最小细菌基因组的设计与合成

翻译自：
Hutchison et al., 2016. Design and synthesis of a minimal bacterial genome. Science 351 (6280): 1414；aad6253

译者：
浙江大学农业与生物技术学院应用生物科学专业（1501 ）
陶天怡，龚杭荻，蒋知妍

*该译文来自《生物信息学》课程作业（樊龙江主讲）

【论文概述】

1984 年，为了理解生命的基本原则，可自主生长的最小细胞即支原体被用作模型探究。1995 年，我们报告了首个完整的细胞基因组序列（嗜血杆菌，1815 个基因；生殖支原体，525 个基因）。通过比较这些序列，我们发现了一组保守的、大约 250 个必要基因的核心成分，而这一核心组件比其中任一基因组都要小。1999 年，我们引入了全局转座子诱变方法并且通过实验证明生殖支原体中含有许多对于生长并不必要的基因，尽管它是在自然条件下发现的能够自主增殖、且具有已知最小基因组的细胞。这表明，人工合成一个比自然条件下都要小的细胞使有可能的。现在，全基因组可通过化学合成寡核苷酸得到，且将其导入一个受体细胞环境中即可获得可存活细胞。我们使用了全基因组设计与合成的方法来解决最小化细胞基因组的问题。

1）基本原理

从获得首个基因组序列开始，大量工作一直通过比较基因组的方法，在各种细菌模型中鉴别非必要基因和明确保守基因功能核心组。通常，有多个基因产物可提供一种特定的必要功能。在这种情况下，承担这些功能的基因都不是必要的，并且任一基因并非都必需被保留下来。因此，这些方法并不能仅靠它们的结果来鉴明足以组成一个可生活基因组的一组基因。我们从设计并建造一个基因组出发，继而测试其可行性来实验性地简明一个最小细胞基因组。我们的目标仅为单个简单细胞，因此我们可以确定每个基因的分子和生物学功能。

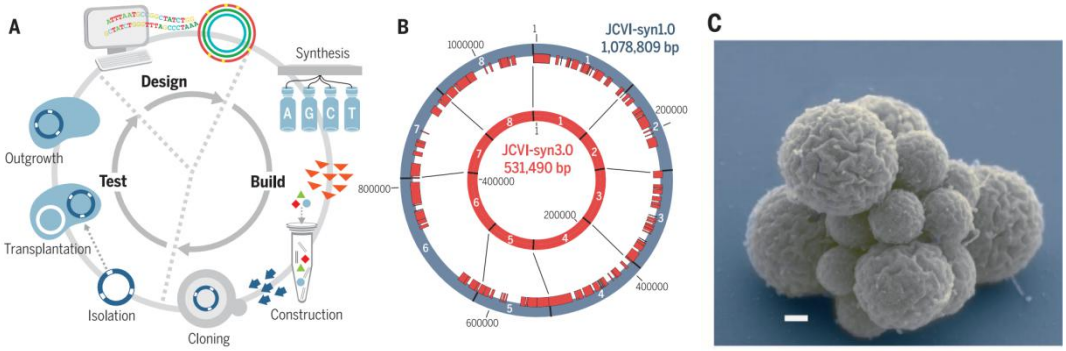
2）结果

我们通过全基因组设计与合成，将丝状支原体 JCVI-syn1.0 的 1079 个碱基对的基因组最小化。根据现存分子生物学知识与有限的转座子诱变数据建立的首个设计并不能产生可生存细胞。优化后的转座子诱变揭示了一组健壮生长所需要的准必要基因，这恰恰解释了我们首个设计的失败。再经由三个设计、合成和检测的循环并保留准必要基因，我们得到了

JCVI-syn3.0（531kbp，473 个基因）。该细胞的基因组比任何天然可自主复制的细胞都要小。JCVI-syn3.0 增殖一次所需的时间约为 180 分钟，且产生的菌落表型与 JCVI-syn1.0 的相似，并且在纤维检测中表现出多态性。

3）结论

最初最小细胞的概念看起来是相当简单的，但如深究，则它的概念变得更为复杂。除了必要基因和非必要基因外，细胞中同时也存在准必要基因。这些准必要基因对于细胞生活性并非绝对必要，但对于健壮生长则是不可或缺的。因此，在基因组最小化的过程中，我们需要在基因组的大小和生长速率中找到一个平衡点。JCVI-syn3.0 是一个接近可行的最小细胞基因组，在小基因组与可行的实验生物生长速率中达到了一定的平衡。它保留了几乎所有涉及到大分子合成与加工的基因。意外地，它依旧含有 149 个未知生物学功能的基因，表明仍有一些生存必需的基因功能未被发现。JCVI-syn3.0 是一个可被用于调查生命的核心功能和探究全基因组设计的通用（versatile）的平台。



四个产生 JCVI-syn3.0 的设计-合成-检测循环

(A) 基因组设计，在酵母中进行合成、克隆以完成建造，通过基因组移植检测可行性的循环。在每个循环后，基因的必要性被通过全局转座子诱变极性重新评估。

(B) JCVI-syn1.0（外侧蓝环）与 JCVI-syn3.0（内侧红环）的比较，通过相分离的 8 个片段显示。外侧环内部的红色条形代表的是 JCVI-syn3.0 中保留的区段。

(C) 一簇 JCVI-syn3.0 细胞，显示其不同大小的球形结构（比例尺，200nm）。

【论文正文】

摘要

我们用全基因组设计和完整的化学合成来减小最小的含有 1079kb 碱基对的合成基因组，丝状支原体 JCVI-syn1.0。最初基于分子生物学与转座子诱变体有限数据信息的设计未能成功产生可存活的细胞。改善后的转座子诱变技术揭示了一类旺盛生长所必须的准必要基因（quasi-essential），这类基因解释了我们最初设计失败的原因。保留准必要基因后，三轮设计、合成和测试的循环创造出了 JCVI-syn3.0（包括 531kb 碱基对，473 个基因），它的基因

组比自然界中为人所知的任何可以自主复制的细胞都要小。出乎意料的是，它还包含了 149 个功能未知的基因。JCVI-syn3.0 提供了一个研究生命核心功能和探究全基因组设计的多功能性平台。

前言

细胞是生命体的最基本单位。可以认为基因组序列是细胞的操作系统，它包含了编码所有细胞功能的基因，而这些又决定了细胞的化学、结构、复制以及其他性质。每个基因组既包含了所有生命体中普遍存在的基因，也包含了物种自身特有的基因片段。基因组的表达依赖于细胞壁的功能，同时，细胞壁的性质又由基因组编码的结构来决定。基因组可以被看作软件，而 DNA 序列让软件序列能够被读取。1984 年，Morowitz 提出了能够自主生长的最小细胞——支原体，并认为它是研究生命基本原理的模型（1）。他的提出的一个关键的早期步骤是对支原体基因组进行测序。而在 1995 年，人们第一次完成了对生殖支原体基因组的测序（2）。即使已经知道了这一序列，破解操作系统仍然是一项艰巨的任务。

长期以来，我们对在实验室的理想条件下通过删除对细胞生长不重要的基因来简化细菌细胞的基因软件很感兴趣，这有助于了解所有生命体必要的基因的分子与生物学功能。大多数细菌细胞为了能够存活，必须要能够适应不同的环境。典型的且研究较多的细菌，如枯草芽孢杆菌和大肠杆菌，携带 4000-5000 个基因。它们因为携带了许多在一些特殊环境下能够提供特殊功能的基因而有很强的适应性。然而，有些细菌生长环境严格，并在进化过程中已经经历了基因组减少，所以他们已经丢失了在这一稳定环境下所不必要的基因。支原体仅在动物寄主营养丰富的环境下生存，其拥有已知的能够自主复制细胞的最小基因组。比较最早两个被测得的基因组序列，流感嗜血杆菌[1815 个基因(3)]和生殖支原体[已知的最小支原体基因组; 525 个基因(2)]，可知共同的核心基因只有 256 个，比两者的基因组都要小得多。这被认为是生命体所需的最小基因。

1999 年，为了将这一比较研究进行实验测试，我们引入了全局转座子诱变法(5)。通过这一方法，我们分出了生殖支原体中的 150 个非必要基因，并预测了 375 个必要基因。这些结果表明创造一个比自然界中任何基因组都要小的最小基因组是可能成功的，但是这一最小基因组比 256 个共同基因要大。在那时，我们打算构造并验证一个基于盒式的最小人造基因组(5)。在那之后，我们为了达到这一目的建造了工具，开发了一种化学合成生殖支原体基因组的方法。但是生殖支原体的生长速度很慢，所以我们转用生长更快速的丝状支原体来设计最小基因组。我们优化了基因组转移的方法，使丝状支原体的基因组可以以独立 DNA 分子的形式转入到一个新的种，山羊支原体，的细胞中(8,9)。这一过程中，山羊支原体的基因组丢失，细胞中仅含有转入的外源基因组。在 2010 年，我们报道了 *M. mycoides* JCVI-syn1.0 基因组[1,078,809bp(10); 以下缩写为 syn1.0]的完整的化学合成以及组装合成途径。除了几个水印标记与靶标序列以外，这个基因组几乎与野生型丝状支原体的基因组完全相同。

之前的一系列连续的序列缺失实验成功减小了细菌基因组，比如大肠杆菌和枯草芽孢杆菌(11,12)。每次缺失后，他们的生存能力、生长速度和一些其他表型都会发生改变。与这种方法相反，我们开始设计一个最小基因组，并对之进行构建与测试。最初，基于现有的转座子诱变和缺失数据以及已有文献中报道的分子生物学相关知识，我们设计了一个最小基因组（hypothetical minimal genome, HMG）。

我们将基因组分为八个片段设计，其中的每个片都都可以被独立的在其他 7/8 个 syng1.0 基因组（即，完整的的 7/8 的 syn1.0 基因组）中测试其存活性。最初设计的 HMG 八个片段中，只有一个片段的基因组区域是可存活的。在对全局转座子诱变方法进行改进之后，我们能够确切地将基因分类为必要或不必要基因，并确定旺盛生长所必需的但不是绝对必要的准必要基因[见图 S1-S4(9); 肺炎支原体中也有类似的结果(15)]。我们还建立了从基因组设计中除去基因而不干扰其他基因表达的规则。在构建 syn1.0 的过程中，我们创建了一种将新的基因组构建为酵母的着丝粒质粒的新方法，并将它们转入山羊支原体受体细胞中来检测它们的生存力和其他表型性状。这里所提到的改进方法以及实验方法组成了设计-构建-测试(DBT)循环（见图 1）。

这篇文章中，我们报道了一个由 531kbp 的合成基因组控制的新细胞，并命名为 JCVI-syn3.0（缩写为 syn3.0），它编码 438 个蛋白质和 35 个注释的 RNA。这是最接近于最小细胞的结构，基因组大小比生殖支原体更小，倍增速度大约是生殖支原体的 5 倍。

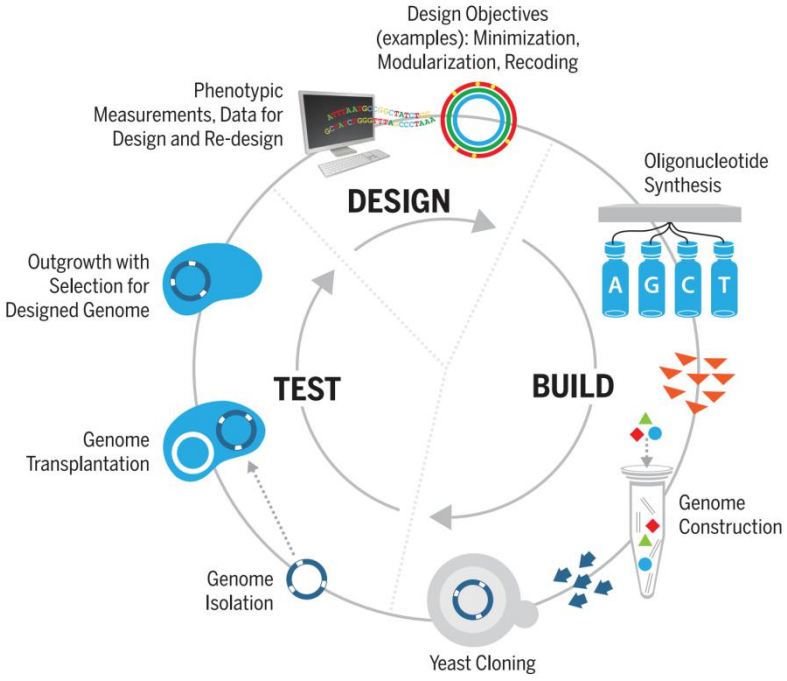


图 1 细菌基因组的 JCVI DBT 循环。在每个循环中，就像构建一个酵母菌的着丝粒质粒一样来构建基因组，然后通过移植将基因组导入一个山羊支原体受体。在这个研究中，我们最主要的设计目标是基因组的最小化。从 syn1.0 开始，我们通过移除非必要基因（用全局 Tn5 插入法鉴定）设计了一个简化基因组。八个减小了的片段均在一个 7/8 的 syn1.0 基因组环境内进行检验，然后和其他减小的片段结合。在每个循环中，

被减少了的最小可存活基因组以及可健壮生长的 syn1.0 的片段上基因的必要性都会用 Tn5 诱变进行重新评估。

结果

1. 基于原有知识设计的 HMG 不能产生可存活细胞

第一次试验，我们以 syn1.0(10)为基础，结合生物化学论文中的知识以及转座子诱变数据来设计一个合理可行的最小细胞。被转座子插入后不会影响细胞存活力的基因被称为不必要基因。向 syn1.0 基因组中插入约 16,000 个 Tn4001 和 Tn5 转座子后，我们发现 440 个不必要基因并将它们从 syn1.0 的基因组中敲除。最后得到的 HMG 有 483kbp，其中包括 432 个蛋白基因和 39 个 RNA 基因（数据库 S1 罗列了具体的基因列表）。

在设计 HMG 的过程中，我们制订了一套在整个项目中使用的删除规则：（i）总的来说，每个不必要基因的整个密码子区域都被删除，包括起始与终止密码子（例外情况如下所述）。（ii）当含有多于一个连续基因的基因簇被删除时，则基因簇内的基因间区也同时被删除。（iii）保留被删除的基因或者基因簇两端的基因间隙。（iv）如果待删除的基因与要保留的基因部分重复，则保留该部分重复基因。（v）如果待删除的基因部分包含保留基因的核糖体结合位点或者启动子，则该段基因部分被保留。（vi）当两个基因单独发生转录时，我们假设两者之间的基因间区包含两者转录的启动子。（vii）当删除基因会导致转录融合，且不存在双向终止子时，插入双向终止子。

由于设计可能存在缺陷，我们将基因组分为 8 个可以独立合成、测试，且含有重叠部分的片段。先前我们用这种方法找到了 syn1.0 构建体中的一个致死点突变(10)。如前所述，这八个设计合成的片段在 syn1.0 中都有对应的片段。这样我们可以将未测试片段与可存活的 syn1.0 片段进行对应单片段或特定组合混合匹配(16, 17)。此外，八个靶标片段中的每一个都通过重组酶介导的盒式交换法[RMCE; (18)](图 S10)转入到 7/8 的 syn1.0 环境中。在所得的酵母菌株（图 S9 和表 S12）中，特殊的限制性位点（NotI 位点）位于每一个 HMG 或者 syn1.0 片段的两端（9）。转入后，我们获得了带有所有可存活 HMG 片段（侧接 NotI 位点）的支原体菌株的和其他八个支原体菌株，每个都携带一个 syn1.0 片段（侧接 NotI 位点）。这加快了 1/8 的基因组片段的产生，因为它们可以从细菌培养物中回收，而细菌培养物产生的 DNA 比酵母产生的产量更高且质量更高（9）。全部 8 个 HMG 片段在 syn1.0 环境中进行了测试，但是仅有一个片段设计能够得到可存活、生长能力弱的菌落（HMG 片段 2）。

可能我们从 HMG 的实验中得到的最有价值的，是通过错误纠正程序改进了半自动 DNA 合成方法。之前我们开发了多种从寡核苷酸衍生到整个染色体的 DNA 合成和组装方法。在这项工作中，我们优化了方法，可以从半自动 DNA 合成途径中的重叠寡核苷酸开始，快速生成无差错的大型 DNA 构建体。这一方法严格遵守以下协议：（i）从重复的寡核苷酸开始的 1.4kbp DNA 碎片的大量单次反应，（ii）去除综合误差、简化大量单次反应和无错 7-kbp 盒克隆（iii）能够在一次反应中同时鉴定数百个无错克隆的盒序列检验，以及（iv）酵母大

质粒 DNA 的滚环扩增法（RCA）。以上这些方法都大大提高了 DBT 循环的运行速度（9）。

图 2 以 HMG 为例，说明了我们用来进行全基因组合成与组装的常用方法。从 DNA 序列设计开始，为了构建重复的寡核苷酸序列，我们建立了自动化基因组合成协议（9）。简言之，软件参数包括了装配阶段的数目、重复序列的长度、最大的寡核苷酸大小和促进聚合酶链反应（PCR）扩增或克隆和分层 DNA 组装的附加序列。在单个反应中，将 48 个寡核苷酸合并、组装并扩增从而产生 1.4kbp 的 DNA 片段（图 S12 和 S13）（9）。这些 1.4kbp 的 DNA 片段随后进行错误修复、再扩增、每次装配 5 个片段并将之转入到大肠杆菌中。在 DNA 测序仪（Illumina MiSeq）上检测无错 7kbp 盒，并将 15 个盒在酵母中组装合成 1/8 的分子。超螺旋质粒 DNA 从结果呈阳性的酵母克隆中被筛选出来，RCA 被用于酵母中全基因组组装所需的微量 DNA 的合成（图 S14 至 S16）。该全基因组合成工作过程可以在三周内完成，比 2008 年第一次报道合成的细菌基因组（由我们团队完成）快两个数量级（7）。

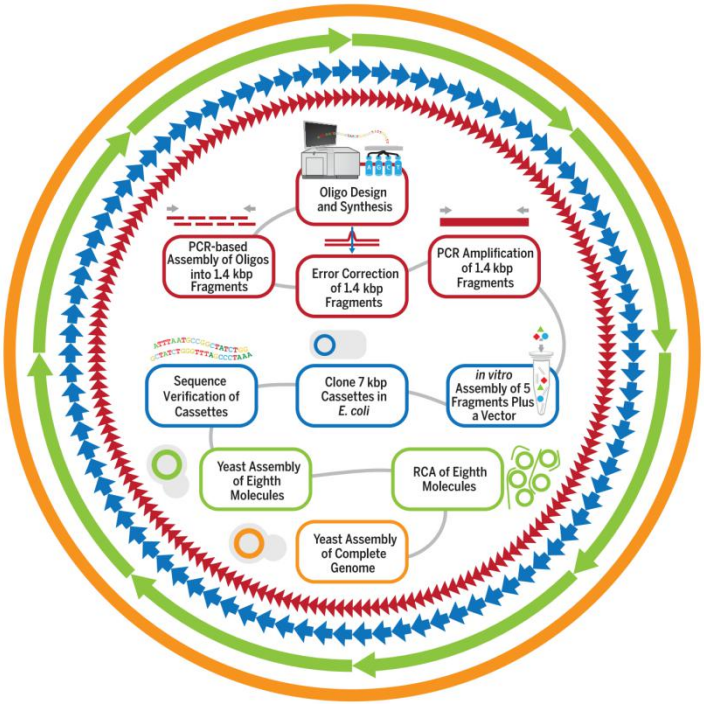


图 2 全基因组合成策略。设计有重叠的寡核苷酸，进行化学合成，并且组装成 1.4kbp 的片段（红色）。在误差校正和 PCR 扩增之后，五个片段被组装成 7kbp 的盒（蓝色）。序列检测盒后，将其在酵母菌内组装得到 1/8 个分子（绿色）。RCA 扩增这八个分子，然后在酵母菌内重组得到完整的基因组（橙色）。

2. Tn5 转座子诱变鉴定了必要、准必要和不必要基因

从我们的 HMG 成功的设计中可以清楚地看到，我们需要更好地了解哪些基因是必要的、哪些是不必要的。为了达到这一目的，我们用了 Tn5 转座子诱变（图 S1）。最初的 Tn5 诱变图谱是通过将 988bp 微型 Tn-5 嘌呤霉素抗性转座子（图 S1）（9）转入 JCVI-syn1.0 的△RE △IS 细胞[它们全部的限制酶（RE）基因和 6 个插入（IS）元件被除去；表 S6]的方法建立的。在含有 10mg/ml 嘌呤霉素的琼脂平板上选择转化的细胞。平板上总共生长出约 80,000 个单

Tn5 转入的转化子。对这个"P0"库中提取的 DNA 样品用反向 PCR 以及 DNA 测序技术进行机械剪切及 Tn5 插入位点的分析。P0 的数据建立了约 30,000 个独立的插入位点。为了去除生长缓慢的诱变细胞，将 P0 库中的细胞连续传代培养 40 多代，为了构建"P4"数据库，制备 DNA 并测序，这个数据库包含约 14, 000 插入位点。(图 S2)

基因可以分为三大类：（i）从未被命中的基因，或在 20%的 3'末端或 5'末端前几个碱基中被偶尔命中的基因，被分类为必需基因 （"e-genes"）（5）（ii）被 P0 和 P4 转座子频繁命中的靶标基因，被归类为非必需基因（"n-genes"）（iii）主要由 P0 插入而不是 P4 插入的基因，被归类为半必要或准必要基因，删除这类基因会导致生长损失（"i-genes"）。 准必要基因被破坏的细胞会经历连续的不同程度的生长损伤，从最小到最严重不等。为了区分这种生长损伤连续性，我们将缺失后生长损伤最小的准必要基因定义为"in-genes"，而将生长缺陷严重的那些定义为"ie-genes"。在 syn1.0 基因组的 901 个注释蛋白和 RNA 编码基因中，最初，432 个被分类为 n-基因，240 个为 e-基因以及 229 个为 i-基因（图 3 的 A 和 B，和图 S3）。

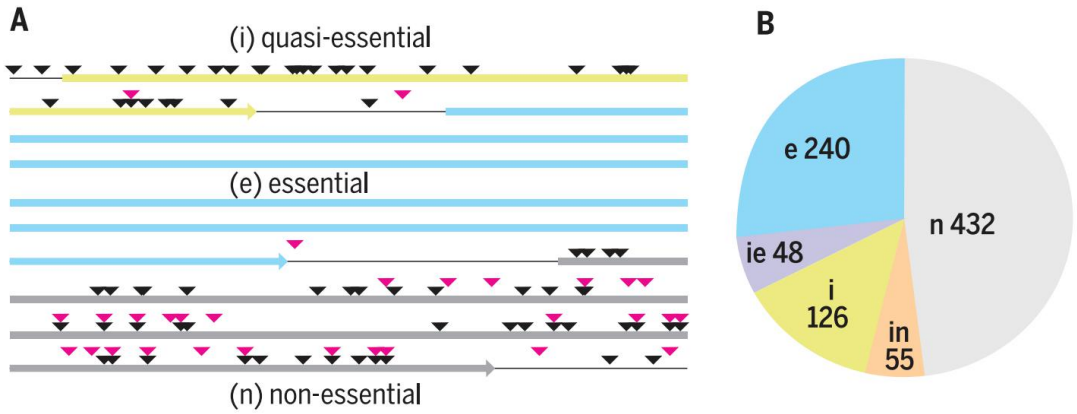


图 3 用转座子诱变进行基因必要性分类。（A）基于 Tn5 诱变数据的三种基因分类实例。syn1.0 序列第 166,735 位到 170,077 位被表达。基因 MMSYN1_0128（黄绿色箭头）有很多 P0 Tn5 插入（黑色三角形），是一个准必要基因（i-gene）。下一个基因即 MMSYN1_0129（亮蓝色箭头）无插入，是必要基因（e-gene）。最后一个基因即 MMSYN1_0130 (灰色箭头)，在 P0（黑色三角形）和 P4 代中（品红色三角形）都有插入，是非必要基因（n-gene）。基因间的序列由黑色线条表示。（B）每个 Tn5 诱变分类类别中 syn1.0 基因的数量。非必要基因（n-gene）和 in-gene 都是候选的待删除基因。

P4 插入的 syn1.0 的图（图 S4）清楚地表明，非必要基因在基因簇中发生的频率远高于我们预期的偶然发生频率。我们通过删除分析，证实了大部分非必要基因簇的删除不会使其丢失存活力或者影响其生长速度（9）。单独的基因簇（或者有时单个基因）按照下述规律被酵母 URA3 标记物代替。两端带有待删除基因的 50bp 序列通过 PCR 添加到 URA3 标记物的末端，并将该段 DNA 插入到带有 syn1.0 基因组的酵母细胞中。在不含尿嘧啶的平板上筛选克隆酵母，PCR 进行验证并通过转接验证其存活力。删除可以分为三类：(i)转接不成功，表明删除的是一个必要基因；(ii)转接后生长率正常或接近正常，表明删除的是非必要基因；(iii)

转接后生长缓慢，表明被删除的是准必要基因。

包括所有 HMG 缺失在内的大量缺失都分别进行了生存力测试并为后续的基因组缩减设计提供了重要信息。在 HMG 设计时可获得的转座子插入数据全部都是从 P0 库中收集得到的。因此，带有插入基因的基因包括了随后被判定为是准必要基因的非必要基因，所以一些 HMG 缺失体产生的菌落非常小或者不能存活。

与下述的 DBT 循环迭代的同时，我们还采用传统的序列删除方法来减小基因组。我们对中等到较大的基因簇进行了逐步到无痕缺失（图 S8 和表 S11）（9），从而构造一系列更多基因被去除的菌株。去除了 255 个基因和 357kbp DNA 的菌株 D22 生长速度类似于 syn1.0（表 S6）。 当我们发现每个 DBT 循环后合成的，重新设计的片段都可以更快的得到更小的细胞时，我们便停止了这种方法。这些删除研究也有助于验证我们的最简删除规则。

3. 保留准必要基因可产生功能片段，但无法产生完整的功能基因组

为了优化 HMG 的设计，我们利用 Tn5 和上述的删除数据重建了一个基因组。这个重组基因组（RGD1.0）是 syn1.0 的 50%删除简化版本，即是通过删除 90%非必要基因（表 S1）得到的。在少数情况下，非必要基因被保留了下来：比如，如果这些基因的生化功能必要，或它们是两个必要基因或准必要基因簇之间的唯一基因即被保留。为了保护删除区段上下游的基因表达，我们沿用了 HMG 设计的规则。

RGD1.0 的 8 个片段以上文的方法进行化学合成得到，通过 RMCE 方法，在酵母体内将每一个删减后的合成片段均插入至 7/8 syn1.0 背景中（图 S10 和表 S13）。每一个 1/8 RGD- 7/8 syn1.0 基因组随后便从酵母中移植出来，以测定其可行性。每一个删减后的片段均可以产生可行的移植体；然而，片段 6 在最初的 6 天里仅能产生很少的菌落（colony）。在后续 6 天的培养过程中，出现了部分生长加速的细胞（图 S18）。我们对其中几个独立的快速生长细胞进行了测序，并发现这些细胞中由于删除过程产生的转录终止子变异是不稳定的，且这些转录终止子是由于其上游的非必要基因被删除后被迫与一个必要基因相连的（图 S19 和 S21）。另一种突变则是在必要基因前产生了连续的 TATAAT 框（图 S20）。这阐明了基因删除可能会导致表达错误，但同时，它也证明这些错误有时可通过后续自发的突变而修正。最终，我们鉴别出了一个在先前设计的过程中被忽视且被错误地删除的启动子。当这一启动子根据设计规则被补充进去后，具有设计合成的片段 6 的细胞即可快速生长。这一解决方法也被整合入后续的设计中。

与含有各设计片段的细胞的生长情况相反，将所有 8 个 RGD1.0 片段（其中片段 6 为自修复版本）整合入单一基因组中，如将其导入山羊支原体（*M.capricolum*），依然无法产生可存活细胞（9）。随后，我们将 8 个 RGD1.0 片段与 8 个 syn1.0 片段混合，在酵母内形成组合装配的基因组。我们获得了一部分完全重组的基因组，并且其中的 RGD1.0、syn1.0 的片段组合是不同的。将其转移时，部分组合产生了可存活细胞（表 S7）。其中一个含有 RGD1.0

的片段 2、6、7、8 和 syn1.0 的片段 1、3、4、5（RGD2678）具有可接受的生长速度（105min 增殖一倍，而 syn1.0 的增殖速度为 60min），并在后续过程中被更详尽地分析。

4. 为了获得一个功能基因组，避免删除有必要功能的冗余基因对

在细菌中，多个基因行使某一特定必要或准必要功能是非常常见的。这些基因可能是同源也可能不是。假定基因 A 和 B 均提供了某一必要功能 E1。当这两个基因被分别单独删除时，并不会引起 E1 功能的丧失，因此通过单敲除研究，这两个基因均会被定义为非必要基因。然而，如果它们被同时删除，细胞则会因为缺失 E1 功能而死亡。这种致死突变组合被称为是合成致死对（synthetic lethal pair）（19）。必要功能的冗余基因在细菌基因组内是十分常见的，尽管它们在诸如支原体一类的经历长期进化删减的生物基因组中的含量较少。因此，我们在设计过程中遇到的最大挑战即为合成致死对，其中，基因 A 在一个片段中被删除，基因 B 在另一个片段中被删除；这两个单一片段在 7/8 的 syn1.0 背景下均是可行片段，但将其拼装时所得到的细胞即不可存活，如这两个基因仅共享某一准必要功能，则该细胞生长十分缓慢。我们并不知晓在这八个片段中分别有多少个必要功能的冗余基因，但当 RGD1.0 的片段 2、6、7、8 相组合时，该细胞为可存活的。

我们对 RGD2678 进行了 Tn5 检测，并发现在此背景下，一些 syn1.0 片段 1、3、4、5 中曾被认为是非必要的基因此时变成了准必要或必要基因。我们推断，这可能是由于这些基因行使的功能与 RGD2678 中删去的基因成冗余关系。

此外，我们检测了 39 个在 RGD1.0 片段 1、3、4、5 中被删去的基因簇和单基因（表 S8）。这些基因在 RGD2678 的背景下每次仅被删去一个（表 S8 和 S14），并且通过移植体测试其可存活性。某些情况下，所生成的移植体生长十分缓慢，甚至根本无移植体产生，表明这些删去的片段中包含了一个或多个与片段 2、6、7、8 中删去的基因呈冗余关系的基因。

结合 Tn5 和删除数据，我们鉴别了 26 个基因（表 S2 和 S9），并可能可以加回 RGD1.0 片段 1、3、4、5 中产生新的 RGD2.0 设计候选（图 S5 和表 S1、S2）。利用这些新设计并合成的 RGD2.0 片段 1、3、4、5，联合 RGD1.0 片段 2、6、7、8，我们在酵母体内完成了一次新的装配（表 S7、S15）。最初，这个装配并不可行，但我们发现用 syn1.0 的片段 5 来替换 RGD2.0 的片段 5 即可产生可行的移植。在这一情况（strain）下进行研究，我们删除了 syn1.0 片段 5 上的一簇基因（*MMSYN1_0454 至 MMSYN1_0474*），并用基因 *MMSYN1_0154* 来替换另一簇基因（*MMSYN1_0484 至 MMSYN1_0492*）（图 S6、S11 和表 S10）（9）。基因 *MMSYN1_0154* 最初从 RGD1.0 的片段 2 上被删除，但将其添加回 RGD2678 时，发现其有促进生长速度的功能。经过上述 RGD2.0 对 syn1.0 片段 5 的修改，这一版本的基因组可产生可存活的细胞，我们将其称作 JCVI-syn2.0（简称 syn2.0；图 4）。根据 syn2.0，我们首次合成了比自然存在的最小细菌即生殖支原体（*M.genitalium*）基因组更小的基因组。syn2.0 在实验室条件下，每增殖一倍的时间为 92 分钟。其全基因组长度为 576kbp，包含 478 个蛋白质编码基因和 38 个

来自丝状支原体的 RNA 基因，并含有用于基因组选择和在酵母和大肠杆菌内繁殖所必须的 12kbp 载体序列。

5. 再移除 42 个基因获得仅次于最小基因组的 syn3.0

我们对 syn2.0 进行了新一轮 Tn5 转座子诱变。在这个新的遗传背景下，我们预计准必要基因向非必要基因的转变是存在的。此处，经连续传代的 P4 细菌群体已经耗尽了原有的非必要基因；影响迅速生长的准必要基因敲除占主导，且根据我们的规则，被归为非必要基因。我们将 90 种基因归为显然非必要基因，且它们被分为了三个亚组。第一组含有 26 个基因，这些基因在先前的诱变循环中常被鉴定为准必要基因或必要基因。第二组含有 27 个基因，在前先的几轮 Tn5 研究中被鉴别为准必要基因或临界准必要基因。第三组组含有 37 个基因，在先前不同基因组环境中，有部分 Tn5 转座子诱变重复显示其为非必要基因。为了完成 RGD3.0 的设计，我们将 syn2.0 中的这 37 个基因删除，包括其中的两个载体序列 *bla* 和 *lacZ* 以及片段 6 中的核糖体 RNA 起始子（图 4 和表 S3）。

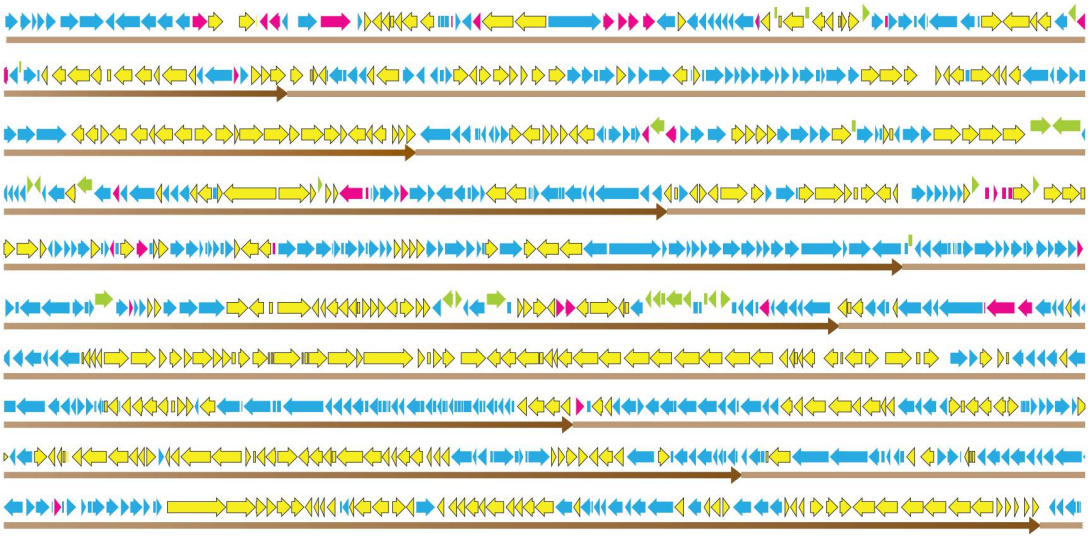


图 4 构建 syn3.0 时的三个 DBT 循环。这张细节图展示了从 syn1.0 到 syn2.0 最终到 syn3.0 过程中，在 syn1.0 基础上，在 DBT 循环中被删除或是加回的基因（相较于 fig.S7）。长棕色箭头表明八个 NotI 组装片段。蓝色箭头代表在这个过程中保持不变的基因。黄色表示在 syn2.0 和 syn3.0 中都被删除的基因。绿色箭头（略有偏移）代表加回的基因。一开始的 RGD1.0 设计无法存活，但是 syn1.0 的片段 1，3，4，5 加上设计的片段 2，6，7，8 后整合的基因组，即 RGD2678 就能存活。在 syn2.0 中另外附加的基因表示为绿色，有八个设计的片段。另外删除的区段用品红色表示，在此基础上产生了 syn3.0（531560bp，473 个基因）。箭头与转录和翻译的方向相同。

新合成的 8 个 RGD3.0 片段经酵母质粒合成并增殖。这些质粒通过 RCA 在体外进行扩增（9）。这 8 个片段随后在酵母中被组装以获得不同组合版本的 RGD3.0 基因组。这些组装所获的 RGD3.0 基因组随后从酵母内转入山羊支原体（*M.cap*）受体细胞，且其中的一些组合

是可行的。其中一个，RGD3.0 菌落 g-19（表 S4）被选作详细分析，并命名为 JCVI-syn3.0。

在 syn3.0 中，我们进行了最后一轮 Tn5 转座子诱变以明确哪些基因在连续传代后依然出现 Tn5 插入。最频繁的插入位点为非必要的载体基因和基因间的序列。与预期一致，连续传代后所得的、在支原体基因上有插入的细胞多在原先被定义为准必要基因上有插入 syn3.0 中的基因在 syn1.0 中多被分类为必要或准必要基因。其中，仅有准必要基因能够发生不致死的 Tn5 插入。最显著的准较比要（ie-gene）、准必要、准不必要（in-gene）参见表 S5。在第四次 DBT 循环中，将其中一些基因删除是有可能的，但这可能会进一步损伤生长速率。此外，一部分非必要基因在此轮检测中依旧被归为非必要基因（表 S5 和数据 S1）。

6. syn3.0 有 149 个基因无法明确其生物功能

Syn3.0 具有 438 个蛋白编码和 35 个 RNA 编码基因。我们根据他们功能的准确程度将这 473 个基因归为 5 类：功能保守的同源蛋白（equivalog），大概率的（probable），推定的（putative），类属的（generic），未知（unknown）（图 5 和数据 S1）。其中的许多基因已经被详尽研究，且它们最基本的生物功能已知。

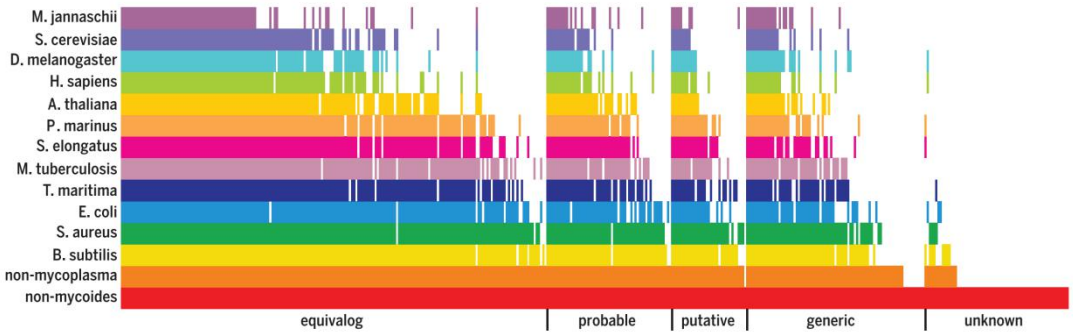


图 5 syn3.0 与其他有机体内同源蛋白基因谱图。用 BLASTP 软件将 syn3.0 中的蛋白编码基因和从非丝状支原体到古菌的 14 种有机体进行比对。以 1e−5 作为相似性截止值。从左到右是从基因功能近完全匹配到未知、缺乏功能信息的五个类别（equivalog, 232 genes; probable, 58 genes; putative, 34 genes; generic, 84 genes; unknown, 65 genes）。白色空位表明 syn3.0 和其他有机体没有同源组。

我们利用基于隐马尔可夫模型的 TIGRfam 同源家族（equivalog family）（20）对 eqiuvalog 基因（大约占总基因的 49%）进行注释。较为不确定的类别则通过逐步方式来定义（图 5）。probable 组所包括的基因在明确（unambiguous）TIGRfam 数学模型中得分较好，但它们的得分并没有达到置信范围内。这些基因同时也被其他证据所支持。基因组背景与晶体结构的线性比对也符合这一分类。putative 组包含的基因也被多种证据所支持；同时，它们的得分、基因组背景或已知生理活动相关结构的联配结果均不可信。generic 组中的基因能够编码明确的蛋白质（如激酶），但对于它们所对应的底物与生物学功能的线索还十分匮乏。unknown 基因是那些虽然可以推断其生理活性，但未能进行明确归类的基因。

因此，占比约 31%的 generic 和 unknown 组别的基因其生物学功能并未被确定。然而，

在其他的有机体中，我们找到了其中一些的可能同源基因，且其大部分编码了通用的、但功能并未被确认的蛋白质。这五个类别在其他生物——从支原体到人类，均存在同源组。但是，每组中均有一些基因的注释是空白的，这表明在我们选取的生物中（参见图 5），这些基因不存在同源组。因为支原体进化十分迅速，图 5 中的一些空白区域与那些较为失败的序列联配结果是相关的，因其在此过程中与其他生物产生了较大的分歧。

在表 1 中，我们将 syn1.0 的基因中的基因归入 30 个功能组，并标明其中有多少个在 syn3.0 中被删除或保留。在 428 个被删除的基因中，最大的组为未明确功能的基因；213 个 unknown 基因中，134 个被删除。所有的 73 个移动元件、DNA 修饰与限制基因均被删除，而大多数脂蛋白编码基因（87 中的 72 个）也被删除。这三类就占据了被删除基因的 65%。此外，因为实验所用的富化培养基提供了几乎所有必须的小分子，许多与转运、分解代谢、蛋白质水解和其他代谢过程相关的基因都变得非必需了。例如，因为葡萄糖在培养基中含量丰富，许多其他碳源的转运、分解基因均被删去（36 个中的 34 个），而涉及葡萄糖转运和糖酵解的 15 个基因全部被保留。

表 1，根据功能分类以及 syn3.0 中是否保留或被删除的 syn1.0 基因列表。

Functional category	Kept	Deleted
Glucose transport and glycolysis*	15	0
Ribosome biogenesis*	14	1
Protein export*	10	0
Transcription*	9	0
RNA metabolism*	7	0
DNA topology*	5	0
Chromosome segregation*	3	0
DNA metabolism*	3	0
Protein folding*	3	0
Translation*	89	2
RNA (rRNAs, tRNAs, small RNAs)*	35	4
DNA replication*	16	2
Lipid salvage and biogenesis*	21	4
Cofactor transport and salvage*	21	4
rRNA modification*	12	3
tRNA modification*	17	2
Efflux*	7	3
Nucleotide salvage	19	8
DNA repair	6	8
Metabolic processes	10	10
Membrane transport	31	32
Redox homeostasis	4	4
Proteolysis	10	11
Regulation	9	10
Unassigned	79	134
Cell division	1	3
Lipoprotein	15	72
Transport and catabolism of nonglucose carbon sources	2	34
Acylglycerol breakdown	0	4
Mobile elements and DNA restriction	0	73
Total	**473**	**428**

*标注星号的类别大多在 syn3.0 中被保留，没有标注的在 syn3.0 中被删除。用于选择基因组和在其他宿主中繁殖的靶标序列没有被包括在这个目录里。

相反的，几乎所有参与到阅读并表达基因组遗传信息、确保在传代过程中遗传信息保存

的基因均被保留了下来。在这两个基本生命过程中的前者，即诸如蛋白质等遗传信息的表达，需要的转录、调节、RNA 代谢、翻译、蛋白质折叠、RNA（rRNA，tRNA 和小 RNA）、核糖体生物合成、rRNA 修饰、tRNA 修饰的 195 个基因全部被保留。而后者，保存基因组序列信息所需要的 DNA 复制、DNA 修复、DNA 拓扑化、染色体分离和细胞分裂的 34 个基因也被保留。这两个生物过程总计占据了 syn3.0 中 473 个基因中的 229 个（48%）（图 6）。

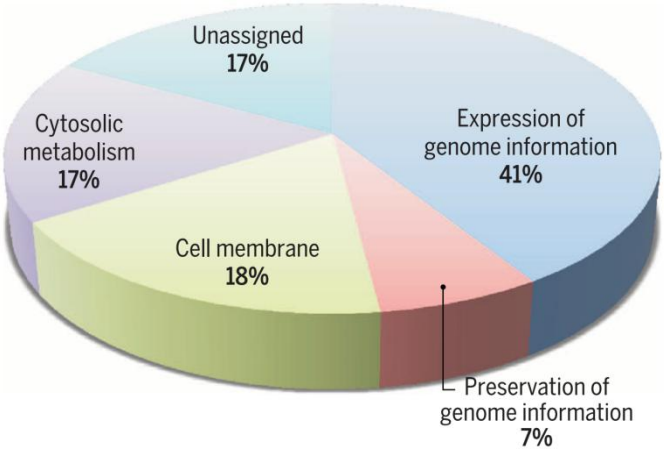


图 6 四种主要功能类别的基因占比。Syn3.0 中有 473 个基因。其中，79 个基因没有指定的功能分类（表 1）。其余可以指定的四种主要功能分类的如下：基因表达（195 个基因）；(ii) 基因组遗传信息保持（34 个基因）；(iii) 细胞膜结构和功能（84 个基因）；(iv)胞质代谢（81 个基因）。每个组别的百分比在图上标出。

除上述两个必要的过程之外，另一个活细胞的重要组分为细胞膜，它使得胞质与外界培养基相分离，并负责跨膜分子运输。它是一种分隔性的结构，syn3.0 中的许多基因编码了它的蛋白质组分。因为我们所合成的最小细胞在氨基酸、脂质、核酸和维生素的生物合成能力等方面都十分匮乏，所以它只能依赖富化培养基提供这些必需小分子。这就强调了膜上众多运输系统的必要性。此外，细胞膜脂蛋白的含量十分丰富。膜相关的基因在 473 个 syn3.0 基因中有占 84 个（18%）。表 1 中包括了脂蛋白、协同运输、流出系统、蛋白质转运和其他转运系统的相关基因。最后，81 个主要参与胞质代谢的基因（17%）也被保留了下来，它涵盖了核酸补救、脂质补救和生物合成、蛋白质水解、代谢过程、氧化还原平衡、非葡萄糖碳源转运和代谢及葡萄糖转运和糖酵解的相关基因。

我们推测这 79 个未能确定其功能组的基因大部分实际上也属于同样的四个功能组之一（基因表达、基因组信息保护、膜结构与功能、胞质代谢）。在这 79 个基因中，有 55 个的功能完全未知，而其他 24 个则被定性为 generic，比如有一种水解酶，它的底物和生物学功能都未鉴别。84 个 generic 基因中其他的 60 个已经根据它们的类属划到了特定功能目录下。比如，一个 ABC 转运蛋白被归为膜转运这一组，尽管它的底物是未知的。其中一些未归类的必要基因具有一些未知功能的结构域，而这些结构域在其他很多有机体内都曾被发现。

7. Syn3.0 增殖一倍的时间为 3 小时，且在表型上表现出多态性

当我们将 syn3.0 与起始细胞 syn1.0（图 7A）作比较时，我们发现它们具有相似的菌落表型，且其所显示的特征与 syn1.0 最初合成构建的来源：野生无壁的丝状支原体的亚种 capri 是一致的（10）。而 syn3.0 较小的菌落外形表明它具有较为缓慢的生长速率，且在固态培养基中可能有变化的菌落结构。与之对应的是 syn3.0 在静态液体培养基（图 7B）中生长速率也有所下降，其增殖一次的时间从 syn1.0 的约 60 分钟延长至约 180 分钟，表明 syn1.0 内源繁殖速率更低。然而这一速率远远超过了生殖支原体 16 个小时的增殖时间。（21）。

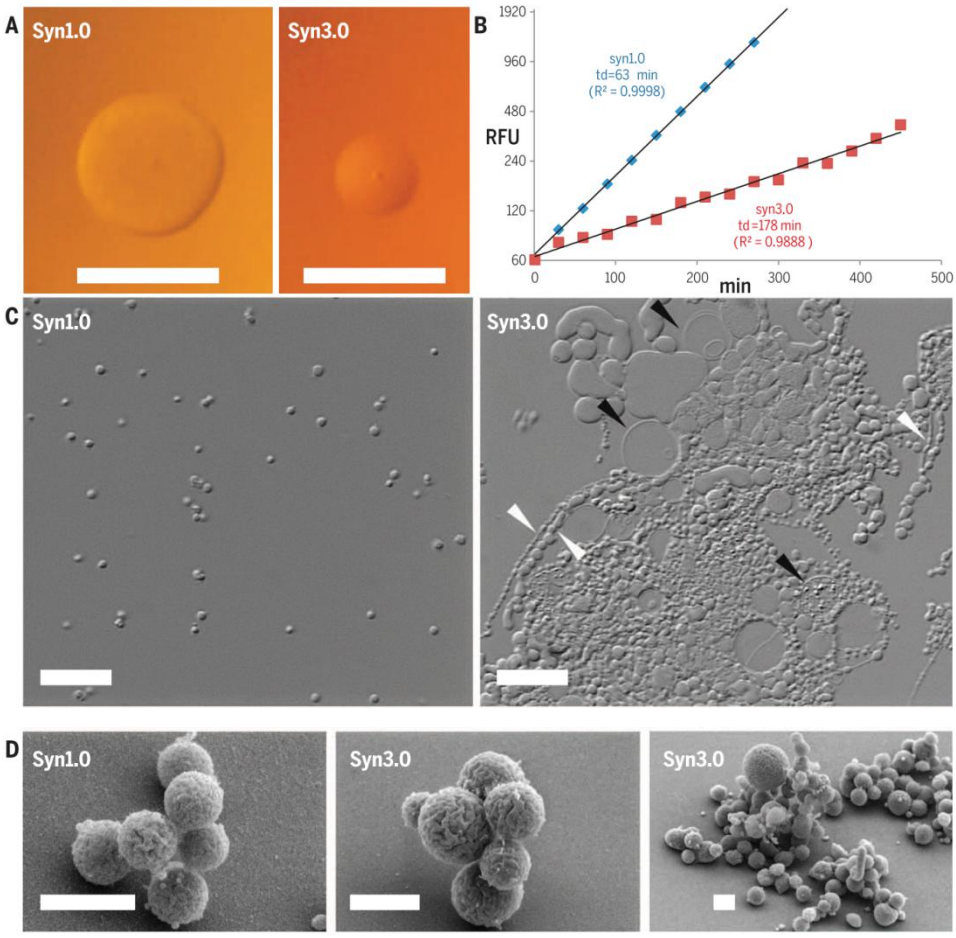


图 7 syn1.0 和 syn3.0 生长特征的比较。（A）以 0.2mm 过滤器过滤培养液中得到细胞，并涂布在琼脂培养基上以比较 96 小时后的菌落大小和形态（刻度尺，1.0mm）。（B）使用荧光量度（相对荧光单位，RFU）随时间（分钟）累积的双链 DNA 计算倍增时间（td），确定液体静态培养物中的生长速率。决定系数（R2）如图。（C）通过差示干涉显微镜术显示液体培养基中湿法制备的天然细胞表型（刻度尺，1.0mm）。箭头表明混杂的分段丝状形态（白色）或是大囊泡（黑色）。（D）扫描电子显微镜下的 syn1.0 和 syn3.0（刻度尺，1.0mm）。右侧的图像为 syn3.0 观察到的多样结构。

与预期内的生长速率削减相反，我们发现了意料之外的 syn3.0 细胞的宏观、微观生长性质。Syn1.0 在静态培养基中培养时，形似不相互粘附的浮游生物悬浮（suspension）在培养基中，且大多数为直径约 400nm 的单细胞。与之相反，在相同的培养条件下，syn3.0 细胞形成缠绕的沉淀物。这些未被扰动细胞的显微图片显示，它们具有由长且分节的丝状结构组

成的巨大网状系统，并伴有大型泡状体（图 7C），这在终末生长期中尤为常见。这两种结构都很容易被物理扰动破坏，但是这种悬浮体也是由小型重复单位构成的。这些单位可以穿过 0.2 微米的过滤器，并且可产生菌落形成单位（CFU）。游离的 syn1.0 培养中有 99.9%的 CFU 也有同样的程式。

8. 探索基因组重组设计

为了进一步优化我们的基因组设计规则，我们也研究了理论组建基因组的前景，并且在核苷酸水平重建了它们。这些过程是为了研究基因的排列顺序与编码序列是否是影响细胞可存活性的主要因素。许多基因组均能够转入长序列的 DNA，这表明总体来说基因的排列顺序是不重要的。我们发现微小的基因顺序也并非影响细胞存活的主要因素。大约 1/8 的基因组序列被重新配置、插入到其他七个连续的 DNA 盒中；在这七个 DNA 盒中，有六个已经处于已知生物系统中，第七个盒中的一些基因在系统水平的联配上具有一定同源性。图 8 中右侧纵轴标明了特定生物系统。独立基因（有色水平线条）和基因间区域（黑色线条）依照左右两侧之间的线条可以将他们原先和新的位置联系起来。具有交叉的线条代表了两个遗传因子相对位置的改变。尽管具有高度重组性，但是比较两者的菌落大小后发现生成的细胞与 syn1.0 生长得一样快。因此，基因组织的微小变化可能会使细胞在高强竞争的自然环境下受到侵害，但极微小的变化对于生命并非很重要。

9. 重编码和 rRNA 替换为基因组可塑性提供了方案

我们设计细菌基因组的 DBT 循环让我们能够根据序列性和功能性来评估基因组的可塑性，这包括检测那些对基本生存所必须的基因修饰方式。我们检测了一个属于必要基因的 16S rRNA 基因（rrs）的修改版本是否能够支持生存（图 9A）。Syn3.0 中 rrs 基因的单拷贝在设计和合成过程中与山羊支原体（*M.capricolum*）的 rrs 基因相比有 7 个单核苷酸变异。此外，我们将螺旋 h39（35 个核苷酸组成）用系统发生上远源的大肠杆菌 rrs 基因对应的部分进行了替换。这个独特的 16S 基因成功地整合到了 syn3.0 中，且并未对生长速率产生显著的影响。我们也构建了其他一些 rrs 基因的变异体，但那些都不可行。这表明我们现在可以检测基因序列的可塑性，且同时提供了一个让我们可以来分辨这些序列的水印。

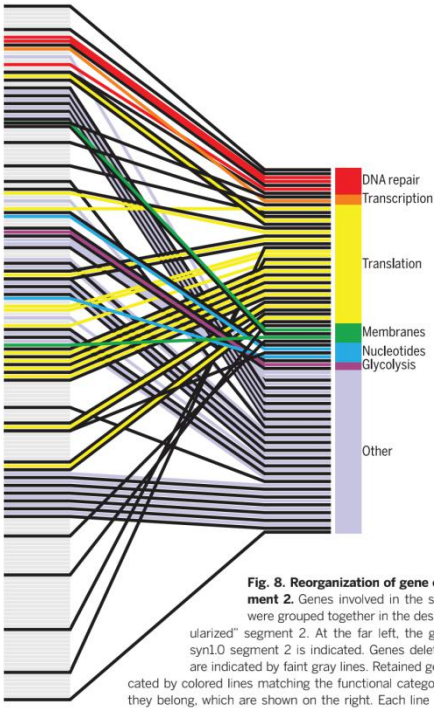


Fig. 8. Reorganization of gene ment 2. Genes involved in the s were grouped together in the des ularized" segment 2. At the far left, the g syn1.0 segment 2 is indicated. Genes dele are indicated by faint gray lines. Retained g cated by colored lines matching the functional catego they belong, which are shown on the right. Each line

图 8 片段 2 的基因序列重组。参与相同过程的基因会组合在一起用来设计“模块化”的片段 2。最左侧的为 syn1.0 的片段 2 基因顺序。Syn3.0 中删除的基因用浅灰色标明。保留的基因用不同颜色的线条代表它们所属的不同功能，对应到右侧。每一条线连接了模块化片段 2 中基因和其相对应的 syn1.0 中的基因位置。黑色的线代表了含有启动子或是转录终止子的基因间序列。

我们同样也检测了丝状支原体（*M.mycoides*）基因组的基本密码子使用规则。丝状支原体有极高含量的腺嘌呤和胸腺嘧啶，且其将 TGA 作为色氨酸的密码子而不是用作终止密码子，它有时也并不适用标准的起始密码子。此外，密码子使用也多为高 AT 含量。我们在含有 3 个必要基因（*era, recO, glyS*）的 5kbp 上根据特异的密码子使用规则进行了修改，以检测各规则在此背景下的功能。特别地，我们对这些区域进行修改时，使之包括①丝状支原体密码子适应指数（codon adaptation index, CAI），但其中起始密码子和色氨酸都用 TGG 进行编码，而非 TGA；②大肠杆菌 CAI，其中的色氨酸仍旧被 TGA 所编码；③大肠杆菌 CAI，且其各项密码子使用都是标准的（TGG 编码色氨酸）（图 9B）。我们意外发现这三种版本都是有功能的，且其生长情况并没有明显差异。然而，如密码子使用规则发生大量变化，我们可能需要修改 tRNA 用量以确保高效的翻译。

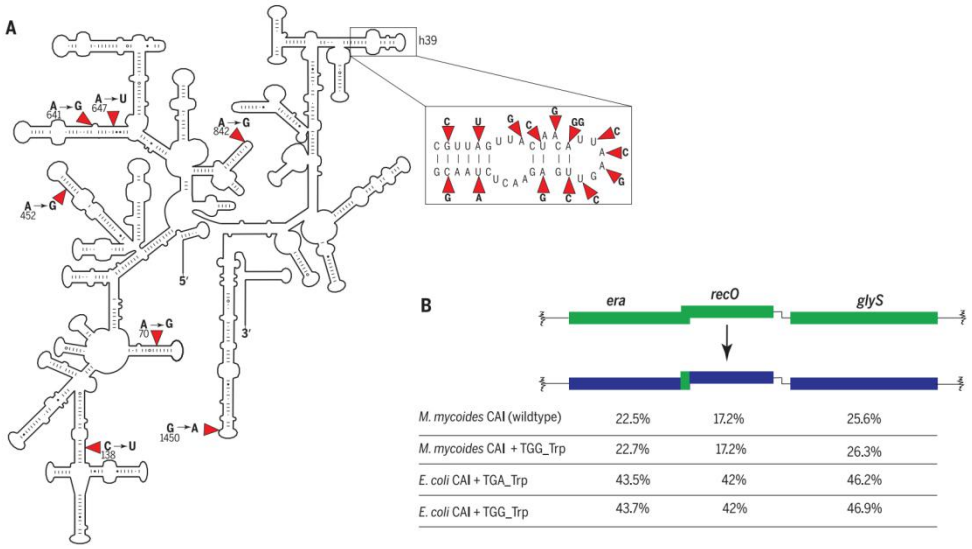


图 9 用 DBT 循环检测基因内容和密码子使用原则。（A）经修改的 rrs 基因的二级结构成功整合入 syn3.0 基因组中；这个基因携带了山羊支原体突变并将 h39（插入）与大肠杆菌的 h39 交换。核苷酸位点的变异用红色箭头标明，大肠杆菌的编号是为了标明山羊支原体的突变位点。（B）必要基因 era, recO, 和 glyS 的序列从三个不同方面进行了修改：用丝状支原体 CAI 进行编码，且 TGG 编码色氨酸；用大肠杆菌 CAI 进行编码，TGG 编码色氨酸；用大肠杆菌 CAI 进行编码，且 TGA 编码色氨酸。野生型和修改后的基因 GC 含量十分显著。JCat 密码子适应工具(www.jcat.de)被用来优化除去重叠基因片段后的三种开放式阅读框架。绿色和紫色分别标明了野生型和密码子优化型序列。

讨论与结论

基因组学这个词原先是一个描述性的词汇，用来描述测序、分析基因组，而后转变为一个合成学的概念，即整个基因组都可以由化学手段合成。鉴于基因需求的种种细节被发掘，从零开始设计一整个基因组将成为可能，用化学手段进行构建基因组，而后将其置于一个合适的受体细胞环境即可。我们通过这种基因组的设计和合成来解决最小细胞基因组的问题。

最小细胞常被定义成一个全部基因都为必要基因的细胞。这个定义是不完整的，因为一个最小基因组能够存活的基因需求和由此决定的最小基因组大小取决于其生长的环境。我们的研究中的这个培养基提供了理论上生命所需的所有小分子物质。一个由如此宽松的环境决定的最小基因组应该揭示了一组不受环境影响的、对生命充分必要的关键功能。如果在一个不够宽松的环境下的话，我们预计构建一个最小基因组还会需要额外的基因。

目前有大量关于最小细胞的概念和有机体的最小必要基因组的文献（回顾参见 22）。这个领域的研究大多集中于比较基因分析和通过插入转座子单个敲除或扰乱基因的实验。这些研究确定了一组核心的必要基因，通常有 250 个左右。但是这并不是一组足以构成一个活细胞的基因，因为在这些研究中，冗余基因被评定为非必要基因。

相反的，我们构建最小基因组是为了通过实验找到一组能使细胞独立复制的基因。我们用丝状支原体（M. mycoides） JCVI-syn1.0 的基因设计了一个基因组（10）。支原体细胞对于达到这个目的有许多优点。首先，支原体的基因组已经非常小了。它从含有大量基因的革兰

氏阳性菌进化而来，丢失了成为哺乳动物寄生菌所不必要的基因。它已经经历了通往最小基因组的漫长进化道路，并且它们因此可能拥有比其他细菌更少的冗余基因。我们还有一系列高度成熟的工具用于构建这个基因组并且将它作为一个额外的酵母染色体来操纵。

我们一开始试图基于现有的分子生物学知识来构建最小基因组，并结合了有限的转座子干扰基因的数据，这些数据提供了基因必要性的额外信息。这些信息在基因未知功能的细节上有极大价值。对于最小基因构建的具体实验仅基于已有的涉及基本生物学过程的相关基因知识（14）。我们的 HMG 由八个片段组合而成，但是被验证为无法存活，虽然其中一个片段（片段 2）在另外七个 syn1.0 片段的环境下是有功能的。这个结果证明从一开始，我们就没有足够的知识从零开始来设计一个运作正常的的最小基因组。因此，为了得到更好的有关基因必要性的信息，我们在转座子诱变方法上做了重大的改进。

为了创造一个包括了所有必要基因和准必要基因的基因组，我们设计了用于细菌基因组的 DBT 循环（图 1），并检测其作为可存活细菌的基因组是否起作用。四次 DBT 循环之后（基因组设计 HMG，RGD1.0，RGD2 和 RGD3.0），我们得到了一个八个片段均减小了的可存活基因组，syn3.0。表 2 总结了得到 syn3.0 的过程。前三个设计并没有得到完整的可存活细胞基因组。但是每一次，都可以得到一个或多个可存活的片段，这些片段和 syn1.0 另外片段结合时可以形成活细胞。表格中列举了部分过程菌株的基因组成。可存活的 syn3.0 细胞是我们得到的近似最小细胞的最佳细胞。我们获得了另外的一个比任何在自然环境下无拘束存活的细胞更加小的一个中间体（syn2.0）。这个细胞包括了七个 RGD2.0 片段加上一个剔除了 31 个基因的 syn1.0 片段 5。

表 2 通过 DBA 循环得到简化基因组，最终得到 syn3.0。

1. Genome design	2. Design size	3. Number of design genes	4. Cellular genome segment composition for key viable strains	5. Cellular genome size	6. Growth rate
-		901	syn1.0: all eight syn1.0 segments	1079 kbp	Doubling time, td = 60 min
HMG	483 kbp	460	HMG segment 2 + 7/8 syn1.0	1003 kbp	Slow-growing
RGD1.0	544 kbp	483	RGD1.0 segments 2,6,7,8 + syn1.0 segments 1,3,4,5	758 kbp	Slow-growing
"	"		RGD1.0 segments 1,2,4,6,8 + syn1.0 segments 3,5,7	718 kbp	Slow-growing
RGD2.0	575 kbp	512	RGD2.0 segments 1,2,3,4,6,7,8 + syn1.0 segment 5	617 kbp	?
"	"		syn2.0: RGD2.0 segments 1,2,3,4,6,7,8 + syn1.0 segment 5 with genes *MMSYN1_0454* to *-0474* and *MMSYN1_0483* to *-0492* deleted	576 kbp	td = 92 min
RGD3.0	531 kbp	473	syn3.0: all eight segments of RGD3.0	531 kbp	td = 180 min

*第 1 列表明了基因组设计的轮次（破折号表示起始基因组，syn1.0），第 2 列标明了设计基因组的大小（以 kbp 为单位），第 3 列是设计出的支原体基因数量。第 4 列表明了活细胞中关键基因组组成；对于无法存活的设计，标明了设计中一个含有最大数量的片段可存活的菌株，并标明了它是一个更加健壮生长的 RGD1.0 被择体（第四行）还是一个更小的 RGD2.0 衍生物（第六行，syn2.0）。第 5 列描述了第 3 列相应的基因组大小，第 6 列则表明了在第 4 列基因组的构建下细胞生长速率的数量或质量评估。

Syn3.0 含有编码 473 个基因、长 531kbp 的基因组。这远小于在纯培养环境下含有自然中最小基因组的丝状支原体（580kbp）。syn3.0 的基因组包含了一系列细胞生命所需的核心基因，但是它仅有 syn1.0 的一半。比较 HMG 和其后导出的可存活 syn3.0 基因组，我们确定了 329 个基因的剔除和 365 个基因的保留。但是，HMG 中被删除的 111 个基因在 syn3.0 中

却被保留，100 个在 syn3.0 中被删除的基因在 HMG 中被保留。这种差异其实是因为初始转座子数据的稀缺以及其质量的低下。数据中有大量被证伪的必要基因和不必要基因，并且没有找到能够影响生长速率的准必要基因（详见下文）。为了说明分类准必须基因的重要性，

举一个用来组成核苷的运输系统的四个基因（MMSYN_0008 至 MMSYN_0011）的准必要基因的例子。原先认为这个系统是核酸 / 半乳糖的 ABC 转运蛋白，这使我们倾向于将它从 HGM 中删除。我们最初的转座子数据显示，4 个基因均多被命中 ，并且由此证实了这些基因是不必要的。然而在后续转座子定位的实验中，P0 代转座子数据证实它们确实有高命中率 ，而在连续传代，生长缓慢的细胞均被剔除后，四个基因都未被命中，说明它们是准必要基因并且应该被保留。

1）Syn3.0 的基因内容

Syn3.0 是最接近最小细胞的成果。我们第一次的合成细胞，syn1.0，含有 901 个支原体基因加上一些水印和靶标序列 。其中，syn3.0 移除了 428 个基因，留下 438 个蛋白编码基因和 35 个 RNA 基因。在保证存活的条件下，更多的基因可能还可以被移除，但是其生长速度可能会受影响。Syn3.0 较为缓慢的生长速度并不是因为一个 rRNA 启动子被移除了。我们同时也用相同的基因组成构建了一个除保留该 rRNA 启动子外基因组成完全相同的中间菌株，而其生长速率还是和 syn3.0 相近。

用于表达的基因是 Syn3.0 里最大的基因类别（195 个基因，41%）。细胞膜（84 个基因, 18%）和新陈代谢(81 个基因, 17%)的基因数量基本相同。在支原体基因逐渐减少的进化过程中，很多生物合成基因被删减并被膜中的转运蛋白取代，这使得这两种基因数量相当。控制基因复制和在细胞分离时保持遗传物质稳定的基因(36 个基因, 7%)则相对少一点。意外的是，有 79 个基因（17%）我们无法将其归类到功能类别中。其中，19 个基因属于必要基因，36 个在细胞快速生长中被需要（i- or ie-genes 准必要基因或准较必要基因），24 个是非必要基因或者准不必要基因（n- or in-genes）。我们推测这些基因中大部分可以被分在上面四个大类别中（基因表达，膜结构和功能，胞浆新陈代谢和基因组的遗传保持），但是有些基因可能会表现出上述未提及的功能。具体而言，19 个无法分类必要基因中有 13 个的功能完全未知。它们中有一些可以在别的细菌甚至真核细胞中被同源基因，而这是编码新功能蛋白质基因的最佳候选。在 syn3.0 中的那些被需要并且在大部分有机体中存在的未知功能的基因一定代表了基本相似的功能，由此可以加深生物学方面的见解。同样的，没有同源基因的未知基因也可能是新兴的，或者它们可能是功能已被熟知的特殊基因序列。相较于那些完全未知的基因，如果一个基因有功能性的分类的话，就很容易过分简化一个基因功能推测的作用。举例来说，大量水解酶和激酶中的某些就毫无疑问的在核苷酸和辅因子补救中起到作用。问题在于，是所有未知基因的功能都这么的常见，还是说它们中的某些代表了根本上的新过程？虽然有些基因对于存活是必要的，但它们的注解却让人费解。举例来说，有六个不同的

外排系统，由基因 MMSYN1_0034, MMSYN1_0371 及 MMSYN1_0372, MMSYN1_0399, MMSYN1_0531, MMSYN1_0639, 和 MMSYN1_0691 编码。除了可能是翻转酶的 MMSYN1_0371 和 MMSYN1_0372 的异二聚体对外，这些蛋白的底物和功能都不明确。想象一下，如果这些基因都是用于排出或降解有毒物质的，似乎有些令人不安。类似的，需要一个相当复杂的用于生产和输出糖甘油酯的途径（23）。虽然一些证据表明呋喃半乳糖残留对于膜的完整性是重要的，但是对糖甘油脂类所实现的生物学作用的详细解释仍然不清楚。

2）Syn3.0 细胞表型

基因信息的复制和其向各个分离的细胞膜隔间的协调分配是现有生命体系的特点，且被普遍认为能用于定义细胞（25）。尽管我们并不知道最小细胞对于这个进程的要求，但是不同领域的证据表明细胞生命所需的机制比大多数真菌类复杂的分类器官要简单的多。

首先，一些细菌细胞类型，野生的（26）或是实验操作的（27，28），在核心细胞骨架成分缺失时都能够分离，尤其显著的是 FtsZ 的细胞骨架支架和产生力的组分。在我们基于经验的设计过程中，一个 syn1.0 中的非必要基因簇（MMYSYN1_0520 到 MMYSYN1_0522）在构建 syn3.0 细胞的时候被移除了。这保留了 ftsZ 和 sepF[编码一个和 FtsZ 互作的膜锚定组分]的同源基因。一个邻近的、在别的系统中和 sepF 有一些冗余功能基因 ftsA，在缺少 ftsZ 时，在逐步减少的构建体中保持了其必要性。

其次，完全由化学合成得到的脂质囊泡在没有大分子搭建或催化反应时表现出自发的分离（31）。在繁殖缺壁细菌时，质膜脂质含量和性状的改变也会导致相似的囊泡膜分离（32）。在一些缺少细胞壁的支原体种类中，丝状的和大囊泡的形态类型和在特定环境下已经被长期观测到的 syn3.0 很像，这部分取决于可用于这些细胞的脂质前体的性质。最终，理解 syn3.0 繁殖表型的遗传和功能基础可能揭示活细胞必需的膜相关细胞区室化的最小要求。

3）DBT 循环除基因组最小化以外的应用

我们专注于将整个基因组 DBT 循环应用于一个特定的问题，即构建一个最小的细胞基因组。但是，我们描述的这个方法还可以应用于构建任何预想的细胞。比方说，可以设计一个有附加新陈代谢通路的细胞（34），或是改变遗传编码（35），亦或是显著的改变基因编排。我们已经开始设计一个带有改良版 16S rDNA 序列的基因组并且评估这个显著改变对于密码子使用的影响。我们 DBT 循环的应用只会受限于我们在合理的成功机会下生产设计的能力。随着对目前未知的必要基因功能愈发深入的了解，并且随着对重构基因组经验的积累，我们期望设计的能力会越来愈强。设计一个所有基因的功能都已知的细胞的能力应该使完整的计算建模变得更轻松（36）。这会使得计算给附加通路生产有用产品（药物或工业化学品）的结果成为可能，并且可以得到更高的开发效率。

方法概述

我们用于识别非必要基因的方法是全局 Tn5 诱变，通过无痕 TREC 删除技术（核酸内切酶切割的串联重复偶联）来操纵酵母菌中的细菌基因组、合成并组装简化基因组、基因组移植、显微分析简化基因组的细胞、观察到的它们的生长特性列举在补充材料中。有关我们方法的大体信息和详细参考资料在全文中均有提及。

参考文献：

1. H. J. Morowitz, The completeness of molecular biology. Isr. J. Med. Sci. 20, 750–753 (1984). pmid: 6511349
2. C. M. Fraser et al., The minimal gene complement of Mycoplasma genitalium. Science 270, 397–404 (1995). doi: 10.1126/science.270.5235.397; pmid: 7569993
3. R. D. Fleischmann et al., Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. Science 269, 496–512 (1995). doi: 10.1126/science.7542800; pmid: 7542800
4. A. R. Mushegian, E. V. Koonin, A minimal gene set for cellular life derived by comparison of complete bacterial genomes. Proc. Natl. Acad. Sci. U.S.A. 93, 10268–10273 (1996). doi: 10.1073/pnas.93.19.10268; pmid: 8816789
5. C. A. Hutchison III et al., Global transposon mutagenesis and a minimal Mycoplasma genome. Science 286, 2165–2169 (1999). doi: 10.1126/science.286.5447.2165; pmid: 10591650
6. J. I. Glass et al., Essential genes of a minimal bacterium. Proc. Natl. Acad. Sci. U.S.A. 103, 425–430 (2006). doi: 10.1073/pnas.0510013103; pmid: 16407165
7. D. G. Gibson et al., Complete chemical synthesis, assembly, and cloning of a Mycoplasma genitalium genome. Science 319, 1215–1220 (2008). pmid: 18218864
8. C. Lartigue et al., Genome transplantation in bacteria: Changing one species to another. Science 317, 632–638 (2007). doi: 10.1126/science.1144622; pmid: 17600181
9. Materials and methods are available as supplementary materials on Science Online.
10. D. G. Gibson et al., Creation of a bacterial cell controlled by a chemically synthesized genome. Science 329, 52–56 (2010). pmid: 20488990
11. M. Juhas, D. R. Reuß, B. Zhu, F. M. Commichau, Bacillus subtilis and Escherichia coli essential genes and minimal cell factories after one decade of genome engineering. Microbiology 160, 2341–2351 (2014). doi: 10.1099/mic.0.079376-0; pmid: 25092907
12. G. Pósfai et al., Emergent properties of reduced-genome Escherichia coli. Science 312, 1044–1046 (2006). doi: 10.1126/science.1126439; pmid: 16645050
13. Y. Suzuki et al., Bacterial genome reduction using the progressive clustering of deletions via yeast sexual cycling. Genome Res. 25, 435–444 (2015). doi: 10.1101/gr.182477.114; pmid: 25654978
14. A. C. Forster, G. M. Church, Towards synthesis of a minimal cell. Mol. Syst. Biol. 2, 45 (2006). doi: 10.1038/msb4100090; pmid: 16924266
15. M. Lluch-Senar et al., Defining a minimal cell: Essentiality of small ORFs and ncRNAs in a genome-reduced bacterium. Mol. Syst. Biol. 11, 780 (2015). pmid: 25609650
16. A. Ramon, H. O. Smith, Single-step linker-based combinatorial assembly of promoter and gene cassettes for pathway engineering. Biotechnol. Lett. 33, 549–555 (2011). doi: 10.1007/s10529-010-0455-x; pmid: 21107654
17. C. Merryman, D. G. Gibson, Methods and applications for assembling large DNA constructs. Metab. Eng. 14, 196–204 (2012). doi: 10.1016/j.ymben.2012.02.005; pmid: 22629570

18. V. N. Noskov, L. Ma, S. Chen, R. Y. Chuang, Recombinase-mediated cassette exchange (RMCE) system for functional genomics studies in Mycoplasma mycoides. Biol. Proced. Online 17, 6 (2015). doi: 10.1186/s12575-015-0016-8; pmid: 25774095
19. T. Dobzhansky, Genetics of natural populations. XIII. Recombination and variability in populations of Drosophila pseudoobscura. Genetics 31, 269–290 (1946).
20. D. H. Haft, J. D. Selengut, O. White, The TIGRFAMs database of protein families. Nucleic Acids Res. 31, 371–373 (2003). doi: 10.1093/nar/gkg128; pmid: 12520025
21. J. S. Jensen, H. T. Hansen, K. Lind, Isolation of Mycoplasma genitalium strains from the male urethra. J. Clin. Microbiol. 34, 286–291 (1996). pmid: 8789002
22. R. Gil, F. J. Silva, J. Peretó, A. Moya, Determination of the core of a minimal bacterial gene set. Microbiol. Mol. Biol. Rev. 68, 518–537 (2004). doi: 10.1128/MMBR.68.3.518-537.2004; pmid: 15353568
23. J. Romero-García, C. Francisco, X. Biarnés, A. Planas, Structure-function features of a Mycoplasma glycolipid synthase derived from structural data integration, molecular simulations, and mutational analysis. PLOS ONE 8, e81990 (2013). doi: 10.1371/journal.pone.0081990; pmid: 24312618
24. E. Schieck et al., Galactofuranose in Mycoplasma mycoides is important for membrane integrity and conceals adhesins but does not contribute to serum resistance. Mol. Microbiol. 99, 55–70 (2016). pmid: 26354009
25. J. C. Xavier, K. R. Patil, I. Rocha, Systems biology perspectives on minimal and simpler cells. Microbiol. Mol. Biol. Rev. 78, 487–509 (2014). doi: 10.1128/MMBR.00050-13; pmid: 25184563
26. R. Bernander, T. J. Ettema, FtsZ-less cell division in archaea and bacteria. Curr. Opin. Microbiol. 13, 747–752 (2010). doi: 10.1016/j.mib.2010.10.005; pmid: 21050804
27. R. Mercier, Y. Kawai, J. Errington, Excess membrane synthesis drives a primitive mode of cell proliferation. Cell 152, 997–1007 (2013). doi: 10.1016/j.cell.2013.01.043; pmid: 23452849
28. M. Lluch-Senar, E. Querol, J. Piñol, Cell division in a minimal bacterium in the absence of ftsZ. Mol. Microbiol. 78, 278–289 (2010). doi: 10.1111/j.1365-2958.2010.07306.x; pmid: 20735775
29. S. Gola, T. Munder, S. Casonato, R. Manganelli, M. Vicente, The essential role of SepF in mycobacterial division. Mol. Microbiol. 97, 560–576 (2015). doi: 10.1111/mmi.13050; pmid: 25943244
30. R. Duman et al., Structural and genetic analyses reveal the protein SepF as a new membrane anchor for the Z ring. Proc. Natl. Acad. Sci. U.S.A. 110, E4601–E4610 (2013). doi: 10.1073/pnas.1313978110; pmid: 24218584
31. J. C. Blain, J. W. Szostak, Progress toward synthetic cells. Annu. Rev. Biochem. 83, 615–640 (2014). doi: 10.1146/ annurev-biochem-080411-124036; pmid: 24606140
32. R. Mercier, P. Domínguez-Cuevas, J. Errington, Crucial role for membrane fluidity in proliferation of primitive cells. Cell Reports 1, 417–423 (2012). doi: 10.1016/j.celrep.2012.03.008; pmid: 22832271
33. S. Razin, B. J. Cosenza, M. E. Tourtellotte, Filamentous growth of mycoplasma. Ann. N. Y. Acad. Sci. 143, 66–72 (1967). doi: 10.1111/j.1749-6632.1967.tb27645.x; pmid: 4861145
34. M. J. Smanski et al., Functional optimization of gene clusters by combinatorial design and assembly. Nat. Biotechnol. 32, 1241–1249 (2014). doi: 10.1038/nbt.3063; pmid: 25419741
35. M. J. Lajoie et al., Genomically recoded organisms expand biological functions. Science 342, 357–360 (2013). pmid: 24136966

36. J. R. Karr et al., A whole-cell computational model predicts phenotype from genotype. Cell 150, 389–401 (2012). doi: 10.1016/j.cell.2012.05.044; pmid: 22817898