

简化基因组测序在群体遗传进化分析中的应用

Application of Reduced-representation Genome Sequencing in Population Genetic and Evolutionary Studies

吕佳^{1,2}, 刘平平², 王师^{1,2,3,*}

¹方宗熙-萨斯海洋分子生物学研究中心, 中国海洋大学, 青岛, 山东省; ²海洋生物遗传学与育种教育部重点实验室, 中国海洋大学, 青岛, 山东省; ³海洋生物学与生物技术功能实验室, 青岛海洋科学与技术国家试点实验室, 青岛, 山东省

*通讯作者邮箱: swang@ouc.edu.cn

引用格式: 吕佳, 刘平平, 王师. (2021). 简化基因组测序在群体遗传进化分析中的应用. **Bio-101 e1010636. Doi: 10.21769/BioProtoc. 1010636.**

How to cite: Lv, J., Liu, P. P. and Wang, S. (2021). Application of Reduced-representation Genome Sequencing in Population Genetic and Evolutionary Studies. **Bio-101 e1010636. Doi: 10.21769/BioProtoc. 1010636. (in Chinese)**

摘要: 简化基因组测序技术是利用限制性内切酶等手段来降低基因组的复杂度, 通过对特定的酶切片段进行高通量测序从而实现全基因组范围遗传分型的技术。其中, **2b-RAD (type IIB restriction endonucleases site-associated DNA)** 是唯一可实现等长标签测序分型的 RAD 类技术。该技术具有能够对所有酶切片段分型, 可通过选择性碱基调节标签密度, 建库流程简便快速, 可处理高度降解或者痕量样品等优势。本方案介绍了一种串联 **2b-RAD** 标签测序技术, 其兼容多种测序平台、具有灵活的标签组合方式, 可实现高效全基因组遗传变异和表观遗传变异同步分析, 在大规模遗传变异筛查、群体遗传进化分析以及分子育种等研究领域具有广泛的应用前景。串联 **2b-RAD** 实验方案主要包括如下步骤: 1. 采用 **IIB** 型限制内切酶将基因组 DNA 进行酶切, 获得长度为 30-35 bp 的等长片段; 2. 将酶切产物与设计的接头进行连接反应; 3. 将得到的连接产物进行扩增并纯化; 4. 混合串联标签文库通过酶切连接反应, 使独立的标签文库的粘性末端互补依次连接形成串联文库; 5. 将串联后的标签文库再次扩增, 引入 **barcode** 序列完成建库。本方案还提供了串联测序文库的数据拆分方法以及基因组分型流程, 以便于将获得的分型数据应用于群体遗传进化分析。

关键词：串联等长标签，简化基因组测序，群体遗传分析

研究背景

简化基因组测序技术通过利用限制性内切酶对约 1-10%的基因组进行测序分型，能够大幅降低基因组的复杂度，具有适用于大量样本低成本基因分型，可不依赖于参考基因组分析等优势 (Campbell *et al.*, 2018)。近些年来，以 RAD (restriction site-associated DNA sequencing) 为代表的简化基因组测序技术已成为广泛应用于动植物遗传领域基因组分型的重要手段。2b-RAD (type IIB restriction endonucleases site-associated DNA) 技术是唯一可对等长标签 (约 30-35 bp) 进行测序分型的 RAD 类技术 (Wang *et al.*, 2012)，该技术具有可对所有酶切片段分型，可通过选择性碱基调节标签密度，建库流程简便快速，可处理高度降解或者痕量样品等优势。该技术已经被广泛应用于高精度遗传图谱绘制、重要性状遗传解析、群体遗传学分析以及全基因组选择等研究方向。Multi-isoRAD (Multi-isolength restriction site-associated DNA) 是在 2b-RAD 技术基础上发展起来的串联标签测序技术 (Wang *et al.*, 2016)，在一个测序片段 (read) 实现了对多达 5 个标签同时分析，并首次实现了全基因组 SNP 分型和 DNA 甲基化的同步联合分析。该技术解决了 2b-RAD 技术无法应用于双末端测序平台的局限，使得简并基因组分析成本大大降低 (单标签测序成本仅为技术改进前的十分之一)。Multi-isoRAD 技术能够使研究者以较低的成本获得基因组大量的变异信息，为开展群体多样性和遗传结构分析以及群体进化和适应性机制解析提供了重要的技术手段。

材料与方法

1. 0.2 ml PCR管 (Axygen, catalog number: PCR-02-L-C)
2. 1.5 ml EP管 (Axygen, catalog number: MCT-150-C)
3. 各种型号的枪头 (10 μ l, 100 μ l, 200 μ l, 1000 μ l, Axxygen)
4. BsaXI (2 U/ μ l) (NEB, catalog number: R0609, -20 °C 储存)
5. CutSmart Buffer (10x) (BsaXI试剂盒配套)
6. T4 DNA Ligase (5 U/ μ l) (NEB, catalog number: M0202, -20 °C储存)
7. T4 DNA ligase buffer (10x) (T4 DNA ligase试剂盒配套)
8. ATP (NEB, catalog number: P0756S, -20 °C储存)

9. Phusion High-Fidelity DNA Polymerases (2 U/μl) (NEB, catalog number: M0530, -20 °C 储存)
10. dNTP Mix(10 mM each) (Vazyme, catalog number: P031-01, -20 °C储存)
11. Lgul (5 U/μl) (Thermo Fisher, catalog number: ER1931, -20 °C储存)
12. QIAquick PCR Purification Kit (QIAGEN, catalog number: 28106, -20 °C储存)
13. 链霉亲和素磁珠 (南通维莱纳米科技有限公司, catalog number: BCS8047, 2-8 °C 储存)
14. 普通DNA产物纯化试剂盒 (天根, DP204)
15. 丙烯酰胺/甲叉双丙烯酰胺 30%溶液 (29:1) (生工, B546017-0500)

仪器设备

1. 96 孔磁力架
2. PCR仪 (SensoQuest)
3. 涡旋振荡器 (VORTEX-6)
4. 台式高速冷冻离心机 (Eppendorf 5427R)
5. Qubit Flex荧光计
6. 0.1-2.5 μl, 0.5-10 μl, 2-20 μl, 10-100 μl, 20-200 μl, 100-1000 μl 移液枪 (Eppendorf系列)
7. 水平电泳槽 (Mupid-2plus)
8. ATTO聚丙烯酰胺电泳槽 (ATTO, 6220)

实验方法

一、串联 2b-RAD 简化基因组文库制备

1. 酶切基因组 DNA

本研究中采用 **BsaXI** 限制性内切酶对基因组 DNA 进行酶切实验，用于酶切的样本 DNA 用量建议在 100-200 ng。可以同时设置不加 **BsaXI** 酶的对照组。

1.1 配制下述反应体系。

表 1. 酶切反应体系

试剂	体积
----	----

DNA (100 ng/μl)	1 μl
10x CutSmart Buffer	1.5 μl
BsaXI (2 U/μl)	1 μl
ddH ₂ O	11.5 μl

- 1.2 将上述体系至于 37 °C PCR 仪中反应 3 h。酶切结束后将体系放置于-80 °C 使酶酶活。
- 1.3 随后利用 1%琼脂糖凝胶电泳检测酶切是否完全。使用 BsaXI 内切酶酶切基因组 DNA 检测结果如图 1 所示，对照组主带仍然清晰可见 (泳道 1, 3, 5, 7, 9)，实验组主带 DNA 消失，泳道内呈现 smear，表明已酶切完全 (泳道 2, 4, 6, 8, 10)。



图 1. 酶切组与对照组检测 (对照组的泳道为 1、3、5、7、9，实验组的泳道为 2、4、6、8、10)

2. 接头连接

2.1 双链接头的制备

分别将接头的长链和短链稀释至 10 μM，等体积 (各加 50 μl) 加入到 0.2 ml PCR 管中，混匀后置于 95 °C PCR 仪中，变性 5 min 后放置室温自然冷却，制备成双链接头。接头长链和短链序列如表 2 所示，例如 Ada1a 接头的长短链分别是 Ada1a-L 和 Ada1a-S, Ada1b 接头的长短链分别是 Ada1b-L 和 Ada1b-S。

表 2. 接头和引物序列表

接头及引物名称	序列 (5'-3')
Ada1a-L	ACACTCTTTCCCTACACGACGCTGTTCCGATCTNNN
Ada1a-S	AGATCGGAACAGC (AminoC6)
Ada1b-L	GTGACTGGAGTTCAGACGTGTGCTCTTCACGANNN
Ada1b-S	TCGTGAAGAGCAC (AminoC6)
Ada2a-L	ACACTCTTTCCCTACACGACGCTCTTCATCGNNN
Ada2a-S	CGATGAAGAGCGT (AminoC6)
Ada2b-L	GTGACTGGAGTTCAGACGTGTGCTCTTCAGCANNN
Ada2b-S	TGCTGAAGAGCAC (AminoC6)
Ada3a-L	ACACTCTTTCCCTACACGACGCTCTTCATGCNNN
Ada3a-S	GATGAAGAGCGT (AminoC6)
Ada3b-L	GTGACTGGAGTTCAGACGTGTGCTCTTCAGACNNN
Ada3b-S	GTCTGAAGAGCAC (AminoC6)
Ada4a-L	ACACTCTTTCCCTACACGACGCTCTTCAGTCNNN
Ada4a-S	GACTGAAGAGCGT (AminoC6)
Ada4b-L	GTGACTGGAGTTCAGACGTGTGCTCTTCACAGNNN
Ada4b-S	CTGTGAAGAGCAC (AminoC6)
Ada5a-L	ACACTCTTTCCCTACACGACGCTCTTCACTGNNN
Ada5a-S	CAGTGAAGAGCGT (AminoC6)
Ada5b-L	GTGACTGGAGTTCAGACGTGTGCTGTTCCGATCTNNN
Ada5b-S	AGATCGGAACAGC (AminoC6)
Prim 1	ACACTCTTTCCCTACACGACGCT
Prim 2	GTGACTGGAGTTCAGACGTGTGCT
BioPrim1	(biotin)-ACACTCTTTCCCTACACGACGCT
BioPrim2	(biotin)-GTGACTGGAGTTCAGACGTGTGCT
Primer 3	AATGATACGGCGACCGAGATCTACACXXXXXXACACT CTTTCCCTACACGACGCTC
Index Primer	CAAGCAGAAGACGGCATACGAGATXXXXXXGTGACTGGAG TTCAGACGTGT

X 代表 index 序列

2.2 以五份样品为单位进行串联文库的构建，向 5 份酶切产物中分别加入 5 组不同的接头，连接体系见表 3，根据五组样品所在的串联标签位置加入不同的接头组合，在下一步扩增反应中需使用对应的引物组合（见表 4）。

表 3. 连接体系

试剂	体积
酶切产物	10 μ l
T4 ligase buffer	2 μ l
AdaA (5 μ M)	0.8 μ l
AdaB (5 μ M)	0.8 μ l
T4 DNA ligase (400 U/ μ l)	0.5 μ l
ddH ₂ O	5.9 μ l

表 4. 各串联标签使用的接头和引物对组合

标签位置	AdaA	AdaB	PrimerA	PrimerB
1	Ada1a	Ada1b	Prim1	BioPrim2
2	Ada2a	Ada2b	BioPrim1	BioPrim2
3	Ada3a	Ada3b	BioPrim1	BioPrim2
4	Ada4a	Ada4b	BioPrim1	BioPrim2
5	Ada5a	Ada5b	BioPrim1	Prim2

2.3 将连接体系放置于 PCR 仪中，设置 4 °C 过夜连接

3. PCR 扩增，富集标签

3.1 参考表 5 配制扩增体系：将上述步骤所得到的五份连接产物分别利用不同的引物对组合（见表 4）进行 PCR 扩增，富集连接有接头的酶切片段。反应体系如下：

表 5. 扩增体系

试剂	体积
Ligated DNA	14 μ l
5x HF buffer	8 μ l
10 mM dNTP	1.2 μ l
PrimerA (10 μ M)	0.8 μ l
PrimerB (10 μ M)	0.8 μ l
Phusion DNA polymerase (2 U/ μ l)	0.4 μ l
ddH ₂ O	14.8 μ l

PCR 程序如下 [98 °C 5 s, 60 °C 20 s, 72 °C 10 s] × 16 cycles, 72 °C 10 min, 4 °C ∞

3.2 切胶回收产物: PCR 产物用 8% 非变性聚丙烯酰胺琼凝胶电泳检测, 扩增产物大小约为 100 bp, 将含有目标片段的胶块切下, 随后用研磨棒碾碎, 12 000 rpm 离心 3 min 后, 加入 45 µl ddH₂O, 室温静置溶胶 2 h。

3.3 PCR 产物回收

将吸附柱置于新的 1.5 ml 离心管中, 将上一步得到的溶胶液加入到柱中央, 13 000 rpm 离心 5 min, 收集切胶产物。

4. 再次扩增富集产物

4.1 以胶回收产物为模板, 再次进行扩增, 如步骤 3.1。

4.2 取 3.5 µl PCR 产物, PAGE 胶电泳检测。观察到清晰均一目的条带后, 将得到的 5 份扩增产物 (标签 1, 2, 3, 4, 5) 进行混合后 (体积为 200 µl), 利用 QIAGEN QIAquick PCR purification kit 进行纯化, 最后用 30 µl 水洗脱。

5. 酶切连接

5.1 对上述纯化后的混合样本 PCR 产物进行酶切, 酶切体系如下表所示:

表 6. 酶切体系

试剂	体积
PCR 产物	15 µl
10x Thermo Tango buffer	3 µl
Lgyl 酶 (5 U/µl)	1 µl
ddH ₂ O	11 µl

37 °C PCR 仪 (或金属浴) 中放置 3 h。

5.2 平衡磁珠: 将磁珠轻轻摇匀, 吸出 10 µl 至 0.2 ml PCR 离心管中, 放在磁力架上静置 2 min, 弃掉上清; 用 20 µl 1x CutSmart buffer 清洗磁珠, 反复轻柔吹吸 5-10 次, 尽量避免产生气泡, 每次洗涤结束时在磁力架上静置 2 min, 弃掉上清。

5.3 将步骤 5.1 得到的酶切产物加入管中, 室温放置 30 min, 期间用枪头反复吹吸

混匀，随后放置在磁力架上静置 3 min，将上清转移至新的微量离心管中，加入 1 μ l 的 T4 DNA 连接酶和 2 μ l 的 10 mM ATP，16 $^{\circ}$ C 反应 45 min。

- 5.4 利用 8% 聚丙烯酰胺凝胶电泳检测串联标签产物，连接产物大小约为 244 bp (如图 2 所示)，切胶回收连接产物。(回收方法同步骤 3.2 和 3.3，溶胶时加 50 μ l ddH₂O 溶胶)

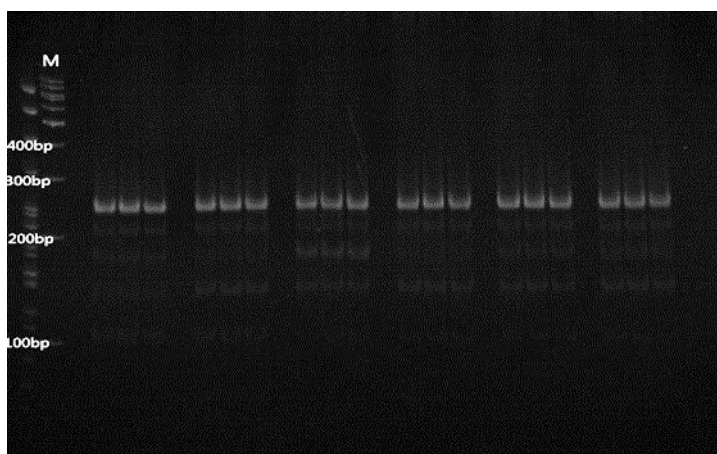


图 2. 标签串联的电泳检测结果

6. 二轮 PCR 扩增，引入文库特异性 barcode 序列

- 6.1 按照表 7 所示配制反应体系 40 μ l，每个样品平行扩增 2 管

表 7. 串联标签的扩增体系

试剂	体积
回收产物	8 μ l
5x HF buffer	8 μ l
10 mM dNTP	1.2 μ l
10 μ M Primer 3	0.4 μ l
2 μ M index primer	2 μ l
Phusion DNA polymerase (2 U/ μ l)	0.4 μ l
ddH ₂ O	20 μ l

PCR 程序: [98 $^{\circ}$ C 5 s, 60 $^{\circ}$ C 20 s, 72 $^{\circ}$ C 10 s] $\times$ 7 cycles, 72 $^{\circ}$ C 10 min, 4 $^{\circ}$ C ∞

- 6.2 取 3.5 μ l PCR 产物 PAGE 胶电泳检测，观察到清晰的目的条带 (300 bp) 后，将两管 PCR 产物合并后利用 QIAGEN QIAquick PCR purification kit 试剂盒纯化，使用 30 μ l ddH₂O 洗脱。
- 6.3 取 1.5 μ l 纯化产物利用 PAGE 胶电泳检测，扩增产物大小约为 299 bp，图 3 电泳图上显示目的条带清晰单一，背景干净，无引物二聚体。同时，Qubit 定量结果在 3 ng/ μ l 以上即可认为获得质量合格的文库。

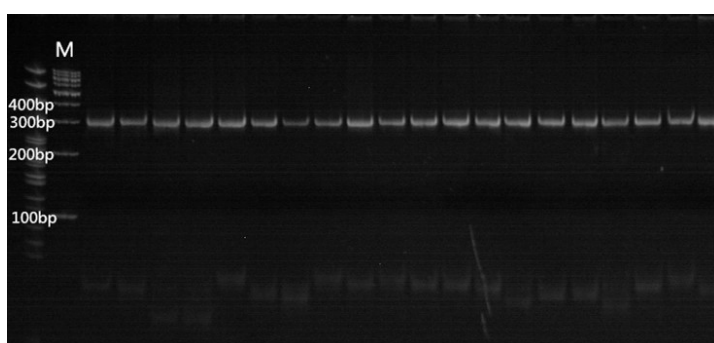


图 3. 串联文库的检测结果

二、串联文库的测序数据分析

1. 基因分型

参考 [Wang et.al\[3\]](#)等发表的串联 RAD 标签数据分析方法进行数据的质量控制及数据拆分分型等步骤，其中 SNP 分型步骤是利用 [Fu et.al \[4\]](#)等发表的 RAD-typing 软件进行分析。

请核实文献[1]和文献[2]，并按正确格式引用

1.1 数据质量过滤

对 Illumina 测序得到的原始数据进行质量过滤，去除含有 N 的序列以及大于 5 个碱基的质量值小于 10 的 reads。

```
$perl extractN.pl -d rawdata
$perl goodquality.pl -d extractNdata
```

1.2 测序 Reads 拼接及标签序列的拆分

对获得的高质量双端测序 reads 利用 PEAR 软件进行拼接，获得五标签串联的完整 reads，然后根据酶切位点所在的位置及五个标签的长度对串联序列进行拆

分并提取含有酶切位点的高质量 reads，获得所有个体的单标签 2b-RAD 数据。使用软件 PEAR 及脚本 common-find.pl、f_cut.pl、Extract_cut_site.pl 实现，具体命令如下：

- a. 将 pair-end 数据进行对齐处理

```
$perl common-find.pl -f Lb1_R1.fastq.ATCG.GoodQuality.fq -b
Lb1_R2.fastq.ATCG.GoodQuality.fq
```

- b. 利用 pear 软件对双端序列进行拼接

```
$PEAR_installed_directory/pear-0.9.6-bin-64 -f
Lb1_R1.fastq.ATCG.GoodQuality.fq-filter -r
Lb1_R2.fastq.ATCG.GoodQuality.fq-filter -m 192 -n 162 -o Lb1
```

- c. 根据 5 个标签所在的位置信息提取标签序列

```
$perl f_cut.pl -i 5tagData/Lb1 -s 1 33 66 99 132 -l 50 50 50 50 60
```

- d. 从标签序列中提取还有酶切位点的序列

```
$perl Extract_cut_site.pl -r 5tagData/Lb1_T1.fasta -c
[AGTC]{9}AC[AGTC]{5}CTCC[ATGC]{7}-
[ATGC]{7}GGAG[AGTC]{5}GT[ATGC]{9} -l 27-27 -o
5tagData/Lb1_T1wCS.fasta
```

1.3 基于 RAD-typing 程序进行 SNP 分析

- a. 利用脚本从基因组中提取 BsaXI 的酶切标签，构建高质量 Reference。

```
perl Extract_cut_site_v1.pl -r genome.fasta -c
[AGTCN]{12}AC[AGTCN]{5}CTCC[ATGCN]{10}-
[ATGCN]{10}GGAG[AGTCN]{5}GT[ATGCN]{12} -l 33-33 -o
Ref_BsaXIsites
```

- b. 序列比对

调用 SOAP 软件将群体数据与上步所得高质量 Reference 进行比对，允许两个碱基错配，使用脚本 reads_map.pl 实现。

```
$ perl reads_map.pl -l 27
```

- c. 共显性标记分型。根据比对结果，利用最大自然法对每个个体的共显性标记进行分型。使用脚本 codom_calling.pl 实现。

```
$ perl codom_calling.pl -a 0.05 -p 0.8
```

- d. 群体 SNP 标记筛选条件

为筛选群体内可靠的多态性标记位点，根据分型结果，依据下列条件筛选群体多态性 **SNP** 标记：(a) 只留取标签内最多有 3 个 **SNP** 的标签位点；(b) 在个体内标签深度在 500 以上的位点不进行分型；(c) 剔除群体中基因型无多态性的位点；(d) 要求群体中至少 80% 的个体可以分型；

```
$ perl codom_filter.pl -max_snp 3 -depth 500 -maf 0.05 -p 0.8
```

- e. 通过上述分析流程获得群体多态性 **SNP** 标记信息可以应用于群体多样性分析、群体结构分析以及系统发生树的构建等。

三、群体遗传进化分析

群体遗传分析主要使用 **POPGENE**、**PLINK1.9**、**Structure** 和 **Mega** 等软件工具包实现。

1. 基本参数计算

利用 **POPGENE** 软件统计分析多态性位点比例 (**P**)、**Nei's** 基因多样性指数 (**H**)、**Shannon** 氏指数 (**I**)、群体间基因分化系数 (**G_{ST}**)、基因流 (**N_m**)、**Nei's** 群体间遗传距离 (**D**) 等。采用 **Arlequin 2.0** 软件群体内核群体间的遗传变异进行分子变异分析 (**AMOVA**)，根据计算出的参数，可以分析群体的遗传变异规律。使用 **PLINK1.9** 软件可以计算群体之间的遗传分化系数 (**F_{st}**)、群体内分化系数 (**F_{is}**) 以及哈温平衡检验等。

2. 群体结构分析

2.1 PCA 分析

基于群体内获得的 **SNP** 位点数据使用 **plink** 进行主成分分析，计算出群体中 **PC1~PC_n** 的数值，使用主成分 1 (**PC1**) 和主成分 2 (**PC2**) 使用 **R** 软件绘制散点图对样本群体进行区分。

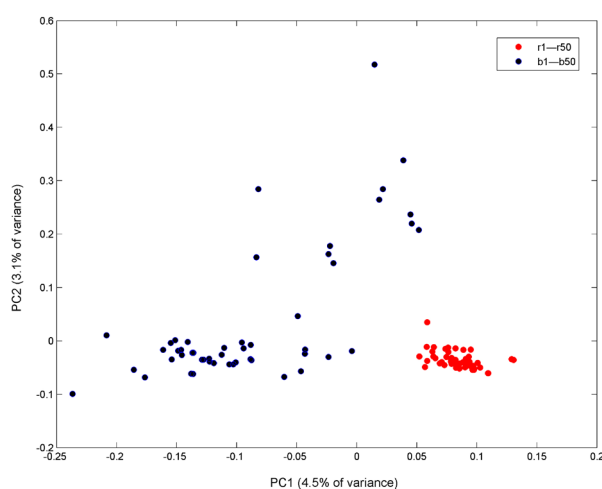


图 4. 两个群体结构主成分分析显示

2.2 群体结构分析

用 **structure** 软件 (version 2.3.4) 进行群体结构聚类分析。首先预设群体亚群数 $K = 1 \sim 5$ ，基于贝叶斯模型的计算方法，根据模拟算法的到的 K 值分布，选择最优 K 值。下图所示为 k 值为 4 时，两种不同壳色群体的海湾扇贝 (100 只) 的 **structure** 群体结构。

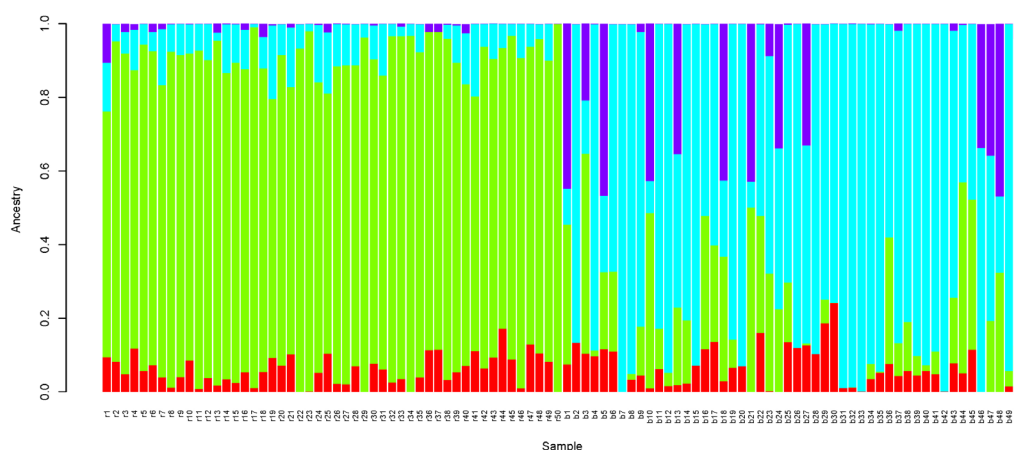


图 5. 来自两个群体的 100 只海湾扇贝的 **Structure** 群体结构

3. 系统发生分析

选取待分析的 **SNP** 位点所在的标签序列，将每个个体的序列首尾相连，如果缺失相

应的位点，则用“-”代替。获得的序列生成 **fasta** 文件导入 **MEGA** 软件中，分别构建 **NJ** (**Neighbor-joining**) 树和 **ML** (**Maximum likelihood**) 树，树的可靠性通过 **bootstrap** 法进行检验 (重复 1000 次)。下图所示为 100 只海湾扇贝 2b-RAD 数据所构建的 **NJ** 树，整棵树分成主要的两大支，其中海湾扇贝红壳群体 50 个个体 (r1-r50) 主要聚为一支，黑壳群体大多数个体 (b1-b50) 聚为另外一支。

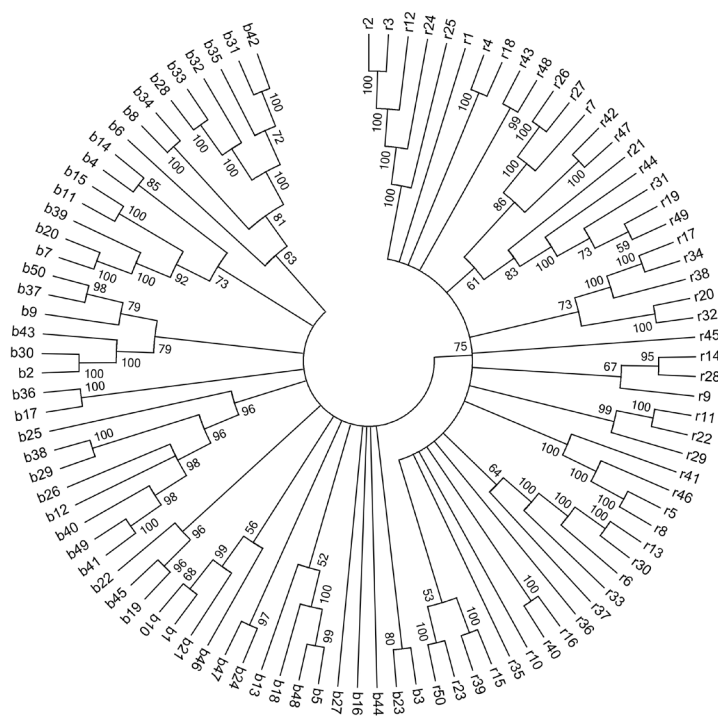


图 6. 基于 **SNP** 位点序列构建的两个群体的 **NJ** 系统树

方法运用

本实验方案介绍了一种串联 **2b-RAD** 标签测序技术，该技术兼容多种测序平台、串联的标签组合可灵活设置。其中，标签类型可采用不同样品、不同内切酶以及不同应用 (**SNP** 分型或 **DNA** 甲基化水平检测) 的组合。串联标签测序技术是一种高效、灵活的全基因组遗传变异和表观遗传变异筛查和检测的手段，在大规模遗传变异筛查、群体遗传进化分析以及分子育种等研究领域具有广泛的应用前景。

致谢

本研究受山东省泰山学者计划项目和三亚崖州湾科技城管理局项目 (**SKJC-KJ-**

2019KY01) 资助。

参考文献

1. Campbell, E. O., Brunet, B. M. T., Dupuis, J. R. and Sperling, F. A. H. (2018). [Would an RRS by any other name sound as RAD?](#) *Methods Ecol Evol* 9(9): 1920-1927.
2. Wang, S., Meyer, E., McKay, J. K. and Matz, M. V. (2012). [2b-RAD: a simple and flexible method for genome-wide genotyping.](#) *Nat Methods* 9: 808-810.
3. Wang, S., Liu, P., Lv, J., Li, Y., Cheng, T., Zhang, L., Xia, Y., Sun, H., Hu, X. and Bao, Z. (2016). [Serial sequencing of isolength RAD tags for cost-efficient genome-wide profiling of genetic and epigenetic variations.](#) *Nat Protoc* 11(11): 2189-2200.
4. Fu, X., Dou, J., Mao, J., Su, H., Jiao, W., Zhang, L., Hu, X., Huang, X., Wang, S. and Bao, Z. (2013). [RADtyping: an integrated package for accurate de novo codominant and dominant RAD genotyping in mapping populations.](#) *PLoS One* 8(11): e79960.