bio-101 www.bio-protocol.org/e1010616 DOI:10.21769/BioProtoc.1010616

远缘袖蝶属杂交和基因渗入的研究方法

The Research Methods to Detect Hybridization and Introgression in the *Heliconius* Butterflies

曾华 [1]，张蔚 [1, 2, *]

[1] 蛋白质与植物基因研究国家重点实验室，生命科学联合中心，前沿交叉学科研究院，北京大学， 北京；

[2] 生命科学学院，北京大学，北京

*通讯作者邮箱：weizhangvv@pku.edu.cn

引用格式：曾华， 张蔚. (2021). 远缘袖蝶属杂交和基因渗入的研究方法. Bio-101 e1010616. Doi: 10.21769/BioProtoc. 1010616.

How to cite: Zeng, H. and Zhang, W. (2021). The Research Methods to Detect Hybridization and Introgression in the *Heliconius* Butterflies. Bio-101 e1010616. Doi: 10.21769/BioProtoc. 1010616. (in Chinese)

摘要：种间杂交是发生在近缘物种或分化不完全的物种之间互换遗传物质的现象，在推动物种适应性辐射中有着重要作用。前期相关研究报道大多在近缘物种中展开，本课题以袖蝶属中的两个远缘物种为研究对象，在全基因组层面对两者间的杂交和基因渗入现象进行了分析探讨，鉴定得到基因渗入片段并判断其方向。实验主要流程如下：(1) 提取组织基因组 DNA；(2) 构建双末端 Illumina 测序文库；(3) 使用 Illumina HiSeq 测序平台进行测序。数据分析主要包含以下四步：(1) 回贴数据和获得基因型信息；(2) 全基因组系统发生分析；(3) 检验种间基因流；(4) 渗入位点功能注释。本研究方法流程简单，是一种针对基于基因组数据集进行渗入位点鉴定的有效手段。

关键词：基因渗入，杂交，第二代测序，基因组，袖蝶属

研究背景

基因渗入 (introgression) 是种间杂交产生的基因交换，是向受体物种引入供体基因组片段的一种有效途径 (Mallet, 2005; Harrison and Larson, 2014)。相对于不适应的 (maladaptive) 或中性 (neutral) 的遗传片段，对生物生存和繁殖有益的基因组片段有更大的几率被固定下来，形成适应性的基因渗入 (adaptive introgression)，是基因组中适

Copyright © 2021 The Authors; exclusive licensee Bio-protocol LLC.

应性遗传变异的重要来源 (Martin and Jiggins, 2017)。其研究难点在于区别基因渗入与不完全谱系分选。随着测序技术的迅速发展，已经在多个研究体系中展开基因组层面的基因渗入研究，例如在现代人类与尼安德特人之间 (Fu *et al.*, 2015)、三刺鱼近缘种间 (Jones *et al.*, 2012)、疟蚊近缘种间 (Fontaine *et al.*, 2015) 等。本研究方法利用进化基因组学理论和最新技术，整合系统发生和群体遗传方法，在非模式动物南美袖蝶远缘种间建立了一套分析流程，能够鉴定得到局部的基因渗入片段并判断其方向，且能够有效区分基因渗入和不完全谱系分选，以此深入研究远缘物种间的杂交现象。

本文提供了详细的基于全基因组数据集，检测远缘物种间基因渗入的实验流程和数据分析方法，该方法的有效性已在鳞翅目类昆虫中得以验证，详细结果请参阅 Zhang *et al.*, 2016.

仪器设备

注：如不需要样品制备，直接对已有数据进行分析，可跳过设备 1-5，只配备和使用设备 6。

1. PCR 仪
2. 电泳仪
3. 超声波 DNA 破碎仪 (Covaris S220)
4. 生物分析仪 (Agilent 2100)
5. Illumina HiSeq 2500 型测序仪
6. 服务器 (catalog number: ThinkSystemSR650；系统：Gentoo；CPU：Xeon 6230R 2.1GHz 26 核；内存：512 Gb)

软件

注：本方法所使用的软件需要在服务器预先安装或下载直接使用，具体的下载和安装流程请参考软件对应的链接。

1. bcl2fastq2 Conversion Software (v2.20) https://support.illumina.com/sequencing/sequencing_software/bcl2fastq-conversion-software.html
2. Trimmomatic (v0.38) http://www.usadellab.org/cms/?page=trimmomatic
3. Bowtie2 (v2.3.4.3) http://bowtie-bio.sourceforge.net/bowtie2/index.shtml
4. SAMtools (v1.9) http://www.htslib.org/

Copyright © 2021 The Authors; exclusive licensee Bio-protocol LLC.

5. Picard (v1.84) http://broadinstitute.github.io/picard/
6. GATK (v3.3) https://gatk.broadinstitute.org
7. VCFtools https://vcftools.github.io/index.html
8. Phylogenomic https://github.com/npchar/Phylogenomic
9. RAxML https://cme.h-its.org/exelixis/web/software/raxml/index.html
10. iTOL https://itol.embl.de/
11. PhyTime https://www.stat.auckland.ac.nz/~guindon/phytime/
12. PhyML (v3.1) http://www.atgc-montpellier.fr/phyml/
13. TreeAnnotator https://beast2.blogs.auckland.ac.nz/treeannotator/
14. Genomics_general https://github.com/simonhmartin/genomics_general
15. R (v3.6.2) https://www.r-project.org
16. Perl (v5.32) https://www.perl.org/get.html
17. Python (v3.6) https://www.python.org/downloads/

实验步骤

一、样品制备和测序

注：如不需要样品制备，直接对已有数据或本方法提供的测试数据进行分析，可跳过步骤一。

1. 基因组 DNA 提取

   以袖蝶属（*Heliconius*）蝴蝶样品为例，收集成虫个体的胸部组织，采用 DNeasy Blood & Tissue Kit (catalog number: 69504，Qiagen)，参照产品说明书的步骤（https://www.qiagen.com/us/shop/pcr/dneasy-blood-and-tissue-kit/#resources)，提取基因组 DNA，采用常规琼脂糖凝胶电泳检验提取的 DNA 产物，保存于-20 °C。

2. 测序文库构建

   采用已制备的基因组 DNA，使用 TruSeq DNA Library Prep Kits (catalog number: FC-121-2001，Illumina)，参照产品说明书的步骤（https://support.illumina.com/downloads/truseq_dna_sample_preparation_guide_15026486.html），构建双末端 Illumina 测序文库。

3. 上机测序

   所建文库使用 Illumina HiSeq2500 测序平台进行测序。Illumina 测序仪下机所得的原始数据通常为 bcl 格式文件（basecall file)，但下游分析一般需要 fastq 格式文件。

Copyright © 2021 The Authors; exclusive licensee Bio-protocol LLC.

因此，在进行下游分析前，使用 bcl2fastq2 Conversion Software (v2.20) 将 bcl 文件根据之前添加的索引 (index) 分出，并转为 fastq 格式文件。命令行格式为：

/usr/local/bin/bcl2fastq –runfolder-dir <运行文件夹及路径> --output-dir <输出文件夹及路径>

二、数据回贴和获得基因型信息

1. 数据获得

本方法可以使用由步骤一获得的测序数据，或者使用发表于 *Genome Biology* (2016 年) 的文章 (Genome-wide introgression among distantly related *Heliconius* butterfly species) (Zhang *et al*., 2016) 的部分数据作为测试数据。该数据可直接在 NCBI 的 SRA 数据库下载获得 (数据登录号：PRJNA308754)。

2. 数据质量控制

使用 Trimmomatic (v0.38) 去除原始序列数据中的低质量碱基和接头序列。去除序列首尾碱基质量低于 3 的碱基，滑窗去除平均碱基质量值低于 10 的 90 bp 窗口，去除长度低于 36 bp 的 reads，得到高质量的基因组测序数据。命令格式为：

java -jar trimmomatic-0.38.jar PE <输入 fastq 压缩文件 1 及路径> <输入 fastq 压缩文件 2 及路径> <输出 fastq 压缩文件 1 及路径> <输出未配对 fastq 压缩文件 1 及路径> <输出 fastq 压缩文件 2 及路径> <输出未配对 fastq 压缩文件 2 及路径> ILLUMINACLIP:TruSeq3-PE.fa:2:30:10:2:keepBothReads SLIDINGWINDOW:90:10 LEADING:3 TRAILING:3 MINLEN:36

3. 数据回贴至参考基因组

使用 Bowtie2 (v2.3.4.3) 及参数 -very-sensitive-local 将序列数据比对至 *H. melpomene* v2 参考基因组 (下载地址：http://ensembl.lepbase.org/Heliconius_melpomene_melpomene_hmel2/Info/Index)。首先需生成参考基因组的索引文件，命令格式为：

bowtie2-build <参考基因组 fasta 文件名及路径> <索引文件基名及路径>

之后对每个个体基因组重测序数据，进行回贴并生成 sam 文件，命令格式为：

bowtie2 --very-sensitive-local -p 8 -x <索引文件基名及路径> -1 <fastq 文件 1 及路径> -2 <fastq 文件 2 及路径> -S <输出 sam 文件名及路径>

Copyright © 2021 The Authors; exclusive licensee Bio-protocol LLC.

4. 数据格式转换

使用 SAMtools (v1.9) 将 sam 文件转为 bam 文件。首先需生成参考基因组的索引文件，命令格式为：

samtools faidx <参考基因组 fasta 文件名及路径>

之后对于需转为 bam 的 sam 文件进行操作，命令格式为：

samtools view -bST <参考基因组 fasta 文件名及路径> -o <输出的 bam 文件名及路径> <输入的 sam 文件名及路径>

5. 添加数据信息和去除 PCR 重复

使用 Picard (v1.84) 的 AddOrReplaceReadGroups.jar 命令对 bam 文件的 reads 添加信息，并使用 MarkDuplicates.jar 去除数据中的 PCR 重复。每一步均会生成一个新的 bam 文件作为后续的输入文件，并且每一步生成新的 bam 文件之后需采用 BuildBamIndex.jar 生成 bam 文件的索引，即 bai 文件。

使用 AddOrReplaceReadGroups.jar 的格式为：

java -jar /usr/share/picard-tools-1.84/AddOrReplaceReadGroups.jar INPUT=<输入 bam 文件名及路径> OUTPUT=<输出 bam 文件名及路径> SORT_ORDER=coordinate RGID=<用户添加 ID 信息> RGLB=<用户添加建库信息> RGPL=<用户添加平台信息> RGSM=<用户添加样品信息> RGPU=none VALIDATION_STRINGENCY=LENIENT

使用 MarkDuplicates.jar 的格式为：

java -jar /usr/share/picard-tools-1.84/MarkDuplicates.jar INPUT=<输入 bam 文件名及路径> OUTPUT=<输出 bam 文件名及路径> METRICS_FILE=<输出 metrics 文件名及路径> REMOVE_DUPLICATES=true ASSUME_SORTED=true VALIDATION_STRINGENCY=LENIENT

使用 BuildBamIndex.jar 的格式为：

java -jar /usr/share/picard-tools-1.84/BuildBamIndex.jar INPUT=<输入 bam 文件名及路径> VALIDATION_STRINGENCY=LENIENT

6. 使用 GATK 流程获得基因型数据集

在使用 GATK 流程处理个体基因组重测序 bam 文件之前，先要使用 Picard (v1.84) 的 CreateSequenceDictionary.jar 命令对参考基因组 fasta 文件生成 dict 文件。

Copyright © 2021 The Authors; exclusive licensee Bio-protocol LLC.

使用 Picard (v1.84) 的 CreateSequenceDictionary.jar 命令的格式为：

java -jar /usr/share/picard-tools-1.84/CreateSequenceDictionary.jar R=<参考基因组 fasta 文件名及路径> O=<参考基因组 dict 文件名及路径>

之后再使用 GATK (v3.3) 中的 RealignerTargetCreator 和 IndelRealigner 命令重新比对每一个 bam 文件序列中的插入缺失区域。

使用 RealignerTargetCreator 的格式为：

java -jar /usr/share/GenomeAnalysisTK-3.3-0/GenomeAnalysisTK.jar –T RealignerTargetCreator -nt 10 -I <输入 bam 文件名及路径> -R <参考基因组 fasta 文件名及路径> -o <输出 intervals 文件名及路径>

使用 IndelRealigner 的格式为：

java -jar /usr/share/GenomeAnalysisTK-3.3-0/GenomeAnalysisTK.jar -T IndelRealigner -I <输入 bam 文件名及路径> -R <参考基因组 fasta 文件名及路径> -targetIntervals <输入 intervals 文件名及路径> -o <输出 bam 文件名及路径> -maxReads 100000

最后并使用 UnifiedGenotyper 命令对预处理过的所有个体基因组重测序数据的 bam 文件进行基因分型并生成 vcf 格式的数据集，本方法使用了以下参数：heterozygosity 0.01, stand_call_conf 50, stand_emit_conf 10, dcov 250。以包括四个个体 bam 文件为例，具体的使用格式为：

java -jar /usr/share/GenomeAnalysisTK-3.3-0/GenomeAnalysisTK.jar -T UnifiedGenotyper -nt 10 -R <参考基因组 fasta 文件名及路径> -I <个体 1bam 文件名及路径> -I <个体 2bam 文件名及路径> -I <个体 3bam 文件名及路径> -I <个体 4bam 文件名及路径> --heterozygosity 0.01 -stand_call_conf 50.0 -stand_emit_conf 10.0 -dcov 250 -o <输出 vcf 文件名及路径>

7. 筛选高质量的单核苷酸多态性 (SNP) 位点

使用 VCFtools 筛选高质量的 SNP 位点 (Qual > 30) 用于后续分析。具体使用格式为：

vcftools --vcf <输入 vcf 文件名及路径> --minQ 30 --out <输出 vcf 文件名及路径> --recode

Copyright © 2021 The Authors; exclusive licensee Bio-protocol LLC.

三、全基因组系统发生分析

基于全基因组 SNP 数据的系统发生分析适用于分析分歧时间相对较短的物种，可以获得基因组水平的界定物种关系的系统发生树，为研究物种间杂交和基因渗入时选取合适的系统发生拓扑结构提供依据。

1. 采用 vcftools 保留并对齐在所有个体中都得到基因型信息的高质量 SNP 位点，具体的使用格式为：

   vcftools --vcf <输入 vcf 文件名及路径> --minQ 30 --max-missing-count 0 --out <输出 vcf 文件名及路径> --recode

2. 采用 GATK (v3.3) 中的 FastaAlternateReferenceMaker 提取 SNP 信息并转换为 fasta 文件，并进一步采用 Phylogenomic 软件包中的 fasta2relaxedPhylip.pl 脚本转换为 phylip 格式文件。具体使用格式为：

   java -Xmx2g -jar GenomeAnalysisTK.jar -R<参考基因组 fasta 文件名及路径> -T FastaAlternateReferenceMaker -o <输出 fasta 文件名及路径> -L <输入 intervals 文件名及路径> -V <输入 vcf 文件名及路径>

   perl ./fasta2relaxedPhylip.pl -f <输入 fasta 文件名> -o <输出 phylip 文件名>

3. 使用 RAxML 软件构建全基因组最大似然树。使用的核苷酸替换模型为 GTRGAMMA，进行 100 次快速自展 (bootstrap) 重复。具体使用格式为：

   raxmlHPC -f a -m GTRGAMMA -p 12345 -x 12345 -# 100 -T 20 -s <输入 phylip 文件名> -n test

4. 使用在线工具 iTOL (https://itol.embl.de/) 对系统发生树进行可视化。

5. 使用 PhyTime 软件对基因组水平的拓扑结构进行时间校正，以 *H. cydno* 和 *H. melpomene* 以及和 *H. pachinus* 之间的平均分歧时间 (1.4 Mya 及 0.43 Mya) 作为校准点，估算其他类群之间的分歧时间，并利用 TreeAnnotator 对结果进行统计。具体使用格式为：

   phytime -i <输入 phylip 文件名> -d nt -q -m GTR --calibration <输入校正节点文件名> -t e -v e -u <输入系统发生树文件名> --r_seed 35

四、检验种间基因流

为了更好的在全基因组水平鉴定潜在的基因渗入位点，建议使用帕特森 *D* 统计量

Copyright © 2021 The Authors; exclusive licensee Bio-protocol LLC.

(Patterson's *D*-statistic) 及 *f* 统计量 (modified *f*-statistic，f_d) 对染色体和局部基因组区域进行统计检验，并进一步整合序列分歧水平等信息，以获得可靠的渗入位点。其中：

1. *D* 统计量通过比较满足 ABBA 和 BABA 模式的衍生等位基因分布情况，检测基因流存在与否。以含有四个分类群的拓扑结构 (((P1, P2), P3), O) 为例，ABBA 和 BABA 模式分别代表了 P3 与 P2 和 P3 与 P1 之间共享衍生等位基因。在零假设下，两种模式的数量应该相同。如两者之间存在显著差异，即表明对应类群之间发生了基因渗入。对于含有多个个体的群体数据，使用等位基因频率代替替换数量进行计算，计算公式如下：

$$D(P_1, P_2, P_3, O) = \frac{\sum_{i=1}^{n}[(1-\hat{P}_{i1})\hat{P}_{i2}\hat{P}_{i3}(1-\hat{P}_{i4}) - \hat{P}_{i1}(1-\hat{P}_{i2})\hat{P}_{i3}(1-\hat{P}_{i4})]}{\sum_{i=1}^{n}[(1-\hat{P}_{i1})\hat{P}_{i2}\hat{P}_{i3}(1-\hat{P}_{i4}) + \hat{P}_{i1}(1-\hat{P}_{i2})\hat{P}_{i3}(1-\hat{P}_{i4})]}$$

其中，P1、P2、P3 和 P4 为分析中使用的四个类群，$\hat{P}_{ij}$ 为 SNP i 在种群 j 中观察到的等位基因频率。对于基因组水平的基因流估计，使用长度为 50 kb 的滑动窗口计算 *D* 统计量，并利用 R 软件包 bootstrap (v. 2012-04) 以刀切法 (jackknife) 计算染色体水平的均值和标准误。

其中，*D* 统计量可以采用 Genomics_general 软件包进行计算，首先采用 parseVCF.py 转换文件格式，进而采用 ABBABABAwindows.py 计算，具体使用格式为：

python parseVCF.py -i <输入压缩 vcf 文件名> --skipIndels --minQual 30 --gtf flag=DP min=5 | bgzip > <输出压缩 geno 文件名>

python ABBABABAwindows.py -g <输出压缩 geno 文件名及路径> -f phased -o <输出 csv 文件名> -w 50000 -m 100 -s 50000 -P1 A -P2 B -P3 C -O D -T 10 --minData 0.5 --popsFile <输入种群信息文件名> --writeFailedWindows --polarize &

为了鉴定不同尺度的局部渗入位点，也使用 5 kb、10 kb 和 50 kb 等多种不同的窗口大小计算 *D* 统计量。当窗口长度较小，处于连锁不平衡的范围内时，相邻窗口之间可能存在相关性。在这种情况下，利用 R 软件包 wwntests (v.1.0.1) 中的 block_bootstrap.R 以滑动分块自助法 (moving-block bootstrap) 计算窗口内部的均值和标准误，其中分块大小设置为 $n^{1/3}$，n 为数据集大小。若均值在双尾 z 检验中显

Copyright © 2021 The Authors; exclusive licensee Bio-protocol LLC.

著偏离 0，则认为该位点存在基因流。由于对基因组内的大量窗口进行统计检验，使用 Benjamini-Hochberg 方法对概率值进行多重比较校正，以错误发现率 (false discovery rate, FDR) 0.01 作为存在显著差异的阈值。

2. f_d 统计量较 *D* 统计量有更高的分辨率，并且能够更好的估计渗入的比例，因此选择使用 f_d 统计量进行进一步的筛选。f_d 统计量计算公式如下：

$$\hat{f} = \frac{S(P_1, P_2, P_3, O)}{S(P_1, P_D, P_D, O)}$$

其中，P_1、P_2、P_3 和 O 为分析中使用的四个类群，P_D 代表了渗入的供体种群，为 P_2 和 P_3 中具有更高衍生等位基因频率的群体。f_d 计算方法同 *D* 统计量，可以采用 Genomics_general 软件包进行计算。

同样以 5 kb、10 kb 和 50 kb 的窗口大小进行统计检验。检验方法同 *D* 统计量。

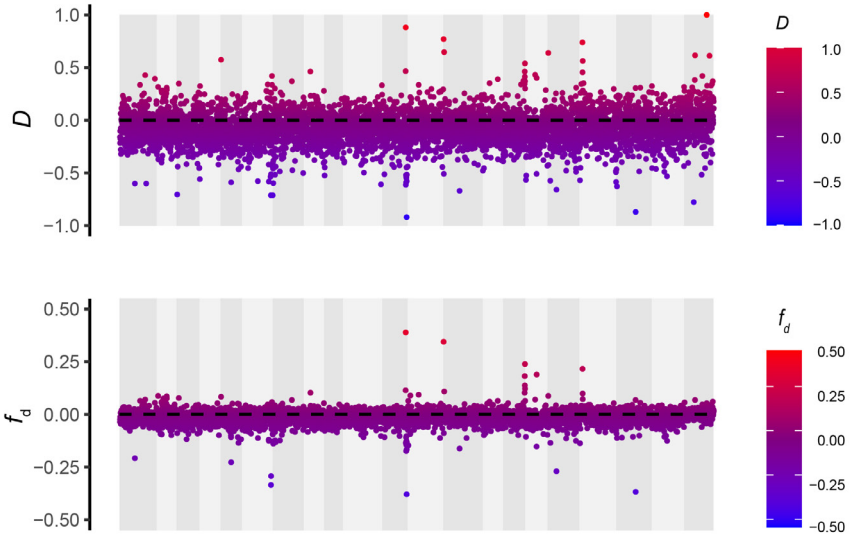

图 1. 基因组范围内帕特森 *D* 统计量和 f_d 统计量示意图

3. 为了排除由于不完全谱系分选造成的假阳性信号，计算类群之间的序列绝对分歧 (sequence divergence, d_{xy})，作为 ABBA-BABA 分析的补充。对于发生基因渗入的区域，序列分歧时间小于两个类群之间的分歧时间，因此通常具有较低的 d_{xy} 水平。使用 50 kb 的滑动窗口计算染色体水平的 d_{xy} 均值，对于潜在的渗入位点，使用 100

Copyright © 2021 The Authors; exclusive licensee Bio-protocol LLC.

bp 窗口计算 d_{xy}，并利用 R 软件包的曼-惠特尼秩和检验 (Mann-Whitney U-test) 与所在染色体进行比较，去除分歧水平高于染色体均值的位点。对于群体数据，d_{xy} 的计算公式如下：

$$d_{xy} = \frac{1}{n}\sum_{i=1}^{n} \hat{p}_{ix}\left(1-\hat{p}_{iy}\right) + \hat{p}_{iy}\left(1-\hat{p}_{ix}\right)$$

其中，p_x 和 p_y 分别为 x 和 y 类群中参考等位基因的频率，各等位基因频率可使用 vcftools 计算获得，具体使用格式为：

vcftools --vcf <输入 vcf 文件名及路径> --keep <待计算的样品名文件> --freq --out <输出文件名及路径>。

4. 进一步检查潜在渗入位点的测序深度，以排除序列错误回贴造成的假阳性信号。由于与参考序列差异较大的重测序数据更加难以有效回贴，因此在基因型鉴定过程中，可能会基于与参考序列相似的数据得到对等位基因频率的错误估计。这种偏差也可能导致显著的渗入信号，特别是从参考基因组 *H. melpomene* 向其他物种的基因流。这些受影响的区域也会表现出序列覆盖深度的降低，因此可以通过位点的深度进行筛选。另一方面，具有异常高覆盖的区域也可能代表了序列的比对错误，并导致对等位基因频率的错误估计。因此，过滤去除测序深度低于 5 和高于 40 的位点。可以使用 vcftools 的参数--site-depth 以及--site-mean-depth 对测序深度进行评估；使用 vcftools 的参数--minDP，--maxDP 以及--min-meanDP，--max-meanDP 按照测序深度对位点进行过滤。
5. 对候选的渗入位点，使用 PhyML (v3.1) 软件构建最大似然树，并根据系统发育关系推断基因流的方向。使用的核苷酸替换模型为 GTR，进行 100 次自展分析。具体使用格式为：

/usr/share/PhyML-3.1/PhyML-3.1_linux64 -i <输入 phylip 文件名及路径> --sequential -n 1 -b 100 -m GTR -t e -v e -o tlr --print_site_lnl --run_id 1

五、渗入位点功能注释

1. 基于 *H. melpomene* 基因组注释，获取渗入位点内部的蛋白编码基因。可在 Lepbase (http://blast.lepbase.org/) 数据库检索基因已知的结构和功能。

Copyright © 2021 The Authors; exclusive licensee Bio-protocol LLC.

2. 通过 blastx 将基因编码序列比对到 NCBI nr (https://blast.ncbi.nlm.nih.gov/Blast.cgi)、Uni-Prot (https://www.uniprot.org/) 等蛋白数据库，基于蛋白序列的同源性，推断渗入基因的潜在功能。
3. 基于 GO (http://geneontology.org)、KEGG (http://www.genome.jp/kegg) 和 Pfam (v33.1,http://pfam.xfam.org) 等数据库对渗入基因进行富集分析。

致谢

感谢张宇博对本文撰写提供的帮助。感谢审稿人在本文修改中提出的建议和意见。本研究得到北京市自然科学基金杰出青年项目（编号 JQ19021）、国家自然科学基金面上项目（编号 31871271）、启东创新基金支持。

参考文献

1 Mallet, J. (2005). Hybridization as an invasion of the genome. *Trends Ecol Evol* 20: 229-237.

2 Harrison, R. G. and Larson, E. L. (2014). Hybridization, introgression, and the nature of species boundaries. *J Hered* 105: 795-809.

3 Martin, S. H. and Jiggins, C. D. (2017). Interpreting the genomic landscape of introgression. *Curr Opin Genet Dev* 47: 69-74.

4 Fu, Q., Hajdinjak, M., Moldovan, O. T., Constantin, S., Mallick, S., Skoglund, P., Patterson, N., Rohland, N., Lazaridis, I., Nickel, B., Viola, B., Prüfer, K., Meyer, M., Kelso, J., Reich, D. and Pääbo, S. (2015). An early modern human from Romania with a recent Neanderthal ancestor. *Nature* 524: 216-219.

5 Jones, F. C., Grabherr, M. G., Chan, Y. F., Russell, P., Mauceli, E., Johnson, J., Swofford, R., Pirun, M., Zody, M. C., White, S., Birney, E., Searle, S., Schmutz, J., Grimwood, J., Dickson, M. C., Myers, R. M., Miller, C. T., Summers, B. R., Knecht, A. K., Brady, S. D., Zhang, H., Pollen, A. A., Howes, T., Amemiya, C., Broad Institute Genome Sequencing Platform & Whole Genome Assembly Team, Baldwin, J., Bloom, T., Jaffe, D. B., Nicol, R., Wilkinson, J., Lander, E. S., Palma, F. D., Lindblad-Toh, K. and Kingsley, D. M. (2012). The genomic basis of adaptive evolution in threespine sticklebacks. *Nature* 484: 55-61.

6 Fontaine, M. C., Pease, J. B., Steele, A., Waterhouse, R. M., Neafsey, D. E.,

Copyright © 2021 The Authors; exclusive licensee Bio-protocol LLC.

Sharakhov, I. V., Jiang, X., Hall, A. B., Catteruccia, F., Kakani, E., Mitchell, S. N., Wu, Y. C., Smith, H. A., Love, R. R., Lawniczak, M. K., Slotman, M. A., Emrich, S. J., Hahn, M. W. and Besansky, N. J. (2015). Extensive introgression in a malaria vector species complex revealed by phylogenomics. *Science* 347: 1258524.

7 Zhang, W., Dasmahapatra, K. K., Mallet, J., Moreira, G. R. P. and Kronforst, M. R. (2016). Genome-wide introgression among distantly related *Heliconius* butterfly species. *Genome Biol* 17: 25.

Copyright © 2021 The Authors; exclusive licensee Bio-protocol LLC.