



ELSEVIER

Contents lists available at ScienceDirect

Engineering

journal homepage: www.elsevier.com/locate/eng



Research
Genetic Engineering—Article

超高通量、灵活靶向的全基因组分型技术 HD-Marker

刘平平^{a,b,#}, 吕佳^{a,*}, 马岑^a, 张天琦^a, 黄晓文^a, 杨志辉^a, 张玲玲^{a,b}, 胡景杰^{a,c}, 王师^{a,b,c}, 包振民^{a,c,d,*}

^a MOE Key Laboratory of Marine Genetics and Breeding and Sars-Fang Center, Ocean University of China, Qingdao 266003, China

^b Laboratory for Marine Biology and Biotechnology, Pilot Qingdao National Laboratory for Marine Science and Technology, Qingdao 266237, China

^c Laboratory of Tropical Marine Germplasm Resources and Breeding Engineering, Sanya Oceanographic Institution, Ocean University of China, Sanya 572000, China

^d Laboratory for Marine Fisheries Science and Food Production Processes, Pilot Qingdao National Laboratory for Marine Science and Technology, Qingdao 266237, China

ARTICLE INFO

Article history:

Received 26 January 2021

Revised 5 June 2021

Accepted 6 July 2021

Available online 22 December 2021

关键词

HD-Marker

靶向分型

全基因组

非模式生物

摘要

在生物、医学领域,靶向分型技术是检测已知变异位点的有效方法。然而,在非模式生物上,如何高效、低成本地进行大规模靶向位点分型仍然是一大挑战。为了解决这一问题,本文提出了一种基于液相分子杂交的超高通量 HD-Marker 技术,该方法可在单管内实现全基因组中 86 000 个位点的同时靶向分析。与以往的 Illumina GoldenGate 技术和低通量 HD-Marker 技术相比,单管内分析位点的数目分别提升了 27 倍和 6 倍。本研究对多种量级的 HD-Marker 技术(30 k、56 k 和 86 k)进行了综合评价,结果显示所有量级均具有较高的捕获率(约 96%)和基因分型准确率(约 96%)。因在成本(单位点分型成本低至 0.0006 美元)和技术灵活性等方面的优势,超高通量的 HD-Marker 技术在非模式生物的遗传、生态和进化研究中具有广阔的应用潜力。

© 2021 THE AUTHORS. Published by Elsevier LTD on behalf of Chinese Academy of Engineering and Higher Education Press Limited Company. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

1. 引言

鉴定表型相关的遗传变异是现代遗传学研究的核心。遗传多态性,特别是单核苷酸多态性(SNP),已被广泛应用于生态、农业、医学等领域[1–3]。近年来,随着高通量测序平台的发展,全基因组 SNP 位点的筛查和分型得以实现,为基因组学研究提供了前所未有的契机[4]。全基因组重测序,虽能获取最全面的基因组变异信息,但若应用于具有较大基因组的物种或上百乃至上千样本量的大规模分析时,成本仍然过高[5]。近年来,基于限制性内切酶的简化基因组测序方法被广泛应用[6],这一类技

术(如 RAD [7]、GBS [8]、2b-RAD [9–10])通过酶切手段降低基因组的复杂度,能够以较低成本获得全基因组范围内的部分 SNP 位点。因其只能针对酶切位点附近的 SNP 进行测序和分型,使得这些方法更适用于大规模标记的开发。随着基因组学研究的深入,丰富的“功能性”标记资源得以深度挖掘和积累,大量与表型性状相关的标记定位于基因或基因附近区域,对这类遗传标记的开发和研究在生态、农业和医学等领域具有重要的意义[11–12],因此靶向基因分型是后基因组时代的必然需求[13–14]。基因芯片技术(如 Illumina Infinium 和 Axiom Affymetrix 平台)是一种高效的靶位点分型技术,已广泛应用于人和

* Corresponding authors.

E-mail addresses: lvjia@ouc.edu.cn (J. Lv), zmbao@ouc.edu.cn (Z. Bao).

These authors contributed equally to this work.

模式生物[15–18]。但非模式生物仍缺乏成熟的标准化商业芯片[17,19]，若要应用芯片技术，通常需要进行额外定制。然而，固相芯片在定制后难以更新位点，若在有限的群体资源中开发固定阵列易使芯片存在固有偏差[20]。

近年来，基于测序平台的液相杂交捕获技术逐渐兴起，有望克服上述芯片平台的局限性。迄今为止，已发展出多种基于测序平台的靶向分型技术，主要包括基于聚合酶链反应（PCR）的方法和液相杂交捕获的方法[21–22]。其中，基于PCR的方法，如微滴PCR[23]和Ion AmpliSeq[24]等易于自动化处理大量样品，但依赖于专业仪器的检测（如RainStorm或者Ion Proton systems），且体系内可能存在引物间的竞争，导致非特异性扩增产物的产生[21]。多重PCR中复杂的引物互作也给通量的进一步提升带来难度[21,25]。液相分子杂交方法（如NimbleGen和SureSelect）可以捕获较长（250 Kb~5 Mb）的目标区域，是一类有效的捕获分型技术[26–28]。通量可超过50 000个靶向位点[21,27,29]，并且可以应用于痕量DNA（小于10 ng）甚至是降解的DNA样品，这一特性使液相捕获技术优于固相芯片平台[30–31]。通常这类方法更适合用于检测较长基因组区域内的变异，并不只针对单位点的靶向检测[21–22,30]。这类技术的靶向特异性为40%~60% [26]，较低的靶向特异性往往需要更高的测序深度来实现目标位点的均匀覆盖；这也使得测序成本较高[15,27,32–35]。

目前，对于非模式生物，仍迫切需要开发以高效、低成本和灵活的方式对全基因组位点靶向分型的技术。本文作者团队在前期研究中开发了一种基于液相分子杂交的方法HD-Marker，可以在单管内对12 472个位点同时进行靶向基因分型，在通量级和标记类型[36]的选择上具有很高的灵活性。该方法的原理是基于位点特异性探针（LSP）与目标位点的侧翼序列进行杂交，通过延伸、连接和扩增步骤，完成高通量文库的构建。HD-Marker技术有效结合了GoldenGate技术的高特异性、灵活性和测序平台的成本优势，可检测数百到12 472个位点，具有较高的捕获率（超过98%）和分型准确率（超过97%），对于非模式生物的大规模靶向基因分型是一种有前景和有吸引力的工具。然而，HD-Marker的检测性能还没有被充分挖掘，以往的12 000通量还不能满足全转录组基因覆盖的需求。因此在本研究中，进一步提升了单管内容纳的检测位点通量，使之超过86 000个位点，并且综合评估了三种通量级别（30 k、56 k和86 k）的检测性能，结果表明各通量级别下均具有较高的捕获率（约96%）和基因分型准确率（约96%），且单位点的分型成本可低至0.0006美元。鉴于超

高通量级、高灵活性和高稳定性的检测性能，超高通量的HD-Marker技术有望成为非模式生物大规模靶向基因分型的理想工具。

2. 方法

2.1. 探针设计与制备

首先对取自辽宁省不同地理位置的30只虾夷扇贝（*Patinopecten yessoensis*）样本进行全基因组重测序，这些个体来自东港、庄河和大长山三个群体，以及海大金贝和獐子红两个选育品种，每个群体选取6只个体。在获得的高质量SNP位点中，选取最小等位基因频率为0.2~0.5的SNP用于HD-Marker探针设计。针对每个SNP位点分别设计特异性探针LSP1和LSP2，探针包含SNP位点的侧翼序列和通用引物。LSP1的侧翼序列来自SNP位点的上游-22~-1位置，LSP2序列来自SNP位点的下游+5~+26位置（SNP坐标为0）[36]。探针序列需满足GC含量为40%~60%，退火温度为55~65℃，并且探针序列在基因组中无冗余的条件。附录A中的表S1提供了所有位点的LSP1和LSP2序列。合成探针前需将LSP1和LSP2与通用序列以及Nt.AlwI、Nb.BsrDI、Nt.BsmAI的特异性酶切位点序列进行组合，形成长度为126 bp的寡核苷酸序列。寡核苷酸池由美国CustomArray公司合成。随后，通过阵列合成的寡核苷酸池，经过扩增、酶切和链霉亲和素磁珠分离，得到LSP1和LSP2探针。详细步骤如下。

2.1.1. PCR扩增

将原始寡核苷酸池稀释200倍，然后扩增，反应体积为60 μL：包含1.8 μmol·L⁻¹的生物索引物（Oligo_F和Oligo_R）、0.6 mmol·L⁻¹脱氧核苷酸（dNTP）混合溶液、1×Phusion HF缓冲液和0.8 units Phusion高保真DNA聚合酶（美国NEB公司）。PCR反应条件为：98℃ 30 s；98℃ 15 s、60℃ 10 s、72℃ 15 s、24个循环；72℃、5 min。两管PCR产物混合后使用QIAquick PCR纯化试剂盒（德国Qiagen公司）纯化，并用32 μL纯水洗脱。

2.1.2. 酶切消化

纯化后的产物经限制性内切酶酶切后分离得到LSP1和LSP2。约2 μg的产物（约20 μL）在60 μL体系内进行酶切。首先加入3 μL的Nt.AlwI（美国NEB公司）在37℃下消化3 h，然后80℃热灭活20 min；随后加入3 μL BsrDI（美国NEB公司），65℃孵育3 h，80℃热灭活20 min。最后在体系中加入4 μL Nt.BsmAI（美国NEB

公司), 65 °C 孵育 3 h, 80 °C 热灭活 20 min。

2.1.3. 磁珠分离探针

链霉亲和素磁珠用于分离有生物素标记的探针互补链。首先, 用缓冲液 (0.5 mol·L⁻¹ NaCl、20 mmol·L⁻¹ Tris-Cl、1 mmol·L⁻¹ EDTA) 洗涤 50 μL 的链霉亲和素磁珠 (美国 NEB 公司)。随后, 将上一步的酶切产物 (67 μL) 加入单管中, 与磁珠混合 20 min。将混合物置于 95 °C 变性 5 min, 迅速放置在冰上, 保持 5 min。用磁铁吸附磁珠, 吸取上清液并将上清液转移至新管中, 利用 Nucleotide Removal Kit (美国 Qiagen 公司) 纯化, 使用 30 μL 的洗脱缓冲液 (10 mmol·L⁻¹ Tris-Cl, pH = 8.5) 洗脱分离的探针池, 用于下一步杂交。

2.2. 文库制备和测序

2.2.1. 生物素标记基因组 DNA 的制备

采用苯酚/氯仿提取法[37]从虾夷扇贝的闭壳肌组织提取基因组 DNA。取 3 μg 基因组 DNA 样本, 按照 PHOTO-PROBE 生物素标记试剂盒 (美国 Vector Labs 公司) 的操作说明, 通过热偶联作用进行生物素标记。

2.2.2. 杂交

为保证超高通量的探针进行有效的杂交反应, 本研究对杂交步骤进行了优化。首先将 5~10 μL 生物素标记基因组 DNA 加入含有 10 μL 磁珠的体系中, 用 50 μL 的 Ultra-HybOligo 杂交缓冲液 (美国 Ambion 公司) 洗涤两次。室温放置 5 min 后, 用磁铁吸附, 弃掉上清液。随后, 在体系内加入 15~30 μL 的制备探针以及 UltraHybOligo 杂交缓冲液 (美国 Ambion 公司), 杂交总体积为 100 μL。将杂交反应体系置于 PCR 仪 (美国 Bio-Rad 公司) 中, 设置 70~30 °C 的梯度降温条件, 进行杂交, 杂交时间约为 8 h。

2.2.3. 延伸和连接

杂交完毕后, 分别用缓冲液 1 [2 × 盐水柠檬酸钠 (SSC) 缓冲液、0.5% 十二烷基硫酸钠 (SDS) 缓冲液] 和缓冲液 2 (2 × SSC) 洗涤两次, 去除非特异性以及未结合的探针。然后制备延伸连接体系 (总体积为 25 μL), 该体系包含: 0.4 ~ 0.8 units Phusion 高保真 DNA 聚合酶 (美国 NEB 公司)、40 ~ 80 units Taq DNA 连接酶 (美国 NEB 公司)、1 mmol·L⁻¹ NAD (β-烟酰胺腺嘌呤二核苷酸) (美国 NEB 公司)、0.1 mmol·L⁻¹ dNTP 和 1 × Phusion HF 缓冲液。将延伸连接反应液加入洗涤好的磁珠中, 45 °C 孵育 20 min, 然后用洗脱缓冲液 (10 mmol·L⁻¹ Tris-Cl, pH = 8.5) 洗涤磁珠, 最后用洗脱缓冲液 35 μL 重悬磁

珠, 在 95 °C 下加热 1 min 后, 磁分离吸取上清液作为模板。

2.2.4. 文库制备和测序

参照 HD-Marker 测序文库的制备方法[36], 50 μL 的扩增体系包括: 上一步的连接产物 (约 30 μL)、0.8 units Phusion 高保真 DNA 聚合酶 (美国 NEB 公司)、0.1 μmol·L⁻¹ 通用 PCR 引物、dNTP 和 1 × Phusion HF 缓冲液。PCR 条件为: 98 °C 10 s、60 °C 20 s、72 °C 10 s 扩增 26 个循环, 最后 72 °C 延伸 5 min。用 8% 聚丙烯酰胺凝胶电泳检测目标产物并切胶, 产物大小为 116 bp。切胶产物回收后进行 PCR 扩增, 程序与上一步相同, PCR 循环数为 7 个。用 QIAquick PCR 产物纯化试剂盒 (美国 Qiagen 公司) 纯化扩增产物, 最后使用 32 μL 纯水洗脱。纯化后的产物利用 Qubit 进行定量, 利用 Bioanalyzer (美国 Agilent 公司) 检查文库质量。文库质量合格后, 在 Illumina HiSeq 平台上用 PE150 模式测序。

2.3. 数据处理与分析

对所有样品的 reads 1 (R1) 进行预处理, 每条序列截取第一个碱基到第 50 个碱基进行后续分析。去除低质量的序列, 即包含 N 的序列、有 10 个以上的相同连续碱基以及超过 20% 的碱基质量值小于 10 的序列。提取目标位点所在的 50 bp 的基因组区域作为参考序列, 使用 BWA 软件 (Burrows-Wheeler Alignment tool) [38] 将高质量 reads 与参考序列进行比对。随后使用 SAMtools [39] 将输出文件转换为 bam 文件并排序。使用 Varscan 软件[40]对位点进行分型, 要求位点的测序深度大于或等于 8 条 reads, 参数为 “--min-coverage 8 --min-reads 2 2 --min-var-freq 0.01 --min-freq-for-hom 0.99 -p-value 99e-2”。

为评估 HD-Marker 的分型准确性, 对同一个个体进行基因组重测序。使用 Next-Ultra DNA Library Prep Kit for Illumina (美国 NEB 公司) 构建技术重复文库。文库在 Illumina HiSeq X-Ten 平台进行测序, 测序深度约 21 ×。使用 BWA 软件将测序 reads 与虾夷扇贝参考基因组 (GenBank 登录号: GCA_002113885.2) 进行比对[38]。利用 Varscan 软件[40]做位点分型, 参数为 “--min-coverage 3 --min-reads 2 1 --min-var-freq 0.01 --min-freq-for-hom 0.99 --p-value 99e-2”。将两个重测序文库中分型一致的位点用于验证 HD-Marker 位点分型的准确性。测序数据已提交美国国家生物技术信息中心 (NCBI) Sequence Read Archive (<https://www.ncbi.nlm.nih.gov/sra>), 登录号为 PRJNA669118 和 PRJNA669126。

3. 结果

3.1. 位点通量及文库的设置

虾夷扇贝具有高质量的参考基因组和丰富的 SNP 资源[41–45]，因此以虾夷扇贝为实验对象，对超高通量 HD-Marker 进行技术验证。从 30 个扇贝个体的重测序数据中，获得满足探针设计标准的 SNP 位点共 2 044 646 个。基于这些高质量的 SNP，设计的位点均匀分布于基因组中的三个通量探针池（30 k-plex、56 k-plex、86 k-plex）（图 1）。大部分 SNP 来源于基因区域，在 30 k-plex、56 k-plex 和 86 k-plex 中的占比分别为 65.78%、71.81% 和 70.08% [见附录 A 中的图 S1 (a)]。所设计的位点覆盖了 20 100 个基因，覆盖了虾夷扇贝基因组[45]中 87% 的 Swis-sprot 注释基因和 90% 的 GO 注释基因。在 30 k-plex、56 k-plex 和 86 k-plex 量级中每个基因对应的 SNP 数量为 1~3 个（图 1）。在位于基因区的 SNP 位点中，有 52.30%~56.12% 来自外显子区域，有 8.17%~12.24% 来自 3'-5'-UTR 区域，其他 32.06%~39.53% 的位点来自内含子区[见附录 A 中的图 S1 (b) 和表 1]。为了比较不同通量级之间的捕获分型结果，更高通量级的位点池需覆盖低通量池的所有位点。即 86 k-plex 包含 30 k-plex 和 56k-plex 的所有位点，56 k-plex 包含 30 k-plex 中的所有位点。利用三个通量级探针池分别制备两个技术重复文库，总共有 6 个 HD-Marker 文库用于 Illumina 测序。

原先 12 k-plex 量级的 HD-Marker 文库制备方案[36]并不适用于超高通量的位点杂交（如 86 k-plex）。本研究通过在杂交前使用磁珠去除未标记的基因组 DNA，以及调整探针和生物素标记的基因组 DNA 配比等方式，优化了杂交反应条件。凝胶电泳结果表明，经体系优化后制备的文库质量显著提升（见附录 A 中的图 S2）。

3.2. 特异性、捕获率和均匀性

首先，分析 HD-Marker 的靶向特异性，统计比对到目

标区域的 reads 占比（这一指标将直接影响测序的成本）。分别在 30 k-plex、56 k-plex 和 86 k-plex 的文库中获得了超过 20 M、30 M 和 49 M 的 reads，其中高质量 reads 占比为 98.53%~99.80%（表 2）。30 k-plex、56 k-plex 和 86 k-plex 中分别有 81.24%、79.98% 和 79.72% 的高质量 reads 可以比对到目标区域（表 2）。尽管与 30 k-plex 相比，56 k-plex 和 86 k-plex 的特异性略低（约 1%），但总体三个通量级别均具有较高的特异性。其次，对 HD-Marker 的捕获率进行分析，绝大多数目标位点（96.65%~96.94%）在三个量级文库中都被检出（表 3 和图 2），并且在两个技术重复文库中检测到的位点重复性较高，均超过 97.57% 和 97.64%（表 3）。此外，位点在不同量级之间的重现性也很高，在 30 k-plex 和 56 k-plex 中有 99.65%~99.71%（表 4）的位点都能在更高的通量中被检出。最后，评估了位点测序深度的均匀性，位点在不同量级之间的测序深度具有较高的一致性（技术重复之间的 Pearson 相关系数为 0.92，通量级之间为 0.91~0.92）。目标位点的测序深度倍数变化为 2~4 个数量级，三个量级下分别有 94.63%、93.49% 和 93.24% 的位点位于 100 倍深度变化范围内（图 4），并且捕获位点的均匀度不受 GC 含量的影响，Pearson r 的范围为 0.060~0.098（见附录 A 中的图 S3）。

3.3. 基因分型率和分型准确率

本文进一步评估了位点的覆盖深度，发现超过 98% 的检出位点的测序深度大于 8（30 k-plex、56 k-plex 和 86 k-plex 分别为 98.44%、98.43% 和 98.44%）（表 3）。所有量级下都有较高的分型率（97.94%~98.94%）（表 3）。从三个方面对分型的准确率进行评价。首先，技术重复之间的分型结果比较显示，在三个量级中，有 95.57%~95.73% 的位点具有一致的分型结果（表 3）。其次，在不同量级之间的比较中，基因型的一致性为 95.87%~96.31%（表 4）。最后，与重测序结果相比较，三个量级中的分型准确率均大于 96%（表 5），表明 HD-Marker 在所有量级

表 1 靶位点的基因区分布

Genic regions	HD-Marker SNP genotypes					
	30 k-plex		56 k-plex		86 k-plex	
	No. of target SNPs	Percentage (%)	No. of target SNPs	Percentage (%)	No. of target SNPs	Percentage (%)
Exon	11 075	55.70	22 748	56.12	31 636	52.30
Intron	6375	32.06	13 806	34.06	23 915	39.53
5'_UTR	1339	6.73	2203	5.43	2624	4.34
3'_UTR	1095	5.51	1778	4.39	2317	3.83
Total	19 884	100	40 535	100	60 492	100

No.: number.

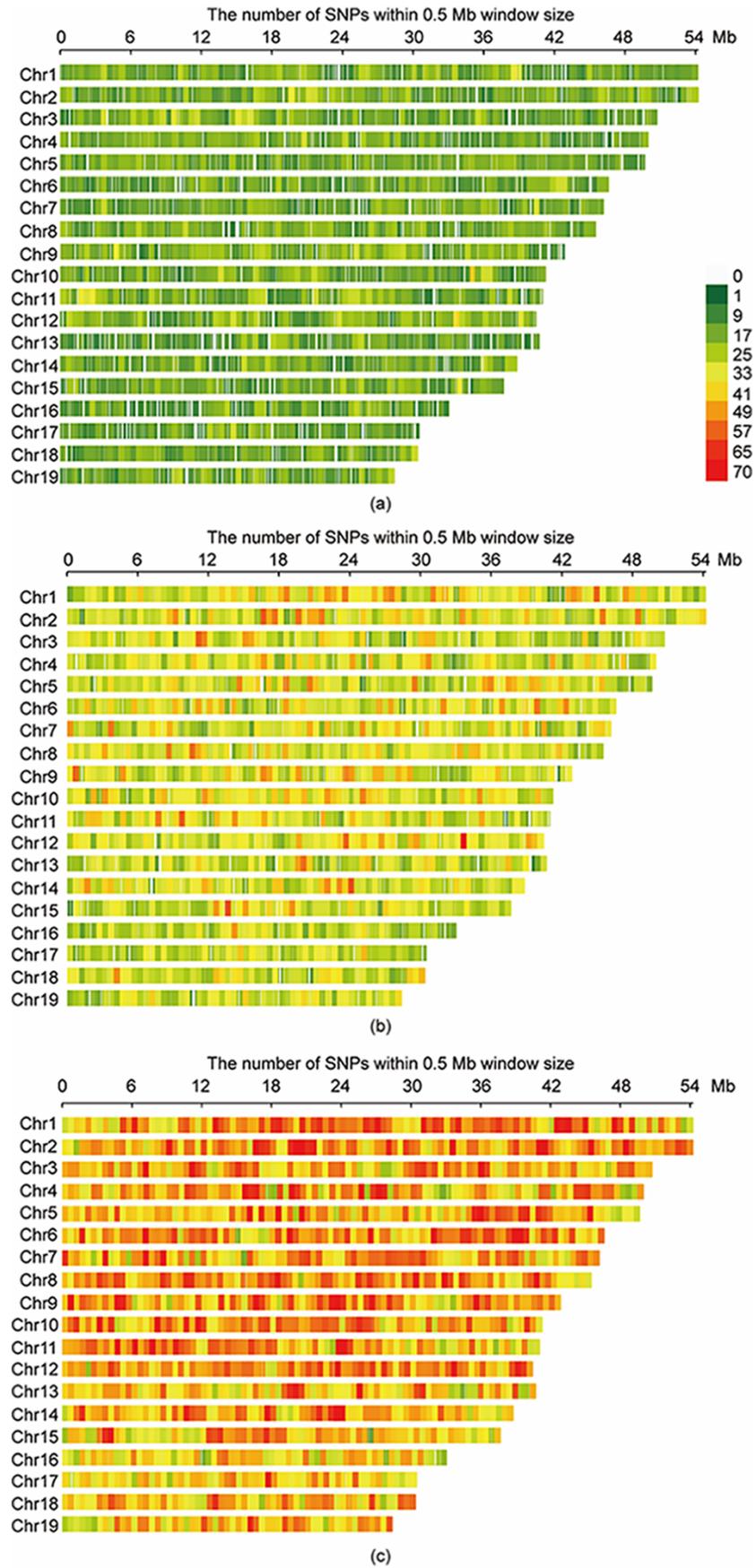


图1. 三个通量级的SNP位点的基因组分布。(a) 30 k-plex; (b) 56 k-plex; (c) 86 k-plex。

表2 测序reads与目标区域的比对情况

Multiplex level	Technical replicate	Read processing				Aligned to target regions		
		Raw reads (M)	HQ reads (M)	Efficiency (%)	Ave. efficiency (%)	Aligned reads (M)	Efficiency ^a (%)	Ave. efficiency (%)
30 230	Replicate 1	20,31	20.01	98.53	98.58	16.05	80.24	81.24
	Replicate 2	20,57	20.28	98.63		16.68	82.24	
56 445	Replicate 1	33.58	33.36	99.34	98.94	26.89	80.62	79.98
	Replicate 2	33.40	32.91	98.53		26.11	79.34	
86 025	Replicate 1	49.93	49.83	99.80	99.79	39.93	80.13	79.72
	Replicate 2	49.52	49.40	99.77		39.18	79.31	

^a Mapping efficiency was calculated by the number of aligned reads divided by the total number of HQ reads.

Ave.: average; M: millions; Rep: replicate.

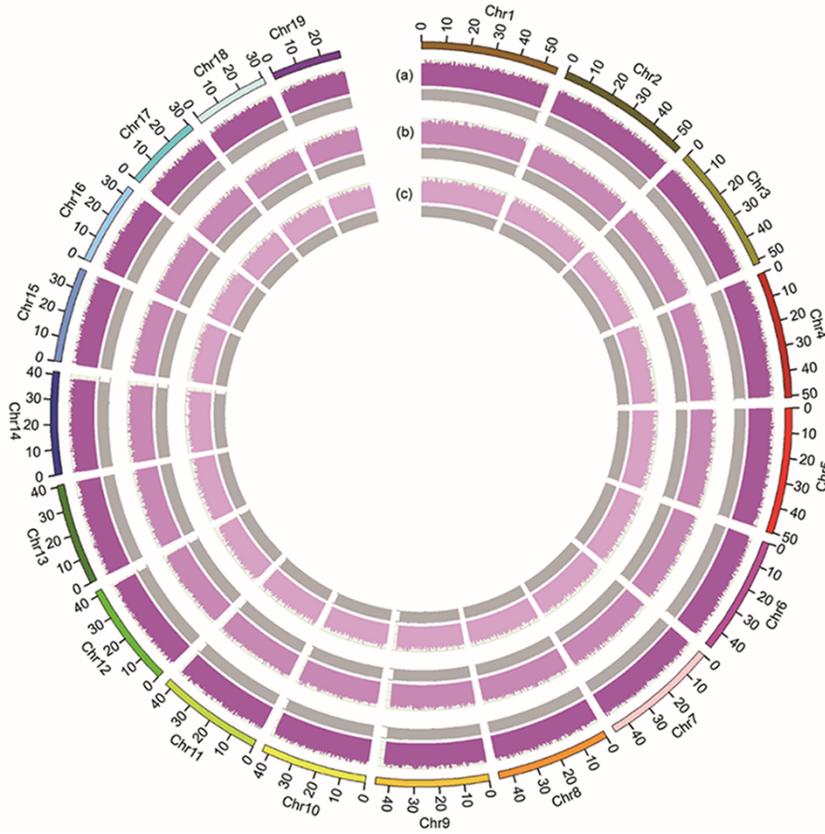


图2. 三个量级下SNP位点的测序深度分布图。在所有量级下都具有较高的捕获率（96%~98%）和均匀的测序覆盖度。（a）30 k-plex、（b）56 k-plex、（c）86 k-plex。

表3 位点检出率、分型率以及技术重复性

Multiplex level	Replicate	Loci detection			Genotype calling			Concordance between replicates			
		No. of loci	Rate (%)	Ave. rate (%)	No. of loci	Rate (%)	Ave. rate (%)	Common detected	Common calling	Consistent genotyping	Consistent rate (%)
30 230	Rep1	28 950	95.77	96.81	28 354	97.94	98.44	28 864	28 172	26 923	95.57
	Rep2	29 582	97.86		29 269	98.94					
56 445	Rep1	54 694	96.90	96.65	53 879	98.51	98.43	53 712	52 434	50 195	95.73
	Rep2	54 411	96.40		53 517	98.36					
86 025	Rep1	84 260	97.95	96.94	83 354	98.92	98.44	82 272	80 347	76 894	95.70
	Rep2	82 523	95.93		80 841	97.96					

表4 不同量级之间共有SNP的基因型评价

	For 30k common loci				For 56k common loci				
	30 k-plex	56 k-plex	86 k-plex	Common (percentage ^a)	Consistent (percentage ^b)	56 k-plex	86 k-plex	Common (percentage ^a)	Consistent (percentage ^b)
Detected	29 582	29 072	29 533	28 970 (99.65%)	—	54 694	55 439	54 537 (99.71%)	—
Calling	29 269	28 597	29 159	28 396 (99.30%)	27 222 (95.87%)	53 879	54 825	53 549 (99.39%)	51 574 (96.31%)

^a Percentage was calculated by dividing the number of loci that were commonly detected or called across multiplex levels by the number of loci that were detected or called in the lowest multiplex levels (30 230 or 56 445).

^b Percentage was calculated by dividing the number of consistently genotyped loci by the number of commonly called loci across multiplex levels.

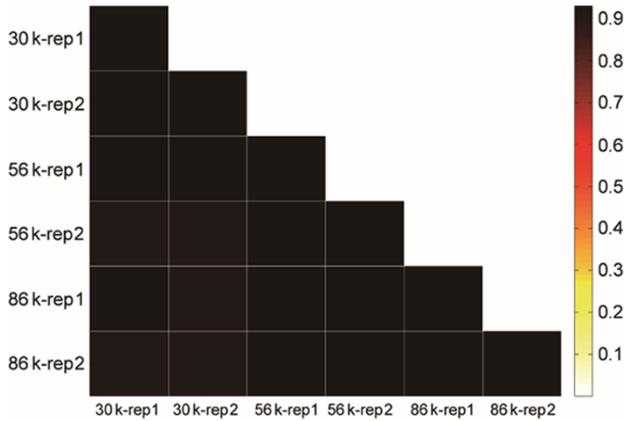


图3. Pearson相关性热图。在技术重复和多量级水平之间检出位点的测序深度一致性较高（技术重复之间的 r 为0.92，多量级之间的 r 为0.91或0.92）。

水平都具有较高的基因分型准确率。通常杂合位点要比纯合位点的检测分析更困难，进一步将两类位点分开统计，发现在所有量级下杂合位点的一致性均超过96.29%（表5），并且在不同量级的重复样品中杂合子位点的等位基因深度比例趋近于0.5，纯合位点的等位基因深度比例趋近于1（图5）。

3.4. 成本分析

为获得具有成本效益的最佳测序量，本研究合并了每个量级水平的技术重复数据，然后进行抽样分析。三个量级水平下，随着测序数据量的增加，位点检出率、分型率以及分型准确性最开始急剧上升，随后进入平台期。在平台期随着测序数据的增加，各个指标提升的幅度均较低（图6）。当30 k-plex、56 k-plex和86 k-plex的测序reads分

别达到5 M（M代表million，下同）、10 M和13.5 M时，位点检出率达到饱和，有95.8%~96.5%的位点可以被检出（图6）。在最佳测序量下，30 k-plex、56 k-plex和86 k-plex的基因分型准确率分别可达到96.40%、96.01%和96.15%。通过抽样分析可以计算在既定测序量下位点的检出率和分型准确性等，从而估算出对目标位点进行基因分型所需的最小测序深度，进而平衡位点检出率、准确性以及检测成本。进一步估算各量级下不同样本规模的基因分型成本（包括探针合成、文库制备和测序成本）。在基于饱和度曲线估算的最适宜测序量下，30 k-plex、56 k-plex和86 k-plex的每个样本的成本分别为29.4~92.1美元、44.1~106.7美元和58.5~121.2美元（表6）。此外，由于较大的样本量可以分担每个样本使用的探针成本，因此随着样本数量的增加，单样本和单位点的成本随之降低。如在86 k-plex下，对于100个样本规模，单样本的成本为121.20美元，当样本量扩大到1000个样本规模时，单样本的成本为64.23美元（表6），而在10 000个样本规模时，86 000量级下单位点的基因分型成本可以低至0.0006美元。

4. 讨论

靶向基因分型技术是检测目标遗传变异的有效工具。然而对于非模式生物，以较低的成本实现大规模靶位点（如数万到数十万个位点）的分型仍然具有挑战性。目前已有的靶向基因分型技术存在一定局限性，例如，基于PCR的方法（如微滴PCR和AmpliSeq）只能对数千个位

表5 基于基因组重测序的基因分型结果验证

Resequencing-based genotype	HD-Marker SNP genotypes								
	30 k-plex			56 k-plex			86 k-plex		
	Same	Different	Validation rate (%)	Same	Different	Validation rate (%)	Same	Different	Validation rate (%)
Homozygote	13 652	510	96.40	24 735	1021	96.04	36 413	1755	95.40
Heterozygote	11 379	406	96.55	22 034	849	96.29	34 926	1039	97.11
Total	25 031	916	96.47	46 769	1870	96.16	71 339	2794	96.23

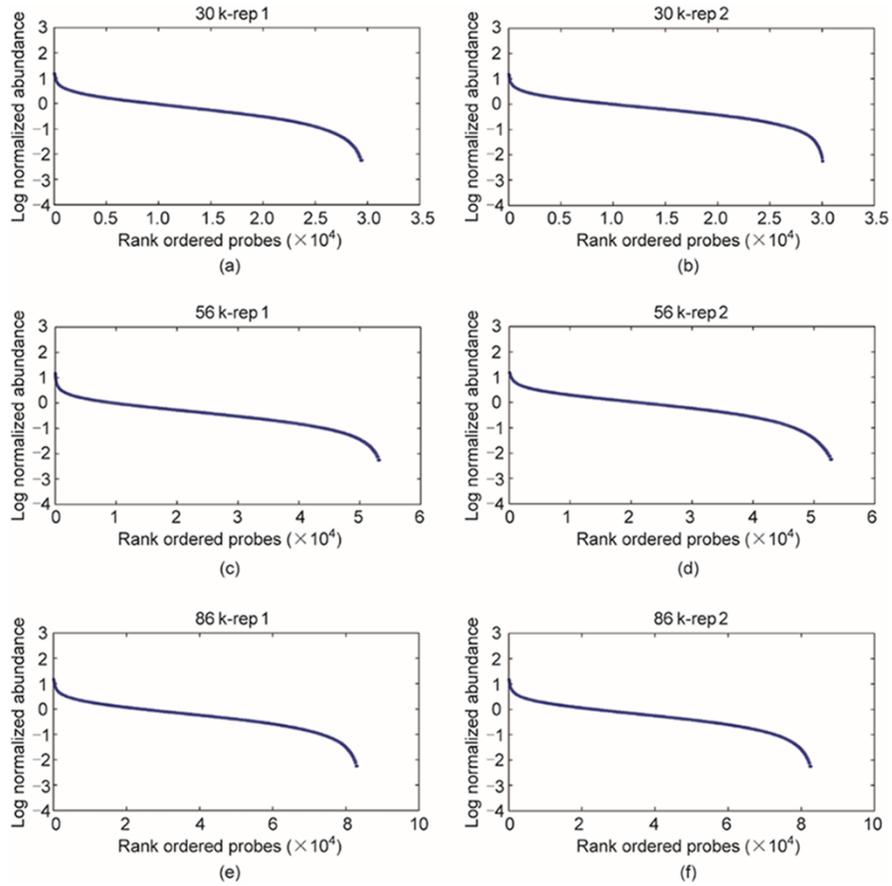


图4. 不同量级下的捕获均匀性。30 k-plex、56 k-plex 和 86 k-plex 的捕获均匀度在2~4个数量级之间变化，且93.24%~94.63%的位点深度在100倍范围内变化。

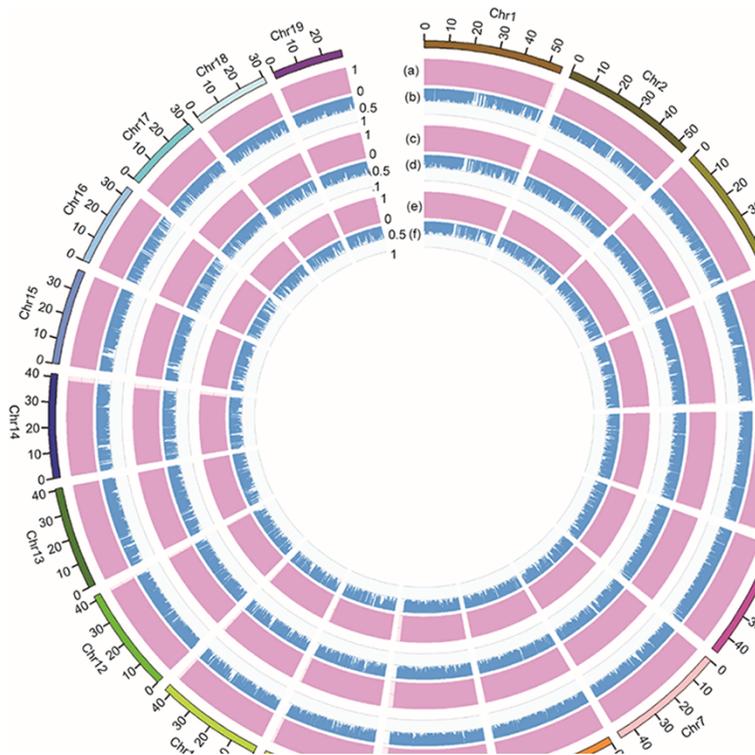


图5. 三个量级水平等位基因深度抽样分析。等位基因抽样分布对于杂合位点[(b)、(d)、(f)]基本收敛于0.5，纯合位点[(a)、(c)、(e)]收敛于1。(a)、(b) 30 k-plex；(c)、(d) 56 k-plex；(e)、(f) 86 k-plex。

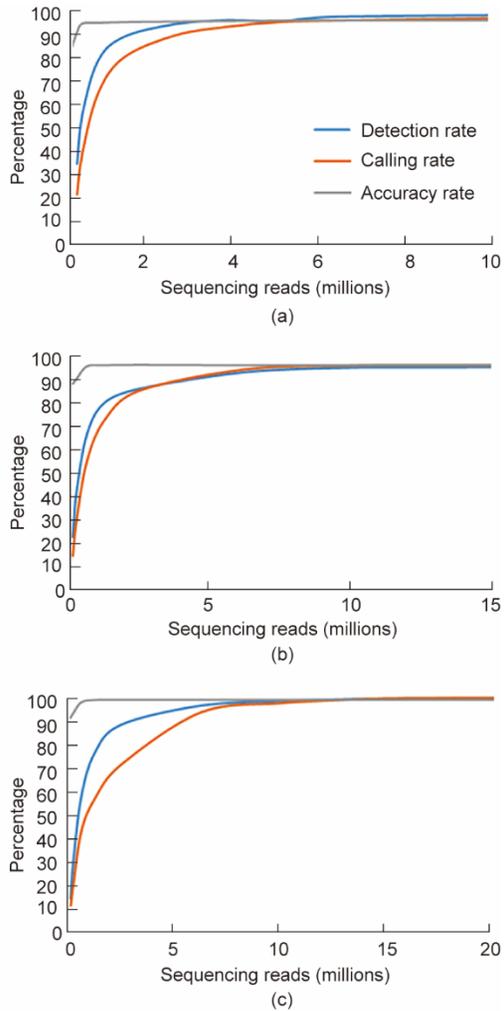


图6. 不同测序量下的检出率、分型率及准确率的饱和度曲线。对于30 k-plex (a)、56 k-plex (b)、86 k-plex (c)，位点检出率分别在5 M、10 M和13.5 M reads时达到饱和，最优测序深度下基因分型准确率分别为96.40%、96.01%和96.15%。

点进行分型[21,25]；基于固相芯片的分型方法（如Affymetrix阵列）需要较高的定制成本[5,46]；还有一类区域捕获技术（如Agilent SureSelect）针对较大的基因组区域

捕获而不针对单个位点[21–22]。Illumina的GoldenGate技术由于具备多重通量级以及较高的灵活性，被认为是一种具有潜力的靶位点分型工具[47–50]。但最初的GoldenGate检测技术需要使用荧光标记引物以及特殊的仪器检测荧光。在前期研究中，将该技术与高通量测序平台相结合，可以实现单管内超过12 000个位点的同时靶向分型[36]。

本研究提出了一种超高通量的HD-Marker方法，通过在杂交前使用磁珠去除未标记的基因组DNA，以及调整探针和生物素标记的基因组DNA配比等方式，优化了杂交反应条件，实现单管内86 000个位点的靶向分型，进而实现全基因库的覆盖。通量较Illumina GoldenGate技术和原有的HD-Marker技术分别提升了27倍和6倍。本研究从特异性、捕获率、均一性、基因型重现性以及准确性等方面充分验证了HD-Marker在不同量级水平（30 k-plex、56 k-plex和86 k-plex）下的稳健性和检测性能。与目前主流的捕获分析技术的特异性（约52%~57%）相比，HD-Marker表现出更高的靶标特异性（79.72%~81.24%）[26]。成本方面，30 k-plex到86 k-plex的每个样本为29~121美元，与传统的靶向基因分型方法相比，成本降低了40%~60% [53]。此外，与基因组重测序的金标准相比，HD-Marker表现出较高的基因分型准确率（所有量级的基因分型一致性均大于96%）。全基因组关联分析表明，高密度的SNP将有助于覆盖群体范围的连锁不平衡[51]，也可提高基因组选择的预测准确性[52]，因此更高的通量以及对稀有变异的有效检测是未来技术发展的重要需求。后续检测通量的提升可以考虑通过混合多个86 k-plex探针组来实现，使之达到与固相芯片相当的检测通量。本研究中，86 000个SNP位点的探针设计主要针对群体中的常见遗传变异进行检测（即基于具有不同地理位置背景的30个个体重测序数据），而若要对常见变异和稀有变

表6 各量级下不同样本规模的基因分型成本

No. of samples	No. of targeted loci					
	30 k-plex		56 k-plex		86 k-plex	
	per sample (USD)	per genotype (USD)	per sample (USD)	per genotype (USD)	per sample (USD)	per genotype (USD)
100	92.07 (81.28/10.79)	0.0031 (0.0027/0.0004)	106.74 (85.16/21.58)	0.0019 (0.0015/0.0004)	121.20 (92.07/29.13)	0.0014 (0.0011/0.0003)
1000	35.10 (24.31/10.79)	0.0012 (0.0008/0.0004)	49.78 (28.20/21.58)	0.0009 (0.0005/0.0004)	64.23 (35.10/29.13)	0.0007 (0.0004/0.0003)
10 000	29.40 (18.61/10.79)	0.0010 (0.0006/0.0004)	44.08 (22.50/21.58)	0.0008 (0.0004/0.0004)	58.53 (29.40/29.13)	0.0006 (0.0003/0.0003)

The estimated costs (USD) include both library preparation and NGS sequencing (optimal sequencing determined by rarefaction analysis; see Fig. 6; separate costs are shown in brackets (library preparation/Illumina sequencing); probe costs are calculated based on array-synthesized probes).

异位点的同时靶向分析，则需对大量个体进行重测序以获取可靠的稀有变异位点。

HD-Marker技术提供了一种高通量、灵活可扩展的多量级的靶向分析方法。HD-Marker的建库涉及常规分子生物学实验操作，无需任何昂贵的专业仪器，而且易于开展。该技术在标记数量和标记类型方面提供了较大的灵活性，研究人员可以根据需要选择不同量级的探针池以满足不同的研究需求。通常，中低通量的芯片是家系鉴定等育种应用中的首选，在构建高密度连锁图谱、估计性状遗传力、全基因组关联分析以及基因组选择应用时，研究人员倾向于选择更高通量的位点[54]，以提高标记分析的精度。此外，可以考虑将液相芯片技术与基因型填充技术结合[55–56]，这样可以在不增加成本的情况下显著提高位点数目，更具成本效益。近期也有研究表明，当使用500~2000个与性状显著关联的SNP位点做基因组预测时，预测准确性与使用全部位点相当，甚至具有更好的预测准确性[57]。在这个量级水平，HD-Marker方法的单样本成本低于10美元。作为一种高效、经济的靶位点分型技术，超高通量的HD-Marker技术有望成为非模式生物遗传、生态和进化研究的重要支撑工具。

致谢

本研究得到国家自然科学基金项目(32130107、32002446、32102778)、三亚崖州湾科技城管理局项目(SKJC-KJ-2019KY01)、国家现代农业产业技术体系、山东省泰山学者项目资助。

Compliance with ethics guidelines

An early version of HD-Marker has been granted a Chinese patent (ZL201310549040.7). Pingping Liu, Jia Lv, Cen Ma, Tianqi Zhang, Xiaowen Huang, Zhihui Yang, Lingling Zhang, Jingjie Hu, Shi Wang, and Zhenmin Bao declare that they have no other conflict of interest or financial conflicts to disclose.

Appendix A. Supplementary data

Supplementary data to this article can be found online <https://doi.org/10.1016/j.eng.2021.07.027>.

References

- [1] Stapley J, Reger J, Feulner PGD, Smađja C, Galindo J, Ekblom R, et al. Adaptation genomics: the next generation. *Trends Ecol Evol* 2010;25(12):705–12.
- [2] Shafer ABA, Wolf JBW, Alves PC, Bergström L, Bruford MW, Brännström I, et al. Genomics and the challenging translation into conservation practice. *Trends Ecol Evol* 2015;30(2):78–87.
- [3] Helyar SJ, Hemmer-Hansen J, Bekkevold D, Taylor MI, Ogden R, Limborg MT, et al. Application of SNPs for population genetics of nonmodel organisms: new opportunities and challenges. *Mol Ecol Resour* 2011;11(Suppl 1):123–136.
- [4] Davey JW, Hohenlohe PA, Etter PD, Boone JQ, Catchen JM, Blaxter ML. Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nat Rev Genet* 2011;12(7):499–510.
- [5] Jiang Z, Wang H, Michal JJ, Zhou X, Liu B, Woods LCS, et al. Genome wide sampling sequencing for SNP genotyping, methods: challenges and future development. *Int J Biol Sci* 2016;12(1):100–8.
- [6] Andrews KR, Good JM, Miller MR, Luikart G, Hohenlohe PA. Harnessing the power of RADseq for ecological and evolutionary genomics. *Nat Rev Genet* 2016;17(2):81–92.
- [7] Baird NA, Etter PD, Atwood TS, Currey MC, Shiver AL, Lewis ZA, et al. Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLoS One* 2008;3(10):e3376.
- [8] Elshire RJ, Glaubitz JC, Sun Q, Poland JA, Kawamoto K, Buckler ES, et al. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS One* 2011;6(5):e19379.
- [9] Wang S, Meyer E, McKay JK, Matz MV. 2b-RAD: a simple and flexible method for genome-wide genotyping. *Nat Methods* 2012;9(8):808–10.
- [10] Wang S, Liu P, Lv J, Li Y, Cheng T, Zhang L, et al. Serial sequencing of isolength RAD tags for cost-efficient genome-wide profiling of genetic and epigenetic variations. *Nat Protoc* 2016;11(11):2189–200.
- [11] De Wit P, Pespenti MH, Palumbi SR. SNP genotyping and population genomics from expressed sequences - current advances and future possibilities. *Mol Ecol* 2015;24(10):2310–23.
- [12] Jiao W, Fu X, Li J, Li L, Feng L, Lv J, et al. Large-scale development of gene-associated single-nucleotide polymorphism markers for molluscan population genomic, comparative genomic, and genome-wide association studies. *DNA Res* 2014;21(2):183–93.
- [13] Jones MR, Good JM. Targeted capture in evolutionary and ecological genomics. *Mol Ecol* 2016;25(1):185–202.
- [14] Zenger KR, Khatkar MS, Jones DB, Khaliliasamani N, Jerry DR, Raadsma HW. Genomic selection in aquaculture: application, limitations and opportunities with special reference to Marine Shrimp and Pearl Oysters. *Front Genet* 2019;9:693.
- [15] Asan Y, Xu Y, Jiang H, Tyler-Smith C, Xue Y, Jiang T, et al. Comprehensive comparison of three commercial human whole-exome capture platforms. *Genome Biol* 2011;12(9):R95.
- [16] Fan B, Du Z, Gorbach DM, Rothschild MF. Development and application of high-density SNP arrays in genomic studies of domestic animals. *Asian. Austral J Anim* 2010;23(7):833–47.
- [17] Rasheed A, Hao Y, Xia X, Khan A, Xu Y, Varshney RK, et al. Crop breeding chips and genotyping platforms: progress, challenges, and perspectives. *Mol Plant* 2017;10(8):1047–64.
- [18] Mangal M, Bansal S, Sharma SK, Gupta RK. Molecular detection of foodborne pathogens: a rapid and accurate answer to food safety. *Crit Rev Food Sci Nutr* 2016;56(9):1568–84.
- [19] Guppy JL, Jones DB, Jerry DR, Wade NM, Raadsma HW, Huerlimann R, et al. The state of “Omics” research for farmed penaeids: advances in research and impediments to industry utilization. *Front Genet* 2018;9:282.
- [20] Albrechtsen A, Nielsen FC, Nielsen R. Ascertainment biases in SNP chips affect measures of population divergence. *Mol Biol Evol* 2010;27(11):2534–47.
- [21] Mertes F, Elsharawy A, Sauer S, van Helvoort JMLM, van der Zaag PJ, Franke A, et al. Targeted enrichment of genomic DNA regions for next-generation sequencing. *Brief Funct Genomics* 2011;10(6):374–86.
- [22] Mamanova L, Coffey AJ, Scott CE, Kozarewa I, Turner EH, Kumar A, et al. Target-enrichment strategies for next-generation sequencing. *Nat Methods* 2010;7(2):111–8.
- [23] Tewhey R, Warner JB, Nakano M, Libby B, Medkova M, David PH, et al. Microdroplet-based PCR enrichment for large-scale targeted sequencing. *Nat Biotechnol* 2009;27(11):1025–31.
- [24] Damiati E, Borsani G, Giacomuzzi E. Amplicon-based semiconductor

- sequencing of human exomes: performance evaluation and optimization strategies. *Hum Genet* 2016;135(5):499–511.
- [25] Kozarewa I, Armisen J, Gardner AF, Slatko BE, Hendrickson CL. Overview of Target Enrichment Strategies. *Curr Protoc Mol Biol* 2015;112:7.21.1–7.21.23.
- [26] Teer JK, Bonnycastle LL, Chines PS, Hansen NF, Aoyama N, Swift AJ, et al., and the NISC Comparative Sequencing Program. Systematic comparison of three genomic enrichment methods for massively parallel DNA sequencing. *Genome Res* 2010;20(10):1420–31.
- [27] Clark MJ, Chen R, Lam HYK, Karczewski KJ, Chen R, Euskirchen G, et al. Performance comparison of exome DNA sequencing technologies. *Nat Biotechnol* 2011;29(10):908–14.
- [28] Schott RK, Panesar B, Card DC, Preston M, Castoe TA, Chang BSW. Targeted Capture of Complete Coding Regions across Divergent Species. *Genome Biol Evol* 2017;9(2):398–414.
- [29] Sulonen AM, Ellonen P, Almusa H, Lepistö M, Eldfors S, Hannula S, et al. Comparison of solution-based exome capture methods for next generation sequencing. *Genome Biol* 2011;12(9):R94.
- [30] Gasc C, Peyretaillade E, Peyret P. Sequence capture by hybridization to explore modern and ancient genomic diversity in model and nonmodel organisms. *Nucleic Acids Res* 2016;44(10):4504–18.
- [31] Chung J, Son DS, Jeon HJ, Kim KM, Park G, Ryu GH, et al. The minimal amount of starting DNA for Agilent’s hybrid capture-based targeted massively parallel sequencing. *Sci Rep* 2016;6:26732.
- [32] Zhang Y, Li B, Li C, Cai Q, Zheng W, Long J. Improved variant calling accuracy by merging replicates in whole-exome sequencing studies. *BioMed Res Int* 2014;2014:319534.
- [33] Yi X, Liang Y, Huerta-Sanchez E, Jin X, Cuo ZXP, Pool JE, et al. Sequencing of 50 human exomes reveals adaptation to high altitude. *Science* 2010;329(5987):75–8.
- [34] Yigit E, Zhang Q, Xi L, Grilley D, Widom J, Wang J, et al. High-resolution nucleosome mapping of targeted regions using BAC-based enrichment. *Nucleic Acids Res* 2013;41(7):e87.
- [35] Cao H, Wu J, Wang Y, Jiang H, Zhang T, Liu X, et al. An integrated tool to study MHC region: accurate SNV detection and HLA genes typing in human MHC region using targeted high-throughput sequencing. *PLoS One* 2013;8(7):e69388.
- [36] Lv J, Jiao W, Guo H, Liu P, Wang R, Zhang L, et al. HD-Marker: a highly multiplexed and flexible approach for targeted genotyping of more than 10,000 genes in a single-tube assay. *Genome Res* 2018;28(12):1919–30.
- [37] Sambrook J, Fritsch EF, Maniatis T. *Molecular cloning, a laboratory manual*. 2nd ed. Now York: Cold Spring Harbor Laboratory Press; 1989.
- [38] Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 2009;25(14):1754–60.
- [39] Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al., and the 1000 Genome Project Data Processing Subgroup. The sequence alignment/map format and SAMtools. *Bioinformatics* 2009;25(16):2078–9.
- [40] Koboldt DC, Chen K, Wylie T, Larson DE, McLellan MD, Mardis ER, et al. VarScan: variant detection in massively parallel sequencing of individual and pooled samples. *Bioinformatics* 2009;25(17):2283–5.
- [41] Liu F, Li Y, Yu H, Zhang L, Hu J, Bao Z, et al. MolluscDB: an integrated functional and evolutionary genomics database for the hyper-diverse animal phylum Mollusca. *Nucleic Acids Res* 2021;49(D1):D1556.
- [42] Yang Z, Zhang L, Hu J, Wang J, Bao Z, Wang S. The evo-devo of molluscs: Insights from a genomic perspective. *Evolution & Development* 2020;22(6):409–424.
- [43] Hou R, Bao Z, Wang S, Su H, Li Y, Du H, et al. Transcriptome sequencing and de novo analysis for Yesso scallop (*Patinopecten yessoensis*) using 454 GS FLX. *PLoS One* 2011;6(6):e21560.
- [44] Wang S, Hou R, Bao Z, Du H, He Y, Su H, et al. Transcriptome sequencing of Zhikong scallop (*Chlamys farreri*) and comparative transcriptomic analysis with Yesso scallop (*Patinopecten yessoensis*). *PLoS One* 2013;8(5):e63927.
- [45] Wang S, Zhang J, Jiao W, Li J, Xun X, Sun Y, et al. Scallop genome provides insights into evolution of bilaterian karyotype and development. *Nat Ecol Evol* 2017;1(5):0120.
- [46] Thomson MJ. High-throughput SNP genotyping to accelerate crop improvement. *Plant Breed Biotechnol* 2014;2(3):195–212.
- [47] Syvänen AC. Toward genome-wide SNP genotyping. *Nat Genet* 2005;37(S6 Suppl):S5–10.
- [48] Fan JB, Chee MS, Gunderson KL. Highly parallel genomic assays. *Nat Rev Genet* 2006;7(8):632–44.
- [49] Perkel J. SNP genotyping: six technologies that keyed a revolution. *Nat Methods* 2008;5(5):447–54.
- [50] Paux E, Sourdil P, Mackay I, Feuillet C. Sequence-based marker development in wheat: advances and applications to breeding. *Biotechnol Adv* 2012;30(5):1071–88.
- [51] Hayes B, Goddard M. Genome-wide association and genomic selection in animal breeding. *Genome* 2010;53(11):876–83.
- [52] Goddard ME, Hayes BJ, Meuwissen THE. Using the genomic relationship matrix to predict the accuracy of genomic selection. *J Anim Breed Genet* 2011;128(6):409–21.
- [53] Ballester LY, Luthra R, Kanagal-Shamanna R, Singh RR. Advances in clinical next-generation sequencing: target enrichment and sequencing technologies. *Expert Rev Mol Diagn* 2016;16(3):357–72.
- [54] Robledo D, Palaiokostas C, Bargelloni L, Martínez P, Houston R. Applications of genotyping by sequencing in aquaculture breeding and genetics. *Rev Aquacult* 2018;10(3):670–82.
- [55] de Oliveira AA, Guimaraes LJM, Guimaraes CT, de Oliveira Guimaraes PE, de Oliveira Pinto M, Pastina MM, et al. Single nucleotide polymorphism calling and imputation strategies for cost-effective genotyping in a tropical maize breeding program. *Crop Sci* 2020;60(6):3066–82.
- [56] Tsairidou S, Hamilton A, Robledo D, Bron JE, Houston RD. Optimizing low-cost genotyping and imputation strategies for genomic selection in Atlantic Salmon. *G3-Genes Genom Genet* 2020;10(2):581–90.
- [57] Luo Z, Yu Y, Xiang J, Li F. Genomic selection using a subset of SNPs identified by genome-wide association analysis for disease resistance traits in aquaculture species. *Aquaculture* 2021;539:736620.