水产学报
JOURNAL OF FISHERIES OF CHINA
DOI: 10.11964/jfc.20230213920
Science Press

水产群体基因组重测序数据分析软件包的开发

徐清腾[1,2]，吴昊天[1,2]，江丽华[3*]，陆　颖[1,2*]
(1. 上海海洋大学，水产种质资源发掘与利用教育部重点实验室，上海　201306；
2. 上海海洋大学，海洋生物科学国际联合研究中心，上海　201306；
3. 浙江海洋大学，国家海洋设施养殖工程技术研究中心，浙江 舟山　316022)

摘要：为了帮助水生动物学研究者解决群体遗传学基础分析的困难，本实验在调研现有的水生动物重测序数据分析研究和成果的基础上，基于水生动物群体遗传学的常用方法和通用分析软件，构建可于本地运行的、能够完成大部分基因组重测序数据基础计算的软件包。软件包首先将质控过滤后的重测序数据与参考基因组序列进行比对，利用比对结果检测基因组的遗传变异，对群体进行系统发育分析、群体结构分析、主成分分析、遗传多样性重要量化指标的计算和选择性消除分析等，并通过 R 或 Python 语言工具包对分析结果进行可视化。根据该软件包对来自 3 个群体、共约 30 尾大黄鱼个体的简化基因组测序数据进行分析测试的结果，完成了软件包携带的测序数据比对、单核苷酸多态性 (SNP) 鉴定、系统进化树构建、群体结构预测、连锁不平衡检测和多样性指标等计算功能，并且较好地图形可视化了分析结果。该群体基因组重测序分析的简易软件包可用于野生和自然群体的群体遗传学分析的大部分基础统计、计算和绘图，适合包括水生生物学在内的相关领域的生物学研究者进行群体基因组学研究。本研究为水生动物重测序数据分析提供便利，节约科研时间，减少人力物力成本。相关源码和使用说明文档已公开上传至 GitHub: https://github.com/xqteng/Re-seq_analysis。

关键词: 水生动物重测序；单核苷酸多态性；群体遗传学；软件；基因组
中图分类号: Q 785；S 917.4　　**文献标志码:** A

随着测序技术的发展，测序逐渐地成为科研的利器。尤其是第二代高通量基因组测序技术 (Next generation sequencing, NGS) 的出现，使得研究人员对物种从基因组层面进行全面且细化的分析成为可能[1]。基因组重测序是指在某物种存在已经测序或发表的参考基因组的情况下，重新对该物种的不同个体进行测序、并进行序列比对，得到测序和分析结果的过程[2]。重测序数据与参考基因组序列的比对结果，通过一系列软件工具的计算，能够发现单核苷酸多态性位点 (SNPs)、拷贝数变异 (CNV)、插入/缺失 (indel) 等序列变异类型[3]。将单个参考基因组信息映射到整个生物群体的遗传特征上，是研究野生群体或养殖群体内部的遗传多样性、群体之间的分化程度，以及发现物种染色体上受人工或环境选择痕迹的主要手段[4]。近年来，各种重测序数据分析的流程和相应工具已经日渐成熟，具体过程：从测序公司拿到质控数据 (clean data) 之后，进行 Illumina 短序列 (short reads) 的序列比对，然后在标记重复序列之后进行遗传变异检测，在这过程中进行 SNP 位

收稿日期：2023-02-23　　**修回日期：**2023-04-05
资助项目：国家重点研发计划 (2022YFD2400804，2018YFD0900101)
第一作者：徐清腾 (照片)，从事生物信息学研究，E-mail：xqteng123@163.com
通信作者：江丽华，从事海洋生物生理生态学研究，E-mail：florencejiang@126.com；
陆颖，从事动植物基因组学与遗传育种研究，E-mail：yinglu@shou.edu.cn

点检测及结果的注释[5]；遗传变异检测之后可以得到群体变异集文件；利用这种格式的文件可以进行系统发育分析、主成分分析和群体结构分析，从而了解群体结构[6]；对各群体进行连锁不平衡检测、遗传多样性计算和选择消除分析，探究群体的遗传多样性情况。就比较水生野生动物群体和养殖群体而言，二者在食物结构、生存环境以及活动状况等方面存在差异[7]，后者因受到驯化导致相关基因区域的遗传多样性呈现出显著降低的特征，因此可以利用选择消除分析的方法鉴定这些群体的染色体上受选择的区域[8]。例如，吴昊天等[9]对同一种源的新疆和西藏棕鳟 (*Salmo trutta*) 群体基于群体分化指数 (F_{ST}) 和核苷酸多态性 (θ_{π}) 计算的选择消除分析发现西藏棕鳟免疫相关基因受到的环境选择压力更加明显。此外，Song 等[10]对俄罗斯鲟 (*Acipenser gueldenstaedtii*) 不同繁殖能力的群体计算 F_{ST} 结合跨群体复合似然比 (XP-CLR) 鉴定出鱼子酱产量相关的候选基因。鉴于大多数生物学研究者不具备必要的生物信息学技能，这些工作往往只能求助于商业公司或商业平台来进行。这样不仅费用较高、分析周期长，而且在商业公司的分析过程中一般使用的都是默认参数分析，他们不会从群体和物种的特点出发，制定个性化的分析流程和参数调整，经常发生因使用不合理的计算模型而导致的计算错误。当分析结果不合理而需要重新计算时，研究者还要支付额外的费用。随着群体基因组学研究在水产动物分子育种中的使用日益广泛，迫切需要一个能进行基础计算的本地化的工具来进行重测序数据的基本处理。因此，实验构建了一个免费开放的流程化软件包，用于处理群体基因组重测序数据的基础计算，提供更友好的操作方式和更灵活的参数设定。研究者只需要根据说明文档，通过命令行输入，并按需要改变相应参数进行计算即可。这样，大多数缺乏生物信息学技能的研究者也能够顺利完成群体基因组重测序数据的基础分析。

1　材料与方法

1.1　软件包开发环境与工具

本软件包是在 Linux 环境下使用 Python、Perl、R 语言, 利用 IDEA 工具开发的一套重测序数据分析软件包。软件包调用的软件及版本如表 1 所示。

1.2　软件包功能

经过调研，我们总结了重测序数据的分析需求，参考已发表的重测序数据分析方法，设计了一套基因组重测序数据处理分析流程，软件包采用模块化设计，主要有序列比对、遗传变异检测、变异过滤、SNP 信息统计、系统发育分析、主成分分析、群体结构分析、遗传多样性测度、连锁不平衡检测、选择消除分析 10 个功能模块 (图 1)。其主要功能模块包括：

表 1　软件最低版本和环境要求

Tab. 1　Minimum version and environment requirements of the softwares

软件/工具 software / tools	最低版本/要求 minimum version / requirements
操作系统 OS	Linux
编译环境 compilation environment	Rscript (3.6.0), perl (v5.16.3), Python (3.8.5)
Burrows-Wheeler Alignment Tool (BWA)	0.7.16a-r1181
Genome Analysis Toolkit (GATK)	v4.3.0.0
Samtools	1.6 (using htslib 1.6)
VCFtools	v0.1.13
PLINK	v1.90b6.17
ADMIXTURE	1.3.0
RAxML	8.2.12
PopLDDecay	v3.40
xpclr (Python)	1.1.2

原始数据的比对　　这个模块的主要功能是将低差异度的短序列，通过比对软件与参考基因组进行比对[11]。BWA (Burrows-Wheeler Alignment Tool) 是将测序序列和参考基因组进行快速比对的开源软件[12]，在二代或者三代测序中运用的相对比较多，尤其是其 BWA-MEM 算法对 Illumina 的数据效果更好[13]，因此序列比对工具使用的是 BWA。序列比对时需要准备合理命名的 clean data (fq.gz 格式) 和相对应物种的参考基因组文件 (.fna 格式)，软件包会首先检查数据是否成对存在，检查无误后对参考基因组构建索引 (BWA-index 功能)，将 clean data 中的序列同索引进行比对，得到 SAM (sequence alignment/map) 格式的序列比对文件[14]。

遗传变异的检测　　该模块的主要功能是对全基因组中的变异位点进行扫描，以测序质量值 (quality scores)、测序深度和重复性序列等影响因素作为参考指标进行筛选和过滤，最终生成 GVCF 格式的 SNP 位点数据集[15]。遗传变异检测所采用的工具是基因组分析工具包 (Genome Ana-

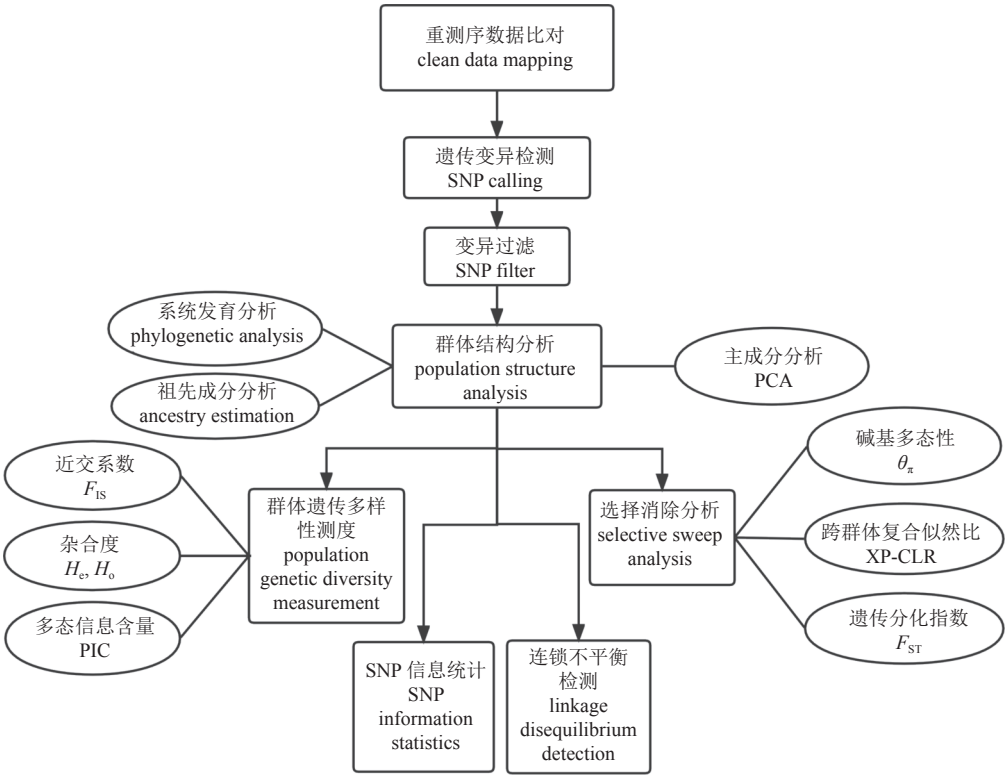

图 1　数据分析流程及模块

Fig. 1　Modules and processes of data analysis

lysis Toolkit, GATK)，它主要用于从测序数据中处理变异信息的系统规范性软件[16]。Samtools 是主要用于处理 SAM/BAM(SAM 的二进制格式，用于压缩空间) 格式的比对文件的工具集[17]，它能够输入和输出 SAM 格式文件，对其进行排序、合并、建立索引等处理[18]。在这一过程中软件包使用 Samtools 软件对上一步的结果文件 (SAM 格式) 进行排序，然后调用 GATK MarkDuplicates 将测序时因 PCR 扩增导致的测序序列重复进行去除，以降低后续变异检测中的假阳性。在调用 Samtools 建立索引后，通过 GATK HaplotypeCaller 进行变异检测，得到记录有全部 SNP 信息的 VCF 文件。

遗传变异计算结果的过滤　在这个模块中，将检测到个体间基因组序列的变异储存至记录标准变异内容的 VCF 文件中，其信息包括变异位点的突变类型 (genotype)、基因型的质量值 (genotype quality scores)、reads 的覆盖深度等。在此基础上，根据突变质量值、基因型及群体遗传学参数等多项指标进行筛选，获得最终的变异位点信息用于后续分析。在变异过滤模块的主要处理过程中，首先使用 GATK SelectVariants 从 VCF 文件中提取 SNP，得到只包含 SNP 位点信息 VCF 文件，然后进行两次 SNP 过滤：根据测序深度、变异质量值等因素对 SNP 进行第一次过滤；考虑最大缺失率 (max-missing) 和次等位基因频率 (minor allele frequency, MAF) 等信息进行第二次过滤。变异过滤功能模块调用的 GATK 工具既能进行 SNP 检测，还能通过各项参数对变异信息进行质量检测和过滤，因此第一次过滤部分采用 GATK 工具，而 VCFtools 在拥有转换 VCF 文件格式的功能的同时还可以根据 MAF、max-missing 等信息对变异信息进行第二次过滤[19]。软件包中，变异过滤过程的参数可以在命令行界面按需更改。变异过滤结束后，得到符合要求的 SNP 变异信息文件 (VCF 格式)。

SNP 信息的统计　这一模块使用 R 语言工具包 CMplot[20] 统计 SNP 在每条染色体上的分布，即 SNP 分布密度统计。首先，将记录 SNP 位点的变异集文件 (VCF 格式) 和用户提前准备的采样群体信息文件 (sampleID 文件，一列含有采样个体信息的文本文件) 进行整合，得到各采样群体的 SNP 变异集文件 (VCF 格式)；然后，通过命令行界面设置参数——窗口大小 (window size) 和参数图例颜色最大值 (当 SNP 数量大于 bin.max 时，将

使用相同的颜色)，绘制 SNP 密度图 (png/pdf 格式)。

个体的系统发育分析　系统发育分析模块利用鉴定得到的 SNP 信息，构建所有个体的系统进化树，用于描述群体的演化过程。模块调用 RAxML[21] 软件进行系统发育树分析、构建进化树。期间通过改变 RAxML 软件的 bootstrap value 可以提高各分支的可信度[22]。简单说来，使用 Perl 工具将 SNP 变异信息 (VCF 格式) 挑选并制作成 RAxML 软件能够识别的格式 (.seq)，然后设置输入参数 bootstrap values，利用最大似然法构建最佳的系统发育树 (.bestTree.nwk 格式)。

群体结构的分析　这一模块利用遗传标记推断群体遗传结构，从而观察不同来源的各个群体之间所包含的亚群或共同祖先的个数以及相似率。一般来说，进行群体结构分析所使用的软件为 STRUCTURE，但是 STRUCTURE 的运行速度较慢，如今 Admixture 等[23] 凭借其高速的运算速度逐渐成为群体遗传结构分析的主流软件。本模块优先选用 Admixture 软件，它的输入格式由开源的全基因组关联分析工具集 Plink 计算变异信息文件 (VCF 格式) 生成，包括 Plink(.bed) 和 Plink(.ped) 文件，同时 Plink 还可以帮助完成后续分析数据的可视化、VCF 文件的注释以及分析结果的暂时存储[24]。然后通过输入预设参数亚群数 (K values) 计算 CV errors (cross-validation error) 来挑选合适的 K 值，最后使用 R 脚本绘制 Admixture 结果图，用来展示群体遗传结构和祖先背景。

主成分分析　主成分分析 (PCA) 模块可以根据变异位点识别并调整样本的群体分层[25]。软件包会自动识别 SNP 信息统计功能模块所分离出的各群体 SNP 的 VCF 文件，并生成以.map,.nosex 和.ped 为后缀的 Plink 格式的文件；然后继续调用 Plink 计算 PCA。这一步会得到 2 个结果文件，分别是以.eigenval 和.eigenvec 为后缀的结果文件，前者记录每个 PC 所占的比重作为特征值，后者记录特征向量作为坐标轴；最后使用 R 工具包 scatterplot3d[26] 绘制 PCA 图，将分析结果可视化。

连锁不平衡的检测　连锁不平衡 (linkage disequilibrium，LD) 是指不同基因位点 (loci) 的等位基因 (allele) 之间非随机 (nonrandom) 的关联[27]，通过 LD 衰减 (LD decay) 在不同物种间或同物种的不同亚群间的衰减速度快慢来判断基因位点间由连锁不平衡到连锁平衡的演化过程，能可靠可信地描述群体遗传过程[28]。这一模块调用软件 PopLDdecay，该软件基于 VCF 格式文件，能够以较快的速度和更高的效率检测群体间的连锁不平衡情况[29]，通过提供含有变异信息的 VCF 文件和群体信息文件 (sampleID)，无需设置参数，计算 VCF 文件中的各群体的 LD 衰减，得到结果文件 (LDdecay.stat 格式)，并使用 Perl 脚本绘制 LD decay 图。

群体遗传多样性重要指标参数的计算　衡量群体多样性的主要指标性参数包括期望杂合度 (expected heterozygosity，H_e)[30]、观测杂合度 (observed heterozygosity，H_o)[31]、近交系数 (Wright's inbreeding coefficient，F_{IS})[32] 和多态信息含量 (polymorphism information content，PIC)[33]。其中，期望杂合度和观测杂合度可以用来衡量群体的遗传多样性的高低[34]；近交系数是种群内个体之间的近亲交配程度指标[35]；多态信息含量是指通过检测一个标记的等位基因数和它们的分布频率，得到该标记在一个群体中的多态性大小值，从而反映基因的多样性程度[36]。软件包会自动识别所有群体和各群体的变异信息文件 (VCF 格式)，通过调用 Plink 软件和 Python 工具进行计算和统计，得到 H_o 和 H_e 结果 (.hwe 格式)、F_{IS} 结果 (.het 格式)、PIC 结果 (.pic 格式)，最后汇总各项信息。

群体间的选择消除分析　选择消除 (selective sweep) 指由于生存环境、食物结构和活动状况等因素，该物种的某一相关区染色体域会因自然或人工选择而出现遗传多样性降低的现象，从而其所在群体的基因区域中出现高频的或低频的等位基因[37]。在选择消除分析的过程中，F_{ST} 结合 θ_π，能有效检测群体间的分化程度，从而鉴定选择消除区域的方法，结合群体等位基因频率谱分析 (XP-CLR 法)[38]，帮助鉴定物种适应性进化或人工驯化过程中受选择的染色体区域。其中 F_{ST} 表示群体的分化程度，其取值范围为 0~1，数值越大，分化程度越高[39]，相关染色体区域受到选择就越强；θ_π 反映了群体基因组碱基多样性水平，θ_π 越低，该群体的核苷酸多态性就越低，相关染色体区域受选择程度也就越高[40]。跨群体复合似然比法 (XP-CLR) 是一种利用群体间等位基因频率的差异，对覆盖扩展连锁区域末端的等位基因变异频率进行扫描，从而有效定位受正选择的位点的方法。选择消除分析模块计算这些相关数据时，在命令行界面输入参数窗口 (window size) 和步长 (window step) 之后，软件包会自动识别含有变异

信息的 VCF 文件和群体信息文件 (sampleID)，调用软件 VCFtools 计算 F_{ST} 和 θ_{π}，生成 F_{ST} 和 θ_{π} 的结果文件 (文本文件)，使用 Perl 以及 R 工具包 qqman 合并 F_{ST} 和 θ_{π} 结果汇集成 F_{ST}-θ_{π} 表并将其可视化；专用于计算 XP-CLR 的软件 xpclr (Python 版) 由于是按染色体计算，所以需要输入参数染色体数 (CHR)，软件包会计算并整合生成结果文件 (xpclr 格式)，并可视化 XP-CLR 的计算结果。

此外，在使用基因组重测序分析软件包时，使用者需在同一文件目录中准备好相应文件，在命令行执行指令设置参数，软件包便会依次执行各个功能模块。在软件包调用程序时，上一步骤的处理完成的结果文件可能会作为下一步骤的输入数据，因此本软件包会对各个计算模块的输入输出文件格式进行适应化处理 (图 2)。

2 结果

为了测试软件包的各模块能否正常工作，实验采用东海养殖 (YZ)、东海野生 (DH)、湛江野生 (ZJ) 的 3 个群体共 30 尾大黄鱼 (*Larimichthys crocea*) 个体的简化基因组测序数据对软件进行功能测试。30 个个体的 30 对 clean data fastq 数据量共约 101.46 Gb。大黄鱼参考基因组文件：Large-YellowCroaker_genomic_chr.fna (GenBank 序列登录号：GCA_000972845.2)。

这 30 对 clean data 测序数据与参考基因组进行比对和遗传变异检测，软件包采用双向队列，每次对 10 对 reads 并行处理的方式，得到结果文件 all_raw.vcf 共含有 3 820 531 个 SNP(表 2)。其中第一次过滤参数采用默认参数：QD>2.0、MQ>40.0、MQRankSum>−12.5、ReadPosRankSum>−8.0；第二次过滤参数则通过 Linux 命令行界面输入：–MM (max missing): 0.2, –MF(maf): 0.05，变异过滤结果文件 all_snp.vcf 共保留 948 845 个 SNP 位点。根据各采样信息文件 (YZ.sampleID, DH.sampleID, ZJ.sampleID) 分离 SNP 得到 YZ.vcf 含有 780 356 个 SNP 位点，DH.vcf 含有 840 223 个 SNP 位点，ZJ.vcf 含有 569 310 个 SNP 位点 (表 2)。

在运行软件包之前，Linux 命令行输入 SNP 信息统计模块参数：-s(窗口大小 (window size)) : 0.1Mb，-M(图例最大值 (the max value of legend of SNP_density plot))：500。软件包利用 R 脚本计算变异集文件 (all_snp.vcf) 后得到结果如图 3 所示。通过改变窗口大小和图例颜色的最大值便能绘制符合不同需求的 SNP 密度图。

输入系统发育分析模块参数：-N (bootstrap): 1 000，软件包使用变异集文件 (all_snp.vcf) 进行两次格式转换 (.seq 和.tbl 格式) 完处理后得到结果文件 (bestTree.nwk 格式)，借助外部工具 FigTree 绘图以展示结果 (图 4)。在 bootstrap 参数选择 1 000 次自举法检验时，用时 3.4 h。

输入群体结构分析模块参数：-K (预设亚群数量) 1，2，3，4，5，6，7 (此处参数格式是以逗号分隔)，软件包使用 Plink 格式文件计算出 CV errors 最小值对应的 K 值为 1 (图 5-a，根据输出文件 (.plink.Q 格式) 绘制群体结构图 (图 5-b)。

主成分分析模块无需参数，软件包自动识别过滤后的所有 SNP 位点的 VCF 文件 (all_snp.vcf)，调用 Plink 软件进行三步处理 (recode、make-bed、pca 三次格式转换) 得到 PCA 结果文件 (.eigenvec 和.eigenval 格式)，最后使用 R 脚本绘制两种 PCA 分析结果图 (无视采样信息和结合 sampleID) (图 6)。

连锁不平衡检测无需参数，软件包自动识别群体变异集文件 (all_snp.vcf)、3 个采样居群 (东海养殖、东海野生、湛江野生) 的样品 ID 文件 (YZ.sampleID, DH.sampleID, ZJ.sampleID)，自动生成 PopLDdecay 需要的 draw.list 文件，调用 PopLDdecay 计算 LD 衰减，并使用 Perl 脚本将结果文件 (stat.gz) 绘制 LD decay 图 (图 7)。

群体遗传多样性的测度同样无需设置参数，软件包自动识别所有群体 VCF 文件 (all_snp.vcf) 和各群体变异信息的 VCF 文件 (YZ.vcf，DH.vcf，ZJ.vcf)，通过 Plink 软件并使用 Python 工具得到群体遗传多样性结果，使用外部工具可以将结果可视化 (图 8)。

选择消除分析模块需要在命令行界面输入参数-WP (window size)：50 000，-WPS (window step)：2 000，-CHR(染色体数目)：01，02，03，…，23，24 (此处参数格式是以逗号分隔)，软件包自动识别含有全部变异信息的 VCF 文件 (all_snp.vcf) 和群体信息文件 (YZ.SampleID, DH.SampleID, ZJ.SampleID)，调用软件 VCFtools 计算 F_{ST} 和 θ_{π}，生成 F_{ST} 和 θ_{π} 的结果文件 (.pi 和.*fst* 格式)，使用 Perl、R 脚本合并 F_{ST} 和 θ_{π} 结果汇集成 F_{ST}-θ_{π} 表 (YZ_vs_DH.*fst*_pi,YZ_vs_ZJ.*fst*_pi,DH_vs_ZJ.*fst*_pi) 并将其可视化 (图 9)；软件包调用软件 xpclr(Python 版) 按照

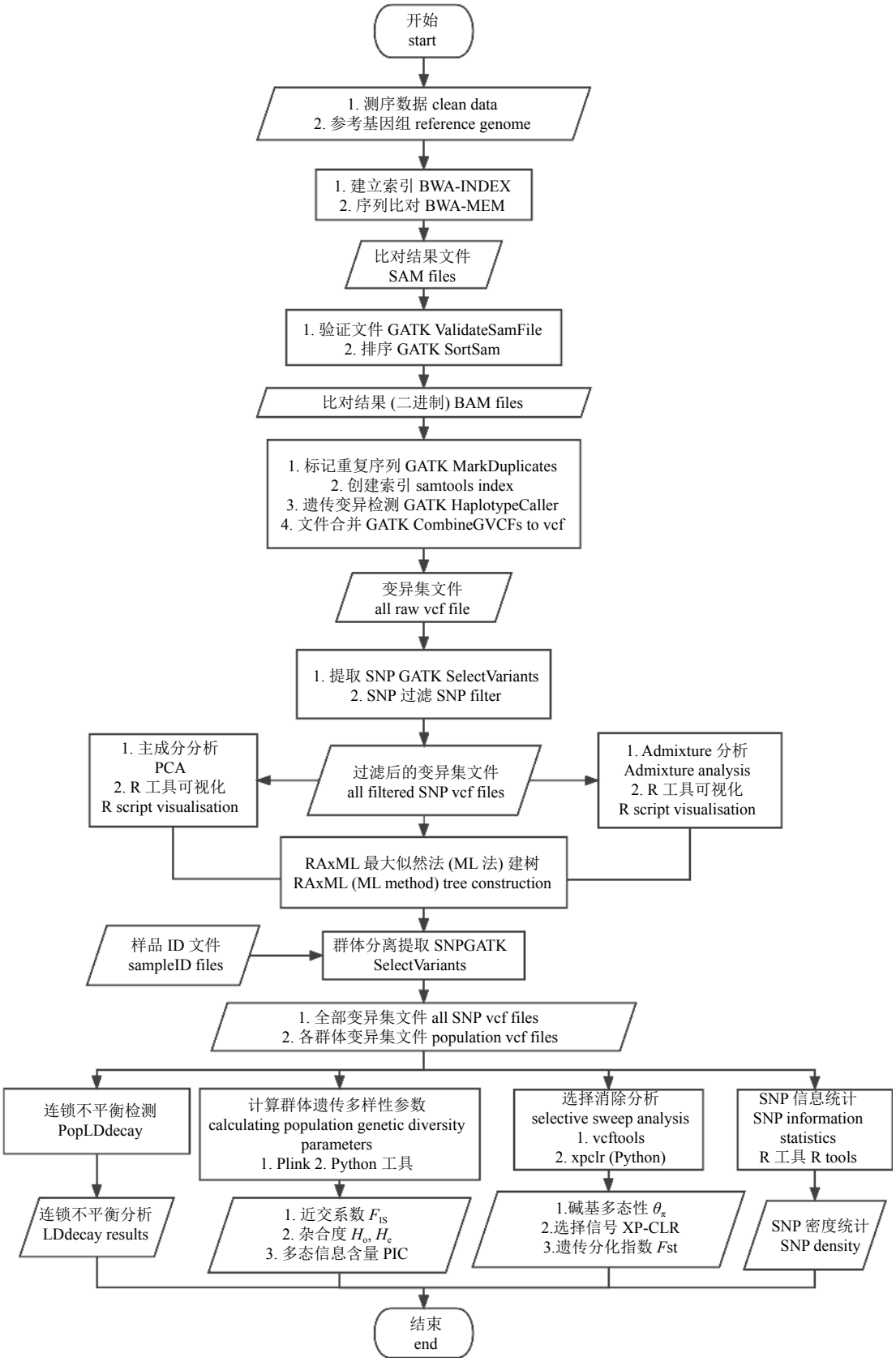

图 2　软件包运行流程图

Fig. 2　Flowchart of platform operation

表 2　群体样本及测序信息

Tab. 2　Information of collected samples and the sequencing

群体 population	采样地点 sampling site	个体数量/尾 sample no.	平均测序深度/x ave. seq. depth	个体平均SNP数量/个 ave. SNP no.
东海养殖 YZ	东海养殖场 Donghai aquaculture farm	11	43.6	780 356
东海野生 DH	东海海域 Donghai wild seas	12	68.3	840 223
湛江野生 ZJ	湛江附近海域 sea off Zhanjiang	7	13.5	569 310
所有样品 all samples		30	41.8	948 845

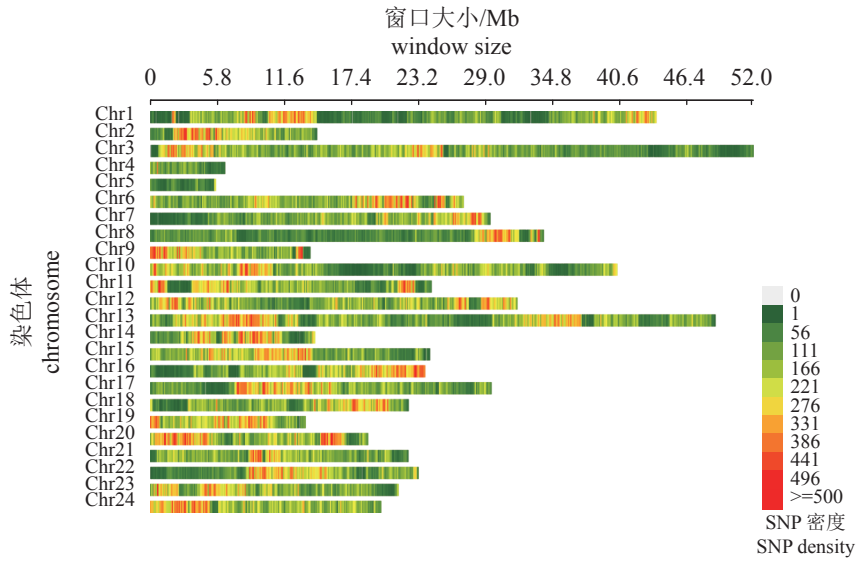

图 3　SNP 密度图 (窗口大小：0.1 Mb)

Fig. 3　SNP density (window size 0.1 Mb)

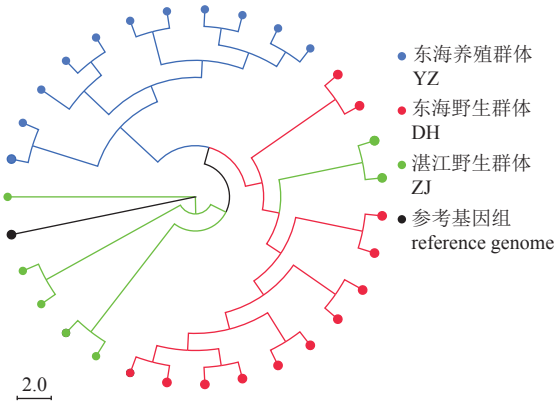

图 4　系统进化树 (Bootstrap = 1 000)

Fig. 4　Phylogenetic tree
(Bootstrap = 1 000)

输入的染色体参数，计算两两群体的 XP-CLR，同时对每条染色体的结果进行整合生成最终结果文件 (YZvsDH.xpclr, YZvsZJ.xpclr, DHvsZJ.xpclr)，最后软件包会使用自编 R 脚本将 XP-CLR 结果可视化 (图 10)。

3　讨论

基因组重测序分析软件包依托超算服务器和 Linux 操作系统，在重测序数据分析的过程中调用多个规范性软件，对软件之间的输入输出文件格式进行约束，整个过程展现了较好的流畅度和可靠的稳定性。使用过程中，在命令行修改参数重新运行本软件包即可再次进行数据分析，体现了软件包的可重复利用性。软件包对于分析数据的输出绝大多数使用了可视化工具进行绘图，分析结果具有良好的可读性。该软件包也存在一些不足，例如，因意外程序终止，软件包不能继续从中断处运行只能重新运行；部分结果的可视化只能借助系统之外的外部工具。各项结果表明，该软件包能够帮助大多数缺乏生物信息学技能的使用者通过简单的操作完成群体基因组重测序数据的基础分析。在后续工作中，软件包还将加入基因注释等更多功能，不断提高数据分析的可靠性和分析结果的可读性，更好地服务于基因组重测序数据分析。该软件包的相关源码和详细安装、

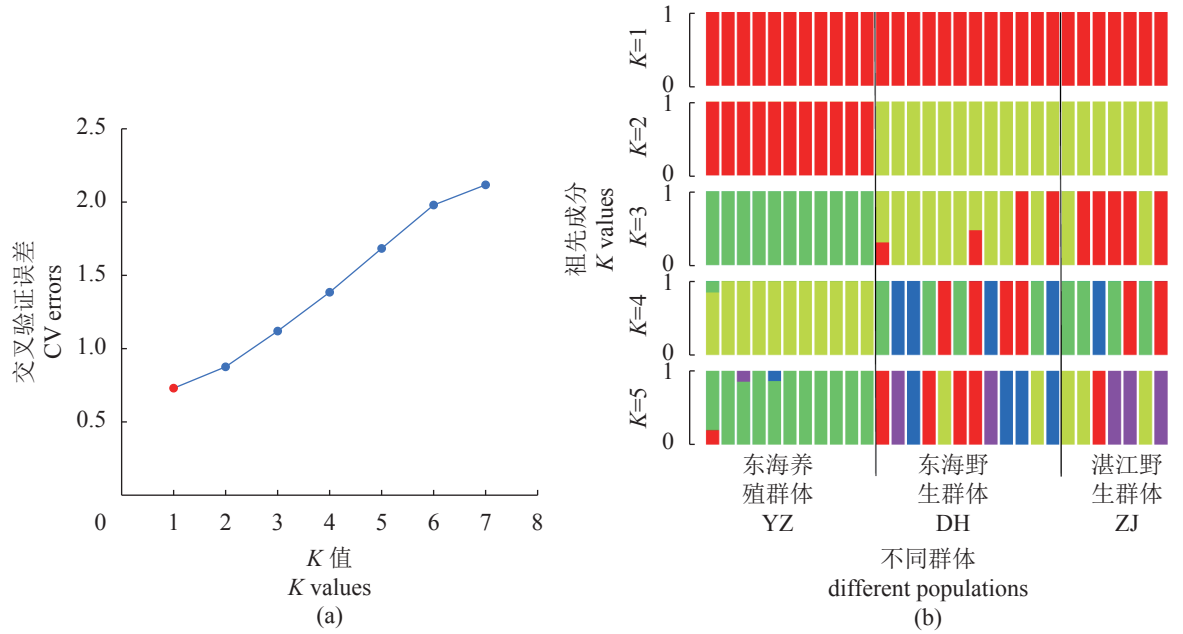

图 5　群体的结构分析

Fig. 5　Admixture analysis of the populations

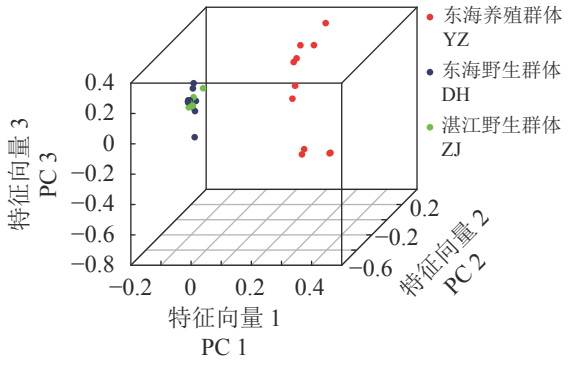

图 6　群体 PCA 分析

Fig. 6　PCA analysis

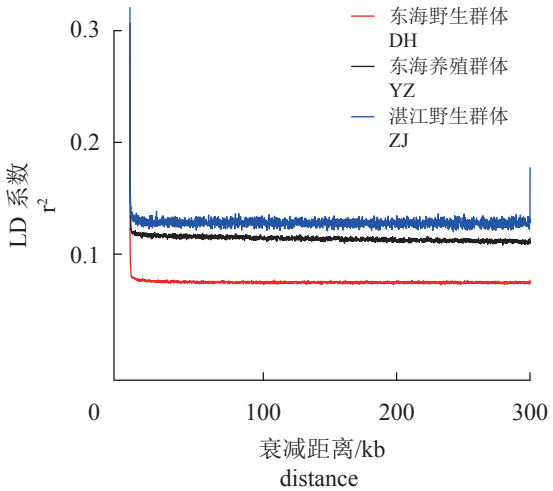

图 7　连锁不平衡检测

Fig. 7　LD decay analysis

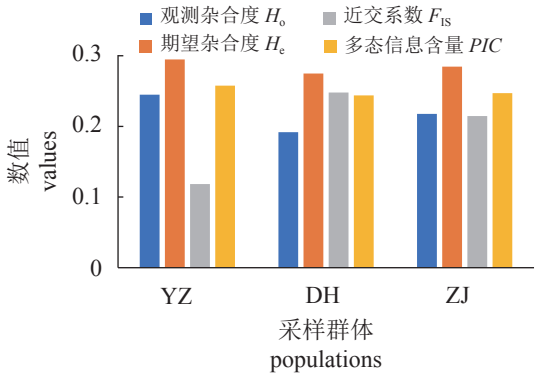

图 8　各群体遗传多样性分析

Fig. 8　Diversity of the populations

使用说明文档已公开上传至 GitHub: https://github.com/xqteng/Re-seq_analysis，同时我们指派了专员定期更新、维护此软件包，使用者在使用过程中有疑问和报错可以随时通过网站留言板或网站邮箱联系相关人员，后续我们会根据内嵌程序的更新换代情况定期同步至该软件包，并在第一时间上传到同一网址，用户重新下载便可以继续使用，从而确保其正常运行。

4　结论

基因组重测序软件包使得重测序的数据分析完全流程化，使生物学研究者能够通过命令行端改变输入参数，进行多次运行，选择最佳的结果。这个完全开放的软件包整合了测序数据比对、遗

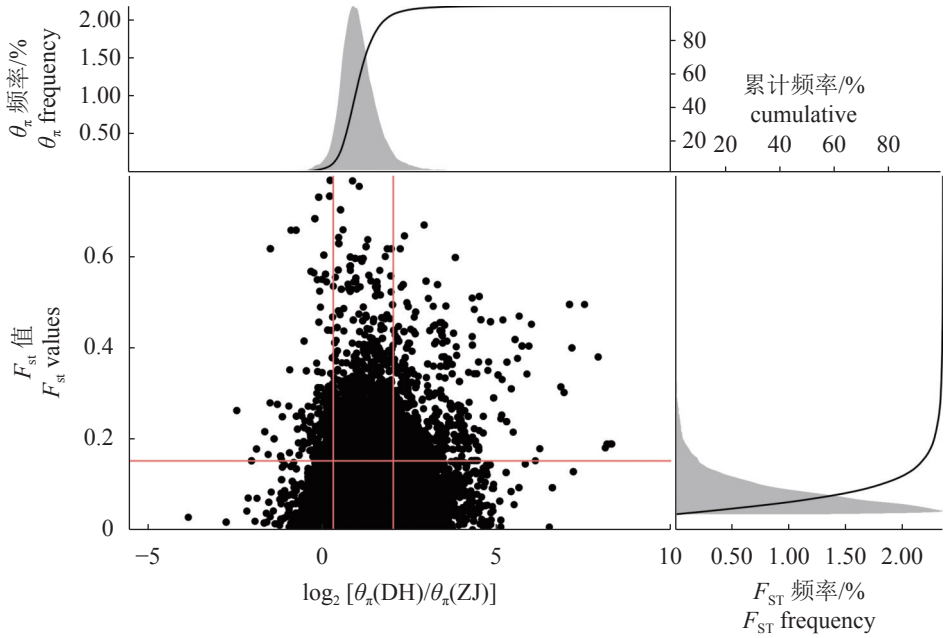

图 9　两两群体 F_{st}-θ_π 的联合图

图中是东海养殖群体 *vs*.湛江野生群体的 F_{st}-θ_π 的联合图。每张图中的右上角红框区域代表 θ_π 比值 (两两群体的 θ_π 之比) 最大的前 1% 与 F_{st} (与 θ_π 比值对应的两群体) 最大的前 1% 的交集，而左上角红框区域代表 θ_π 比值最小的前 1% 区域与 F_{st} 最大的前 1% 区域交集。

Fig. 9　Joint plot of F_{st}-θ_π for two populations

The figure shows the Joint plot of F_{st}-θ_π for East China Sea farmed population (YZ) *vs*. the Zhanjiang wild population(ZJ). The red boxed area on the top right in each graph represents the intersection of the top 1% with the largest θ_π ratio (the ratio of θ_π between the two populations) with the top 1% with the largest F_{st} (the two populations corresponding to the θ_π ratio), while the red boxed area on the top left represents the intersection of the top 1% area with the smallest θ_π ratio with the top 1% area with the largest F_{st}.

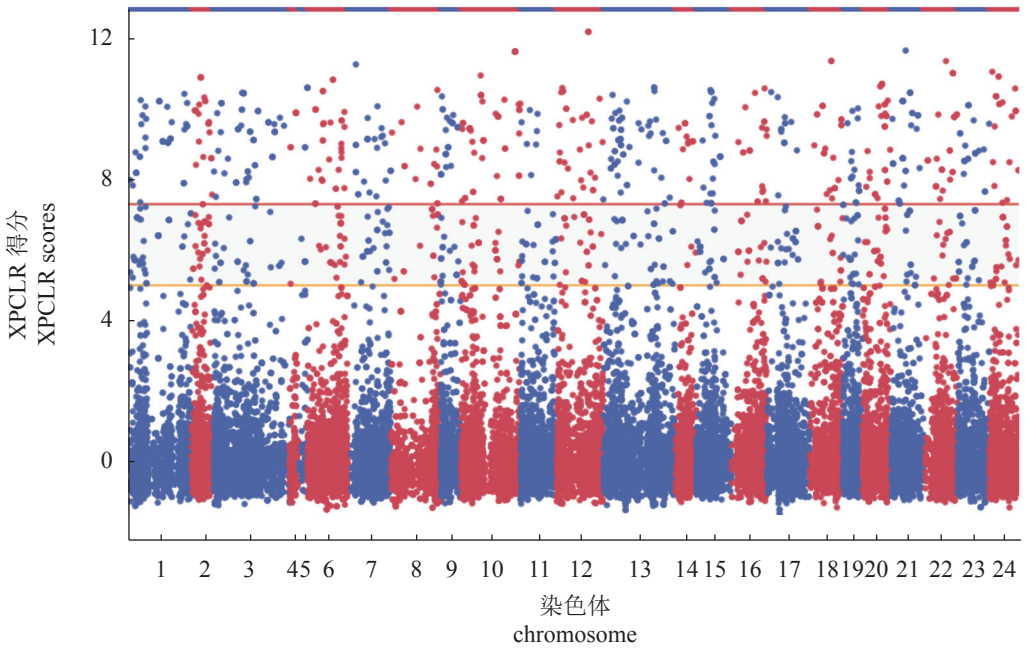

图 10　XP-CLR 得分图

图中黄线表示 XP-CLR < −log(1e-05) 为临界显著性，红线表示 XP-CLR < −log(5e-08) 为高可信显著性。

Fig. 10　XP-CLR scores

The yellow line indicates the critical significance [XP-CLR < −log(1e-05)], the red line indicates the highly credible significance [XP-CLR < −log(5e-08)].

传变异检测、系统进化分析、群体结构分析、主成分分析、遗传多样性分析和选择消除分析等大部分基因组重测序数据分析的基础功能，并将分析结果以绘图展示，适合大部分没有或只有很少生物信息学基础的生物学研究者使用。

（作者声明本文无实际或潜在的利益冲突）

参考文献 (References):

[1] 刘毅, 余新桥, 张安宁, 等. 高通量基因组测序在农作物基因定位与发掘中的应用[J]. 上海农业学报, 2016, 32(6): 171-175.
Liu Y, Yu X Q, Zhang A N, *et al*. Application of high-throughput genome sequencing in crop gene discovery and mapping[J]. Acta Agriculturae Shanghai, 2016, 32(6): 171-175 (in Chinese).

[2] 汪文强, 赵生国, 马利青, 等. 动物基因组学重测序的应用研究进展[J]. 畜牧兽医学报, 2016, 47(10): 1947-1953.
Wang W Q, Zhao S G, Ma L Q, *et al*. The research progress and application of resequencing based on animal genomics[J]. Acta Veterinaria et Zootechnica Sinica, 2016, 47(10): 1947-1953 (in Chinese).

[3] 韩彦卿, 郑洁, 武彩娟, 等. 基于重测序的晋汾52及其突变体抗性差异分析[J]. 山西农业大学学报 (自然科学版), 2018, 38(7): 39-44.
Han Y Q, Zheng J, Wu C J, *et al*. Variation analysis of disease resistance of Jinfen 52 and its mutants by genome resequencing[J]. Journal of Shanxi Agricultural University (Natural Science Edition), 2018, 38(7): 39-44 (in Chinese).

[4] 刘恩民. 中国地方山羊品种群体遗传结构和选择信号分析 [D]. 兰州: 兰州大学, 2018.
Liu E M. Analysis of population genetic structure and select signal in Chinese indigenous goat breeds[D]. Lanzhou: Lanzhou University, 2018 (in Chinese).

[5] 李文轲, 李丰余, 张思瑶, 等. 基因组二代测序数据的自动化分析流程[J]. 遗传, 2014, 36(6): 618-624.
Li W K, Li F Y, Zhang S Y, *et al*. Automatic analysis pipeline of next-generation sequencing data[J]. Hereditas, 2014, 36(6): 618-624 (in Chinese).

[6] 王统苗, 郭其新, 白皓, 等. 基于全基因组重测序对不同鸭遗传资源进行群体结构分析[J]. 中国畜牧杂志, 2021, 57(11): 78-81.
Wang T M, Guo Q X, Bai H, *et al*. Analysis of population structure of different duck genetic resources based on whole-genome resequencing[J]. Chinese Journal of Animal Science, 2021, 57(11): 78-81 (in Chinese).

[7] 王映. 大黄鱼野生群体和养殖群体鉴定技术研究 [D]. 上海: 上海海洋大学, 2017.
Wang Y. Germplasm identification of wild and cultured stock of large yellow croaker (*Larimichthys crocea*)[D]. Shanghai: Shanghai Ocean University, 2017 (in Chinese).

[8] 冯韶华, 李尚, 刘刁, 等. 基于重测序的深县猪群体结构和选择信号研究[J]. 黑龙江畜牧兽医, 2021(16): 35-37,42,148.
Feng S H, Li S, Liu D, *et al*. Study of population structure and selection signal of Shenxian pigs based on genome resequencing[J]. Heilongjiang Animal Science and Veterinary Medicine, 2021(16): 35-37,42,148 (in Chinese).

[9] 吴昊天, 李亚琳, 王军, 等. 新疆和西藏棕鳟群体的遗传多样性分析[J]. 南方农业学报, 2022, 53(5): 1425-1433.
Wu H T, Li Y L, Wang J, *et al*. Genetic diversity analysis on brown trout populations in Xinjiang and Tibet[J]. Journal of Southern Agriculture, 2022, 53(5): 1425-1433 (in Chinese).

[10] Song H L, Zhu B Y, Dong T, *et al*. Whole-genome resequencing reveals selection signatures for caviar yield in Russian sturgeon (*Acipenser gueldenstaedtii*)[J]. Aquaculture, 2023, 568: 739312.

[11] 高静, 焦雅, 张文广. 高通量测序序列比对研究综述[J]. 生命科学研究, 2014, 18(5): 458-464.
Gao J, Jiao Y, Zhang W G. Overview of sequence alignment for high-throughput sequencing data[J]. Life Science Research, 2014, 18(5): 458-464 (in Chinese).

[12] 林晶晶. 基于 Hadoop 的人类全基因组重测序数据处理技术的研究及实现 [D]. 大庆: 东北石油大学, 2015.
Lin J J. Research and implementation on the human whole-genome sequencing data processing technology based on hadoop[D]. Daqing: Northeast Petroleum University, 2015 (in Chinese).

[13] 王文雅, 庞尔丽. 利用人类全基因组二代测序数据比较BWA-MEM和NovoAlign[J]. 北京师范大学学报(自然科学版), 2021, 57(3): 337-344.
Wang W Y, Pang E L. Comparison of BWA-MEM and

NovoAlign using human whole-genome next-generation sequencing reads[J]. Journal of Beijing Normal University (Natural Science), 2021, 57(3): 337-344 (in Chinese).

[14] 孙继鹏. 高通量测序数据集的短读序列比对算法研究[D]. 西安: 西安电子科技大学, 2019.
Sun J P. The study on read alignment algorithm for high-throughput sequencing datasets[D]. Xi'an: Xidian University, 2019 (in Chinese).

[15] 连帅彬, 郭东亮, 戴宪华. 基因结构变异检测方法综述[J]. 现代生物医学进展, 2012, 12(18): 3577-3580,3588.
Lian S B, Guo D L, Dai X H. Review on algorithms of detecting genome structural variants[J]. Progress in Modern Biomedicine, 2012, 12(18): 3577-3580,3588 (in Chinese).

[16] 李泽锋, 卢鹏, 张剑锋, 等. 群体遗传变异鉴定工具系统比较[J]. 烟草科技, 2018, 51(1): 8-14.
Li Z F, Lu P, Zhang J F, *et al*. Systematic comparison of population genetic variation calling tools[J]. Tobacco Science & Technology, 2018, 51(1): 8-14 (in Chinese).

[17] 张雨豪. 基于 Split Mapping 的结构变异检测方法的模拟环境的研究和开发 [D]. 哈尔滨工业大学, 2013.
Zhang Y H. Research and development of thesimulated platform for detectinggenome variation based on split mapping[D]. Harbin: Harbin Institute of Technology, 2013 (in Chinese).

[18] 李坦. 玉米高通量测序数据 SNP 检测流程的优化及应用 [D]. 南京: 南京农业大学, 2015.
Li T. The pipeline for SNP calling based on maizehigh-throughput sequencing data and its application[D]. Nanjing: Nanjing Agricultural University, 2015 (in Chinese).

[19] Danecek P, Auton A, Abecasis G, *et al*. The variant call format and VCFtools[J]. Bioinformatics, 2011, 27(15): 2156-2158.

[20] Yin L L, Zhang H H, Tang Z S, *et al*. rMVP: a memory-efficient, visualization-enhanced, and parallel-accelerated tool for genome-wide association study[J]. Genomics, Proteomics & Bioinformatics, 2021, 19(4): 619-628.

[21] Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies[J]. Bioinformatics, 2014, 30(9): 1312-1313.

[22] Soltis P S, Soltis D E. Applying the bootstrap in phylogeny reconstruction[J]. Statistical Science, 2003, 18(2): 256-267.

[23] Alexander D H, Lange K. Enhancements to the ADMIXTURE algorithm for individual ancestry estimation[J]. BMC Bioinformatics, 2011, 12(1): 246.

[24] Purcell S, Neale B, Todd-Brown K, *et al*. PLINK: a tool set for whole-genome association and population-based linkage analyses[J]. The American Journal of Human Genetics, 2007, 81(3): 559-575.

[25] Price A L, Patterson N J, Plenge R M, *et al*. Principal components analysis corrects for stratification in genome-wide association studies[J]. Nature Genetics, 2006, 38(8): 904-909.

[26] Ligges U, Maechler M. Scatterplot3d-an R package for Visualizing Multivariate Data[J]. Journal of Statistical Software, 2003, 8(11): 1-20.

[27] 游光霞, 张学勇. 基于选择牵连效应的标记/性状关联分析方法简介[J]. 遗传, 2007, 29(7): 881-888.
You G X, Zhang X Y. Identification of important genes by marker-trait association analysis based on hitchhiking mapping[J]. Hereditas, 2007, 29(7): 881-888 (in Chinese).

[28] 董新宇. 利用全基因组测序数据解析藏鸡群体分层和遗传混杂现况 [D]. 武汉: 华中农业大学, 2021.
Dong X Y. Identifying the subpopulations and dissecting patterns of admixture in the Tibetan chickens by whole-genome resequencing data[D]. Wuhan: Huazhong Agricultural University, 2021 (in Chinese).

[29] Zhang C, Dong S S, Xu J Y, *et al*. PopLDdecay: a fast and effective tool for linkage disequilibrium decay analysis based on variant call format files[J]. Bioinformatics, 2019, 35(10): 1786-1788.

[30] 池炳杰, 常玉梅, 闫学春, 等. 瓦氏雅罗鱼达里湖群体和乌苏里江群体的遗传多样性和遗传结构分析[J]. 中国水产科学, 2010, 17(2): 228-235.
Chi B J, Chang Y M, Yan X C, *et al*. Genetic variability and genetic structure of *Leuciscus waleckii* Dybowski in Wusuli River and Dali Lake[J]. Journal of Fishery Sciences of China, 2010, 17(2): 228-235 (in Chinese).

[31] 王星火, 常雪晴, 程方平, 等. 基于2b-RAD简化基因组测序的三门湾海域3种优势鱼类群体遗传多样性分析[J]. 海洋渔业, 2021, 43(5): 513-520.
Wang X H, Chang X Q, Cheng F P, *et al*. Population genetic diversity of three dominant fishes in the Sanmen Bay area based on 2b-RAD simplified genome sequen-

cing[J]. Marine Fisheries, 2021, 43(5): 513-520 (in Chinese).

[32] 张会永, 李国辉, 薛倩, 等. 基于RAD-Seq技术的大围山微型鸡遗传进化分析[J]. 中国畜牧兽医, 2022, 49(4): 1393-1401.
Zhang H Y, Li G H, Xue Q, *et al*. Study on genetic evolution mechanism of Daweishan mini chicken based on RAD-Seq technology[J]. China Animal Husbandry & Veterinary Medicine, 2022, 49(4): 1393-1401 (in Chinese).

[33] 吕宝忠. 多态信息量(PIC)等于杂合度吗?[J]. 遗传, 1994, 16(4): 31-33.
Lv B Z. Does the polymorphism information content equal the heterozygosity?[J]. Hereditas, 1994, 16(4): 31-33 (in Chinese).

[34] 樊翠平. 中国圈养白头鹤及丹顶鹤种群遗传多样性分析 [D]. 哈尔滨: 东北林业大学, 2017.
Fan C P. Genetic diversity of the hooded craneand the red-crowned crane in China in captivity[D]. Harbin: Northeast Forestry University, 2017 (in Chinese).

[35] 张于光. 应用微卫星 DNA 标记进行东北虎亲子鉴定和遗传多样性研究 [D]. 长沙: 湖南农业大学, 2002.
Zhang Y G. Parentage and genetic diversity studies in amur tiger (*Panthera tigris altaica*) using microsatellite DNA markers[D]. Changsha: Hunan Agricultural University, 2002 (in Chinese).

[36] Serrote C M L, Reiniger L R S, Silva K B, *et al*. Determining the polymorphism information content of a molecular marker[J]. Gene, 2020, 726: 144175.

[37] Stephan W. Selective sweeps[J]. Genetics, 2019, 211(1): 5-13.

[38] Chen H, Patterson N, Reich D. Population differentiation as a test for selective sweeps[J]. Genome Research, 2010, 20(3): 393-402.

[39] 杨宇昕, 邹枨. 基于温带和热带玉米群体全基因组F_{ST}和XP-EHH的选择信号检测[J]. 中国农业科学, 2019, 52(4): 579-590.
Yang Y X, Zou C. Genome-wide detection of selection signal in temperate and tropical maize populations with Use of F_{ST} and XP-EHH[J]. Scientia Agricultura Sinica, 2019, 52(4): 579-590 (in Chinese).

[40] Li M Z, Tian S L, Jin L, *et al*. Genomic analyses identify distinct patterns of selection in domesticated pigs and Tibetan wild boars[J]. Nature Genetics, 2013, 45(12): 1431-1438.

Development of a software package for the analysis of genome resequencing data in aquatic populations

XU Qingteng [1,2], WU Haotian [1,2], JIANG Lihua [3*], LU Ying [1,2*]

(1. *Key Laboratory of Exploration and Utilization of Aquatic Genetic Resources*, *Ministry of Education*, *Shanghai Ocean University*, *Shanghai* 201306, *China*;
2. *International Research Center for Marine Bioscience*, *Ministry of Science and Technology*, *Shanghai Ocean University*, *Shanghai* 201306, *China*;
3. *National Engineering Research Center of Marine Facilities Aquaculture*, *Zhejiang Ocean University*, *Zhoushan* 316022, *China*)

Abstract: In recent decades, genome resequencing has been widely applied to study the genetic diversity in wild populations or cultured populations, such as estimation of divergence of populations and detection of artificial or environmental selection on the chromosomes. However, owing to lack of bioinformatics skills, most of the researchers have to resort to commercial platforms to analyze resequencing data, which needs high costs and long analysis period. The commercial platforms usually use the universal pipelines without personalized analysis, which sometime causes the erroneous results due to unreasonable parameters or reference genome data. Owing to the increasing demands of genome resequence data analysis in aquatic animals, we developed a user-friendly software package to facilitate population genetics analysis of genome resequencing data for aquatic biologists who may lack bioinformatics skills. By surveying the current research and achievements in the analysis of resequencing data in aquatic animals, the constructed software package integrated different bioinformatic tools, encompassing mapping the quality-controlled reads to the reference genome, detecting genetic variations, performing phylogeny and principal component analysis, clarifying population structure, calculating quantitative indicators of genetic diversity, and completing selective sweep analysis. All of the resulting data were finally visualized with an R or a Python language package. The present package was tested by analyzing the resequencing data of 30 *Larimichthys crocea* individuals of 3 populations, which successfully completed all of the designed tasks, such as alignment of the reads, identification of single nucleotide polymorphism (SNP), construction of phylogenetic tree and population structure, illustration of linkage disequilibrium decay and calculation of main diversity indexes. The generated outputs were well-visualized. The software package functionally integrated most of the basic statistics, calculation, and plotting for the analysis of wild and natural populations, which enabled most of the researchers to locally perform the data mining of genome resequence data to save time and costs. The corresponding source codes and instruction manuals have been uploaded to GitHub: https://github.com/xqteng/Re-seq_analysis.

Key words: aquatic animal resequencing; single nucleotide polymorphism, SNP; population genetics; software; genome

Corresponding authors: JIANG Lihua. E-mail: florencejiang@126.com;
LU Ying. E-mail: yinglu@shou.edu.cn

Funding projects: National Key Research and Development Program of China (2022YFD2400804, 2018YFD0900101)