate the development of features of software. Issue tracking systems could be readily adapted to coordinate complex tasks involved in designing genomes and learning from tests of

genome designs that require expert analysis. Furthermore, the PROV ontology [103] could be adopted to capture the provenance of these tasks, and to link information throughout the design-build-test-learn cycles, as is now possible in SBOL [45].

3.6 Sharing data through common repositories

Once links are established across different portions of a workflow, unified access to information in databases for various institutions and stages of the workflow can be accomplished using standard federation methods and any of the various mature open tools for database management systems (DBMS).

To support collaborative sharing, genome-engineering consortia should adopt principles of FAIR (findable, accessible, interoperable, reproducible) data management. FAIR principles put specific emphasis on enhancing the ability of machines to automatically find and use the data, in addition to supporting its reuse by individual people. Some software supporting such principles for biological data sharing already exists, such as FAIRDOMHub [104] and EDD [105], and may be adopted for this purpose.

Ultimately, the community should incorporate every aspect of the design-build-test-learn cycle into integrated, reproducible workflows. Such integrated workflows would help teams scale to gigabase genomes, develop extensible workflows that can be adapted for similar engineering projects, and conduct genome engineering transparently and reproducibly. Achieving this will require integrating lab automation systems for systematically executing build and test tasks, workflow engines for systematically executing design and learn tasks, and potentially issue tracking systems for managing more sophisticated tasks that require expert input.

3.7 Laws and Contracts

Few resources are available that can be adopted for coordinating the legal and contractual interactions of a large-scale genome-engineering project. However, some precedents could serve as the foundation for such resources. In the realm of copyright law, the Open Source Initiative (OSI) and other software organizations have systematized licensing to provide compatible families of licenses, and these are used widely to promote collaboration and composability of software. These licences can be classified into four tiers: public domain, permissive (e.g., BSD), copy-left (e.g., GPL) and proprietary. Similarly, the Creative Commons family of licenses provide a simple framework for media and other content sharing that has been widely adopted and is commonly incorporated into tooling. With just a single badge in such an established system of licensing, a user or a machine can tell quickly if an object can be reused, if its reuse is prohibited, or if more complicated negotiation or determination is required.

For dealing with physical biological materials, the first standardized materials transfer agreement was the NIH's Uniform Biological Materials Transfer Agreement (UBMTA), released in 1995. The Addgene public DNA repository uses this agreement, which is widely supported by universities and enables many legal transactions to be automated, although not for commercial uses. Broader and more compatible MTA systems have been developed by the Science Commons project [106], and more recently by the OpenMTA [107]. No publicly available system yet supports automation for management of protections for proprietary genetic materials and reagents, however, and there are also significant considerations to be managed regarding compliance with local regulatory and legal systems, particularly when materials cross international borders.

Developing a machine-compatible system of protections for scientific data and materials would enable transfers with automated systems and reduce institutional friction. In order for such a system to work, the genome-engineering community should define tiered levels of IP protections that are simultaneously intelligible for the common user, legal experts, and computer systems. Enabling effective use of such systems with automation-assisted workflows will also require recording provenance information about which inputs are involved in the production of results, using mechanisms such as the PROV ontology [103]. Finally, large-scale data sharing will also need to assist researchers and engineers in managing the level of exposure of information, whether due to issues of privacy, safety, publication priority, or other similar concerns. Again, no current systems exist, but a basis for developing them may be found in the cross-domain information sharing protocols that have been developed in other domains [108, 109].

4 Recommendations and Outlook

As discussed above, scaling up to gigabase genomes presents a wide range of immediate and long-term challenges (Table 1). These cluster into four general areas, each with a different set of needs and paths for development:

- **Representing and exchanging designs, plans, data, metadata, and knowledge:** Managing information for gigabase genomic design requires addressing many challenges regarding scale, representation, and standards. Relatively mature technologies exist to address most individual needs, as well as to assist with the integration of workflows. The practical implementation of effective workflows will require significant investment in building infrastructure and tools that adopt these technologies, including domain-specific extensions and refinements.
- **Sharing and integrating data quality and experimental measurements:** Sharing and integrating information arising from measurements of biological material poses significant challenges. It remains unclear what information would be advantageous to share, given the difficulty of obtaining and interpreting measurements of biological systems and the expense and unfavorable scaling of data curation. However, effective integration depends on associating measurement data with well-curated knowledge and metadata in compatible representations. A number of potential solutions exist for each of these, but significant investment will be needed to investigate how the state of the art can be extended to satisfactorily address these needs.
- **Integration of modeling and design at the gigabase genomic scale:** Considerable challenges surround efforts to develop a deeper understanding of the relationship between genotype and phenotype, regarding both the interpretation of experimental data and the application of that data to create and validate models, which may be applied in computer-assisted design. Long-term investment in fundamental research is needed, and the suite of biological systems of varying complexity, from cell-free systems to minimal and synthetic cells to natural living systems, may offer suitable experimental platforms to match genes with specific functions.
- **Technical support for Ethical, Legal, and Societal Implications (ELSI) and Intellectual Property (IP) at scale:** At the gigabase scale, computer-assisted workflows will be necessary to manage contracts, intellectual property, materials transfer, and other legal and societal interactions. Such workflows will need to be developed by interdisciplinary teams involving experts in law, ELSI issues, software engineering, and knowledge representation. Moreover, it will be critical to address these issues early.

In short, engineering gigabase-scale genomes presents significant challenges that will require coordinated investment to overcome. Because many other areas of bioscience face similar challenges, solutions to these challenges will likely also benefit the broader bioscience community.

Importantly, the challenges of scale, integration, and lack of knowledge faced in genomic engineering are not fundamentally different in nature than those that have been overcome previously in other engineering ventures, such as aerospace engineering and microchip design, which required organizing humans and sharing information across many institutions over time. Thus, we expect to be able to adapt solutions from these other fields for genome engineering.

Investment in capabilities for genomic engineering workflows is critical to move from a world in which genome engineering is a heroic effort to one in which genome engineering is routine, safe, and reliable. Investment in workflows for genomic engineering will support and enable a vast number of projects, including many not yet conceived, as was the case for reading the human genome. As workflow technologies improve, we anticipate that the trends of Figure 1B will eventually reverse, enabling high-fidelity whole-genome engineering at a modest cost.

5 Competing Interests

The authors declare no competing interests.

6 Acknowledgements and Disclaimers

This work was supported in part by NIH awards P41-EB023912 and R35-GM119771 and by NSF awards 1548123 and 1522074. The views, opinions, and/or findings expressed are those of the author(s) and should not be interpreted as representing the official views or policies of these funding agencies or the U.S. Government. This document does not contain technology or technical data controlled under either U.S. International Traffic in Arms Regulation or U.S. Export Administration Regulations.

Certain commercial equipment, instruments, or materials are identified to adequately specify experimental procedures. Such identification neither implies recommendation nor endorsement by the National Institute of Standards and Technology nor that the equipment, instruments, or materials identified are necessarily the best for the purpose.

7 Author Contributions

B.A.B., J.B., J.R.K., and E.A.S equally contributed to the article.

8 References

1. Boeke, JD, Church, G, Hessel, A, Kelley, NJ, Arkin, A, Cai, Y, Carlson, R, Chakravarti, A, Cornish, VW, Holt, L, *et al.* The Genome Project-Write. *Science* **353**, 126–127 (2016).
2. Carlson, RH. *Biology Is Technology : The Promise, Peril, and New Business of Engineering Life* (Harvard University Press, 2011).
3. Hughes, RA & Ellington, AD. Synthetic DNA synthesis and assembly: putting the synthetic in synthetic biology. *Cold Spring Harbor Perspectives in Biology* **9**, a023812 (2017).
4. Chari, R & Church, GM. Beyond editing to writing large genomes. *Nature Reviews Genetics* **18**, 749 (2017).
5. Khorana, HG. Total synthesis of a gene. *Science* **203**, 614–625 (1979).
6. Mandeck, W, Hayden, MA, Shallcross, MA & Stotland, E. A totally synthetic plasmid for general cloning, gene expression and mutagenesis in *Escherichia coli*. *Gene* **94**, 103–107 (1990).
7. Stemmer, WP, Cramer, A, Ha, KD, Brennan, TM & Heyneker, HL. Single-step assembly of a gene and entire plasmid from large numbers of oligodeoxyribonucleotides. *Gene* **164**, 49–53 (1995).
8. Cello, J, Paul, AV & Wimmer, E. Chemical synthesis of poliovirus cDNA: generation of infectious virus in the absence of natural template. *Science* **297**, 1016–1018 (2002).
9. Tian, J, Gong, H, Sheng, N, Zhou, X, Gulari, E, Gao, X & Church, G. Accurate multiplex gene synthesis from programmable DNA microchips. *Nature* **432**, 1050 (2004).
10. Kodumal, SJ, Patel, KG, Reid, R, Menzella, HG, Welch, M & Santi, DV. Total synthesis of long DNA sequences: synthesis of a contiguous 32-kb polyketide synthase gene cluster. *Proceedings of the National Academy of Sciences* **101**, 15573–15578 (2004).
11. Gibson, DG, Benders, GA, Andrews-Pfannkoch, C, Denisova, EA, Baden-Tillson, H, Zaveri, J, Stockwell, TB, Brownley, A, Thomas, DW, Algire, MA, *et al.* Complete chemical synthesis, assembly, and cloning of a *Mycoplasma genitalium* genome. *Science* **319**, 1215–1220 (2008).
12. Gibson, DG, Glass, JI, Lartigue, C, Noskov, VN, Chuang, R-Y, Algire, MA, Benders, GA, Montague, MG, Ma, L, Moodie, MM, *et al.* Creation of a bacterial cell controlled by a chemically synthesized genome. *Science* **329**, 52–56 (2010).
13. 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* **353**, 819–822 (2016).

14. Fredens, J, Wang, K, de la Torre, D, Funke, LF, Robertson, WE, Christova, Y, Chia, T, Schmied, WH, Dunkelmann, DL, Beránek, V, *et al.* Total synthesis of *Escherichia coli* with a recoded genome. *Nature* **569**, 514 (2019).
15. Lau, YH, Stirling, F, Kuo, J, Karrenbelt, MA, Chan, YA, Riesselman, A, Horton, CA, Schäfer, E, Lips, D, Weinstock, MT, *et al.* Large-scale recoding of a bacterial genome by iterative recombineering of synthetic DNA. *Nucleic Acids Research* **45**, 6971–6980 (2017).
16. Dymond, JS, Richardson, SM, Coombes, CE, Babatz, T, Muller, H, Annaluru, N, Blake, WJ, Schwerzmann, JW, Dai, J, Lindstrom, DL, *et al.* Synthetic chromosome arms function in yeast and generate phenotypic diversity by design. *Nature* **477**, 471 (2011).
17. Annaluru, N, Muller, H, Mitchell, LA, Ramalingam, S, Stracquadanio, G, Richardson, SM, Dymond, JS, Kuang, Z, Scheifele, LZ, Cooper, EM, *et al.* Total synthesis of a functional designer eukaryotic chromosome. *Science* **344**, 55–58 (2014).
18. 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* **355**, 1040–1044 (2017).
19. Write Leadership Group, TG. *Ultra-safe Cells Resistant to Natural Viruses Announced as First GP-write Grand-scale Project* 2018. <<http://www.engineeringbiologycenter.org/press/may2018.pdf>>.
20. Ostrov, N, Beal, J, Ellis, T, Gordon, DB, Karas, BJ, Lee, HH, Lenaghan, S, Schloss, JA, Stracquadanio, G, Trefzer, A, Bader, JS, Church, GM, Coelho, CM, Efcavitch, JW, Guell, M, Mitchell, LA, Nielsen, AA, Peck, B, Smith, AC, C. Neal Stewart, J & Tekotte, H. A technology roadmap for writing genomes. *in revision* (TBD).
21. Appleton, E, Densmore, D, Madsen, C & Roehner, N. Needs and opportunities in bio-design automation: four areas for focus. *Current Opinion in Chemical Biology* **40**, 111–118 (2017).
22. Carbonell, P, Jervis, AJ, Robinson, CJ, Yan, C, Dunstan, M, Swainston, N, Vinaixa, M, Hollywood, KA, Currin, A, Rattray, NJ, *et al.* An automated design-build-test-learn pipeline for enhanced microbial production of fine chemicals. *Communications Biology* **1**, 66 (2018).
23. Gill, RT, Halweg-Edwards, AL, Clauset, A & Way, SF. Synthesis aided design: The biological design-build-test engineering paradigm? *Biotechnology and Bioengineering* **113**, 7–10 (2016).
24. Poust, S, Hagen, A, Katz, L & Keasling, JD. Narrowing the gap between the promise and reality of polyketide synthases as a synthetic biology platform. *Current Opinion in Biotechnology* **30**, 32–39 (2014).
25. Pouvreau, B, Vanhercke, T & Singh, S. From plant metabolic engineering to plant synthetic biology: The evolution of the design/build/test/learn cycle. *Plant Science* **273**, 3–12 (2018).
26. Hutchison, CA, Chuang, R-Y, 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* **351**, aad6253 (2016).
27. Purnick, PE & Weiss, R. The second wave of synthetic biology: from modules to systems. *Nature Reviews Molecular Cell Biology* **10**, 410 (2009).
28. Galdzicki, M, Clancy, KP, Oberortner, E, Pocock, M, Quinn, JY, Rodriguez, CA, Roehner, N, Wilson, ML, Adam, L, Anderson, JC, *et al.* The Synthetic Biology Open Language (SBOL) provides a community standard for communicating designs in synthetic biology. *Nature Biotechnology* **32**, 545 (2014).
29. Roehner, N, Beal, J, Clancy, K, Bartley, B, Misirli, G, Grnberg, R, Oberortner, E, Pocock, M, Bissell, M, Madsen, C, *et al.* Sharing structure and function in biological design with SBOL 2.0. *ACS Synthetic Biology* **5**, 498–506 (2016).
30. Steen, EJ, Chan, R, Prasad, N, Myers, S, Petzold, CJ, Redding, A, Ouellet, M & Keasling, JD. Metabolic engineering of *Saccharomyces cerevisiae* for the production of n-butanol. *Microbial Cell Factories* **7**, 36 (2008).

31. Sun, W, Starly, B, Nam, J & Darling, A. Bio-CAD modeling and its applications in computer-aided tissue engineering. *Computer-Aided Design* **37**, 1097–1114 (2005).
32. Cochrane, G, Karsch-Mizrachi, I, Takagi, T & Sequence Database Collaboration, IN. The International Nucleotide Sequence Database collaboration. *Nucleic Acids Research* **44**, D48–D50 (2015).
33. Cunningham, F, Achuthan, P, Akanni, W, Allen, J, Amode, MR, Armean, IM, Bennett, R, Bhai, J, Billis, K, Boddu, S, *et al.* Ensembl 2019. *Nucleic Acids Research* **47**, D745–D751 (2018).
34. Kersey, PJ, Allen, JE, Armean, I, Boddu, S, Bolt, BJ, Carvalho-Silva, D, Christensen, M, Davis, P, Falin, LJ, Grabmueller, C, *et al.* Ensembl Genomes 2016: more genomes, more complexity. *Nucleic Acids Research* **44**, D574–D580 (2015).
35. O’Leary, NA, Wright, MW, Brister, JR, Ciuffo, S, Haddad, D, McVeigh, R, Rajput, B, Robbertse, B, Smith-White, B, Ako-Adjei, D, *et al.* Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Research* **44**, D733–D745 (2015).
36. Haft, DH, DiCuccio, M, Badretdin, A, Brover, V, Chetvernin, V, O’Neill, K, Li, W, Chitsaz, F, Derbyshire, MK, Gonzales, NR, *et al.* RefSeq: an update on prokaryotic genome annotation and curation. *Nucleic Acids Research* **46**, D851–D860 (2017).
37. Mashima, J, Kodama, Y, Kosuge, T, Fujisawa, T, Katayama, T, Nagasaki, H, Okuda, Y, Kaminuma, E, Ogasawara, O, Okubo, K, *et al.* DNA Data Bank of Japan (DDBJ) progress report. *Nucleic Acids Research* **44**, D51–D57 (2015).
38. National Center for Biotechnology Information. *Genome Data Viewer* <<https://www.ncbi.nlm.nih.gov/genome/gdv/>> (2019).
39. Liao, Y, Wang, J, Jaehnig, EJ, Shi, Z & Zhang, B. WebGestalt 2019: gene set analysis toolkit with revamped UIs and APIs. *Nucleic Acids Research* (2019).
40. Huang, DW, Sherman, BT & Lempicki, RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature Protocols* **4**, 44 (2009).
41. Frankish, A, Uszczyńska, B, Ritchie, GR, Gonzalez, JM, Pervouchine, D, Petryszak, R, Mudge, JM, Fonseca, N, Brazma, A, Guigo, R, *et al.* Comparison of GENCODE and RefSeq gene annotation and the impact of reference geneset on variant effect prediction. *BMC Genomics* **16**, S2 (2015).
42. McCarthy, DJ, Humburg, P, Kanapin, A, Rivas, MA, Gaulton, K, Cazier, J-B & Donnelly, P. Choice of transcripts and software has a large effect on variant annotation. *Genome Medicine* **6**, 26 (2014).
43. Chibucos, MC, Siegele, DA, Hu, JC & Giglio, M. in *The Gene Ontology Handbook* 245–259 (Humana Press, New York, NY, 2017).
44. Eilbeck, K, Lewis, SE, Mungall, CJ, Yandell, M, Stein, L, Durbin, R & Ashburner, M. The Sequence Ontology: a tool for the unification of genome annotations. *Genome biology* **6**, R44 (2005).
45. Cox, RS, Madsen, C, McLaughlin, JA, Nguyen, T, Roehner, N, Bartley, B, Beal, J, Bissell, M, Choi, K, Clancy, K, *et al.* Synthetic Biology Open Language (SBOL) version 2.2. 0. *Journal of Integrative Bioinformatics* **15** (2018).
46. Swainston, N, Smallbone, K, Hefzi, H, Dobson, PD, Brewer, J, Hanscho, M, Zielinski, DC, Ang, KS, Gardiner, NJ, Gutierrez, JM, *et al.* Recon 2.2: from reconstruction to model of human metabolism. *Metabolomics* **12**, 109 (2016).
47. Monk, JM, Lloyd, CJ, Brunk, E, Mih, N, Sastry, A, King, Z, Takeuchi, R, Nomura, W, Zhang, Z, Mori, H, *et al.* iML1515, a knowledgebase that computes Escherichia coli traits. *Nature Biotechnology* **35**, 904 (2017).
48. 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* **150**, 389–401 (2012).
49. Markowitz, VM, Chen, I-MA, Palaniappan, K, Chu, K, Szeto, E, Grechkin, Y, Ratner, A, Jacob, B, Huang, J, Williams, P, *et al.* IMG: the integrated microbial genomes database and comparative analysis system. *Nucleic Acids Research* **40**, D115–D122 (2011).

50. Overbeek, R, Olson, R, Pusch, GD, Olsen, GJ, Davis, JJ, Disz, T, Edwards, RA, Gerdes, S, Parrello, B, Shukla, M, *et al.* The SEED and the Rapid Annotation of microbial genomes using Subsystems Technology (RAST). *Nucleic Acids Research* **42**, D206–D214 (2013).
51. Venkatachalam, A, Mellichamp, JM & Miller, DM. A knowledge-based approach to design for manufacturability. *Journal of Intelligent Manufacturing* **4**, 355–366 (1993).
52. Abrantes, R, Basto, M, Varajão, J, Magalhães, L, Ribeiro, P & Freitas, LC. *Rule ontology for automatic design verification application to PCB manufacturing and assembly* in *IECON 2017-43rd Annual Conference of the IEEE Industrial Electronics Society* (2017), 3403–3409.
53. Oberortner, E, Cheng, J-F, Hillson, NJ & Deutsch, S. Streamlining the design-to-build transition with build-optimization software tools. *ACS Synthetic Biology* **6**, 485–496 (2016).
54. Engler, C, Gruetzner, R, Kandzia, R & Marillonnet, S. Golden gate shuffling: a one-pot DNA shuffling method based on type II restriction enzymes. *PloS One* **4**, e5553 (2009).
55. Weber, E, Engler, C, Gruetzner, R, Werner, S & Marillonnet, S. A modular cloning system for standardized assembly of multigene constructs. *PloS One* **6**, e16765 (2011).
56. Gibson, DG, Young, L, Chuang, R-Y, Venter, JC, Hutchison III, CA & Smith, HO. Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nature Methods* **6**, 343 (2009).
57. Gibson, DG, Benders, GA, Axelrod, KC, Zaveri, J, Algire, MA, Moodie, M, Montague, MG, Venter, JC, Smith, HO & Hutchison, CA. One-step assembly in yeast of 25 overlapping DNA fragments to form a complete synthetic *Mycoplasma genitalium* genome. *Proceedings of the National Academy of Sciences* **105**, 20404–20409 (2008).
58. Lang, PF, Chebaro, Y & Karr, JR. BpForms: a toolkit for concretely describing modified DNA, RNA and proteins. *arXiv preprint arXiv:1903.10042* (2019).
59. Ham, TS, Dmytriv, Z, Plahar, H, Chen, J, Hillson, NJ & Keasling, JD. Design, implementation and practice of JBEI-ICE: an open source biological part registry platform and tools. *Nucleic Acids Research* **40**, e141–e141 (2012).
60. McLaughlin, JA, Myers, CJ, Zundel, Z, Misrlh, G, Zhang, M, Ofiteru, ID, Goi-Moreno, A & Wipat, A. SynBioHub: a standards-enabled design repository for synthetic biology. *ACS Synthetic Biology* **7**, 682–688 (2018).
61. Reese, MG, Moore, B, Batchelor, C, Salas, F, Cunningham, F, Marth, GT, Stein, L, Flicek, P, Yandell, M & Eilbeck, K. A standard variation file format for human genome sequences. *Genome Biology* **11**, R88 (2010).
62. Wang, L & Hoffman, RA. Standardization, calibration, and control in flow cytometry. *Current Protocols in Cytometry* **79**, 1–3 (2017).
63. Beal, J, Haddock-Angelli, T, Baldwin, G, Gershater, M, Dwijayanti, A, Storch, M, de Mora, K, Lizarazo, M, Rettberg, R, *et al.* Quantification of bacterial fluorescence using independent calibrants. *PloS One* **13**, e0199432 (2018).
64. Stevenson, K, McVey, AF, Clark, IB, Swain, PS & Pilizota, T. General calibration of microbial growth in microplate readers. *Scientific Reports* **6**, 38828 (2016).
65. Lee, H, Pine, PS, McDaniel, J, Salit, M & Oliver, B. External RNA controls consortium beta version update. *Journal of Genomics* **4**, 19 (2016).
66. Wolstencroft, K, Owen, S, Horridge, M, Krebs, O, Mueller, W, Snoep, JL, du Preez, F & Goble, C. RightField: embedding ontology annotation in spreadsheets. *Bioinformatics* **27**, 2021–2022 (2011).
67. Sarma, GP, Lee, CW, Portegys, T, Ghayoomie, V, Jacobs, T, Alicea, B, Cantarelli, M, Currie, M, Gerkin, RC, Gingell, S, *et al.* OpenWorm: overview and recent advances in integrative biological simulation of *Caenorhabditis elegans*. *Philosophical Transactions of the Royal Society B* **373**, 20170382 (2018).
68. Smith, B, Ashburner, M, Rosse, C, Bard, J, Bug, W, Ceusters, W, Goldberg, LJ, Eilbeck, K, Ireland, A, Mungall, CJ, *et al.* The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nature Biotechnology* **25**, 1251 (2007).

69. Malone, J, Holloway, E, Adamusiak, T, Kapushesky, M, Zheng, J, Kolesnikov, N, Zhukova, A, Brazma, A & Parkinson, H. Modeling sample variables with an Experimental Factor Ontology. *Bioinformatics* **26**, 1112–1118 (2010).
70. Goldberg, AP, Szigeti, B, Chew, YH, Sekar, JA, Roth, YD & Karr, JR. Emerging whole-cell modeling principles and methods. *Current Opinion in Biotechnology* **51**, 97–102 (2018).
71. Szigeti, B, Roth, YD, Sekar, JA, Goldberg, AP, Pochiraju, SC & Karr, JR. A blueprint for human whole-cell modeling. *Current Opinion in Systems Biology* **7**, 8–15 (2018).
72. Cooling, MT, Rouilly, V, Misirli, G, Lawson, J, Yu, T, Hallinan, J & Wipat, A. Standard virtual biological parts: a repository of modular modeling components for synthetic biology. *Bioinformatics* **26**, 925–931 (2010).
73. Kirk, P, Thorne, T & Stumpf, MP. Model selection in systems and synthetic biology. *Current Opinion in Biotechnology* **24**, 767–774 (2013).
74. Glont, M, Nguyen, TVN, Graesslin, M, Hälke, R, Ali, R, Schramm, J, Wimalaratne, SM, Kothamachu, VB, Rodriguez, N, Swat, MJ, *et al.* BioModels: expanding horizons to include more modelling approaches and formats. *Nucleic Acids Research* **46**, D1248–D1253 (2017).
75. Crook, SM & Dietrich, S. Model exchange with the NeuroML model database. *BMC Neuroscience* **15**, P171 (2014).
76. Gleeson, P, Cantarelli, M, Marin, B, Quintana, A, Earnshaw, M, Sadeh, S, Piasini, E, Birgiolas, J, Cannon, RC, Cayco-Gajic, NA, *et al.* Open Source Brain: A Collaborative Resource for Visualizing, Analyzing, Simulating, and Developing Standardized Models of Neurons and Circuits. *Neuron* (2019).
77. Yu, T, Lloyd, CM, Nickerson, DP, Cooling, MT, Miller, AK, Garny, A, Terkildsen, JR, Lawson, J, Britten, RD, Hunter, PJ, *et al.* The Physiome Model Repository 2. *Bioinformatics* **27**, 743–744 (2011).
78. Kim, HU, Kim, TY & Lee, SY. Metabolic flux analysis and metabolic engineering of microorganisms. *Molecular BioSystems* **4**, 113–120 (2008).
79. Yu, NY, Wagner, JR, Laird, MR, Melli, G, Rey, S, Lo, R, Dao, P, Sahinalp, SC, Ester, M, Foster, LJ, *et al.* PSORTb 3.0: improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes. *Bioinformatics* **26**, 1608–1615 (2010).
80. Villalobos, A, Ness, JE, Gustafsson, C, Minshull, J & Govindarajan, S. Gene Designer: a synthetic biology tool for constructing artificial DNA segments. *BMC Bioinformatics* **7**, 285 (2006).
81. Cuellar, A, Hedley, W, Nelson, M, Lloyd, C, Halstead, M, Bullivant, D, Nickerson, D, Hunter, P & Nielsen, P. The CellML 1.1 specification. *Journal of Integrative Bioinformatics* **12**, 4–85 (2015).
82. Gleeson, P, Crook, S, Silver, A & Cannon, R. Development of NeuroML version 2.0: greater extensibility, support for abstract neuronal models and interaction with systems biology languages. *BMC Neuroscience* **12**, P29 (2011).
83. Hucka, M, Finney, A, Sauro, HM, Bolouri, H, Doyle, JC, Kitano, H, Arkin, AP, Bornstein, BJ, Bray, D, Cornish-Bowden, A, *et al.* The Systems Biology Markup Language (SBML): a medium for representation and exchange of biochemical network models. *Bioinformatics* **19**, 524–531 (2003).
84. Docker Inc. *DockerHub* <<https://hub.docker.com>> (2019).
85. Myers, CJ, Beal, J, Goroehowski, TE, Kuwahara, H, Madsen, C, McLaughlin, JA, Misirli, G, Nguyen, T, Oberortner, E, Samineni, M, *et al.* A standard-enabled workflow for synthetic biology. *Biochemical Society Transactions* **45**, 793–803 (2017).
86. Goñi Moreno, A, Carcajona, M, Kim, J, Martínez-García, E, Amos, M & de Lorenzo, V. An implementation-focused bio/algorithmic workflow for synthetic biology. *ACS Synthetic Biology* **5**, 1127–1135 (2016).
87. Sadowski, MI, Grant, C & Fell, TS. Harnessing QbD, programming languages, and automation for reproducible biology. *Trends in Biotechnology* **34**, 214–227 (2016).
88. Prasad, PJ & Bodhe, G. Trends in laboratory information management system. *Chemometrics and Intelligent Laboratory Systems* **118**, 187–192 (2012).

89. Miles, B & Lee, PL. Achieving reproducibility and closed-loop automation in biological experimentation with an IoT-enabled Lab of the future. *SLAS Technology* **23**, 432–439 (2018).
90. Keller, B, Vrana, J, Miller, A, Newman, G & Klavins, E. Aquarium: The laboratory operating system version 2.6.0. doi:10.5281/zenodo.2583232 (2019).
91. Synthace. *Antha* <<http://www.antha-lang.org>> (2019).
92. Yang, Y, Nemhauser, JL & Klavins, E. Synthetic bistability and differentiation in yeast. *ACS Synthetic Biology* (2019).
93. Broad Institute. *The Workflow Description Language and Cromwell* <<https://software.broadinstitute.org/wdl>> (2019).
94. Goecks, J, Nekrutenko, A & Taylor, J. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biology* **11**, R86 (2010).
95. Di Tommaso, P, Chatzou, M, Floden, EW, Barja, PP, Palumbo, E & Notredame, C. Nextflow enables reproducible computational workflows. *Nature Biotechnology* **35**, 316 (2017).
96. Vivian, J, Rao, AA, Nothhaft, FA, Ketchum, C, Armstrong, J, Novak, A, Pfeil, J, Narkizian, J, Deran, AD, Musselman-Brown, A, *et al.* Toil enables reproducible, open source, big biomedical data analyses. *Nature Biotechnology* **35**, 314 (2017).
97. Amstutz, P, Crusoe, MR, Tijani, N, Chapman, B, Chilton, J, Heuer, M, Kartashov, A, Leehr, D, Mnager, H, Nedeljkovich, M & *et al.* *Common Workflow Language, v1.0* 2016. doi:10.6084/m9.figshare.3115156.v2. <https://figshare.com/articles/Common_Workflow_Language_draft_3/3115156/2>.
98. O'Connor, BD, Yuen, D, Chung, V, Duncan, AG, Liu, XK, Patricia, J, Paten, B, Stein, L & Ferretti, V. The Dockstore: enabling modular, community-focused sharing of Docker-based genomics tools and workflows. *F1000Research* **6** (2017).
99. Goble, CA, Bhagat, J, Aleksejevs, S, Cruickshank, D, Michaelides, D, Newman, D, Borkum, M, Bechhofer, S, Roos, M, Li, P, *et al.* myExperiment: a repository and social network for the sharing of bioinformatics workflows. *Nucleic Acids Research* **38**, W677–W682 (2010).
100. Bergmann, FT, Adams, R, Moodie, S, Cooper, J, Glont, M, Golebiewski, M, Hucka, M, Laibe, C, Miller, AK, Nickerson, DP, *et al.* COMBINE archive and OMEX format: one file to share all information to reproduce a modeling project. *BMC Bioinformatics* **15**, 369 (2014).
101. Soltesz, S, Pötzl, H, Fiuczynski, ME, Bavier, A & Peterson, L. *Container-based operating system virtualization: a scalable, high-performance alternative to hypervisors* in *ACM SIGOPS Operating Systems Review* **41** (2007), 275–287.
102. GitHub, Inc. *GitHub Guides: Mastering issues* <<https://guides.github.com/features/issues/>> (2019).
103. Missier, P, Belhajjame, K & Cheney, J. *The W3C PROV family of specifications for modelling provenance metadata* in *Proceedings of the 16th International Conference on Extending Database Technology* (2013), 773–776.
104. Wolstencroft, K, Krebs, O, Snoep, JL, Stanford, NJ, Bacall, F, Golebiewski, M, Kuzyakiv, R, Nguyen, Q, Owen, S, Soiland-Reyes, S, *et al.* FAIRDOMHub: a repository and collaboration environment for sharing systems biology research. *Nucleic Acids Research* **45**, D404–D407 (2016).
105. Morrell, WC, Birkel, GW, Forrer, M, Lopez, T, Backman, TW, Dussault, M, Petzold, CJ, Baidoo, EE, Costello, Z, Ando, D, *et al.* The experiment data depot: a web-based software tool for biological experimental data storage, sharing, and visualization. *ACS Synthetic Biology* **6**, 2248–2259 (2017).
106. Nguyen, T. Science commons: Material transfer agreement project. *Innovations: Technology, Governance, Globalization* **2**, 137–143 (2007).
107. Kahl, L, Molloy, J, Patron, N, Matthewman, C, Haseloff, J, Grewal, D, Johnson, R & Endy, D. Opening options for material transfer. *Nature Biotechnology* **36**, 923 (2018).

108. Chandersekaran, C, Simpson, W & Trice, A. *Cross-domain solutions in an era of information sharing* in *The 1st international multi-conference on engineering and technological innovation: IMET2008, Orlando, FL* **1** (2008), 313–318.
109. Sun, J & Fang, Y. Cross-domain data sharing in distributed electronic health record systems. *IEEE Transactions on Parallel and Distributed Systems* **21**, 754–764 (2009).