Advances in Marine Sciences 海洋科学前沿, 2023, 10(4), 272-284
Published Online December 2023 in Hans. https://www.hanspub.org/journal/ams
https://doi.org/10.12677/ams.2023.104028

Hans汉斯

基于全基因组重测序策略对三疣梭子蟹耐副溶血弧菌抗性相关基因进行挖掘

阎德平

海军士官学校，安徽 蚌埠

收稿日期：2023年10月18日；录用日期：2023年12月5日；发布日期：2023年12月13日

摘　要

副溶血弧菌(*Vibrio parahaemolyticus*)是造成三疣梭子蟹批量死亡的重要病原之一。随着测序技术发展，基于Illumina测序及生物信息学技术手段筛选耐病基因已成为可能。本研究利用全基因组重测序的方法对三疣梭子蟹易感群体和耐感群体的肌肉组织进行Illumina测序，过滤获得51.867G的clean reads。将clean reads数据与已有梭子蟹参考基因组进行比对，覆盖比对率高达85%以上，覆盖深度也达到25X，同时还检测到了36,929个单核苷酸多态性位点(single nucleotide polymorohsms, SNP)和145,790个小片段插入缺少位点(insertion/deletion, InDel)。以上位点通过SNP/InDel频率分布，在染色体上作图进行精细定位，获得了257个SNPs和184个InDels。以上位点进一步通过同义突变筛查，获得了55个SNP位点和32个InDel位点。对87个位点设计引物进行验证，利用一代测序技术和ContigExpress软件对变异位点进行SNP分型，结果表明：与参考基因组相比，55个SNP位点和32个InDel位点中，分别有23个SNP标记和10个InDel标记存在碱基变化；继续利用上述引物，分别在易感群体和耐感群体上进行扩增测序，最终筛选出9个SNP标记和2个InDel标记($P < 0.05$)，并瞄定到8个基因上，推测其中有5个基因属于抗逆抗病基因可用于改良三疣梭子蟹优良性状。

关键词

基因组重测序，位点

Mining of Genes Associated with *V. parahemolytica* of the Swimming Crab Resistance Based on the Whole Genome Resequencing Strategy

Deping Yan

The Naval Academy, Bengbu Anhui

文章引用: 阎德平. 基于全基因组重测序策略对三疣梭子蟹耐副溶血弧菌抗性相关基因进行挖掘[J]. 海洋科学前沿, 2023, 10(4): 272-284. DOI: 10.12677/ams.2023.104028

Received: Oct. 18th, 2023; accepted: Dec. 5th, 2023; published: Dec. 13th, 2023

Abstract

***Vibrio parahaemolyticus* is one of the most important pathogens that cause the batch death of the swimming crab. With the development of sequencing technology, it has been possible to screen disease resistance genes based on Illumina sequencing and bioinformatics technology. In this study, high-throughput Illumina sequencing technology was used to perform a whole-genome re-sequencing of muscle tissues in susceptible and resistant populations of *P. trituberculatus*, and filtered to obtain 51.867G clean reads. The sequencing results were compared with the genome of swimming crab. The coverage ratio was above 85% and the coverage depth reached 25X. At the same time, 36,929 single nucleotide polymorohsms and 145,790 Insert/missing mutant fragments were detected. Subsequently, the SNPs/InDels frequency distribution was mapped on the chromosome for fine localization, and finally 257 SNPs and 184 InDels were obtained. We functionally annotated the above results and found that these loci are non-synonymous mutations, mainly concentrated in introns and intergenic regions. We screened the sites obtained from the resequencing of *P. trituberculatus* genomes, selected 55 SNP sites and 32 InDel sites to design primers for verification, and used the first-generation sequencing technology and ContigExpress software to perform SNP genotyping on mutation sites. The results showed that compared with the reference genome, there were base changes in 55 SNP loci and 32 InDel loci, respectively in 23 SNP loci and 10 InDel loci. Continuing to use the above primers to perform amplification and sequencing on susceptible individuals and resistant individuals, respectively, finally, 9 SNP markers and 2 InDel markers were selected ($P < 0.05$), and 8 genes were targeted. It was speculated that 5 genes belonged to resistance genes and could be used to improve the good traits of *Portulus trisulatus*.**

Keywords

The Whole Genome Resequencing, Site

1. 引言

三疣梭子蟹(*Portunus trituberculatus*)是中国重要的养殖蟹类，已经吸引了广泛的研究，这需要越来越多的基因组背景知识。迄今为止，三疣梭子蟹尚没有完整的全基因组信息，而且该物种的转录组信息也很少。因此，利用下一代测序技术对其基因组和转录组进行测序分析，系统地研究三疣梭子蟹在弧菌感染下的抗性基因，不仅可以筛选到一些耐弧菌免疫基因，阐明其代谢途径，还可以挖掘到一些跟弧菌抗性相关的分子标记，为培育抗逆抗病的三疣梭子蟹新品种提供理论基础，使三疣梭子蟹健康可持续发展。

2. 前言

Sanger 测序等一代测序技术曾在解析基因组序列方面发挥了巨大的作用，然而一代测序也存在着成本高、周期长、产出率低等诸多不足之处，一直无法在世界范围内得以大力推广。近些年来，随着科技进步“下一代测序”(Next-Generation Sequencing, NGS)以高通量测序(High-Throughput Sequencing)为标志，凭借其较为低廉的价格、周期短和产出率高的优势在动植物领域得到广泛的应用。目前，下一代测

序的主要测序技术有美国罗氏公司(Roche)的 454 基因组测序仪、Illumina 公司开发的 Illumina 测序仪和 ABI 公司的 SOLID 连接酶测序平台，这些测序平台均采用循环芯片测序法(cyclic-array sequencing)，并被誉为第二代测序技术[1] [2]。

全基因组测序(whole genome sequencing)是对未知基因组序列的物种进行个体的基因组测序，能够快速的鉴定到大量高密度的 SNP 位点，可用于重要候选基因的筛选、遗传变异检测及群体遗传进化分析等，因而全基因组测序广泛应用于群体进化、群体结构、种群历史、遗传定位和连锁图谱的构建，例如王金昌等对海洋贝莱斯芽胞杆菌 Bam-6 基因族进行注释发现了贝莱斯芽胞杆菌代谢合成物的同源物、毛明光等对太平洋鳕鱼线粒体基因族进行测序分析发现一段保守序列[3] [4] [5]。WGS 研究主要包括两方面，一方面为全基因组从头测序(de novo)，另一方面为全基因组重测序(re-sequencing) [6]。

重测序是以物种的参考基因组序列为依据，进行个体或群体间的基因组测序，并对其差异信息进行分析的一种测序方法。相比较于传统的方法，重测序作为二代测序具有许多优点：1) 信息全面，可以获得全基因组的序列信息；2) 信息精确，可以精确挖掘到每个 SNP 位点，直接找到致病位点；3) 产出效率高，可以挖掘到许多性状相关的关键基因[7]。通过全基因组重测序技术，可以获取到大量的遗传变异信息，实现 DNA 分子水平上的遗传分析并筛选到性状相关的候选基因。

关于哺乳动物的研究中，利用 WGS 技术进行遗传分析已经得到广泛应用。Leif Andresson 团队对 9 个群体的鸡进行 WGS 测序筛选到 3 个重要的驯化基因；黄路生教授团队对 11 个中国地方猪种和 3 个野生猪种进行了 WGS 测序，筛选到了 210 个与环境适应性相关的基因；Stothard 团队运用重测序的方法首次在美国荷斯坦牛和黑安格斯牛上开展了拷贝数变异(copy number variation, CNV)检测[8] [9]。尽管重测序在哺乳动物中已经取得了许多成绩，但在水产动物中仍少见报道。

3. 材料和方法

3.1. 实验材料

实验材料均来自于青岛黄海水产研究所，分别来自于健康存活未感染副溶血弧菌的 80 日龄梭子蟹和感染副溶血弧菌 72 h 后存活梭子蟹的肌肉组织。

3.2. 建库、测序及分析

将检验合格的 DNA 样品等量混合为两个混合池，分别命名为易感 DNA 混合池(CG)和耐感 DNA 混合池(CT) [10]。通过 Covaris 破碎机随机打断混合 DNA 样品成 350 bp 片段，使用 Truseq Library Construction Kit 对其建库，建库过程中严格使用说明书中推荐的试剂和耗材。350 bp 片段通过终端修复、ploya 尾、测序接头、纯化、PCR 扩展等一系列步骤后，整个文库就制备完成。

建库具体流程如图 1 所示。文库制备完成后，采用 Qubit2.0 进行初步定量，将文库稀释至 1 ng/ul。然后使用 Agilent 2100 检测文库的 insert size。待检测 insert size 符合标准后，为确保文库质量，还需采用 Q-PCR 法准确定量文库有效浓度(文库有效浓度 > 2 nM)。文库浓度检测合格后，再按有效浓度和目标下机数据量的需求 pooling 对不同文库进行 Illumina HiSeq TM PE150 测序[11]。

由于测序获得的是 raw reads 或 Sequenced Reads，其中可能带有大量低质量的 reads，为了得到 clean reads，需要对 raw reads 进行信息质量分析，其步骤如下[12]：

1) 去除带接头(adapter)的 reads pair；

2) 当单端测序 read 中含有的 N 的含量超过该条 read 长度比例的 10%时，需要去除此对 paired reads；

3) 当单端测序 read 中含有的低质量(Q ≤ 5)碱基数超过该条 read 长度比例的 50%时，需要去除此对 paired reads。

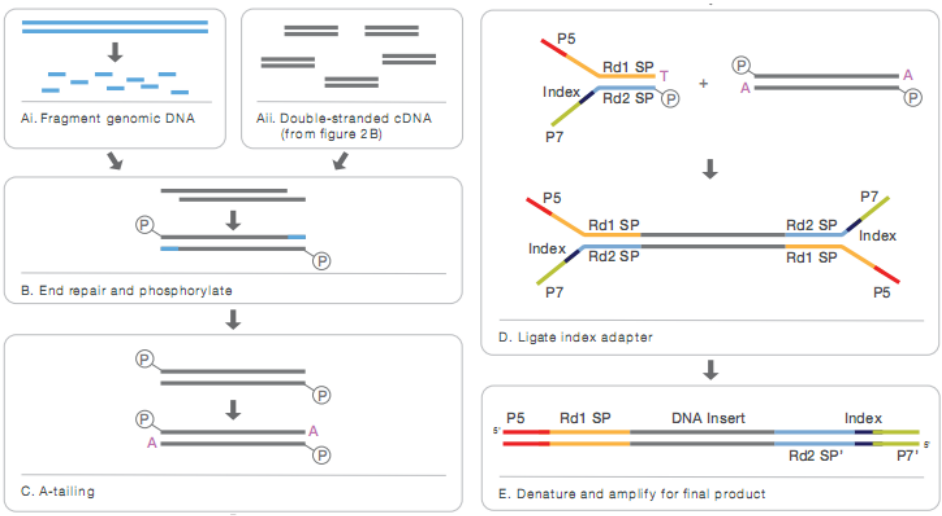


Figure 1. Library construction sequencing process
图 1. 建库测序流程

应使用 Burrows-Wheeler alignment tool (BWA)比对软件将过滤后的有效数据与参考基因组进行比对，参考基因组统计信息见下表 1 所示，使用 SAMTOOLS 软件去除比对结果重复[10]。获得的数据可以用于后续注释分析及 SNP/InDel 检测，可实现 DNA 水平差异功能基因注释及差异基因挖掘，全基因组重测序的生物信息分析流程如图 2 所示。

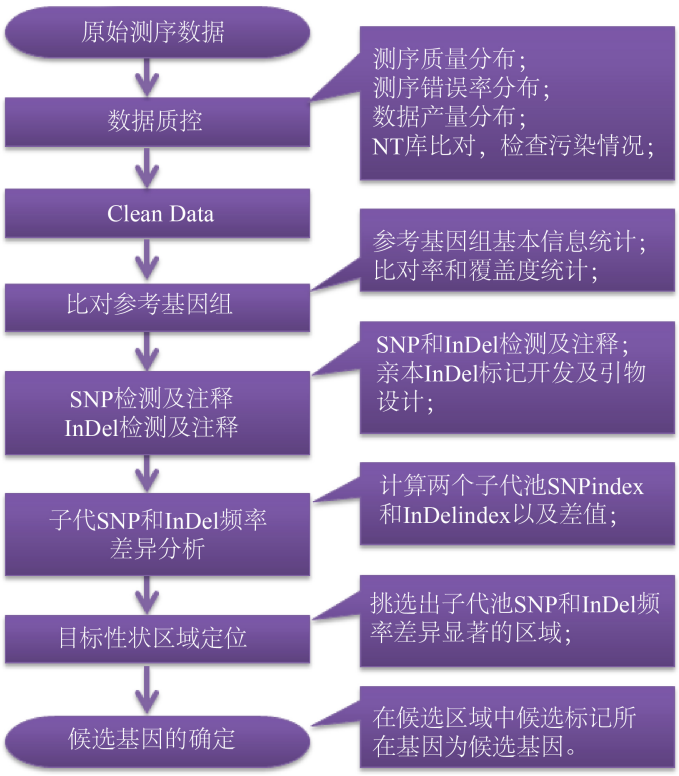


Figure 2. Whole-genome biological information analysis process
图 2. 全基因组生物信息分析流程

Table 1. Statistics of reference genome
表 1. 参考基因组基本情况统计

基因组组装的序列总数	基因组组装结果总长度	碱基 G 和 C 的含量(%)	组装结果中 N 所占的比例	累积长度刚刚超过全部组装序列总长度 50%时的那条 scaffold 的长度	累积长度刚刚超过全部组装序列总长度 90%时的那条 scaffold 的长度
41	850,292,103	36.10	13.17	38,829,317	17,398,227

3.3. QTL-seq 分析

3.3.1. SNP/InDel 检测和注释

SNP (single nucleotide polymorphism，单核苷酸多态性)是指在基因组水平上由于单核苷酸发生变异而引起 DNA 序列产生多态性，包括单碱基的颠换或转换等[9]。而 InDel 则是表示基因组产生小片段的缺失和插入序列。我们采用 Genome analysis toolkit 3.8 (GATK)软件中的 UnifiedGenotyper 模块检测 SNP 和 InDel，SNP 过滤参数设置为：MQ < 40，QD < 4，FS > 60；InDel 过滤参数设置为 QD < 4，FS > 200 [7]。

3.3.2. SNP 频率差异分析

我们以参考基因组作为参考，分析计算易感和耐感个体在每个 SNP 位点中的 SNP-index (SNP 的频率)。如图 3 所示是对个体池中 SNP-index 计算的一种统计方法，其原理是以参考基因组或某一亲本作为参考，通过测序 reads 以便于对碱基位点的碱基进行统计分析。统计在某一个碱基位点处个体池和亲本或者参考基因组是否出现不同或相同的 reads，并统计其中不相同条数占总条数的比例，该比例即为 SNP-index。其中，完全与其参考基因组或亲本完全与其不同的 SNP-index 记为 1，相同的则记为 0。按照此方法计算出敏感池和耐感池的全部 SNP-index。为减少测序错误和比对错误造成的影响，对计算出 SNP-index 后的多态性位点进行过滤，过滤标准如下：

1) 两个个体中 SNP-index 都小于 0.3，并且 SNP 深度都小于 7 的位点，过滤掉；

2) 一个个体 SNP-index 缺失的位点，过滤掉。

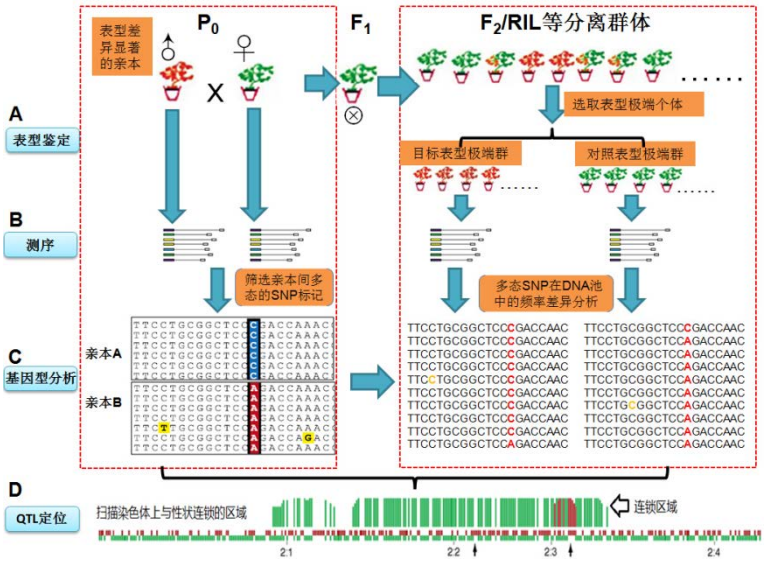


Figure 3. SNP-index calculation method
图 3. SNP-index 计算方法

得到过滤后的多态性位点后，对 SNP-index 在染色体上的分布进行作图。默认选择 1 Mb 为窗口，1 kb 为步长，计算每个窗口中 SNP-index 的平均值来反应个体的 SNP-index 分布。同时对 SNP 频率差异分布进行作图，计算 Δ(SNP-index)，即两个个体 SNP-index 作差：Δ(SNP-index) = SNP-index (耐副溶血弧菌弧菌性状 B) – SNP-index (副溶血弧菌敏感性状 A)。进行 1000 次置换检验，选取 95% (蓝色)置信水平作为筛选的阈值。

3.3.3. InDel 频率差异分析

InDel-index 分析方法同 1.3.2。

3.4. 候选标记挖掘

3.4.1. 候选标记筛选

为了不忽略掉微效 QTL 的影响，在全基因组范围内挑选候选 SNP 和 InDel，如果参考亲本和子代表型相同，则挑选子代池中 All-index 接近 0 的位点；如果参考亲本和子代表型相反，则挑选子代池中 All-index 接近 1 的位点作为候选位点。

SNP 标记筛选原则：

1) 选择个体间 Δ(SNP-index)接近 1 的标记；

2) 选择多个标记位于一个 Conting 上的；

3) 对产生移码突变或 stop gain 或 stop loss 或非同义突变或者可变剪接位点的位点优先进行筛选。

InDel 标记筛选原则：

1) 选择个体间 Δ(InDel-index)接近 1 的标记；

2) 选择多个标记位于一个 Conting 上的；

3) 按照插入片段长度进行排序。

引物设计原则：

1) 避开在标记左右 50 bp 处设计引物；

2) 扩增片段在 300~500 bp 左右；

3) 引物长度在 18~23 bp；

4) 引物退火温度控制在 55℃~65℃之间，上下游引物温度差最好控制在 3℃之间；

5) 引物 GC 含量控制在 40%~60%之间，上下游引物 GC 含量差最好不要超过 5%。

3.4.2. 耐副溶血弧菌标记的验证

采用 PCR 产物测序的方法在 CG 和 CT 群体里对耐副溶血弧菌性状相关候选分子标记进行验证：

1) 首先在标记位点侧翼序列设计引物，其中至少有一条引物距离标记位点 70 bp 以上；

2) 利用设计好的引物分别以 CG 和 CT 混合 DNA 材料为模板进行 PCR 扩增，并将成功扩增的 PCR 产物进行测序，测序引物选择离标记位点较远的引物；

3) 利用 ContigExpress 软件分析测序峰图，挑选 CG 和 CT 两组在对应位置测序峰图有较大差异的标记继续进行个体 DNA 模板的 PCR 扩增和测序分析；

4) 根据测序结果统计每个个体的基因型，并通过 SPSS 软件分析标记与耐副溶血弧菌性状是否相关[13]。

具体的操作步骤如下：

1) PCR 扩增体系和程序

利用全式金公司的高保真酶进行 PCR 扩增。设置 PCR 反应程序为 1，95℃，2 min；2，95℃，20 s；3，55℃，20 s；4，72℃，30 s；2~4，35 个循环；5，72℃，5 min；6，4℃保存。PCR 反应体系见图 4。

PCR扩增体系

试剂	用量
DNA模板	1uL
Pfu Fly Buffer	4uL
dNTP (2.5mM)	1.6uL
引物(上下游)	0.8uL
Pfu酶(5U/ML)	0.4uL
灭菌水	补足至20uL

Figure 4. The PCR amplification system
图 4. PCR 扩增体系

2) 电泳检测

用琼脂糖配置 1%的电泳胶，把制好的琼脂糖凝胶放入水平电泳槽中，进行电泳检测实验，电泳时间设定为 30 min，电泳结束后用凝胶成像系统观察并拍照记录，切割出明亮且条带单一的电泳胶样品送置青岛擎科生物有限公司进行测序。

3) 统计分析

利用 ContigExpress 软件对测序峰图进行分析，选择耐感群体混合模板和易感群体混合模板中碱基位置出现显著差异的引物，并继续使用该引物，以每个个体中进行 PCR 扩增验证，扩增条件和上述一致，然后将明亮且位置大小符合的电泳条带送去测序[10]。

再将返回的个体测序结果用 ContigExpress 软件进行观察并将基因型信息导入 SPSS 软件，利用卡方检验计算 P 值，选出 $P < 0.05$ 的标记为候选标记。

4. 结果与分析

4.1. 全基因组重测序结果

由于现阶段的技术不足，需要在测序前添加上一些接头，导致部分测序结果中也会含有冗余的接头序列信息；此外，测序时也可能会产生一些低质量的序列信息；因此对序列的质量进行评估以及过滤对后续的结果分析极为重要。经测序质量分布检查、测序错误率分布检查以及测序数据过滤后，统计结果表明测序结果极佳，质量都在 Q30 以上，错误率低，有效数据质量高，如图 5 和图 6 所示。

本次测序共产生 Raw data 高达 51.963G，过滤后的 Clean data 也有 51.867G，本次测序质量高(Q20 ≥ 94%、Q30 ≥ 87%)，GC 含量也在 41%左右。因此，本次实验样本的数据量充足，GC 分布正常且测序质量高，符合建库测序成功标准。测序质量数据汇总见表 2。

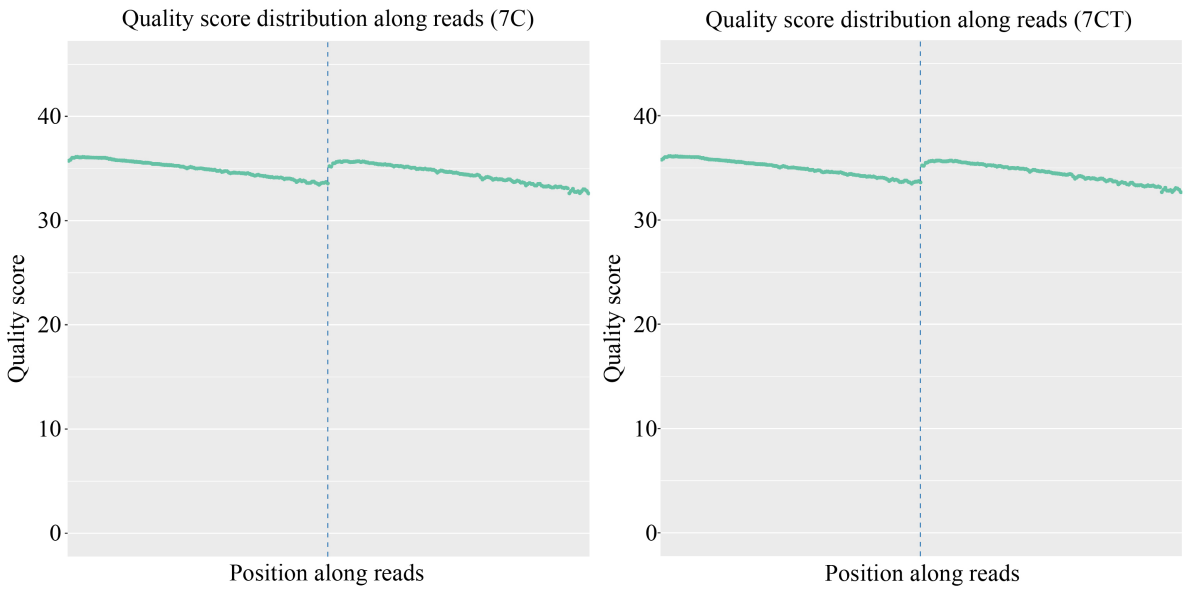
使用 BWA 软件将易感池和耐感池的测序数据和梭子蟹参考基因组进行比对。比对结果表明，所有样本的比对率在 85%以上，对参考基因组(排除 N 区)的平均覆盖深度在 25X 以上，1X 覆盖度(至少有一个碱基的覆盖)在 79%以上。比对结果正常，可用于后续的标记检测分析。具体 Reads 与参考基因组比对情况统计如表 3 所示。

Table 2. Summary of the quality of sequencing data
表 2. 测序数据质量情况汇总

Sample	Raw Base (bp)	Clean Base (bp)	Effective Rate (%)	Error Rate (%)	Q20 (%)	Q30 (%)	GC Content (%)
C	25,917,187,500	25,867,454,700	99.81	0.035	94.22	87.475	41.495
CT	26,046,389,400	25,999,858,800	99.82	0.035	94.255	87.595	40.96

Table 3. Sequencing depth and coverage statistics
表 3. 测序深度及覆盖度统计

样本名	双端比对 reads 条数数	总 reads 条数	比对率	平均测序深度	参考基因组中至少有 1 个碱基覆盖的位点占基因组的百分比	参考基因组中至少有 4 个碱基覆盖的位点占基因组的百分比
C	147,383,735	172,449,698	85.46	25.05	79.15	76.69
CT	147,651,575	173,332,392	85.18	25.32	79.08	76.76

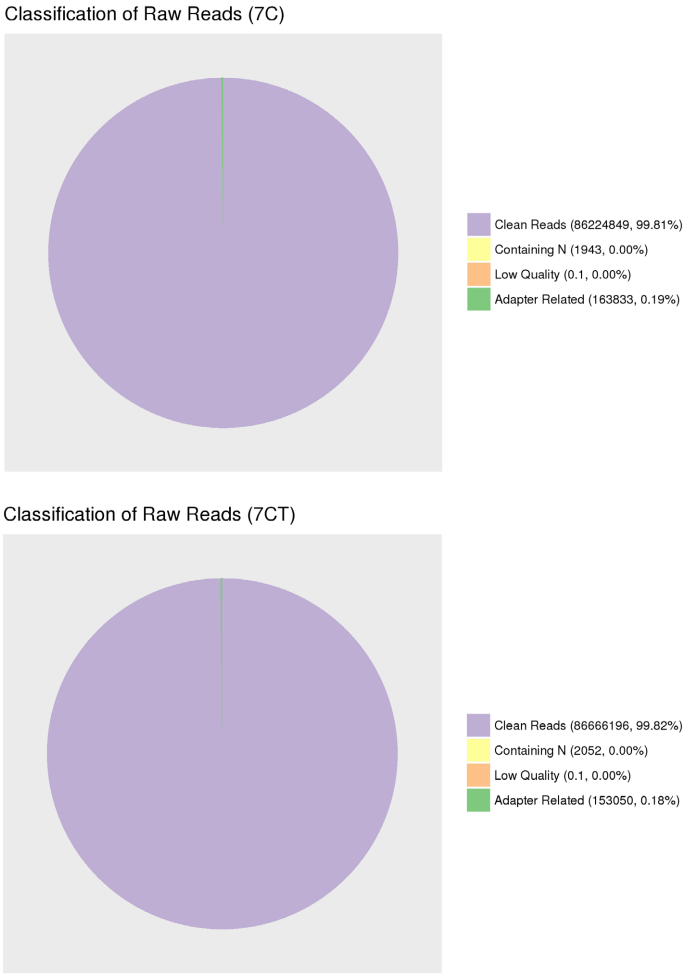


纵坐标为单碱基错误率，横坐标为 reads 的碱基位置；前 150 bp 为双端测序序列的第一端测序 Reads 的质量值分布情况，后 150 bp 为另一端测序 reads 的质量值分布情况。
The abscissa is the base position of reads and the ordinate is the single base error rate; the first 150 bp is the quality value distribution of the first-end sequencing reads of the double-end sequencing sequence, and the last 150 bp is the quality value distribution of the sequencing reads at the other end.

Figure 5. Sequencing quality distribution map
图 5. 测序质量分布图

4.2. SNP/InDel 检测与注释

有效数据经与参考基因组比对，分别检测到 36,929 个 SNP 标记和 145,790 个 InDel 位点。候选位点的注释情况如表 4 和表 5 所示。



(1) Adapter related：因有接头，过滤掉的 reads 数及其占总 raw reads 数的比例。(2) Containing N：因 N 含量超过 10%，过滤掉的 reads 数及其占总 raw reads 数的比例。(3) Low quality：因低质量，过滤掉的 reads 数及其占总 raw reads 数的比例。(4) Clean reads：最终得到的 clean reads 数及其占总 raw reads 数的比例。
(1) Adapter related: Due to the adapter, the number of reads filtered out and its proportion to the total number of raw reads. (2) Containing N: because the N content exceeds 10%, the number of reads filtered out and its proportion to the total number of raw reads. (3) Low quality: due to low quality, the number of reads filtered out and its proportion to the total number of raw reads. (4) Clean reads: The final number of clean reads and their proportion to the total number of raw reads.

Figure 6. Filtering results of raw data
图 6. 原始数据过滤结果

Table 4. Statistics of SNP detection and annotation
表 4. SNP 检测及注释结果统计

类别		SNP 数量
上游		6269
外显子区域	获得终止子变异	25
	失去终止子变异	4
	同义突变	3206
	非同义突变	1767

Continued

内含子区域	78,726
剪切位点	18
下游	5136
基因上游 1 Kb 区域，同时也在另一基因的下游 1 Kb 区域	386
基因之间	270,857
转换	229,986
颠换	139,243
转换和颠换的比率	1651
总计	369,229

Table 5. InDel detection and annotation result statistics
表 5. InDel 检测及注释结果统计

类别		InDel 数量
上游		2548
外显子区域	获得终止子变异	6
	失去终止子变异	8
	缺失造成移码	240
	插入造成移码	114
	非缺失移码	80
	非插入移码	57
内含子区域		29,213
剪切位点		19
下游		2268
基因上游 1 Kb 区域，同时也在另一基因的下游 1 Kb 区域		186
基因之间		110,027
插入		60,375
缺失		85415
总计		145790

4.3. 重测序 SNP/InDel 位点验证分析

经 SNP 和 InDel 合并后的频率差异分析后，得到成功注释的 SNP 标记 257 个，InDel 标记 187 个。在注释结果中选择了 55 个 SNP 标记和 32 个 InDel 标记进行验证，初步验证结果表明 23 个 SNP 标记中有 10 个标记与检测结果一致，阳性率为 41.82%；32 个 InDel 标记中有 10 个标记与检测结果一致，阳性率为 31.25% (InDel 标记由于序列的复杂性导致测序结果较低，可能对阳性率造成了影响)。

继续利用上述位点存在显著差异的引物，分别在易感个体和耐感个体上进行扩增测序，最终筛选出 9 个 SNP 标记和 2 个 InDel 标记，统计结果如表 6 所示。

Table 6. Markers related to *Vibrio resistance* and prediction of their gene functions
表 6. 与弧菌抗性相关标记及其定位基因功能预测

Pose	Mutation type	ΔSNP-index	Variant	Chi-square value	P value	Prediction function
Contig0_13982611	T→G	0.72	intronic	6.162	0.046	dual oxidase maturation factor 1
Contig242_653632	A→G	0.70	intronic	10.667	0.005	WAP four-disulfide core domain protein 1
Contig26_556258	G→A	0.70	intronic	6.044	0.049	GDNF family receptor
Contig405_623426	G→A	0.80	intronic	7.940	0.014	Gustatory receptor trehalose
Contig7_625237	C→T	0.69	intronic	17.143	0.001	Radial spoke head protein
Contig81_1278459	G→A	0.7	downstream	8.640	0.013	DNA-damage-inducible transcript 4
Contig7_625234	G→T	0.69	intronic	11.378	0.003	Radial spoke head protein
Contig104_411267	G→A	0.70	intronic	11.911	0.003	solute carrier family 10
Contig7_4587668	A→T	0.77	intronic	13.972	0.001	Methyltransferase
Contig7_4584002	G→GA	0.7	intergenic	9.6	0.02	Methyltransferase
Contig405_597023	GAA→A	0.76	UTR3	6.096	0.047	Gustatory receptor trehalose
Contig3_6549174	G→GC	0.71	intronic	13.333	0.001	Gap junction beta-5 protein

5. 讨论

由于基因组学研究技术手段的飞快发展，使得检测大规模、高通量的动物基因组内的变异位点变得越来越容易[14]。先前的研究报道表明，不同物种的不同种群 SNP 频率不同，而有关三疣梭子蟹抗弧菌反应相关的免疫相关 SNP 信息鲜有报道[15]。因此，对三疣梭子蟹抗病抗逆相关基因及其 SNPs 进行研究，有助于提高对三疣梭子蟹免疫防御机制的认识，有利于梭子蟹抗病品种的筛选和培养。为得到高质量的检测数据及结果，本研究使用了一种基于 BSA 混合池的全基因组重测序技术对对三疣梭子蟹易感群体和耐感群体之间可能存在的抗病相关的差异基因进行了深入挖掘。相比于传统的抗病 QTL 的定位方法，本研究所用的全基因重测序技术不仅省时省力，并且可以从分子层面直接挖掘与抗病相关的基因。

目前 Illumina 公司的测序平台是应用最为广泛的二代测序平台，本研究即采用 Illumina HiSeq™ PE150 平台进行测序。碱基的质量高低与测序错误率息息相关，鉴于当前测序技术扔存在局限性，因此测序片段前段和末端几个 cycles 的错误率会偏高[16]。测序获得的碱基质量值用一般 Q_{phred} 值表示，如果测序错误率用 e 表示，则 $Q_{phred} = -10\log_{10}(e)$，Q10、Q20、Q30 和 Q40 表示不正确的碱基识别分别 1/10、1/100、1/1000 和 1/10000，即碱基正确识别率分别为 90%、99%、99.9%和 99.99%；根据重测序的结果表明，本研究共产生 51.983G 的数据，Q20 平均值 ≥94.23%，Q30 平均值 ≥87.54%，测序质量较高。重测

序需要将产生的 Clean reads 通过 BWA 软件比对到梭子蟹的参考基因组上。结果表明，两个群体的所有样本跟梭子蟹参考基因组的比对率在 85%以上，与参考基因组具有较高的相似性；同时，所有样本对参考基因组的平均覆盖深度在 25X 以上，1X 覆盖度在 79.08%以上，4X 覆盖度在 76.69%以上，这些结果都表明该测序数据的均一性和参考基因组序列的同源性较高，数据质量高。

SNPs 是基因组中广泛存在的突变，基于基因组中的不同位置，SNP 可以通过不同的机制来影响基因的翻译或转录，基因调控区内 SNP 发生突变可能会影响相关基因的表达像周慧等发现 miR-17-92 基因启动子区 rs1813389 A/G 碱基发生颠换可能于子宫内膜癌症有关；非同义编码区 SNP 突变直接改变基因编码蛋白的氨基酸组成，对蛋白质功能域的出现具有至关重要的作用像黎江溪发现 TNNC1 基因编码区第 44 个碱基由 G 转换成 C 可能造成肥厚型心肌病高风险[17] [18] [19]。先前的研究主要集中在非同义突变 SNP (nonsynonymous single nucleotide polymorphism, nsSNP)，因为这些 SNP 最有可能直接影响蛋白质的结构和活性；然而近些年来，越来越多的研究表明，内含子和基因间的 SNPs 也可能在性状的变化中起着决定性作用[20]。例如，F-box 和富含亮氨酸重复蛋白 17 (leucine rich repeat protein 17, FBXL17)基因的第三内含子中的 SNP 突变解释了拌花生和他性别逆转中存在了 58.4%的表型变异[21]。因此，我们在测序时对三疣梭子蟹内含子和基因之间区域 SNPs 进行分析。

6. 小结

在本研究中，从两个群体的 WGS 分析中共鉴定出 257 个 SNP 和 184 个 InDel，并成功进行了功能注释，并没有检测到非同义突变，大部分标记处于内含子和基因之间。其中位于内含子区域的 SNP 和 InDel 分别占比 18.29%和 25%；位于基因之间区域的 SNP 和 InDel 分别占比 78.6%和 71.7%。

参考文献

[1] Al Margulies, M., Egholm, M., Altman, W., *et al*. (2005) Genome Sequencing in Microfabricated High-Density Picolitre Reactors. *Nature*, **437**, 158-160.

[2] Ada, V., Shoshan, H., Dana, F., *et al*. (2001) Antifungal Activity of a Novel Endochitinase Gene (chit36) from *Trichoderma harzianum* Rifai TM. *FEMS Microbiology Letters*, **200**, 169-174. https://doi.org/10.1111/j.1574-6968.2001.tb10710.x

[3] 施阳. 基于全基因组重测序获得的具 LRR 结构域基因的抗黄瓜白粉病功能鉴定[D]: [硕士学位论文]. 扬州: 扬州大学, 2023.

[4] 王金昌. 海洋贝莱斯芽胞杆菌 Bam-6 全基因组测序及生物信息分析[J]. 中国生物防治学报, 2023, 39(2): 438-452. https://doi.org/10.16409/j.cnki.2095-039x.2023.02.012

[5] 毛明光. 太平洋鳕线粒体全基因组测序及结构特征分析[J]. 水生生物学报, 2022, 43(1): 17-26. https://doi.org/10.7541/2019.003

[6] 潘章源, 贺小云, 刘秋月, 等. 全基因组测序(WGS)在畜禽群体进化和功能基因挖掘中的应用[J]. 农业生物技术学报, 2016, 24(12): 10. https://doi.org/10.3969/j.issn.1674-7968.2016.12.017

[7] Ai, H., Fang, X., Yang, B., *et al*. (2010) Adaptation and Possible Ancient Interspecies Introgression in Pigs Identified by Whole-Genome Sequencing. *Nature Genetics*, **47**, 217-225. https://doi.org/10.1038/ng.3199

[8] Rubin, C.J., Zody, M.C., Eriksson, J., *et al*. (2010) Whole-Genome Sequencing Reveals Loci under Selection during Chicken Domestication. *Nature*, **464**, 587-591. https://doi.org/10.1038/nature08832

[9] 贾秀苹, 卯旭辉, 岳云, 等. BSA-Seq 方法鉴定向日葵关键耐盐基因[C]//中国作物学会油料作物专业委员会第八次会员代表大会暨学术年会综述与摘要集. 青岛: 中国油料作物学报, 2018: 115-123.

[10] 吕建建, 阎德平, 等. 一种三疣梭子蟹耐副溶血弧菌的分子标记 C7 及其应用[P]. 中国, CN201911198805.0. 2021-06-08.

[11] 程凤, 宋蒙飞, 曹蕾, 等. 黄瓜的一个中短果突变体基因的初步定位[J]. 园艺学报, 2021, 48(7): 1359-1370. https://doi.org/10.16420/j.issn.0513-353x.2021-0194

[12] 吉康娜, 郅俊杰, 林丹妮, 等. 基于茄子基因组重测序的 InDel 标记开发及应用[J]. 植物遗传资源学报, 2019,

20(5): 1278-1288. https://doi.org/10.13430/j.cnki.jpgr.20190130001

[13] 吕建建, 阎德平, 陆璇, 等. 一种三疣梭子蟹耐副溶血弧菌的分子标记 C3 及其应用[P]. 中国, CN201911198803.1. 2021-06-08.

[14] Nielsen, R., Paul, J.S., Albrechtsen, A., *et al*. (2011) Genotype and SNP Calling from Next-Generation Sequencing Data. *Nature Reviews Genetics*, **12**, 443-451. https://doi.org/10.1038/nrg2986

[15] Wang, L., Liu, P., Huang, S., *et al*. (2017) Genome-Wide Association Study Identifies Loci Associated with Resistance to Viral Nervous Necrosis Disease in Asian Seabass. *Marine Biotechnology*, **19**, 255-265. https://doi.org/10.1007/s10126-017-9747-7

[16] 程维晟. 基于二代测序的刚地弓形虫Chinese1基因型不同毒力虫株的变异检测[D]: [硕士学位论文]. 合肥: 安徽医科大学, 2016. https://doi.org/10.7666/d.D01025793

[17] Ragoussis, J. (2009) Genotyping Technologies for Genetic Research. *Annual Review of Genomics and Human Genetics*, **10**, 177-133. https://doi.org/10.1146/annurev-genom-082908-150116

[18] 周岩, 吴慧丽, 郑莉莉, 李萍, 高翔. miR-17-92 启动子区基因多态性与子宫内膜癌易感性及预后的相关性[J]. 现代妇产科进展, 2021, 30(11): 805-809.

[19] 黎江溪, 张世梅, 王玉鑫, 等. 基于生物信息学的肥厚型心肌病易感基因 TNNC1 单核苷酸多态性致病表型分析[J]. 中国生物医学工程学报, 2022, 41(1): 114-118.

[20] Li, J. and He, D. (2017) Single Locus Maintains Large Variation of Sex Reversal in Half-Smooth Tongue Sole (*Cynoglossus semilaevis*). *G3*: *Genes*, *Genomes*, *Genetics*, **7**, 583-589. https://doi.org/10.1534/g3.116.036822

[21] Yang, N., Zhang, D.F., Tao, Z., *et al*. (2016) Identification of a Novel Class B Scavenger Receptor Homologue in *Portunus trituberculatus*: Molecular Cloning and Microbial Ligand Binding. *Fish & Shellfish Immunology*, **58**, 73-81. https://doi.org/10.1016/j.fsi.2016.09.023