群体基因组水平的适应性位点检测

Detecting Adaptive Loci with Population Genomic Data

金铃^{1,2}, 郭宝成^{1,2,*}

¹ 动物进化与系统学院重点实验室, 中国科学院动物研究所, 北京 100101; ² 中国科学院大学, 北京 100049

*通讯作者邮箱: guobaocheng@ioz.ac.cn

引用格式: 金铃, 郭宝成. (2021). 群体基因组水平的适应性位点检测. *Bio-101* e1010627. Doi: 10.21769/BioProtoc.1010627.

How to cite: Jin, L. and Guo, B. C. (2021). Detecting Adaptive Loci with Population Genomic Data. *Bio-101* e1010627. Doi: 10.21769/BioProtoc.1010627. (in Chinese)

摘要: 由于近年来全球水环境的急剧变化, 鱼类多样性正受到严重威胁, 理解鱼类局部适应环境变化的遗传机制已成为鱼类进化研究的热点。随着高通量测序技术的成熟, 群体基因组学快速发展, 适应性进化遗传机制的研究也已在野生鱼类群体中广泛开展。 F_{ST} 离群值分析 (outlier analysis, OA) 是检测鱼类群体适应性位点的主要方法之一, 可通过多种方法实现。本文以 BayeScan 软件为例, 介绍从获得重测序数据、获得全基因组范围内的高质量 SNP (single nucleotide polymorphism, 单核苷酸多态性) 数据集、到利用 BayeScan 软件进行 F_{ST} 离群值分析、以及结果解读和绘图的完整流程, 希望帮助读者开展检测群体基因组水平受选择位点的研究。该流程具有一定的普适性, 可应用于鱼类等不同生物类群的适应性进化机制研究。

关键词: 适应性进化, 选择信号, 群体基因组学, F_{ST} 离群值, BayeScan

研究背景

适应性进化是生物适应环境变化赖以生存的基础。生物能否快速适应环境变化, 对于维护生物多样性、人类健康、粮食安全、自然资源的可持续性 etc 均有重要意义 (Carroll *et al.*, 2014)。特别是近年来全球气候和环境正在发生剧烈变化, 理解生物适应环境的遗传机制, 已成为进化生物学研究的一个核心问题 (Bernatchez, 2016)。鱼类包含约 32,000

个物种，占据全球现生脊椎动物物种数的一半以上，目前也正面临着水环境的急剧变化，如海面温度的升高、海洋酸化、水体污染等，鱼类多样性正受到严重威胁 (Joseph *et al.*, 2016)。因此，理解鱼类对环境变化的遗传响应，在气候和环境问题凸显的当下，具有重要的科学和现实意义。

近年来，随着高通量测序技术的发展，群体基因组学数据和研究方法逐步应用于鱼类的本地适应 (local adaptation) 研究中 (López *et al.*, 2015; Bernatchez, 2016)。本地适应是指某群体相比于其他群体更适应其所在生境的进化过程 (Savolainen *et al.*, 2013)。基于重测序数据，可以获得来自不同环境鱼类群体的遗传变异，进而检测与环境适应相关的选择信号，从而揭示鱼类的适应性遗传机制。 F_{ST} 离群值分析 (outlier analysis, OA) 是目前检测受选择位点的主流方法之一 (Liberles *et al.*, 2020)。 F_{ST} 离群值分析可识别群体间相对于中性进化模型所预期的、有着显著更高的遗传分化的变异位点，由此推断受到选择压力而发生分化的基因组区域 (Ahrens *et al.*, 2018)。在诸多可进行 F_{ST} 离群值分析的软件中，BayeScan 软件使用较为广泛 (图 1, 摘自 Ahrens *et al.*, 2018)。BayeScan 的基本原理是，利用贝叶斯模型估算 F_{ST} ，以度量群体间等位基因频率的差异，而后将 F_{ST} 分解为群体特异的 (population-specific, beta) 和位点特异的 (locus-specific, alpha) 两个组分，并从 alpha 组分中识别偏离中性模型的、受自然选择的位点 (Foll and Gaggiotti, 2008)。本文将 BayeScan 软件为例，介绍使用 BayeScan 进行 F_{ST} 离群值分析的完整流程。该流程主要参考 Guo *et al.* (2015, 2016a) 两篇已发表的工作，该方法也可应用于鱼类以外的其他生物类群 (例如，Guo *et al.*, 2016b 和 Yadav *et al.*, 2021)。

图 1. 2010-2016 年使用 F_{ST} 离群值分析各类方法的文献数目 图中横坐标为文献出版年份，纵坐标为该年份使用对应方法的文献数目，不同颜色代表不同的 F_{ST} 离群值分析方法。仅统计使用超过 1 次的方法。(摘自 Ahrens *et al.*, 2018)

仪器设备

1. 服务器 (型号: Inspur NF5280M5; 操作系统: CentOS Linux release 7.5.1804 (Core); CPU: Intel(R) Xeon(R) Gold 6230 CPU @ 2.10 GHz; 80 核, 512 G 内存)

注: 本文使用的软件对运行设备内存有较高要求。例如在利用 **BWA-MEM** 软件建立参考基因组索引时, 要求可用内存有 $5.37N$ 以上 (N 为参考基因组大小, 参见 **BWA** 使用手册 <http://bio-bwa.sourceforge.net/bwa.shtml>); **PGDSpider** 是基于 **JAVA** 的软件, 若 **VCF** 格式变异文件较大 (几百 Mb), 在运行中调用内存可达几十 Gb。考虑到对内存和运行速度的要求, 一般需用同等配置的服务器完成本套分析流程。

2. 普通个人电脑 (需要已安装好 R)

软件版本信息及下载地址

1. **BWA-MEM** v0.7.17-r1188: <https://github.com/lh3/bwa>

2. SAMtools v1.8: <https://github.com/samtools/samtools/releases/tag/1.8>
3. BCFtools v1.8: <https://github.com/samtools/bcftools/releases/tag/1.8>
4. VCFtools v0.1.13: <https://github.com/vcftools/vcftools/releases/tag/v0.1.13>
5. Java v1.8.0_151: <https://java.com/zh-CN/>
6. PGDSpider 2.1.1.5: <http://cmpg.unibe.ch/software/PGDSpider/>
7. BayeScan 2.1: <https://github.com/mfoll/BayeScan>
8. R v3.6.3: <https://www.r-project.org/>
9. CODA (R package, v0.19-4): <https://cran.r-project.org/web/packages/coda/index.html>
10. ggplot2 (R package, v3.3.3): <https://cran.r-project.org/web/packages/ggplot2/index.html>

实验步骤

1. 产生一个高质量的 SNP 数据集

利用 Illumina 二代测序平台,对所研究群体的个体进行重测序。为了获得可靠的 SNP 数据集,我们建议每个群体对至少 10 个个体测序,且每个个体测序深度为 10x 以上,读者也可依据研究目的增加样本量或测序深度(可参考 Fumagalli, 2013)。选取参考基因组(如所研究物种无自身参考基因组则选择所研究物种近缘种的基因组),使用 BWA、SAMtools、BCFtools、VCFtools 等软件进行重测序数据的比对、SNP (single nucleotide polymorphism, 单核苷酸多态性) 检测、SNP 过滤,以获得高质量的 SNP 数据集。高质量的 SNP 数据集的过滤通常考虑:去除插缺突变 (Insertion and Deletion, Indel) 附近的 SNPs; 去除多等位基因仅保留双等位基因 SNPs; 去除覆盖深度过低或过高的 SNPs; 去除群体中基因型缺失比例过高的 SNPs; 去除质量低的 SNPs; 去除次要等位基因频率过低的 SNPs 等。过滤后的 SNP 数目视测序质量、测序深度、样本量、群体数目以及所研究的类群而可能有较大波动,以鱼类为例,基于重测序数据获得的高质量 SNP 数目在几万到几千万不等 (Xu *et al.*, 2019; Wang *et al.*, 2021),而在鸟类中,SNP 数目相对更多,一般在 1000 万个左右 (Dutoit *et al.*, 2017; Weng *et al.*, 2020)。具体分析过程如下:

1) 为参考基因组序列文件建立索引。

```
$ bwa index reference.fasta
```

- 2) 将通过质量控制 (Quality control) 的原始读段 (read) 比对至参考基因组上，并对生成的 BAM 格式比对文件进行排序和去除 PCR 重复序列。

```
$ bwa mem -t 4 -M -R
'@RG\tID:sample01\tPL:ILLUMINA\tLB:sample01\tSM:sample01'
reference.fasta sample01_1_clean.fq.gz sample01_2_clean.fq.gz | samtools
view -@ 4 -Sbh - -o sample01.bam
$ samtools sort -@ 4 -O BAM -o sample01.sort.bam sample01.bam
$ samtools rmdup sample01.sort.bam sample01.sort.rmdup.bam
```

- 3) 使用 BCFtools 进行 SNP 检测 (可分单条染色体进行)。

```
$ samtools faidx reference.fa
$ bcftools mpileup -C 50 -d 10000 -q 20 -Q 20 -f reference.fasta -a
"AD,ADF,ADR,DP,SP,INFO/AD,INFO/ADF,INFO/ADR" -O u --threads 4 -r chr1
-b bam_list.txt | bcftools call -vmO z --threads 4 -o chr1.vcf.gz
```

- 4) 使用 BCFtools、VCFtools 等软件进行 SNP 过滤，以去除低质量或不可靠的 SNPs。

去掉插缺突变附近 10 bp 的 SNPs，以排除插缺突变附近的比对错误引入的假的 SNPs:

```
$ bcftools filter --Snpgap 10 -O z --threads 4 -o chr1.SnpGap10.vcf.gz
chr1.vcf.gz
```

去掉插缺突变，仅保留 SNPs:

```
$ vcftools --gzvcf chr1.SnpGap10.vcf.gz --remove-indels --recode --recode-
INFO-all --stdout | gzip -c > chr1.SnpGap10.xindels.vcf.gz
```

由于一些生信分析软件仅支持分析双等位基因，为了方便进行其他群体基因组学分析，这里仅保留双等位基因，去掉多等位基因:

```
$ vcftools --gzvcf chr1.SnpGap10.xindels.vcf.gz --min-alleles 2 --max-alleles 2
--recode --recode-INFO-all --stdout | gzip -c >
chr1.SnpGap10.xindels.biallele.vcf.gz
```

标记覆盖深度过低或过高的基因型为缺失，以排除覆盖深度过低导致的不准确的基因型、以及重复序列比对错误造成的高覆盖深度的假的 SNPs:

```
$ vcftools --gzvcf chr1.SnpGap10.xindels.biallele.vcf.gz --minDP 2 --maxDP
20 --recode --recode-INFO-all --stdout | gzip -c >
```

chr1.SnpGap10.xindels.biallele.depth.vcf.gz

基因型缺失比例过高的 SNPs 可能并不可靠，因此去掉基因型缺失的比例超过 20% 的 SNPs:

```
$ vcftools --gzvcf chr1.SnpGap10.xindels.biallele.depth.vcf.gz --max-missing 0.8 --recode --recode-INFO-all --stdout | gzip -c > chr1.SnpGap10.xindels.biallele.depth.maxmiss.vcf.gz
```

去掉质量低于 25 的 SNPs:

```
$ vcftools --gzvcf chr1.SnpGap10.xindels.biallele.depth.maxmiss.vcf.gz --minQ 25 --recode --recode-INFO-all --stdout | gzip -c > chr1.SnpGap10.xindels.biallele.depth.maxmiss.minQ.vcf.gz
```

频率过低的次要等位基因可能是 SNP 检测中的错误，因此仅保留次要等位基因频率大于 0.05 的 SNPs:

```
$ vcftools --gzvcf chr1.SnpGap10.xindels.biallele.depth.maxmiss.minQ.vcf.gz --maf 0.05 --recode --recode-INFO-all --stdout | gzip -c > chr1.SnpGap10.xindels.biallele.depth.maxmiss.minQ.maf.vcf.gz
```

5) 使用 BCFtools 将单条染色体的 VCF 文件连接为一个文件。

```
$ bcftools concat -f vcf_list.txt -O v -o test.vcf
```

2. 格式转换

利用重测序数据获得的 SNP 变异文件通常为 VCF 格式，使用 PGDSpider 软件将 VCF 格式转为 BayeScan 软件的输入文件格式 GESTE/BayeScan 格式。

```
$ java -jar PGDSpider2-cli.jar -inputfile test.vcf -inputformat VCF -outputfile BayeScan_input.txt -outputformat GESTE_BAYE_SCAN -spid VCF_GESTE_BAYE_SCAN.spid
```

3. 使用 BayeScan 软件进行离群值分析

```
$ bayescan_2.1 BayeScan_input.txt -od output -threads 16 -n 5000 -thin 10 -nbp 20 -pilot 5000 -burn 50000 -pr_odds 100 -out_pilot
```

注:

1) `--pr_odds` 参数默认设为 10，随着参与分析的 SNP 数目的增加，该参数一般也需要调大，具体参考 BayeScan 使用手册 (https://github.com/mfoll/BayeScan/blob/master/BayeScan2.1_manual.pdf)。

2) 生成文件: 文件*.sel 记录了 RJ-MCMC (reversible-jump Markov Chain Monte

Carlo) 链每次迭代中模型估算的参数, 该文件可用于后续评估收敛性; 文件 *_fst.txt 记录了每个 SNP 偏离中性模型 (即位点受选择) 的 q -value、 α 组分等。 α 组分为正值, 说明该位点受到歧化选择 (diversifying selection); α 系数为负值, 说明该位点受到平衡选择 (balancing selection) 或纯化选择 (purifying selection)。

4. 评估收敛性

使用 R 包 CODA 评估 RJ-MCMC 链是否收敛。

1) 加载 CODA 包。

```
> library(coda)
```

2) 导入数据, BayeScan 的生成文件*.sel 中列出了 RJ-MCMC 链每次迭代中模型估算的参数, 包括每个群体的 F_{ST} 等参数。

```
> chain <- read.table("BayeScan_input.sel",header=TRUE)
```

3) 创建一个 MCMC 对象。

```
> chain <- chain[-c(1)]
```

```
> chain <- mcmc(chain,thin=10)
```

4) 绘图, 初步判断 RJ-MCMC 链是否收敛, 如图 2 所示。可以看出群体 1 的 F_{ST} 随每次迭代在一定值上下波动, 且 F_{ST} 的后验分布大致符合正态分布, 说明群体 1 的 RJ-MCMC 链已经大致收敛。

```
> plot(chain)
```

图 2. 群体 1 F_{ST} (Fst1) 随迭代的路径图 (左图) 和后验分布图 (右图)

5) 使用 Geweke 收敛性检验 (Geweke's convergence diagnostic) 等方法判断

RJ-MCMC 链是否收敛。运行 `geweke.diag()` 函数，输出的检验统计量为标准 Z-score，若 $|\text{标准 Z-score}| > 1.96$ ，则说明 RJ-MCMC 链收敛（置信度 $\alpha = 0.05$ ）。

```
> geweke.diag(chain,frac1=0.1,frac2=0.5)
```

5. 结果分析及绘图

- 1) 输出受到歧化选择 ($q\text{-value} < 0.05$, α 组分 > 0) 的位点数:

```
$ awk 'NR>1 && $4<0.05 && $5>0' BayeScan_input_fst.txt | wc -l
```

输出受到平衡或纯化选择 ($q\text{-value} < 0.05$, α 组分 < 0) 的位点数:

```
$ awk 'NR>1 && $4<0.05 && $5<0' BayeScan_output_fst.txt | wc -l
```

- 2) 使用 `plot_R.r` 作图 (`plot_R.r` 是 BayeScan 软件安装包中自带的作图 R 脚本)。

加载 `plot_R.r` 脚本:

```
> source("plot_R.r")
```

利用输出文件*_fst.txt 作图 (图 3):

```
> plot_bayescan("BayeScan_input_fst.txt",FDR=0.05,size=0.5,add_text=F)
```

图 3. 基于约 70,000 个 SNPs 检测到的离群位点 竖线代表错误发现率 (false discovery rate, FDR) 的阈值 0.05。

致谢

我们的研究受到国家自然科学基金委员会优秀青年科学基金项目 (32022009)、中国科学院率先行动百人计划项目和第二次青藏高原综合科学考察研究项目 (Grant No. 2019QZKK0501) 的资助支持。

竞争性利益声明

作者声明没有利益冲突。

参考文献

1. Ahrens, C. W., Rymer, P. D., Stow, A., Bragg, J., Dillon, S., Umbers, K. D. L. and Dudaniec, R. Y. (2018). [The search for loci under selection: trends, biases and progress](#). *Mol Ecol* 27(6): 1342-1356.
2. Bernatchez, L. (2016). [On the maintenance of genetic variation and adaptation to environmental change: considerations from population genomics in fishes](#). *J Fish Biol* 89(6): 2519-2556.
3. Carroll, S. P., Jorgensen, P. S., Kinnison, M. T., Bergstrom, C. T., Denison, R. F., Gluckman, P., Smith, T. B., Strauss, S. Y. and Tabashnik, B. E. (2014). [Applying evolutionary biology to address global challenges](#). *Science* 346(6207): 1245993.
4. Dutoit, L., Burri, R., Nater, A., Mugal, C. F. and Ellegren, H. (2017). [Genomic distribution and estimation of nucleotide diversity in natural populations: perspectives from the collared flycatcher \(*Ficedula albicollis*\) genome](#). *Mol Ecol Resour* 17: 586-597.
5. Foll, M. and Gaggiotti, O. (2008). [A genome-scan method to identify selected loci appropriate for both dominant and codominant markers: a Bayesian perspective](#). *Genetics* 180(2): 977-993.
6. Fumagalli, M. (2013). [Assessing the effect of sequencing depth and sample size in population genetics inferences](#). *PLoS ONE* 8(11): e79667.
7. Guo, B., DeFaveri, J., Sotelo, G., Nair, A. and Merilä, J. (2015). [Population genomic evidence for adaptive differentiation in Baltic Sea three-spined sticklebacks](#). *BMC Biol* 13(1): 19.
8. Guo, B., Li, Z. and Merilä, J. (2016a). [Population genomic evidence for adaptive differentiation in the Baltic Sea herring](#). *Mol Ecol* 25(12): 2833-2852.
9. Guo, B., Lu, D., Liao, W. B. and Merilä, J. (2016b). [Genomewide scan for adaptive differentiation along altitudinal gradient in the Andrew's toad *Bufo andrewsi*](#). *Mol Ecol* 25(16): 3884-3900.
10. Liberles, D. A., Chang, B., Geiler-Samerotte, K., Goldman, A., Hey, J., Kacar, B.,

- Meyer, M., Murphy, W., Posada, D. and Storfer, A. (2020). [Emerging frontiers in the study of molecular evolution](#). *J Mol Evol* 88(3): 211-226.
11. López, M. E., Neira, R. and Yanez, J. M. (2015). [Applications in the search for genomic selection signatures in fish](#). *Front Genet* 5: 12.
12. Nelson, J., Grande, T. and Wilson, M. (2016). *Fishes of the world, fifth edition*. Wiley. ISBN: 9781118342336.
13. Salvolainen, O., Lascoux, M. and Merilä, J. (2013). [Ecological genomics of local adaptation](#). *Nat Rev Genet* 14: 807-820.
14. Wang, S., Kuang, Y., Liang, L., Sun, B., Zhao, X., Zhang, L. and Chang, Y. (2021). [Resequencing and SNP discovery of Amur ide \(*Leuciscus waleckii*\) provides insights into local adaptations to extreme environments](#). *Sci Rep* 11: 5064.
15. Weng, Z., Xu, Y., Li, W., Chen, J., Zhong, M., Zhong, F., Du, B., Zhang, B. and Huang, X. (2020). [Genomic variations and signatures of selection in Wuhua yellow chicken](#). *PLoS ONE* 15(10): e0241137.
16. Xu, S., Zhao, L., Xiao, S. and Gao, T. (2019). [Whole genome resequencing data for three rockfish species of *Sebastes*](#). *Sci Data* 6: 97.
17. Yadav, S., Stow, A. J. and Dudaniec, R. Y. (2021). [Microgeographical adaptation corresponds to elevational distributions of congeneric montane grasshoppers](#). *Mol Ecol* 30(2): 481-498.