大规模基因组工程的大科学框架

千兆碱基规模的基因组工程

中国科学院上海营养与健康研究所

生命健康科技智库

上海市生物工程学会

2021 年 2 月

千兆碱基规模基因组工程的科学框架

编者按：基因组工程的规模化给科学、工业、医学和社会等都带来了重大影响。DNA 合成领域已经实现对兆碱基（MB）基因组的调控。但是，协调和整合千兆碱基（gigabase）的基因组工程，对团队规模和工作量仍然是个相当大的挑战。2020 年，美国西奈山伊坎医学院、BBN 科技公司、美国国家标准与技术研究院（NIST）的研究人员合作研究，对未来千兆碱基基因组工程的发展道路提出 4 方面的建议：扩展现有计划和样本、数据及工作流的表征方法；开发新的数据管理和质量控制技术；在基因组建模和设计规模上开展基础研究；支持开发新的法律和合同以促进合作。

对有机体的全基因组进行工程化改造，有望使其组织、功能及与环境的相互作用发生巨大变化，从而对科学、医学、产业和社会产生广泛影响。过去几十年，在合成 DNA 和基因组修饰等方面已经取得显著进展。自 40 年前 Khorana 创造第一个合成基因以来，人们构建 DNA 序列的能力大约每 3 年翻一番（图 1a），从 20 世纪 90 年代初期的质粒、21 世纪初期的病毒、21 世纪中期的基因簇、再到 2008 年首个细菌染色体合成。最近，一些研究小组重新设计了 4Mb 基因组的大肠杆菌和鼠伤寒沙门氏菌，以及合成酵母（Sc2.0）项目，几乎完成了对酿酒酵母 11.4 Mb 基因组的重组。展望未来，学术界和产业界 2016 年成立了基因组编写计划，启动高级真核生物的千兆碱基基因组工程，目标在于设计一种抗病毒、安全的人源性细胞系，用于药品生产。

1. 从工程基因到工程基因组

千兆碱基规模的发展带来重大的技术和科学挑战。DNA 合成和编辑、基因组建模、设计和检测等方面的挑战受到广泛关注，但很少关注如何将这些任务整合到计算机应用环境下自动化的工作流（workflow）所需的技术、存储库、标准等。

工作流整合是千兆碱基规模基因组工程面临的首要问题。在过去 40 年，开创性基因组工程项目的研究人员数量随着基因组的大小而显著增加，基因组工程的复杂性也随着基因组大小而变化（图 1b）。如果这些趋势持续下去，预计到

2050 年，要使千兆碱基基因组的工程化成为可能，大约需要由 500 人组成研究团队合作工作。为了在不使用大规模团队的情况下管理如此复杂的项目，该报告提倡开发由工具、服务、自动化和其他资源组成的生态系统，大力提升生物工程团队的能力。为此，通过研究基因组工程中设计-构建-测试-学习的工作流，确定了关键接口，并为技术、存储库、标准和框架的采用或开发提出建议。

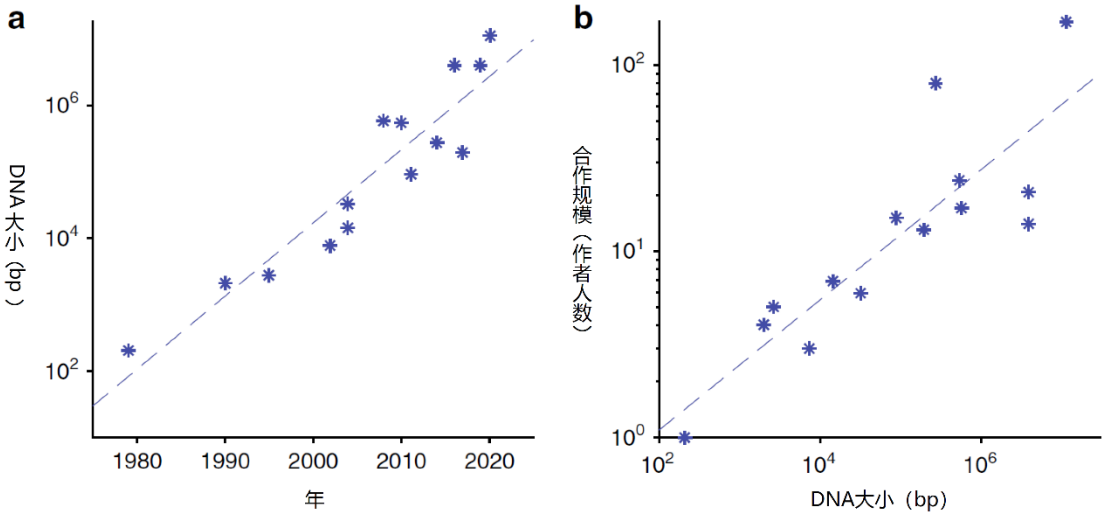

图 1 随着基因组工程能力的快速发展，每个开创性基因组工程项目团队的规模在扩大

a. 1980 年至今，工程基因组的规模呈指数增长，大约每 3 年翻一番。这个趋势表明，到 2050 年，千兆碱基的工程或将变得可行。b. 创造这些基因组的作者数量也呈指数增长。表明千兆碱基基因组工程需要大约 500 名工作人员的团队，他们可以直接作为团队的一部分，也可以是间接地通过工具、服务、自动化和其他资源的生态系统。表 1 提供了图 1 中的数据。

表 1 年份、基因组大小（bp）以及过去三十年间参与开创性基因组工程项目的人员数量

年份	DNA 大小（bp）	合作规模（人数）	参考文献	注释
1979	207	1	Khorana	首个合成基因
1990	2050	4	Mandecki et al.	首个合成质粒
1995	2700	5	Stemmer et al.	合成质粒
2002	7.5000E+03	3	Cello et al.	脊髓灰质炎病毒 cDNA
2004	1.4600E+04	7	Tian et al.	rRNA 基因
2004	3.1656E+04	6	Kodumal et al.	基因簇
2008	5.8297E+05	17	Gibson et al.	生殖支原体
2010	5.3100E+05	24	Gibson et al.	支原体，JCVI 合成细胞
2011	9.1010E+04	15	Dymond et al.	Sc 2.0 synIXR

2014	2.7287E+05	80	Annaluru et al.	酵母染色体 synIII
2016	3.9700E+06	21	Ostrov et al.	部分编码的大肠杆菌，基因组中的 62k 编辑
2017	2.0000E+05	13	Lau et al.	鼠伤寒沙门菌部分基因组
2019	4.0000E+06	14	Fredens et al.	重组大肠杆菌
2020	1.14E+07	172	Richardson et al.	SC2.0 的预计完成日期；参考表 3 的基因组大小；合作规模根据 SC2.0 网站估算

2. 新兴的基因组工程工作流

一些团体近期提出了生物工程的工作流，并会聚到一个由图 2 所示的 4 个阶段组成的通用工程循环，其内容包括：（1）设计：生物工程师使用模型和启发式设计具有特定表型的基因组；（2）构建：基因工程师在目标生物体中构建所需的 DNA 序列；（3）测试：实验人员分析工程生物体的分子和行为表型；（4）学习：建模者分析所需表型和观察到的表型之间的差异，从而开发改进的模型和启发式设计。重复这个过程，直到鉴定出具有所需表型的有机体。尽管人们对生物学的复杂性还不完全了解，但这种渐进式的方法使得工程化成为可能。

图 2 中的内环代表很多当前基因组工程项目使用的工作流，它们专注于现有基因组“自上而下”的重构，例如，通过重新编写密码子或将基因组精简为必需序列。从长远来看，合成生物学的主要目标之一就是通过对分子模块和设备 “自下而上”的设计改造出具有新表型的生物体。有机体工程师已经开始利用这种方法，小规模地商业生产设计高价值化学品的新代谢途径。对于千兆碱基基因组工程来说，这种方法可能需要更复杂的工作流，以利用更复杂的设计工具、表型分析、数据分析和模型等（图 2 外环）。

执行这些多步骤工作流需要在众多的工具、人员、机构和存储库之间进行材料、信息和其他资源的广泛交流。设计阶段必须将基因组设计传递给构建阶段，构建阶段必须将 DNA 结构和细胞系传递到测试阶段，测试阶段则须将测量结果传送给学习阶段，而学习阶段再为设计阶段提供模型和启发式设计，工作流需要协调所有这些阶段中的互动与运行。

除了这些技术挑战，基因组工程还必须解决一些安全、安保、法律、合同和伦理方面的问题。在整个基因组工程中，生物工程师必须关注生物安全、生物安

保和网络安全。为了在多机构中实施基因组工程的工作流，生物工程师必须了解材料转让协议、版权、专利和许可证等方面的规定。

基因组工程工作流的每个环节都必须扩展到能够处理千兆碱基基因组的规模。最终，大部分或全部步骤都应该实现自动化，各步骤之间的接口都实现正规化以便进行机器推理，并尽可能多地删除基因组工程中以人为中心的内容。在许多情况下，可以通过采用或扩展小规模基因组的解决方案来实现，还可以采用系统生物学、基因组学、遗传学、生物信息学、软件工程、数据库工程和高性能计算等相关领域的解决方案。然而，千兆碱基基因组工程的其他挑战可能需要开发新的系统或进一步的基础研究。

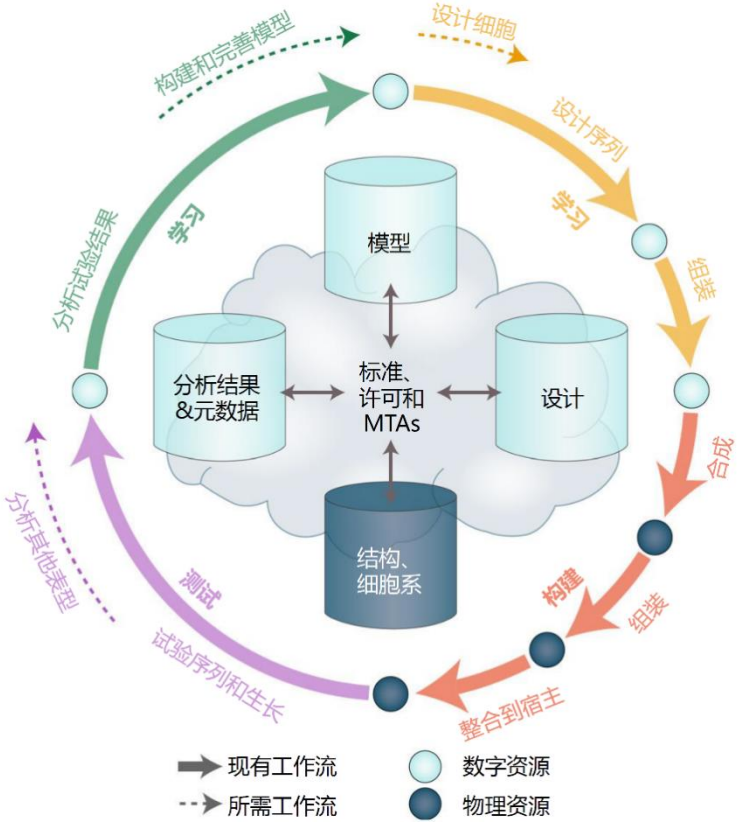

图 2 新兴的基因组工程设计–构建–测试–学习的工作流示意图

当前（实心箭头）以及未来可能的（虚线箭头）任务、接口（圆圈）、存储库（圆柱体），可以是数字的（浅色）或物理的（深色）。

3. 确定并弥补现有技术的差距

该报告讨论了上述挑战，回顾了与千兆碱基基因组工程新兴需求相关的技术和标准的最新进展，但不关注正在迅速发展的特定的协议和方法，而是聚焦使协

议或方法能够有效构成全面的工作流程所必须具备的条件。通过此分析，确定了关键的差距和机遇，其中的技术和标准将促进千兆碱基工程基因组工作流的开发，总结了已有的潜在解决方案（表 2）。

3.1 基因组重构与设计

当前的基因组工程项目主要致力于在保留细胞功能的同时重构基因组，例如，最近的三个项目涉及删除非必需组分、对基因进行重新排序，并将其插入代谢通路。在这个层面上，两个关键的挑战是获取注释良好的源基因组，以及修饰基因组的表征和交换设计。有机体的功能变化也将带来与产生新型细胞功能的组成部分相关的挑战。

目前，基因组设计通常涉及到修饰先前存在的有机体序列，例如，国际核苷酸序列数据库合作组织（INSDC）公开的序列，目前包含约 105 个细菌基因组和数百个真核基因组。功能注释是关键因素，基因组工程师需要以此考虑组织特异性表达模式、调控因素、结构组分、复制起源位点，以及具有临床意义的 DNA 重组位点等。注释的一致性是其中的关键挑战，许多基因组由不同的工具链进行注释，因此得到显著不同的注释信息。例如，由 RefSeq 和 GENCODE 项目产生的人类基因组的注释存在差异，可能会影响其工程化的结果，例如，由于与其他剪接的相互作用而丧失预测功能。尽管可以通过 NCBI Genome Viewer、WebGestal 和 DAVID 等服务来整合注释，但是这些内容大部分也分散在不同的资源中。对于千兆碱基的规模来说，改进的注释将很有价值，对注释可信度和可靠性的评估也一样，例如，RefSeq 数据库处理的证据和结论本体论（Evidence and Conclusion Ontology）。

千兆碱基规模也对基因组的表征和交换设计提出了挑战。由于 GenBank 和 EMBL 等格式在处理序列时是单一的，这使得在多个用户之间难以整合或协调编辑，甚至简单地传输数据也存在困难。目前有两种更适合基因组工程的格式：通用特征格式（GFF）第 3 版和合成生物学开放语言（SBOL）第 2 版。GFF3 允许对序列描述进行分层组织（例如，基因可以组织成簇，将簇聚成染色体），通过 Sequence Ontology 进行序列注释，并且已经在 Sc2.0 基因组工程项目中使用。SBOL 2 也经常用于编辑基因组的分层次描述，并且可以与 GFF3 互操作（尽管 GFF3 仅代表 SBOL 的一个子集）。SBOL 提供了更丰富的、以设计为中心的语

言，包括对变体、文库和部分设计的支持（例如，识别簇中的基因，但还不是特定的变体或簇的排列）、其他元素和细胞功能（例如，蛋白质、代谢途径、调节相互作用）。SBOL 也与系统生物学标记语言（SBML）编码的模型进行互操作。SBOL 支持（并且可以扩展 GFF3 支持）非标准碱基和增强序列编码语言中的序列修饰，例如 BpForms。

基因组设计的表征也需要表达设计所需的固定参数和策略，例如，剔除限制位点、分离重叠特征、替换密码子以及优化 DNA 合成。SC2.0 项目已经通过结合人工编辑指导方针和自定义软件工具来实现这一目标，DNA 合成供应商可以提供接口，以检查可制造性的限制条件。然而，在千兆碱基规模下，采用功能更强大且更具表达力的语言来描述设计或将更有益，例如，基于分子生物学研究生物功能的基础，并在设计表现形式中涵盖组装和转换策略，以简化可制造性的调整。JGI 的 BOOST 工具在这方面提供了原型。GenBank、GFF3 和 SBOL 都很适合完成这项任务，不过 SBOL 在原则上更适合进行扩展以编码此类信息。

随着基因组工程学不局限于重构和重新编码，并且可以进行有机体功能的复杂改造，建模将变得越来越重要。通过结合来自多个数据库（如 BioCyc 和 SEED）的生化和基因组信息，可以构建基因组规模的代谢模型和全细胞模型。模型还可以预测由特定的基因元件、装置、线路和基因组片段组成的生物体的行为，但仍需要进行实质性基础研究，以实现这些模型千兆碱基规模的实用性。

表 2 整合新兴千兆碱基工程工作流的潜在方法

阶段	工作流间的接口	建议	理论基础
设计	输入设计材料	扩展：注释方法，ECO	当前的基因组注释工具存在规模、一致性和未整合等问题，这些问题可以通过更好地利用证据和结论本体之类的本体论来解决
设计	设计序列→计划组装、共享设计	扩展：GFF3 和/或 SBOL 与染色体协调，BpForms	GFF3 和 SBOL 都是基因组设计规范的有效格式，BpForms 允许对带有非标准碱基和修饰的序列进行注释
设计	设计组装→构建	采用：FASTA/GenBank/GFF，迁移到 SBOL（功能和计划灵活的可制造性）	FASTA、GenBank、GFF 和 SBOL 格式都能相对容易地编码序列并进行构建。SBOL 可以编码预期功能和组装计划，允许更大灵活性以应对制造限制条件。

设计	构建和完善模型→设计细胞	研究：跨多个模型的整合 CAD	目前的方法还没有接近所需的能力。
构建	合成→组装→整合到宿主	拓展：带有质量度量和要求的 FASTQ、GVI 和/或 SBOL	FASTQ 和 GVF 编码变异，但无要求；SBOL 可以同时编码这两种形式，但不是当前设备的初始形式。关于度量和需求的非正式实践需进行组织和格式化。
构建	整合到宿主→试验序列		
构建	共享方法、结构和细胞系	采用：生物制造最佳实践和 LIMS	存在很多适用的系统和方法，选择的权衡将取决于所涉及项目和设施的细节。
测试	整合到宿主→生长和表型测试	开发：适用性度量和相关的规范语言	关于度量和需求的非正式实践需进行组织和格式化。
		研究：整合组学、其他测量、表型规范	目前的方法还没有接近所需的能力。
测试	对比试验结果	拓展：过程控制和校准标准	一些实验（例如 RNAseq、流式细胞）已经建立了确保数据有效性的实践，其他实验需要开发类似的方式。
测试	分享试验结果	拓展：本体和管理工具	开放生物和生物医学本体（OBO）铸造厂中提供了标准化词汇；另外还需要一些术语以及更好的管理工具以减少使用中的摩擦。
学习	试验结果→分析试验结果	拓展：SBOL+OBO 本体，额外的元数据和知识管理工具	SBOL 可以使用开放生物和生物医学本体（OBO）铸造厂的标准化词汇连接设计、分析、执行追踪和数据；需要更好的管理工具以减少使用中的摩擦。
学习	分析试验结果→构建和完善模型	研究：自动化和可拓展的模型生成和验证	目前的方法还没有接近所需的能力。
学习	模型组成和共享	拓展：SBML、CellML、生物模型和相关的结合标准	生物网络计算模型（COMBINE）已经开发出一套可操作性的模型共享、组合工具和标准，为复杂模型的建立提供了坚实基础。
		研究：描述多规模模型的标准	目前的方法还没有接近所需的能力。
跨部门	工作流管理	采用：CWL、PROV-O、SBOL 2.2 设计/构建/测试，容器工具，团队合作	SBOL 可以代表周期中所有阶段的元素，PROV-O 可以将它们连接起来；现有的工作流语言和团队协调工具能够

		工具	管理机器和人力工作流；现有容器工具支持可移植性。
		开发：互操作性的实验室自动化语言	现存很多实用的系统和方法，但是捕获工作流和流程的方法不同且彼此不兼容。
跨部门	数据库联盟	采用：现有的开放 DBMS 解决方案、FAIRDOMhub、EDD 或类似解决方案	数据库联合工具得到了很好地开发，包括领域特有工具。
跨部门	IP 追踪和组成	开发：基于 OSI/CC/Science Commons、PROV-O 的框架	先前开源计划、知识共享和科学共享的努力解决了大部分但不是全部挑战；PROV-O 可用于知识产权应用的自动追踪。
跨部门	结构和细胞系的转移	开发：在 OpenMTA 上构建的法律/合同框架	OpenMTA 支持材料转移，但存在一些需要解决的关于法律/法规遵从性和专利的问题。
跨部门	管理敏感和受控信息	开发：基于跨领域信息共享框架、PROV-O	现存信息追踪和共享工具需要领域适应性。
新的工作流中每个接口分为三类：采用或扩展相对成熟的现有方法（绿色）、开发新的解决方案或扩展新型方法（黄色）和进行额外的基础研究（红色）。			

3.2 构建工程化基因组

构建工程化基因组的技术和流程正在迅速发展。根据工程生物体的特定宿主和预期功能，存在许多潜在的有关 DNA 合成、组装的交付方法和流程。目前，尚未实现最佳实践指南的要求，包括工程化基因组和中间样品等相关信息的测量、跟踪和共享。

在组装过程中操纵 DNA，降低了 DNA 序列设计中的产量、断裂、错误和其他来源不确定性的几率。将短的 DNA 片段组装成较大结构所需的商品化试剂盒通常涉及放大、处理、纯化、转换或其他存储和传递步骤，这些步骤增加了 DNA 质量和数量的不确定性。组装的 DNA 也可能包含不具有生物活性的附加序列，例如，对于使用限制性酶而增加的序列，或者使用 Golden Gate Assembly 或 MoClo 后造成的损伤。Gibson Assembly 是无缝克隆，但是产量和具体结果可能还依赖

于 DNA 片段的二级结构。因此，除了序列信息，可能还需要进一步扩展工作流，以实现可以追踪影响组装产品的所有信息，包括 DNA 的二级结构、组装方法、组装所需的序列、在 DNA 分子上的位置（例如，特定大肠杆菌或酵母菌株的兼容性序列或位点），以及预期的表观遗传修饰等。验证中间序列和最终序列结构的结果通常以 FASTQ 格式产生，通常对较小的结构来说是足够的。对大规模基因组进行操作，更全面地描述一个基因组及其变化可以用 GVF 或 SBOL 来表示。

目前，普遍缺乏适合运送大型、组装 DNA 结构和整个基因组的方法。现有方法，例如电和化学转化或基因组移植，可以大大提高有效性，但还应开发更广泛用于所有生物和细胞类型的方法。这可能需要识别新的无细胞环境或用于组装和操纵 DNA 的、基于细胞的骨架，它们与基因组修饰和进入的宿主生物更具兼容性。为了促进这种发展，应该以机器可读格式提供生物和技术重复试验方法、测量以及交付流程等相关信息，其中包括宿主细胞的信息，例如还没有得到充分证实的基因型。在实验室信息管理系统（LIMS）的工业生物制造设置和实施中采用最佳实践，可以为整合测量、过程控制和信息处理以及跟踪和交换样品提供路径。推进利用自动化技术支持工程化基因组工作流的构建步骤，需要评估哪些步骤可以降低成本和速度、自动化方法的可用性，有效地共享这些方法，并在不同平台和制造商中进行调整，更简单地整合和协调自动化的工作流。

3.3 工程化基因组功能的测试

菌株适应性和其他表型可以通过广泛的生化和组学测量进行评估。但是，在所有情况下，合作组织都需就具体测试标准、控制和校准测量方式等达成一致，以确保结果在所有合作实验室中具有可比性和可用性。

通常评估 DNA 结构及其相关生长表型，以确定由于基因组序列修饰而对细胞功能和适应性产生影响的性质和程度。还应该对工程细胞系进行评估，包括其在预期应用过程中可能经历的环境变化的稳健性，以及在相关时间尺度上进化或适应的稳定性，但由于需要对适应性、代谢负担和其他表型特征进行共享定义和测量，因此增加了评估的复杂性。

标准流程、参考细胞系和实验设计的使用都可以提高测试结果的严谨性和可信度，也有助于开发确保工程化基因组测试的标准。这样的基础可以帮助识别基因型和表型之间的关系，或是对被测性状提供生物随机性和测量不确定性贡献，

尽管这类综合方法可能需要重要的基础研究。

生物试验的校准有助于单个实验室内和不同实验室之间的比较结果。最近的研究，例如荧光素、吸光度和 RNAseq 测量，证明在生物测量中实现可伸缩性和成本效益可比性的可能，或将通过开发校准测量方法以及有机体特性的绝对定量来推进生物工程。

建立元数据、过程控制和校准的共享协议及实施也至关重要。实验室之间的数据、元数据、过程控制和校准的自动化整合和比较将有助于利用建模和模拟来促进检测过程和学习。一些现有的本体可以用于此目的，例如实验条件本体论（ECO）、实验因子本体论（EFO）、测量方法本体论（MMO）。此外，实验室信息管理系统（LIMS）的工具和管理辅助软件（例如 RightField）对于需要持续构建的元数据至关重要，可以通过人工手段来限制输入的信息。

3.3 从测试结果中系统学习

由于基因组工程会影响整个生物系统，需要综合的模型来帮助预测和解释基因型与表型之间的关系。尽管有些模型是为整个细胞或有机体构建的，但开发和调整此类模型非常具有挑战性。要扩展到千兆碱基基因组，提高创建、校准和验证模型的能力非常重要。

从数据到学习的第一个挑战是发现并整理所需数据。目前已有部分解决方案，例如在 SBOL2.2 中引入工作流模型、开放生物学与生物医学本体论、实验因素本体论、系统生物学本体论，以及表型本体论等理论，但这些都需要整合和扩展，以便满足基因组工程的所有需求。

然而，大规模自动化辅助模型的生成和验证仍存在许多基础研究挑战，包括解决生物学整合的复杂性以及基因组和有机体行为之间的多尺度、大型模型的高性能模仿、模型验证，以及模型语义和来源表征。

在拥有全面的预测模型之前，工程师们可能更多地依赖生物体部分预测模型、数据驱动模型以及启发式设计规则的组合。例如，基于约束的模型通常用于代谢工程，PSORTb 可以帮助将蛋白质定位于特定的隔间，GC 含量优化可用于改善宿主相容性。千兆碱基规模的基因组工程需要同时应用许多这样的模型，因此也将受益于促进生物模型共享和合成的现有标准格式，例如 SBML、CellML、NeuroML 等其他生物网络的计算建模（COMBINE）。这些大量模型已经可以在

公共数据库中找到，例如 BioModels、NeuroML、Open Source Brain 和 Physiome Model Repository 数据库。此外，Kipoi 和 DockerHub 储存库已经可以用于数据驱动模型共享。这些格式的进一步扩展将有助于学习过程的自动化，包括将语义与模型组件关联起来，获取模型元素的来源（例如，数据源、假设和设计动机），并获取有关其预测能力及适用范围的信息等。

为了提高从数据中学习此类模型的自动化程度，需要开发可能构建整个有机体模型的生物元件的新存储库；产生模型变体的新方法，可以通过添加元件、替代动力学定律或替代参数值等模型来解释新的观察结果；非线性多尺度模型的新模型选择技术。

3.4 复杂工作流中的协调与共享

千兆碱基基因组工程的设计-构建-测试-学习循环的高效操作将需要协调众多以上讨论的任务，形成清晰、有凝聚力、可重复的工作流程，用于软件交互、实验室流程，以及任务和人员的管理。自动化工作流还提供了实施网络安全、网络生物安全和生物安保的最佳实践机会。

为了整合信息任务，计算工作流工具支持特定、可重复的操作，以及涉及多个软件程序和计算环境的复杂工作流的交换。当前的工作流工具包括诸如通用工作流语言（CWL）、Dockstore 和 MyExperiment 共享环境等通用工具，以及用于追踪信息来源的 PROV 本体（已应用在 SBOL 中将设计-构建-测试-学习循环连接起来）。还有许多生物信息学的工具，例如 Cromwell、Galaxy、NextFlow 和 Toil。千兆碱基工程可以通过下面这些步骤采用这些工具，包括在 COMBINE 档案中的 CWL 文件、开发 REST 或其他用于基因组工程数据库的程序性接口、基因组工程计算工具的集成化，并将这些内容存放到 DockerHub 之类的注册表中。其他可能有用的增强功能包括：开发用于基因组工程的图形工作流工具，用于注释工作流任务语义含义的本体，以及跟踪系统的应用，例如 GitHub 问题或 Jira，用以帮助设计过程中需要人为干预的复杂任务时的团队协调。

对于实验流程，已经开发了许多用于自动化和整合实验工作流的技术。实验室自动化系统很大程度上提高了再现性和效率并与 LIMS 整合，从而帮助追踪工作流和试剂库存。已经开发了许多自动化的语言和系统，包括 Aquarium、Antha 和 Autoprotocol。虽然这些方法尚未广泛应用，但它们已成功应用于基因工程，

千兆碱基基因组工程将受益于这些系统的标准化和整合，以用于构建和测试流程。

一旦在工作流的各个环节建立链接，就可以通过标准联合方法和数据库管理系统（DBMS）的各种成熟开放工具，实现各阶段不同机构和工作流程对数据库信息的统一访问。通过采用 FAIR（可找到、可访问、互操作、可复制）数据管理原则，将进一步增强可扩展的共享，这也特别强调了数据共享的自动化友好性。支持这些原则并使其适用于基因组工程的存储库，包括 FAIRDOMHub、实验数据库（EDD）和 SynBioHub 等。

3.5 合同、知识产权和法律

大规模基因组工程为法律和合同的衔接与协调提出了新挑战。使用数字信息时，人和机器都需要知道附带的版权和许可义务。开源倡议（OSI）和其他软件组织已经为软件开发了系统许可制度，并与知识共享（CC）一起开发了媒体和其他内容的系统许可制度，这两种许可都允许用户或机器确定数字对象是否可重用，或是否禁止重用，或是否需要更复杂的协商。这样的系统可以应用于基因组工程中的许多数字信息。在敏感个人信息和欧盟数据库保护权方面需要谨慎，目前，这些问题并未解决。

1995 年，美国国立卫生研究院（NIH）统一的生物材料转移协议（UBMTA）首次将物理生物材料的转移标准化，协议被 Addgene 等组织广泛使用。以科学共享项目和 OpenMTA 形式开发出更广泛、更兼容的系统，但是，在遵守当地法规和法律制度方面，尤其是当材料跨国际边界时，仍然存在重大的开放性问题。此外，材料转让协议一般不涉及材料的知识产权，而知识产权通常受专利法管辖。尚无支持专利授权自动化的公共可用系统。尽管确定哪些材料或用途可以划分为哪些层级可能是法律解释上的难点，但通过定义让普通用户、法律专家和计算机系统都能理解的分层级别，可以支持自动化友好的知识产权管理的发展。对自动化辅助工作流的有效使用还需要诸如 PROV 本体之类的机制用以记录有关生产中涉及哪些输入信息。

最后，由于隐私、安全、发布优先级或其他类似问题，还需要管理信息的公开程度。同样，目前没有现成的系统，但是可以在其他已开发的跨域信息共享协议中找到开发新系统的基础。

4. 建议和展望

综上所述，扩展到千兆碱基基因组存在很多挑战（表 2），主要分为 4 个主题，每个主题都有不同的需求和发展道路。

第一个主题是设计、计划、数据、元数据和知识的表示与交换。千兆碱基基因组设计信息的管理需要解决许多关于规模、表现形式和标准的挑战。相对成熟的技术可以满足大多数需求，也可以帮助工作流的整合。有效工作流程的实施需要大量投资，以构建采用这些技术的基础设施和工具。

第二个主题是数据质量和实验测量的共享与集成。共享和整合由生物材料测量产生的信息带来了巨大的挑战。鉴于难以获取和解释生物系统的测量值，以及数据管理的费用和不适宜的规模，目前尚不清楚共享哪些信息更有利。但是，有效整合取决于将可重复测量数据与有良好管理的知识和元数据以兼容形式进行关联。每种方法都有许多潜在的解决方案，需要大量投资来研究如何扩展现有技术以满足这些需求。

第三个主题是千兆规模的建模和设计的整合。对基因型和表型关系的深入理解也面临挑战，涉及实验数据的解读、应用这些数据构建和验证模型，这些模型可用于计算机辅助设计。需要长期投资基础研究，从无细胞系统到最小细胞系统，从合成细胞到自然生命系统，复杂程度各异的生物系统，可能为基因型和表型之间关系的研究提供合适的实验平台。

第四个主题是对伦理、法律和社会影响（ELSI）以及知识产权（IP）的技术支持。在千兆碱基规模下，计算机辅助的工作流将是管理合同、知识产权、材料转让以及其他法律和社会互动的必要条件。这种工作流需要涉及法律、ELSI 问题、软件工程和知识的跨学科团队共同开发。此外，及早解决这些问题，尽量减少出现资源重复利用等相关的纠纷可能具有关键意义。

简而言之，千兆碱基规模基因组的设计是一项重大挑战，需要协调投资等多个方面。由于生物科学其他领域也面临类似挑战，因此应对这些挑战的解决方案可能会使更广泛的生物科学界受益。重要的是，基因组工程所面临的规模、整合和缺乏知识的挑战与先前曾在其他工程项目中被攻克的挑战没有本质差异，例如航空航天工程和微芯片设计，都需要更多组织人员并实现多机构的信息共享。因此，也期望其他领域为基因组工程提供解决方案。

对提升基因组工程工作流的能力进行投资至关重要，有助于基因组工程发展成为常规、安全和可靠的领域。基因组工程工作流的投资将支持和实现大量项目，包括很多还未设想的项目，例如人类基因组解读。随着工作流技术的改进，预计未来参与团队规模不断扩大的趋势最终会逆转，以适度成本实现高保真的全基因组工程，并支持广泛的医疗和工业应用。

刘晓 张学博 编译自

https://www.nature.com/articles/s41467-020-14314-z

https://www.nist.gov/news-events/news/2020/02/big-science-framework-big-genomes