CENTER *of* EXCELLENCE
for ENGINEERING BIOLOGY

基因组编写技术的挑战和里程碑

合成基因组学需要改进完善的技术

中国科学院上海生命科学信息中心
上海市生物工程学会
2019 年 11 月

基因组编写技术的挑战和里程碑

合成基因组学需要改进完善的技术

编者按：2019 年 10 月，GP-write 工作组在 *Science* 杂志发文，总结了基因组编写技术面临的技术挑战，将其分为基因组设计、DNA 合成、基因组编辑、染色体构建 4 个领域，并提出了在未来 10 年这些技术领域需要改进和完善的方向。

重组 DNA 相关的工程生物学，也被广泛称为合成生物学，在近十年来获得了迅速的发展，原因主要在于 DNA 合成的持续工业化、分子工具与生物体的开发，以及建模分析方法的日益复杂化。然而，由于无法对合成基因组的整体基因组进行编写和测试，人们尚无法理解工程生物学的全部潜能，需要进行实质性的改进以降低成本并提高遗传工具的速度和可靠性。

本文明确了**基因组设计（genome design）、DNA 合成（DNA synthesis）、基因组编辑（genome editing）和染色体构建（chromosome construction）**四大新兴技术，以及需要对这些领域中现有技术所做的改进，以便在未来 10 年内推动合成基因组学发展（表 1）。与其他负责推进创新技术的大规模项目类似，例如人类基因组计划、国际性跨学科项目需要公共和私人部门的共同努力，从而实现投资回报的最大化，并开辟新的研究和生物技术途径。

在活细胞中进行基因组设计和编写的能力，为当前难以处理的问题提供了独特的解决机会。这些变革性技术具有广泛的科学、社会和经济影响；它们的发展和应用需要通过公开讨论来积极识别潜在陷阱。文中的展望内容是国际基因组计划-编写（GP-编写）计划的组成部分[1]，旨在鼓励科学家、律师、伦理学家、教育工作者、环保主义者、利益相关者和公众之间的包容性对话，确保这些新技术能够负责、安全、协调地实施。

合成基因组学是相对比较新兴的学科，大多数编写技术（商业 DNA 合成之外的）仍由学术界研究实验室开发。因此，与其他更成熟、可合理预测工程进度的行业不同（例如半导体业），这些新兴领域对时间线和成本的预测仍然有高度推测性。

[1] J. D. Boeke et al., Science 353, 126 (2016).

表 1 对四大主要领域中现有技术所做的改进

关键技术开发目标	所需的里程碑示例	预计时间（年）
基因组设计		
开发用于基因组规模设计、可视化和质量控制的工具	抗病毒哺乳动物染色体的设计	3
将结构信息（2D 和 3D）整合入基因组设计软件	预测合成酵母的染色体结构	5
建立从序列到表型的全细胞模型	优化代谢曲线，精确到两倍以内，由一个合成的防病毒染色体生产 100 个关键基因产品	10
DNA 合成		
提高寡核苷酸合成的偶联效率	合成长度超过 500 个核苷酸的高保真寡核苷酸	3
提高 20kb 以上片段的体外 DNA 组装效率	以>50%的产率组装 20kb 片段	4
发展合成高难度序列的方法，包括均聚物、高 GC 含量和二级结构的序列	合成着丝粒	5
开发酶法直接合成几千个碱基的 DNA 片段	（在不组装的情况下）合成一个 10kb 片段	7
DNA 合成成本降低 1000 倍	以 1000 美元的成本实现单倍体（即 3.2×10^9 个碱基）人类基因组 DNA 的合成与组装	10
基因组编辑		
拓展 DNA 编辑的多样性和精确性	在单个细菌、哺乳动物或植物细胞中，以 1：10000 个基因组的脱靶频率同时编辑 1000 个不同的目标	2
提高同源定向修复效率（HDR）介导的哺乳动物和植物细胞编辑	在非分化哺乳动物细胞中以>90%的效率进行 HDR 介导的编辑	3
开发能够在任何需要的基因组位点上进行任何核苷酸高效、精确替换的编辑酶	在人体细胞中缺乏 PAM 序列的位点进行效率>95%的等位基因编辑	5
染色体构建		
开发在时间和空间水平调控单染色体的方法，例如染色质状态	工程师分离，稳定的人工染色体（HAC）	2
以高 DNA 组装效率培养高度分化的宿主细胞，特别是难以装配的序列	在寄主腔色链霉菌（72%GC 含量）中建立体内染色体组装方法	5
开发高效、廉价的常规和自动化方法，将整条染色体导入细胞	通过细胞融合，在哺乳动物细胞中展示常规、基于设备的染色体传递	3
开发 Mb 大小染色体的组装和测试方法	从 DNA 片段组装合成的重编码人类 21 号染色体	10

1. 基因组设计

基因组设计旨在为染色体水平的 DNA 序列制定更高层次的设计准则，这将需要计算机辅助设计（CAD）技术：①可靠地获得期望的表型；②最大限度地在实验反馈和技术可行性方面影响设计；③通过设计信息的处理和交换标准以促进协作。目前使用的合成基因组 CAD 软件，例如用于 Sc2.0[2]，使用自动化和协作工具进行从质粒到整个染色体水平的 DNA 设计，但是编辑功能效应（例如基因缺失）还有待人类专家的评估。未来还需要这些工具可以根据基因组设计，预测生存能力和表型。

尽管简单的模型足以处理沉默编辑（silent edits）（例如，水印编码序列的同源编辑），日益复杂和精准的模型是预测基因组序列改变如何影响基因调控和蛋白质功能所必需的。高等真核生物的综合机理模型可能几十年后会消失，机器学习方法可以通过公共数据库和基因组编写项目中高维度、高通量的系统生物学数据来促进表型预测，类似于使用这些技术预测蛋白质结构的方式。

利用这些模型，有必要使用实验设计工具来实现基因组设计所需的高成本迭代次数的最小化。例如，重新设计相对较小的支原体基因组需要 4 次迭代，通过基因组筛选来确定必需基因也是如此[3]。更大的项目需要更多中间阶段的构建和支持试验。为了在后续设计中提供最有价值的反馈并利用产生的数据，需要新的算法来自动设计实验并选择适当的工程技术来实现它们（例如，“编写与编辑”）。其中，相关且未解决的需求是确保设计好的 DNA 与下游合成、组装、传递和分析阶段的相容性（例如，在软件引导下将染色体序列解析为合成相容性片段或引导指定序列进行组装和传递）。工具将需要适应编写技术的预期进展。

所有这些努力都依赖并会产生大型数据集，这些数据集需要简化模型定义并促进成果共享。生物信息整合的两大障碍是数据不兼容，以及缺少足够的描述性元数据。因此，业界也一直鼓励研究人员使用广泛采用的数据交换格式，例如 GenBank 和通用要素格式，并继续建立和坚持实验标准元数据，例如合成生物学开放语言（SBOL）。

资助机构和行业的利益相关者应该优先支持软件和标准数据格式的长期开发，包括协作、可视化和质量控制能力，类似于基因组学领域已经存在的方案，

[2] S. M. Richardson et al., Science 355, 1040 (2017).
[3] C. A. Hutchison 3rd et al., Science 351, aad6253 (2016).

例如，威康信托开放研究基金和陈扎克伯格生物中心。合成基因组学软件不仅会提高计划和执行大规模基因项目的能力，而且促进了基本方法学的进展，从而推进基因型−表型关系从相关性到因果关系分析的发展。

2. DNA 合成

基因组编写项目依赖于大量长[>5000 碱基对（bp）]和精确的合成 DNA 结构[4,5]。不过 DNA 的化学合成仍然局限于短的寡核苷酸（寡聚体）的产生，通常长度为 200 bp。尽管寡聚体推动了重组 DNA 技术的重大进展，更大的 DNA 结构需要组装多个寡聚体，这是费力又有损失的过程。因此，需要实现常规生产长而精确的合成 DNA 片段。

近年来，商业供应商已经实现了并行和小型的工业化，使得 DNA 的获取更加容易，但是酰胺三酯合成法几乎没有任何发展，这限制了 DNA 长度、生产速度和成本。全染色体的构建依然费时费力，例如，寡聚体阵列的合成需要每核苷酸 0.0005 美元，因此，合成 3 千兆碱基的 DNA 需要 150 万美元，这大约是人类基因组的大小。未来需要全新的 DNA 组装、纯化和合成方法，以实现成本降低以及简便程度的实质性进展。

能够减少或消除使用寡聚体进行装配、误差修正和 DNA 片段克隆需求的创新工作可以提高当前 DNA 合成基础设施的生产力。目前，完美序列的产量在 5%到 60%之间[6]，为了提高产量，杂交和错误修正可以利用高保真聚合酶和连接酶工程。这些进展很大程度上受到工业化驱动，将会减少运营成本和生产时间。克隆效率还可以通过利用具有快速划分和/或高重组率的宿主来实现，或是通过使用无细胞克隆和人工细胞进行。这些技术需要先进行基础研究，继而做好商业化准备。

能够合成高质量长 DNA 片段的新技术将从根本上改变染色体工程的水平。最近已经实现了使用不依赖模板的 DNA 聚合酶 TdT（末端脱氧核苷酸转移酶）来合成已知序列短单链 DNA（ssDNA）[7,8]。TdT 提供了以高聚合速率和高耦合效率直接合成多碱基序列的潜力。为了与现有亚磷酰胺化学合成方法竞争，酶学

[4] J. Fredens et al., Nature 569, 514 (2019).
[5] N. Ostrov et al., Science 353, 819 (2016).
[6] N. B. Lubock, D. Zhang, A. M. Sidore, G. M. Church, S. Kosuri, Nucleic Acids Res. 45, 9206 (2017).
[7] S. Palluk et al., Nat. Biotechnol. 36, 645 (2018).
[8] H. H. Lee et al., Nat. Commun. 10, 2383 (2019).

合成方法需要进一步开发以自动化、性价比高的方式合成复杂序列。这些工作，也吸引了初创公司的关注，将受益于基础研究领域的持续投资，以阐明酶末端转移酶反应的分子机制等。

为了支持基因组项目所需的 DNA 的规模和质量，需要增强机电系统和生物工具创新。可能通过进一步的并行化和小型化实现通量的持续增长，例如通过半导体制造或基于液滴的技术。DNA 的质量和生产速度的提高将会受到生物工具应用的影响，例如酶和生物体。

3. 基因组编辑

强大的新型 DNA 编辑工具降低了进行高精密度遗传和表观遗传修饰技术的障碍。完整基因组的复合编辑大大缩短了大规模修饰所需的时间和劳动力，在某些情况下，还可以规避从头开始的合成和染色体组装。

然而，尽管在可编程核酸酶领域取得了相当大的成功，例如 Cas9、TALEN 和 ZFN 在多种细胞类型中可以进行精确时间和组织特异性的调控，但是基因组规模的编辑仍然具有局限性。单个局部、核酸酶诱导的双链断裂可以用来提高每个位点的编辑效率，但是多个同时断裂常常引起细胞毒性。为了避免毒性，对“碱基编辑”酶进行改造，将核酸酶替换为碱基修饰酶[9]，使用少数引导序列同时对人体细胞中 13000 多个 Alu 重复序列进行编辑[10]。其他改造后的 Cas9 工具用于抑制、活化或靶向 DNA 插入（位点）。多个全基因组水平的编辑的主要瓶颈依然是引导 RNA 的传递（gRNA），多个独特的基因组改变需要在同一个细胞中存在多个独特的 gRNA。预计未来这个障碍以及由序列特异性编辑酶引发的脱靶突变和限制[例如原型隔区相邻基序（PAM）序列要求]将被克服，并促进常规多重编辑。

基因组规模的编辑也可以通过寡核苷酸重组[11]来完成，寡核苷酸重组依赖同源重组（HR）并能降低体内毒性。然而，这项技术目前仅限于少数生物体，在这些生物体中重组酶可以使用捐赠者的 ssDNA 作为靶点来催化高效同源重组。为了实现植物和哺乳动物细胞的重组编辑，必须发现或设计新的重组酶，但有必要制定并调节生物体的修复路径从而提高同源重组。

---

[9] H. A. Rees, D. R. Liu, Nat. Rev. Genet. 19, 770 (2018).
[10] C. J. Smith et al., bioRxiv 574020 (2019). https://doi.org/10.1101/574020.
[11] M. J. Lajoie et al., Science 342, 357 (2013).

此外，需要获得整套分子工具以加速基因组编辑的测试和优化。例如，可以获得能够编程 TALEN 或 ZFN 的核酸酶，用于靶向人类所有细胞中的 UAG 终止密码子。类似地，靶向所有 PAM 的 CRISP-Cas9 引导文库能够用于探索植物、人类或真菌细胞中的多个等位基因特异性靶向。获取这些全基因组资源的工作将提供“可访问的地图”实验证据，可以反映出基因组靶点编辑效率变化。这些数据将优化目标序列的选择，建立预测性计算模型，并加深对染色体结构、折叠和修复的认识。

4. 染色体构建

合成基因组学面临的最关键障碍是将合成染色体进行组装并引导其进入宿主细胞。如何将所有需要的 DNA 片段缝合在一起从而构建一个全功能染色体？一旦建成，应该如何控制染色体定位和结构，从而确保细胞存活？如何替换多倍体中的所有染色体拷贝？由于大多数独立生活的生物体基因组大于 2 Mb，需要常规操作大型 DNA 片段的方法。

尽管 DNA 合成和体外克隆最近有所发展，这种方法对于完整染色体的构建不太有效。通过体内同源重组在酵母菌中进行长度至少为 1 Mb 的染色体高级折叠组装，这是一项稳健的技术，可以用于迄今为止所有报道过的合成染色体，包括病毒、细菌、酵母、藻类染色体以及小鼠和人类基因组片段[12,13]。

在其他遗传可追溯的生物体中还没有发现可以比拟酿酒酵母的 DNA 组装效率。为了拓展用于实现酿酒酵母中难以进行的特异性合成染色体编写的工具，需要开发能够耐受高 GC、直接重复并能实现预期转录后修饰的新型同源重组-成熟克隆生物体。能够适应极端环境，例如沙漠、深海或太空旅行的生物体可能为 DNA 组装提供新的途径，例如，极端微生物耐辐射球菌。

一旦完成了染色体的构建，传递和调控就成为大多数宿主主要的工程学瓶颈。为了进行大规模构建，必须在多种跨种属的生物体中开发强有力、高通量的 DNA 转化方法。例如 DNA 传递领域的突破可以彻底改变植物工程，目前该领域受到物种特异性、劳动密集型转化方法的阻碍，并且局限于传统保守投资方式。高风险、高回报资金支持植物研究的现代化，如开发不依赖于组织培养 DNA 传递方法，有助于改善与合成生物学相关的农业生物。酵母、细菌、植物和哺乳动物细

[12] B. J. Karas et al., Chromosome Res. 23, 57 (2015).
[13] L. A. Mitchell et al., Science 355, eaaf4831 (2017).

胞之间染色体转移方法的自动化（例如细胞融合、基因组移植或微注射）需要在微流控与传统细胞和分子生物学等多学科的工作之间搭建桥梁的融资机会。必须有政府支持，以及在基础研究水平期间获得对早期概念验证工作的支持。

塑造基因组结构和功能的许多细胞力量在很大程度上依然是未知的。需要进行基础研究来阐明序列和表观遗传调控染色体内外相互作用及确定基因组结构的机制。新兴的染色质可编程修饰技术，例如，绝缘体引导的染色质重建模、DNA 插入的安全位点以及正交重组酶，是基因疗法发展所必需的[14]。对极难改造的细胞器基因组（质体、线粒体）的深入理解，将会为人工染色体的稳定维护和合并提供新的途径。

引入合成结构的最后一个挑战是快速确定它们是否在目标单元中按需要执行功能。全基因组 DNA 和 RNA 测序将第一步验证染色体完整性。具有表型报告基因的定制细胞系可能开发用于评估大型合成结构的性能。可靠的类器官模型以及清晰认识驱动组织发育变化的相关调控和表达，这些对于从单细胞到多细胞生物的染色体功能设计的推算来说至关重要。

5. 多部门和多学科

新技术可能来自于合成生物化学，例如可编程的合成原细胞；来自硬件和软件的接口，例如固相 DNA 组装平台；或来自基础生物科学研究的发现，例如通过发现有价值的新型酶或传递系统。

创新将由政府拨款和基因组学与癌症研究驱动，新兴的自动化生物工程中心、生物基地也日益发挥作用。政府及私营部门高度跨学科、跨国界的工作将帮助实现和推广这些成果，以影响生物医学、制药、农业和化学工业等。

刘晓 熊燕 编译自 *Science*

[14] F. Ceroni, T. Ellis, Nat. Rev. Mol. Cell Biol. 19, 481 (2018).