基于混合 PCR 产物制备捕获探针的基因组捕获方法

Genome Sequence Capture Based on Pooled PCR-generated Baits

张圆, 梁丹, 张鹏*

生命科学学院, 中山大学, 广州, 广东省 510006

*通讯作者邮箱: zhangp35@mail.sysu.edu.cn

引用格式: 张圆, 梁丹, 张鹏. (2021). 基于混合PCR产物制备捕获探针的基因组捕获方法. **Bio-101 e1010599**. Doi: 10.21769/BioProtoc.1010599.

How to cite: Zhang, Y., Liang, D. and Zhang, P. (2021). Genome Sequence Capture Based on Pooled PCR-generated Baits. *Bio-101* e1010599. Doi: 10.21769/BioProtoc.1010599. (in Chinese)

摘要: 近年来, 目标序列捕获技术已经成为了系统发育基因组学获取数据的有效手段。使用 PCR 产物自制捕获探针, 相比于以往的商业合成探针可以极大地降低成本, 节约时间, 且不需要参考基因组数据, 非常适用于非模式生物类群。分子标记的变异程度影响系统发育关系的解析效果, 需要根据探针与被捕获样本间的遗传距离进行选择。当研究较高阶元的系统发育关系时 (如解析科间关系), 为了保证探针在所有被捕获样本中都有较高的捕获效率, 需要根据探针与最远缘的样本间的遗传距离选择分子标记, 二者的序列相似度需要大于 85%, 否则将难以捕获。这就导致可选的分子标记总是非常保守, 难以解析系统发育关系。本方法是一种基于混合 PCR 产物自制探针的目标序列捕获方法, 从探针制备策略入手解决上述困难问题。实验的主要流程可以概括为: 根据已知的系统发育关系或分类学知识, 混合来自多个支系或分类单元的代表物种 DNA, 以 DNA 混合液为模板 PCR 扩增核编码分子标记, 再混合所有 PCR 产物, 自制捕获探针, 进行目标序列捕获。本方法使用混合 DNA 捕获探针, 相比于只使用单个样本制备的探针, 有效地缩小了探针与被捕获样本间的遗传距离, 保证在使用变异程度高的分子标记时依然有较高的捕获效率, 进而获得具有充足信号的序列来解析系统发育关系。并且, 因为捕获目标区域为常用的 PCR 分子标记, 获得的数据与 PCR 扩增子测序得到的数据十分相似, 易于与已有数据整合。且本方法除了可以捕获到目标区域外, 还可获得大量的侧翼序列, 能为系统发育学研究提供更多的信号。经测试, 本方法对于年代久远, DNA

质量差的博物馆样本也有较高的捕获效率。推荐使用超过 30 个核编码分子标记、种级以上阶元的系统发育学研究使用本方法。

关键词：目标序列捕获，PCR 产物自制探针，混合探针，核编码分子标记，系统发育学

研究背景：

目标序列捕获技术 (Target sequence capture) 用于富集目标基因：通过探针与 DNA 文库杂交，物理的使感兴趣的目标片段结合到探针上，之后洗去未结合的非目标片段，从而实现目标基因的富集。目标序列捕获技术经济高效，对 DNA 的质量要求较低，适用于 DNA 降解较为严重的样本，近年来已经成为了系统发育学获取数据的主流方法。然而，捕获效率会受到探针与被捕获样本间遗传距离的影响，随着遗传距离增大，捕获效率会逐渐降低，当探针与被捕获样本间的序列相似度小于 85% 时，捕获将十分困难。因此，使用目标序列捕获技术解析高阶元（如科间及以上水平）系统发育关系并不容易。

目前已经有学者提出多种策略试图解决上述问题，比如：使用高度保守的基因组区域设计探针，如锚定序列捕获 (anchored enrichment, AHE) 和超保守序列元件捕获技术 (ultraconserved element, UCE)，由于目标区域高度保守，探针在远缘类群中也有较高捕获效率，但与此同时，高度保守的目标区域难以为系统发育研究提供充足信号，信号大多来自捕获过程中随机获得的侧翼序列 (flanking sequence)。另一种策略是直接针对保守度较低的外显子区域设计探针，但使用较为宽松的杂交条件，可以在一定程度上提高对远缘类群的捕获效率。最近，又有学者提出了一种新的策略：根据系统发育关系将所研究类群分为多个进化支系，为每一支系设计一套探针，再将多套探针混合起来进行杂交捕获。这种策略能够缩小了探针与被捕获样本间的遗传距离，提高捕获效率。然而，上述所有方法均使用合成探针，需要参考基因组序列用于探针设计，不适用于缺乏基因组资源的非模式生物类群。

近期有学者提出可以使用 PCR 产物自制捕获探针，该策略将所研究类群已有的分子标记作为捕获的目标基因，很好的解决了合成探针需要参考基因组资源的难题。并且，该方法获得的数据与 PCR 扩增子测序得到的数据十分相似，易于与 NCBI 等公共数据库中的数据进行整合。目前针对不同生物类群的通用分子标记集正不断被开发，例如 102 个脊椎动物通用核基因分子标记集 (Shen *et al.*, 2013)，96 个蛇类通用分子标记集

(Li *et al.*, 2017), 95 个甲虫通用核基因分子标记集 (Che *et al.*, 2018), 94 个鳞翅目通用核基因分子标记集 (Zhang *et al.*, 2019) 等, 这些分子标记资源可以保证在使用 PCR 产物自制捕获探针时具有充足的目标基因数量。然而, 由于该策略只使用单一样本进行 PCR 扩增, 探针与被捕获样本间遗传距离大时捕获效率低的问题仍然存在。

本方法在使用 PCR 产物自制捕获探针的基础上, 结合前文提到的混合多套探针集的思路, 提出了一种新策略, 其核心思想是混合来自不同进化支系的代表物种 DNA 自制 PCR 产物捕获探针。以往使用单一样本制备探针, 捕获效率受限于探针与其最远缘样本间的遗传距离, 而混合不同进化支系的代表物种 DNA 制备探针, 相当于为每一进化支系分别制备了一套探针, 捕获效率仅受限于一个进化支系内部的遗传距离。因此, 即便是高阶元的系统发育学研究, 也可以选择较为高变的分子标记, 获得更多的信号。我们在鳞翅目中对这一策略进行测试: 基于 94 个鳞翅目通用核基因分子标记集, 对来自 7 个总科, 17 个科的 43 种鳞翅目昆虫样本进行杂交捕获, 得到了稳健可靠的系统发育关系。样本中绝大多数为干标本, DNA 有一定程度的降解, 最终所有样本平均 94% 的目标区域可以被捕获, 除此之外, 还得到了大量的核编码侧翼序列, 使数据集总长远超预期。此外, 该策略不仅仅可以应用于高阶元, 由于探针制备过程十分灵活, 可以根据所研究类群的系统发育关系、物种跨度自由选择代表物种, 经测试, 种级以上阶元的研究均适用: 解析凤蛾科各属之间的系统发育关系的研究结果已发表于 Zhang *et al.*, 2020, 解析环蛱蝶属属内种间系统发育关系的研究结果已发表于 Ma *et al.*, 2020。

材料和试剂

1. 1.5 ml 离心管, 200 μ l 离心管, 200 μ l 八联管, 96 孔板 (BBI Life Sciences)
2. 20 μ l, 200 μ l, 1 ml 移液枪枪头 (BBI Life Sciences)
3. 琼脂糖 (Life Technologies)
4. 矿物油 (上海生工)
5. TransTaq-T DNA Polymerase (5U) (北京全式金生物)
6. TransTaq-T buffer (10x) (北京全式金生物)
7. dNTPs (10 mM) (上海生工)
8. AMPure XP beads (Beckman Coulter Inc.)
9. 无水乙醇 (分析纯)

10. ATP (100 mM) (上海生工)
11. T4 Polynucleotide Kinase (10 U/μl) (Fermentas)
12. T4 Polynucleotide Kinase buffer (10x) (Fermentas)
13. T4 DNA ligase (5 U/μl) (Fermentas)
14. T4 DNA ligase buffer (10x) (Fermentas)
15. PEG-4000 (50%) (上海生工)
16. NaCl (5 M) (分析纯)
17. Tris-HCl, pH 8.0 (1 M) (上海生工)
18. EDTA, pH 8.0 (0.5 M) (上海生工)
19. dsDNA Fragmentase (New England BioLabs Inc. 与超声波 DNA 打断仪二选一)
20. dsDNA Fragmentase Buffer (10x) (New England BioLabs Inc. 与超声波 DNA 打断仪二选一)
21. NEBNext Ultra DNA Library Prep Kit (New England BioLabs Inc.)
22. Human Cot1 (1 μg/μl) (Life Technologies)
23. SSPE (20x) (上海生工)
24. Denhardt's (50x) (上海生工)
25. SDS (10%) (上海生工)
26. SSC (20x) (上海生工)
27. Dynabeads MyOne Streptavidin C1 beads (Life Technologies)
28. Tween-20 (10%) (上海生工)
29. 通用型 DNA 纯化回收试剂盒 (天根生化科技 (北京) 有限公司)
30. 0.1x TE 缓冲液 (上海生工)
31. 双蒸水 (广州誉维生物科技仪器有限公司 Unique 超纯水机制)
32. 生物素接头 (见溶液配方)
33. 磁珠结合缓冲液 (见溶液配方)
34. 洗脱缓冲液 1 (见溶液配方)
35. 洗脱缓冲液 2 (见溶液配方)

仪器设备

1. 0.2-2 μl, 2-20 μl, 20-200 μl, 100-1000μl 移液枪

2. PCR 仪 (96 孔, 可进行 Touch down PCR, 博日 BIOER TC-E-96G)
3. 涡旋振荡器 (Scientific Industries Vortex-Genie2 涡旋振荡器)
4. 适配 1.5 ml 离心管的磁力架 (推荐 16 孔, Invitrogen)
5. 适配 1.5 ml 离心管的震荡孵育器 (杭州奥盛恒温混匀仪 MS-100)
6. 适配 1.5 ml 离心管、200 μ l 离心管的小型离心机
7. NanoDrop 2000 (Thermo)
8. 电泳仪
9. 服务器 (Linux 系统, 32G 以上的内存, 500G 的存储)
10. 超声波 DNA 打断仪 (与 DNA 片段化酶二选一, 宁波新芝 SCIENTZ18-A)

软件

1. Trimmomatic: <http://www.usadellab.org/cms/?page=trimmomatic>
2. SPAdes: <https://github.com/ablab/spades/>
3. CD-HIT: <https://github.com/weizhongli/cdhit/>
4. BLAST: <https://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/>
5. Exonerate: <https://github.com/nathanweeks/exonerate.git>
6. Mafft: <https://mafft.cbrc.jp/alignment/software/>
7. Gblocks: <http://molevol.cmima.csic.es/castresana/Gblocks.html>
8. RAxML: <https://github.com/stamatak/standard-RAxML>

图 1. 实验步骤与数据分析总流程图

实验步骤

1. 探针制备 (参考图 1 探针制备部分)

分子标记的选择: 推荐分子标记数量大于 30 个时使用本方法。建议选择长度为 500~2000 bp 的分子标记, 分子标记在所研究类群中的变异程度不宜过低, 以保证样本之间序列存在差异, 为系统发育学分析提供信号。

用于制备探针的样本的选择、DNA 质量及投入量: 本方法在不同支系选取代表物种制备混合 DNA 捕获探针 (参照图 1, 如不清楚系统发育学关系, 可按照分类学选取代表物种。如研究属间关系, 可以从每个属选择一个代表物种)。用于制备探针

的样本的 DNA 质量要好：DNA 的降解程度低，最好有主带。等量混合代表物种的 DNA，使用 0.1x TE 缓冲液调整 DNA 混合液浓度，使其终浓度为 30 ng/μl。具体每个用于制备探针的样本混合多少 DNA，需要根据用于制备探针的样本数量和分子标记数量进行估算。例如，混合 10 个样本的 DNA 用于制备探针，共有 100 个分子标记。每个分子标记的 PCR 扩增需要 30 ng DNA (表 1)，100 个分子标记总共需要 3000 ng DNA，那么每个样本就需要 3000/10=300 ng DNA。

1.1 对所选分子标记进行 PCR 扩增 (表 1-表 4)

- 1) 为保证 PCR 扩增成功率，推荐使用巢式 PCR 分别对每个分子标记进行两轮扩增。如果已知所选分子标记十分容易扩增，如 COI, Cytb 等，也可以只进行一轮 PCR 扩增。如果分子标记数量较多，推荐使用 96 孔板进行反应。需保证 PCR 产物 3'有一个突出 A，这里用的 Taq 酶为全式金的 TransTaq-T。以下 PCR 程序仅为示例，可根据自己实验室的习惯进行修改。

表 1. 第一轮 PCR (准备如下 20 μl 的反应体系)

试剂	体积 (μl)	终浓度
TransTaq-T buffer (10×)	2.0	1×
dNTPs (10 mM)	0.4	200 μM
TransTaq-T DNA Polymerase (5 U/μl)	0.2	0.05 U/μl
引物 F1 (8 μM)	0.5	0.2 μM
引物 R1 (8 μM)	0.5	0.2 μM
DNA混合液	1	
双蒸水	15.4	

表 2. PCR 反应程序

循环阶段	温度	时间	循环数
预变性	94 °C	4 min	1
变性	94 °C	45 s	35
退火	45 °C	40 s	
延伸	72 °C	2 min	
最终延伸	72 °C	10 min	1
保存	4 °C	∞	

表 3. 第二轮 PCR (准备如下 20 μ l 的反应体系)

试剂	体积 (μ l)	终浓度
TransTaq-T buffer (10 $\times$)	2.0	1 $\times$
dNTPs (10 mM)	0.4	200 μ M
TransTaq-T DNA Polymerase (5 U/ μ l)	0.2	0.05 U/ μ l
引物 F2 (8 μ M)	0.5	0.2 μ M
引物 R2 (8 μ M)	0.5	0.2 μ M
一轮的PCR产物	1	
双蒸水	15.4	

表 4. PCR 反应程序

循环阶段	温度	时间	循环数
预变性	94 $^{\circ}$ C	4 min	1
变性	94 $^{\circ}$ C	45 s	30
退火	45 $^{\circ}$ C	40 s	
延伸	72 $^{\circ}$ C	2 min	
最终延伸	72 $^{\circ}$ C	10 min	1
保存	4 $^{\circ}$ C	∞	

- 2) 使用 1%的 TAE 琼脂糖凝胶对第二轮的 PCR 产物进行电泳检测，等体积混合所有成功扩增的分子标记的第二轮 PCR 产物 (完全没有扩增出来才算失败，如果扩增结果是在目标区域外有非特异性扩增，这个分子标记算作扩增成功)。

1.2 使用 AMPure XP 磁珠纯化 PCR 产物混合液

因等体积混合所有成功扩增的分子标记后，PCR 产物混合液的 DNA 总量非常多，如果全部制成探针，通常会远超过杂交反应的需求量。考虑到后续试剂成本和实验工作量，不建议将所有 PCR 产物都制成探针，而是根据样本数取适量的 PCR 产物混合液进行制备。我们提供了一个经验上的参考：4 μ g 磁珠纯化后的 PCR 产物混合液，足够进行 300 个样本的杂交捕获反应。由于未经纯化的 PCR 产物无法测定浓度，因此在磁珠纯化前无法估计 DNA 总量。建议先纯化 100 μ l PCR 产物混合液，之后测定浓度，计算 DNA 总量，再考虑要不要

纯化更多 PCR 产物混合液。剩余的 PCR 产物混合液-20 °C 保存，日后需要时可以重新制备探针。

- 1) 重悬 AMPure XP 磁珠。
- 2) 使用移液枪吸取 100 μ l PCR 产物混合液至 1.5 ml 离心管中，加入 180 μ l 重悬的 AMPure XP 磁珠 (待纯化溶液体积的 1.8x)，用移液枪吸打混匀至少十次。
- 3) 室温放置 5 min。
- 4) 短暂离心 (简易离心机离心 10 s)，然后将离心管置于磁力架上，静置 5 min，使磁珠与上清液分离，可以观察到溶液变得清亮，此时小心弃去上清液。
注：不要丢弃磁珠。
- 5) 保持离心管于磁力架上，向管中加入 200 μ l 新配制的 80%乙醇，室温放置 30 s，然后小心弃去上清。
- 6) 重复步骤 5) 一次。
- 7) 保持离心管于磁力架上，打开离心管盖，风干磁珠 5 min。
注：不要风干过度，否则会影响片段回收。
- 8) 将离心管从磁力架上移下，加入 100 μ l 0.1x TE 缓冲液，用移液枪吸打混匀至少十次，使磁珠重悬，之后室温放置 2 min。
- 9) 短暂离心后将离心管置于磁力架上 5 min，可以观察到溶液变得清亮，使用移液枪将上清液转移至新的 1.5 ml 离心管，使用 Nanodrop 测定浓度。

1.3 末端磷酸化

每个反应体系可以处理 4 μ g 上一步骤获得的纯化的 PCR 产物混合液，足够进行 300 个样本的杂交捕获反应，剩余纯化的 PCR 产物混合液-20 °C 保存，日后需要时可以重新制备探针。

- 1) 使用200 μ l离心管配制如下30 μ l的反应体系 (表5)。

表5. 末端磷酸化

试剂	体积 (μl)	终浓度
ATP (100 mM)	0.3	1 mM
T4 Polynucleotide Kinase (10 U/μl)	1.5	0.5 U/μl
T4 Polynucleotide Kinase buffer (10×)	3	1×
纯化的PCR产物混合液+双蒸水	25.2	

2) 使用移液枪吸打混匀，之后短暂离心。

3) 将离心管置于PCR仪中，运行以下程序：

25 °C 30 min

4 °C短暂保存 (30 min 以内)

注：程序运行结束后尽快进行下一个步骤。

1.4 连接生物素接头

1) 向上一步骤反应后的离心管中加入以下30 μl混合液 (表6)，使连接反应的体积为60 μl。

表 6. 连接生物素接头

试剂	体积 (μl)	终浓度
T4 DNA ligase buffer (10×)	3	0.5×
PEG-4000 (50%)	3	2.5%
生物素接头 (10 μM) *	4	0.667 μM
T4 DNA ligase (5 U/μl)	1	0.083 U/μl
双蒸水	19	

*生物素接头为两条寡核苷酸链 Bio-T 和 TR 退火得到，见配方及附录。

1.5 使用 AMPure XP 磁珠纯化连接产物。

同步骤 1.2，使用 20 μl 0.1x TE 缓冲液洗脱。

1.6 使用带有生物素修饰的引物 Bio-T 对纯化后的连接产物进行 PCR 扩增

进行扩增的目的是确保扩增后的探针都带有生物素。引物 Bio-T 就是步骤 1.4 中组成双链生物素接头中的寡核苷酸链 Bio-T，需稀释到 10 μM。

1) 使用 200 μl 离心管配制如下 25 μl 的反应体系，一个反应处理 2 μl 纯化后的连接产物，因为上一步骤的洗脱体积为 20 μl，建议预留 1 μl 用于后续电

泳比对 PCR 扩增效果，所以共需进行 9 个 PCR 反应 (表 7-表 8)。

表 7. 使用带有生物素修饰的引物 Bio-T 对纯化后的连接产物进行 PCR 扩增

试剂	体积 (μl)	终浓度
TransTaq-T buffer (10×)	2.5	1×
dNTPs (10 mM each)	0.5	200 μM
TransTaq-T DNA Polymerase (5U/μl)	0.25	0.05 U/μl
Bio-T (10 μM)	1	0.4 μM
纯化后的连接产物	2	
双蒸水	18.75	

表 8. PCR 反应程序

循环阶段	温度	时间	循环数
预变性	94 °C	4 min	1
变性	94 °C	30 s	8
退火	45 °C	45 s	
延伸	72 °C	2 min	
最终延伸	72 °C	5 min	1
保存	4 °C	∞	

2) 使用 1% 的 TAE 琼脂糖凝胶对未进行 PCR 扩增的连接产物和 PCR 产物同时进行电泳检测，比较扩增效果。

1.7 使用 AMPure XP 磁珠纯化 PCR 产物

同步步骤1.2，将9个反应的PCR产物混合后一起纯化，使用100 μl 0.1x TE缓冲液洗脱。使用Nanodrop测量浓度，之后使用0.1x TE缓冲液调整至终浓度为50 ng/μl，至此探针制备完成。

2. Illumina 文库构建 (参考图 1 DNA 文库构建部分)

文库构建的插入片段大小为 200-400 bp。DNA 质量及投入量：对 DNA 质量要求低，DNA 的平均长度大于 200 bp 时就能得到较好效果。投入的 DNA 总量在 100 ng~500 ng 均可，建议使用 500 ng。

2.1 DNA 片段化

可以使用超声波破碎仪和 DNA 片段化酶两种方法进行基因组 DNA 片段化。超声波破碎仪相比于 DNA 片段化酶更好控制，不易使 DNA 过度降解，且不需要后续纯化，实验操作更加简便，因此推荐使用超声波破碎仪来进行 DNA 片段化。

参数建议：使用宁波新芝 SCIENTZ18-A 超声波 DNA 打断仪，每个打断循环 20s：打 10s+停 10s，共打 12 个循环。

如果选择使用 DNA 片段化酶的方法进行 DNA 片段化，参考以下步骤：

- 1) 使用 200 μ l 离心管配制如下 20 μ l 的反应体系 (表 9)：

表 9. DNA 片段化

试剂	体积 (μ l)	终浓度
dsDNA Fragmentase Buffer (10 $\times$)	2	1 $\times$
dsDNA Fragmentase	1	
DNA+双蒸水	17	

- 2) 使用移液枪吸打混匀，之后短暂离心。
- 3) 置于 PCR 仪中，运行以下程序：37 $^{\circ}$ C 10-25 min*
*需要先进行预实验，梯度测试，根据片段化效果确定具体运行时间
- 4) 将离心管快速转移至冰上，加入 2 μ l 0.5 M EDTA，快速吸打混匀，终止反应。
- 5) 使用 2% TAE 琼脂糖凝胶进行电泳检测。
- 6) 使用 AMPure XP 磁珠纯化片段化产物：同步骤 1.2，使用 30 μ l 0.1 $\times$ TE 缓冲液洗脱。

2.2 使用 NEBNext Ultra DNA Library Prep 试剂盒构建文库

文库间需要用特异性标签进行区分。文库可以批量构建好后-20 $^{\circ}$ C 保存，待探针和文库都准备好后一起进行杂交捕获。

3. 杂交捕获

推荐将4个带有不同特异性标签的文库混合在一起，进行杂交捕获。一个混合文库的总量为500 ng，每个文库取125 ng。经测试，混合文库的杂交效果很好，可以节约文库和实验成本。

3.1 准备文库混合液和杂交缓冲液

文库混合液和杂交缓冲液共50 μ l，分别配制：

- 1) 推荐使用200 μ l每孔的八联管按如下配方配制31.5 μ l文库混合液，使用移液枪吸打混匀，短暂离心后加入10 μ l矿物油，防止杂交时溶液受热蒸发。之后置于冰上备用 (表10)。

表 10. 配制文库混合液

试剂	体积 (μ l)	终浓度
Human Cot1 (1 μ g/ μ L)	2.5	0.05 μ g/ μ L
BO1.P5.F (100 μ M)*	0.5	1 μ M
BO3.P7.part1.F (100 μ M)*	0.5	1 μ M
BO5.P7.part2.F (100 μ M)*	0.5	1 μ M
500ng混合文库	20	
200ng自制探针+双蒸水	7.5	

*寡核苷酸阻碍物BO1-5的作用是防止文库的接头发生交联，避免非目标区域的捕获。此外，每条寡核苷酸阻碍物的3'末端添加了磷酸化修饰，目的是防止它们在扩增富集文库时作为PCR引物。序列见附录。

- 2) 推荐使用200 μ l每孔的八联管按如下配方配制18.5 μ l杂交缓冲液，使用移液枪吸打混匀，短暂离心后加入10 μ l矿物油，之后置于冰上备用 (表11)。

表 11. 配制杂交缓冲液

试剂	体积 (μ l)	终浓度
SSPE (20x)*	12.5	5x
EDTA (0.5 M)	0.5	0.005 M
Denhardt's (50x)	5	5x
SDS (10%)*	0.5	0.1%

*20x SSPE在使用前要充分涡旋，10% SDS需先置于室温充分溶解。

3.2 杂交反应

- 1) 设置PCR仪程序：95 $^{\circ}$ C 5 min，65 $^{\circ}$ C 5 min，65 $^{\circ}$ C 6 h，60 $^{\circ}$ C 6 h，55 $^{\circ}$ C 6 h，50 $^{\circ}$ C 6 h，45 $^{\circ}$ C 6 h。反应结束后可在45 $^{\circ}$ C存放几个小时。

- 2) 将装有文库混合液的八联管放入PCR仪，运行程序，使文库DNA在95 °C变性5 min。
- 3) 当PCR仪降温至65 °C时，保持文库混合液的八联管在PCR仪中，将装有杂交缓冲液的八联管也放入PCR仪，使杂交缓冲液在65 °C预热5 min，之后使用移液枪将文库混合液转移到杂交缓冲液的八联管中，使用移液枪快速吸打混匀，简短离心后放回PCR仪中继续运行程序，直至程序结束。

3.3 洗脱杂交产物

提前配好磁珠结合缓冲液、洗脱缓冲液 1 和洗脱缓冲液 2 (见配方)，可在杂交程序结束前约 90 min 开始以下步骤：

- 1) 重悬 MyOne 磁珠，使用移液枪吸取杂交反应数量×10 µl 的 MyOne 磁珠到 1.5 ml 离心管中，每管中总磁珠量不超过 180 µl。
- 2) 将磁珠于磁力架上静置 2 min，之后弃上清。
- 3) 加入 200 µl 磁珠结合缓冲液，涡旋 5-10 s，置于磁力架上静置 2 min，弃上清。
- 4) 重复步骤 3)两次。
- 5) 加入杂交反应数量×20 µl 的磁珠结合缓冲液重悬 MyOne 磁珠，之后加入 1 µl 10% Tween-20，涡旋混匀。
- 6) 根据杂交反应数量准备新的 1.5 ml 离心管，编号。向空管中分别加入 180 µl 磁珠结合缓冲液和 20 µl 上一步重悬的 MyOne 磁珠。之后将这些管子放置在震荡孵化器中，45 °C (杂交的最后一个温度) 静置温育 2 min。
- 7) 使用移液枪将杂交产物分别转移至上一步温育好的离心管中。之后将这些管子置于震荡孵化器中，45 °C (杂交的最后一个温度)，500 rpm 震荡温育 30 min。之后简短离心，于磁力架上静置 2 min，弃上清。

注：转移杂交产物时尽量不要吸到矿物油。

- 8) 将离心管从磁力架上取下，向其中加入 200 µl 洗脱缓冲液 1，反复吸打混匀，室温静置 10 min。简短离心后于磁力架上静置 2 min，弃上清。
- 9) 重复步骤 8) 一次，同时将洗脱缓冲液 2 置于 45 °C (杂交的最后一个温度) 预热。
- 10) 将离心管从磁力架上取下，向其中加入 200 µl 预热到 45 °C 的洗脱缓冲液

2, 反复吸打混匀后放入震荡孵化器中, 45 °C 静置温育 10 min。简短离心后于磁力架上静置 2 min, 弃上清。

11) 重复步骤 10) 两次

12) 使用 30 µl 0.1x TE 缓冲液洗脱, 可-20 °C 保存。

3.4 杂交后的 PCR 扩增

1) 使用 200 µl 离心管配制如下 25 µl 的反应体系, 对每个杂交反应洗脱的杂交产物进行 PCR 扩增 (表 12-表 13):

表 12. 杂交后的 PCR 扩增

试剂	体积 (µl)	终浓度
TransTaq-T buffer (10×)	2.5	1×
dNTPs (10 mM)	0.5	200 µM
TransTaq-T DNA Polymerase(5U/µl)	0.25	0.05 U/µl
P5 引物 (10 µM)*	0.5	0.2 µM
P7 引物 (10 µM)*	0.5	0.2 µM
杂交产物	5	
双蒸水	15.75	

*P5 引物, P7 引物序列见附录。

表 13. PCR 反应程序

循环阶段	温度	时间	循环数
预变性	94°C	4 min	1
变性	94°C	30 s	8
退火	65°C	30 s	
延伸	72°C	45 s	
最终延伸	72°C	5 min	1
保存	4°C	∞	

2) 使用 2% TAE 琼脂糖凝胶进行电泳检测。

3.5 使用 AMPure XP 磁珠纯化 PCR 产物

同步步骤1.2, 洗脱30 µl。使用Nanodrop测量浓度。

4. 混合纯化产物，割胶纯化，高通量测序

根据浓度等量混合每个杂交反应的 PCR 产物，使用通用型 DNA 纯化回收试剂盒进行割胶回收，洗脱 25 μ l。之后进行高通量测序。因胶回收试剂盒回收效率偏低，高通量测序的最低总量要求约为 500 ng，根据经验，割胶时投入的纯化 PCR 产物总量应大于 1 μ g。

结果与分析 (参考图 1 生物信息学数据分析部分)

1. 数据预处理与组装

- 1) 使用 Trimmomatic (v 0.39) 对测序得到的原始数据 (raw data) 进行质量控制和筛选，去除 reads 中的接头序列和低质量的 reads，得到高质量的 clean reads。
- 2) 依据各样本的特异性标签对 clean reads 进行分选。
- 3) 使用 SPAdes (v 3.8.1) 对 reads 进行从头拼接，得到 contigs。
- 4) 使用 cd-hit-est (v 4.8.1) 合并相似度大于 95% 的 contigs。
- 5) 保留平均测序深度大于 5 $\times$ 的 contigs 用于后续分析。

命令行：

```
$ java -jar trimmomatic-0.39.jar PE sample1_R1.fq.gz sample1_R2.fq.gz
sample1_R1_paired.fq.gz sample1_R1_unpaired.fq.gz
sample1_R2_paired.fq.gz sample1_R2_unpaired.fq.gz
ILLUMINACLIP:TruSeq3-PE.fa:2:30:10:2:keepBothReads LEADING:3
TRAILING:3 MINLEN:36
```

**sample1_R1.fq.gz 和 sample1_R2.fq.gz 为测序返回的正反向 reads。*

sample1_R1_paired.fq.gz 和 sample1_R2_paired.fq.gz 为软件输出的过滤后的成对 reads。

```
$ spades.py -t 10 --cov-cutoff auto -1 sample1_R1_paired.fq.gz -2
sample1_R2_paired.fq.gz -o sample1_contigs.fasta
```

**-1 sample1_R1_paired.fq.gz 和 -2 sample1_R2_paired.fq.gz，为上一步骤输出的过滤后的成对 reads。-o 为拼接后输出的 contigs 文件名。*

```
$ cd-hit-est -i sample1_contigs.fasta -o sample1_contigs_cd_hit.fasta -c 0.95
*-i sample1_contigs.fasta 为上一步拼接出的 contigs。
```

2. 同源序列的鉴定

使用 **tblastn** 的方法在各样本的 **contigs** 中寻找各分子标记的同源序列：

- 1) **tblastn** 所用参考序列的获得：从 NCBI 等数据库下载各分子标记近源物种的核酸序列，并将其翻译为蛋白参考序列。
- 2) 将蛋白参考序列作为 **query**，各样本拼接得到的 **contigs** 作为 **database** 进行 **tblastn** (**e-value**<1e-5, **identity**>60%, **coverage**>50%)，提取每条参考序列找到的 **score** 值最高的 **contig** 作为该分子标记的同源序列。将同一个分子标记在各样本中找到的同源序列写入一个 **fasta** 文件，获得各分子标记的同源序列组。

命令行：

```
$ tblastn -db sample1_contigs_cd_hit.fasta -query QUERY -outfmt 5 -out
OUTFILE -evalue 1e-5
```

3. 鉴定外显子和内含子的边界

因本方法在获得各核编码分子标记的目标区域的同时，还能额外获得其两端的侧翼序列，这些侧翼序列中可能包含内含子序列，因此需要使用 **exonerate** (v 2.2.0) 软件识别每条 **contig** 的外显子和内含子的边界，只使用所有的外显子序列进行后续分析。**exonerate** 的 **query** 序列为一个分子标记所在基因的完整蛋白序列，**target** 为该分子标记的同源序列组。根据 **exonerate** 的结果，串联每条 **contig** 中的外显子序列。

命令行：

```
$ exonerate --model p2g --query QUERY --target TARGET --showtargetgff T >
OUTFILE
```

4. 序列比对与数据精制，构建系统发育树

- 1) 使用 **mafft** (v 7.450) 对各分子标记的同源序列组进行比对，获得各分子标记的比对矩阵 (**alignment**)，之后使用 **Gblocks** (v 0.91b, **-t = c**, **-b5 = h**) 去除比对矩阵中比对不佳的区域。
- 2) 为了进一步提高数据质量，需要根据单基因树删除数据集中的问题序列。使用 **RAxML** (v 8.0.0) 对各分子标记的比对矩阵构建单基因树，去除单基因树中明显的长枝对应的序列。
- 3) 串联各分子标记的比对矩阵，使用 **RAxML** (v 8.0.0, **GTR+I+Γ model**) 构建系统发育树。

命令行:

```
$ mafft --auto INPUT_FILE > OUTFILE
```

**INPUT_FILE* 为每个分子标记的同源序列组的 *fasta* 文件, *OUTFILE* 为输出文件名。

```
$ Gblocks INPUT_FILE -t=c -b5=h
```

**INPUT_FILE* 为上一步 *mafft* 输出的比对结果。

```
$ raxmlHPC-PTHREADS-AVX -f a -x 12345 -T 3 -p 12345 -# 200 -m  
GTRGAMMA -s INPUT_FILE -q PARTITION_FILE
```

溶液配方

1. 生物素接头:

- 1) 先配制10× oligo hybridization buffer, 以配制10 ml为例 (表14):

表 14. oligo hybridization buffer 配方

试剂	体积	终浓度
NaCl (5 M)	1 ml	500 mM
Tris-HCl, pH 8.0 (1 M)	100 µl	10 mM
EDTA, pH 8.0 (0.5 M)	20 µl	1 mM
双蒸水	8.88 ml	

- 2) 单链 Bio-T、TR 需要合成, 之后按以下配方进行配置。(以配制 100 µl 生物素接头为例) (表 15):

表 15. 生物素接头配方

试剂	体积 (µl)	终浓度
Bio-T (100 µM)	10	10 µM
TR (100 µM)	10	10 µM
Oligo hybridization buffer (10×)	10	1×
双蒸水	70	

**Bio-T*、*TR*序列见附录

- 3) 退火: 涡旋混匀后, 分装于每孔 200 µl 的八联管中, 短暂离心。将八联管放入

PCR 仪，运行程序：95 °C 3 min，之后每秒降温 0.1 °C/s，降温至 10 °C，也可以在 95 °C 3 min 后关闭 PCR 仪，使其自然降温至室温。

2. 磁珠结合缓冲液 (以配制 10 ml 为例) (表 16):

表 16. 磁珠结合缓冲液配方

试剂	体积	终浓度
NaCl (5 M)	4 ml	2 M
Tris-HCl, pH 8.0 (1 M)	100 µl	10 mM
EDTA, pH 8.0 (0.5 M)	20 µl	1 mM
双蒸水	5.88 ml	

3. 洗脱缓冲液 1 (以配制 10 ml 为例) (表 17):

表 17. 洗脱缓冲液 1 配方

试剂	体积	终浓度
SSC (20×)	500 µl	1×
SDS (10%)	100 µl	0.1%
双蒸水	9.4 ml	

4. 洗脱缓冲液 2 (以配制 10 ml 为例) (表 18):

表 18. 洗脱缓冲液 2 配方

试剂	体积	终浓度 (总体积10 ml)
SSC (20×)	50 µl	0.1×
SDS (10%)	100 µl	0.1%
双蒸水	9.85 ml	

致谢

实验方案摘自发表的文章 Zhang, Y., Deng, S., Liang, D., Zhang, P. (2019). Sequence capture across large phylogenetic scales by using pooled PCR-generated baits: A case study of Lepidoptera. *Mol Ecol Resour* 00:1-15. doi: 10.1111/1755-0998.13026

竞争性利益声明

无经济或非经济性竞争性利益。

参考文献

1. Che, L. H., Zhang, S. Q., Li, Y., Liang, D., Pang, H., Ślipiński, A. and Zhang, P. (2017). [Genome-wide survey of nuclear protein-coding markers for beetle phylogenetics and their application in resolving both deep and shallow-level divergences](#). *Mol Ecol Resour* 17(6): 1342-1358.
2. Li, C., Hofreiter, M., Straube, N., Corrigan, S. and Naylor, G. J. (2013). [Capturing protein-coding genes across highly divergent species](#). *Biotechniques* 54(6): 321-326.
3. Li, J. N., He, C., Guo, P., Zhang, P. and Liang, D. (2017). [A workflow of massive identification and application of intron markers using snakes as a model](#). *Ecol Evol* 7(24): 10042-10055.
4. Shen, X. X., Liang, D., Feng, Y. J., Chen, M. Y. and Zhang, P. (2013). [A versatile and highly efficient toolkit including 102 nuclear markers for vertebrate phylogenomics, tested by resolving the higher level relationships of the Caudata](#). *Mol Biol Evol* 30(10): 2235-2248.
5. Peñalba, J. V., Smith, L. L., Tonione, M. A., Sass, C., Hykin, S. M., Skipwith, P. L., McGuire, J. A., Bowie, R.C. and Moritz, C. (2014). [Sequence capture using PCR-generated probes: A cost-effective method of targeted high-throughput sequencing for nonmodel organisms](#). *Mol Ecol Resour* 14(5): 1000-1010.
6. Ma, L., Zhang, Y., Lohman, D. J., Wahlberg, N., Ma, F., Nylin, S., Janz, N., Yago, M., Adu se-poku, K., Peggie, D., Wang, M., Zhang, P. and Wang, H. (2020). [A phylogenomic tree inferred with an inexpensive pcr-generated probe kit resolves higher-level relationships among *Neptis* butterflies \(nymphalidae: limenitidinae\)](#). *Syst Entomol* 45(4): 924-934.
7. Maricic, T., Whitten, M. and Pääbo, S. (2010). [Multiplexed DNA sequence capture of mitochondrial genomes using PCR products](#). *PLoS One* 5(11): e14004.
8. Zhang, Y., Deng, S., Liang, D. and Zhang, P. (2019). [Sequence capture across large phylogenetic scales by using pooled PCR-generated baits: A case study of *Lepidoptera*](#). *Mol Ecol Resour* 00: 1-15.

9. Zhang, Y., Huang, S., Liang, D., Wang, H., and Zhang, P.(2020). [A multilocus analysis of Epicopeiidae \(Lepidoptera, Geometroidea\) provides new insights into their relationships and the evolutionary history of mimicry.](#) *Mol Phylogenet Evol* 149: 106847.

附录

合成序列信息 (5'-3')

Bio-T: biotin-CAAGGACATCCGT

TR: CGGATGTCCTTGC

BO1.P5.F:

AATGATACGGCGACCAACGAGATCTACACTCTTTCCCTACACGACGCTCTTCCG

ATCT-phosphate

BO3.P7.part1.F: AGATCGGAAGAGCACACGTCTGAACTCCAGTCAC-phosphate

BO5.P7.part2.F: ATCTCGTATGCCGTCTTCTGCTTG-phosphate

P5 引物: AATGATACGGCGACCAACGAGATCTA

P7 引物: CAAGCAGAAGACGGCATAACGAGATTG