Standards to Enable Genome-scale Engineering

Bryan Bartley[1], Jacob Beal[1], Jonathan R. Karr[2], and Elizabeth A. Strychalski[3]

[1]Raytheon BBN Technologies, [2]Icahn School of Medicine at Mount Sinai, [3]National Institute of Standards and Technology

Summary

In the near-term, genome-scale engineering will likely focus significantly on rewriting and refactoring existing genomes, while designing novel genomes and phenotypes remains a long-term goal. Substantial interdisciplinary collaboration will likely be required to achieve these goals. To facilitate the collaboration that will be required to realize genome-scale engineering, we summarize the challenges to collaboratively designing and writing genomes and provide recommendations regarding standard formats and protocols that will help enable each aspect of the emerging workflows for large-scale for genome design and writing (Figure 1). In this document, we summarize key challenges related to the integration of workflows for genome engineering, and we recommend standards that may be adopted or developed to enable and advance large-scale, distributed collaboration. We find that near-term genome-scale rewriting and refactoring can likely be supported by adopting or extending existing technical standards and developing new legal and contractual frameworks. In the long term, new technical standards will also likely be needed to support genome-scale design. Table 1 summarizes our current recommendations.

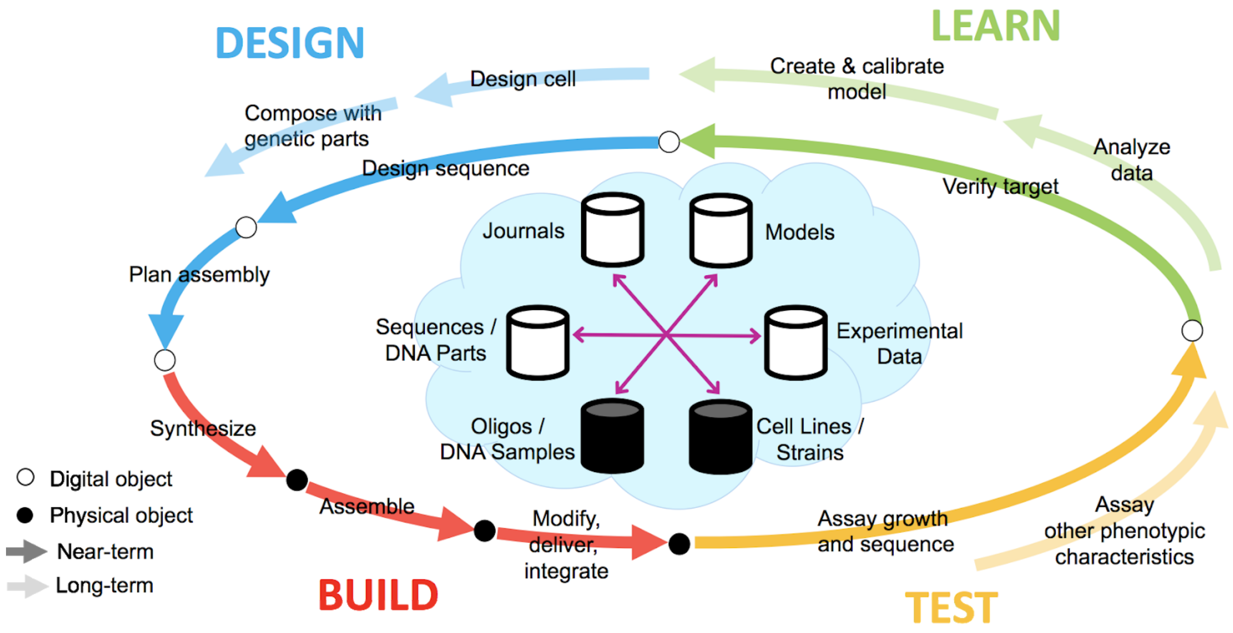

Figure 1. Genome writing and genome-scale design will likely be achieved through a multi-stage process that is likely

to include distributed collaboration. Interfaces between stages and organizations will likely benefit from a combination of new and existing and standards.

Recommendations

Standard reference materials and methods, as well as establishing shared best practices and protocols, will be beneficial for reliable, practical, and scalable genome-scale engineering. Realizing the vision of GP-write will require reproducible and comparable data within and across the laboratories of the GP-consortium. To this end, experimental laboratory practices, such as DNA synthesis, assembly, modification, delivery, and integration should be coupled with standardized protocols and measurements for viability, evolutionary stability, and phenotypic characteristics.

Data exchange standards will help GP-write consortium members integrate efforts across distributed teams. The largest genome engineering project to date, the Sc2.0 synthetic yeast genome project, has employed a divide and conquer approach in which institutions have separately assembled chromosomes in parallel (Richardson 2017). Synthesis of human genomes will be an endeavor roughly 3 orders of magnitude greater in number of engineered base-pairs than this. Therefore, future genome-scale engineering pipelines will likely require an even higher degree of parallelization across organizations**.** Existing standards can be leveraged to support GP-write in its early phase, though consortium members may need assistance incorporating these standards into their tools and training personnel in their use. In its later phase, GP-write may transition from editing genotypes toward rationally designing genomes, and novel standards will likely need to be developed.

GP-write members should adopt FAIR (findable, accessible, interoperable, reproducible) principles for data management (Wilkinson 2016). Computers will be needed to design genomes and process the large amount of data generated from testing genome designs. To leverage computing effectively, GP-write members should utilize machine-readable data formats and ontologies which enable computers to easily interpret models, genome designs, and the results of testing genome designs; adopt common procedures and prioritize the collection of the metadata to make the models, genome designs, and data generated by GP-write comprehensible, reproducible, and reusable; and utilize common repositories which make the model, genome designs, and data generated by GP-write accessible to every member.

FAIR repositories will be critical infrastructure for GP-write. Managing, curating, and transferring collections of DNA designs, constructs, cell lines, experimental data, and models could incur significant logistical overhead for individual labs and hinder genome engineering efforts. These transfers will benefit from being mediated through repositories, particularly with the assistance of automated interfaces. New types of sample and cell-line repositories may need to be developed, as existing infrastructure (e.g., Addgene) is not well-equipped to handle DNA constructs on the order of 100kb or more. Large DNA constructs are more likely to exhibit random mutations and may require more frequent sequence verification. Criteria for periodic resequencing of constructs may need to consider natural rates of genetic drift as well as selection against synthetic sequences. Repositories may need to track pedigrees of cell lines and strains as they proceed through successive stages of sequence modification and variation.

The GP-write Standards Working Group should play an ongoing role in helping consortium members institute FAIR practices. Human habits, inadequate incentives, and a lack of supporting software tools are often significant barriers to FAIR data management. The Standards Working Group should help members overcome these challenges at an organizational level by advising GP-write leadership on standards, as well as at a technical level by soliciting standards needs from GP-write members, observing workflows, and establishing liaisons to the working groups; developing, extending, and integrating standards; and training personnel in their use.

GP-write members should coordinate with the Standards Working Group to address their interoperability needs. Members may begin by considering their likely collaborators and where their institution fits within the general workflows depicted in Figure 1. This will allow the Standards Working Group to plan for the specific workflows and tools that will be used and to propose specific solutions, such as particular repositories and ontologies. Member organizations may wish to appoint representatives to participate in Standards Working Group meetings and email discussions, and/or GP-write can sponsor integration liaisons to observe and study experimental and computational workflows in practice at the front lines of genome engineering.

Table 1. Recommended standards adoption and development. These standards apply to data and materials exchange at each workflow and repository interface represented in Figure 1. Recommended actions are adopting and/or extending existing standards, creating new standards, or monitoring the need for standards.

Interface between workflow stages	Time Frame	Recommendation
Genomic Design → Assembly Plan	Near	Extend: GFF3 and/or SBOL with chromosomal coordinates
Assembly Plan → Short ssDNA	Near	Adopt: FASTA/GenBank
Assembly Plan → Short ssDNA	Medium	Adopt: SBOL (function information for manufacturability flexibility)
Short ssDNA/dsDNA Samples → Long dsDNA samples	Near	Adopt: SBOL and/or GFF3
Long dsDNA samples → Integrated genomic insert	Near	Extend: SBOL and/or GFF3 with integration context
Long dsDNA samples → Modified dsDNA	Medium	Extend: SBOL and/or GFF3 with modifications (e.g., chromatin state)
Strains → Growth phenotype	Near	Create: encoding of viability metric requirements
Strains → Verified or invalid sequence	Near	Extend: FASTQ, GVF, and/or SBOL with quality metric requirements
Strains → Other Phenotypic Assays (e.g., ‘Omics)	Medium	Monitor
Experimental Data → Analysis Pipelines	Near	Extend: SBOL 2.2 Design/Build/Test + OBO & EFO ontologies for experimental design
Analyzed Data → Model Creation	Medium	Monitor
Models → Cell Design Tools	Long	Monitor
Cell Design → Composition of Parts	Long	Monitor

Repositories	Time Frame	Recommendation
Database federation	Near	Adopt: existing open DBMS solutions

Link designs → samples → strains → data → models	Near	Extend: PROV-O, SBOL 2.2 Design/Build/Test
Import from public repositories (e.g., NCBI)	Near	Extend: existing FASTA/GenBank importers/API
Sharing of genome-scale designs	Near	Adopt: SynBioHub / ICE
Sharing inventory of ssNA/dsNA samples, cell strains	Near	Adopt: SBOL 2.2 Design/Build/Test
Sharing of ssNA/dsNA biological material samples	Near	Create: inventory-compatible legal/contractual framework
Sharing of cell strain biological materials	Near	Create: inventory-compatible legal/contractual framework
Sharing of experimental data	Near	Adopt: FAIRDOMHub or other open cloud solution
Cross-laboratory experimental data comparison	Near	Extend: process control and calibration standards
Model sharing and composition	Medium	Adopt: SBML, BioModels and related COMBINE standards

Legal and Administrative:	**Time Frame**	**Recommendation**
IP tracking and composition	Near	Create: based on PROV-O, OSI / CC / ScienceCommons
Privacy management, public release timing	Near	Create: based on PROV-O, cross-domain information sharing protocols

Abbreviations and References: CC: Creative Commons; DBMS: Database Management System; GFF3: Generic Feature Format Version 3 (Stein, 2013); GVF: Genome Variation Format (Reese, 2010); ICE: Inventory of Composable Elements (Ham, 2012); OBO: Open Biomedical Ontologies (Smith, 2007); EFO: Experimental Factor Ontology (Malone, 2010); OSI: Open Systems Interconnection (Zimmerman, 1980); PROV-O: PROVenance Ontology (Lebo, 2013); SBOL: Synthetic Biology Open Language (Roehner, 2016); SBML: Systems Biology Markup Language (Hucka, 2003); SynBioHub (McLaughlin, 2018)

Acknowledgment

We thank Joanne Kamens of Addgene for information on challenges of large DNA curation, and John Wilbanks of Sage Bionetworks and Bryan Bishop of LedgerX for information on challenges related to law and IP.

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

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.

Karr was supported by NIH awards P41-EB023912 and R35-GM119771 and NSF award 1548123.

References

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. 2012 Jun 18;40(18):e141.

Hucka M, Finney A, Sauro HM, Bolouri H, Doyle JC, Kitano H, Arkin AP, Bornstein BJ, Bray D, Cornish-Bowden A, Cuellar AA. The Systems Biology Markup Language (SBML): a medium for representation and exchange of biochemical network models. Bioinformatics. 2003 Mar 1;19(4):524-31.

Lebo T, Sahoo S, McGuinness D, Belhajjame K, Cheney J, Corsar D, Garijo D, Soiland-Reyes S, Zednik S, Zhao J. Prov-o: The PROV ontology. W3C recommendation. 2013 Apr 30;30.

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. 2010 Mar 3;26(8):1112-8.

McLaughlin JA, Myers CJ, Zundel Z, Mısırlı G, Zhang M, Ofiteru ID, Goni-Moreno A, Wipat A. SynBioHub: A standards-enabled design repository for synthetic biology. ACS Synthetic Biology. 2018 Jan 30;7(2):682-8.

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. 2010 Aug;11(8):R88.

Richardson SM, Mitchell LA, Stracquadanio G, Yang K, Dymond JS, DiCarlo JE, Lee D, Huang CL, Chandrasegaran S, Cai Y, Boeke JD. Design of a synthetic yeast genome. Science. 2017 Mar 10;355(6329):1040-4.

Roehner N, Beal J, Clancy K, Bartley B, Misirli G, Grünberg R, Oberortner E, Pocock M, Bissell M, Madsen C, Nguyen T. Sharing structure and function in biological design with SBOL 2.0. ACS Synthetic Biology. 2016 May 4;5(6):498-506.

Smith B, Ashburner M, Rosse C, Bard J, Bug W, Ceusters W, Goldberg LJ, Eilbeck K, Ireland A, Mungall CJ, Leontis N. The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. Nature Biotechnology. 2007 Nov;25(11):1251.

Stein L. Generic feature format version 3 (GFF3). Sequence Ontology Project. 2013;1.

Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, Blomberg N, Boiten JW, da Silva Santos LB, Bourne PE, Bouwman J. The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data. 2016;3.

Zimmermann H. OSI reference model--The ISO model of architecture for open systems interconnection. IEEE Transactions on Communications. 1980 Apr;28(4):425-32.