应用环境关联分析检测本地适应位点

Detecting Loci Underlying Local Adaptation Using Environmental Association Analysis

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). 应用环境关联分析检测本地适应位点. *Bio-101* e1010615. Doi: 10.21769/BioProtoc.1010615.

How to cite: Jin, L. and Guo, B. C. (2021). Detecting Loci Underlying Local Adaptation Using Environmental Association Analysis. *Bio-101* e1010615. Doi: 10.21769/BioProtoc.1010615. (in Chinese)

摘要: 随着二代测序技术的广泛应用, 以及气候、地理、环境观测大数据的积累, 利用环境关联分析 (environmental association analysis, EAA) 检测全基因组范围内的环境选择信号, 已成为研究生物本地适应 (local adaptation) 遗传基础的重要方法。本文将介绍基于重测序数据获得的 SNP (single nucleotide polymorphism, 单核苷酸多态性) 遗传变异数据, 以 BAYENV2 软件为例进行环境关联分析的详细流程, 包括数据准备、格式转换、协方差矩阵估算、环境相关性分析、结果解读和绘图等步骤, 鉴定出与环境变量相关联的遗传变异, 从而揭示环境选择压力驱动的适应机制。该数据分析方案目前已广泛应用于包括鱼类在内的诸多生物类群的景观基因组学研究中。

关键词: 本地适应, 景观基因组学, 环境关联分析, 选择信号, BAYENV2

研究背景

随着全球气候和环境的剧烈变化, 理解生物适应环境变化的遗传响应, 已成为进化生物学研究中的一个重要问题。利用 RAD-seq、全基因组重测序等技术获得 SNP (single nucleotide polymorphism, 单核苷酸多态性), 以 SNP 作为遗传标记进行选择信号分析, 无论对于模式物种还是非模式物种, 均已成为目前适应性进化遗传机制研究的主流方法 (Pardo-Diaz *et al.*, 2015)。基于全基因组范围内 SNP 数据, 在野生群体的适应机制研

究中检测受选择信号主要有两类方法： F_{ST} 离群值检测 (outlier analysis, OA) 和环境关联分析 (environmental association analysis, EAA) (Liberles *et al.*, 2020), 其中 F_{ST} 离群值检测的具体流程已在本手册另一篇文章 (金铃和郭宝成, 2021) 中详细介绍。相较于 F_{ST} 离群值检测, 环境关联分析除需基因型数据外, 还需环境变量数据, 通过特定的方法分析某个环境变量与 SNP 等位基因频率变异间的关联, 从而推断选择压力作用于哪些位点或基因组区域 (Rellstab *et al.*, 2015)。

环境关联分析常用的程序或软件包括 BAYENV、BAYENV2、LFMM、BAYESCENV、VEGAN 等, 常见方法及软件的基本原理、适用的数据类型等参见附表 1 (译自 Rellstab *et al.*, 2015 Table 1)。其中 BAYENV 及 BAYENV2 使用频率较高 (Ahrens *et al.*, 2018)。BAYENV 是一种贝叶斯方法, 它通过指定一组中性位点估算群体间等位基因频率变异的零模型, 而后检验加入环境变量的新模型相对于原模型是否可以更好地拟合数据, 从而推断与环境变量相关联的遗传变异 (Coop *et al.*, 2010; Guenther and Coop, 2013)。相较于其他可进行同类分析的方法, BAYENV 具有可对群体间样本大小的差异和群体间共享的群体历史进行校正的优势 (Guenther and Coop, 2013)。本文将 BAYENV2 软件为例, 向读者介绍基于重测序数据利用 BAYENV2 进行环境关联分析的详细分析方案, 该分析方案主要参考 Guo *et al.* (2015, 2016a) 两篇已发表的工作。该分析方案已在鱼类野生群体的适应性机制研究中广泛应用 (如 Magalhaes *et al.*, 2021), 也可应用于鱼类以外的其他生物类群 (如 Guo *et al.*, 2016b; Fraik *et al.*, 2020)。

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)
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. BAYENV2: https://bitbucket.org/tguenther/bayenv2_public/src/master/
8. R v3.6.3: <https://www.r-project.org/>
9. ggplot2 (R package, v3.3.3): <https://cran.r-project.org/web/packages/ggplot2/index.html>
10. scales (R package, v1.1.1): <https://CRAN.R-project.org/package=scales>

实验步骤

1. 产生一个高质量的 SNP 数据集

利用 BWA、SAMtools、BCFtools、VCFtools 等软件，将二代测序平台产生的对研究群体的重测序数据比对至参考基因组上，并进行 SNP 检测、SNP 过滤等，获得一个高质量的 SNP 数据集（一般为 VCF 格式）。详细操作流程参见本手册另一篇文章（金铃和郭宝成，2021）。

2. 格式转换

使用 PGDSpider 软件将 SNP 数据集由 VCF 格式转换为 BAYENV 格式。

```
$ java -jar PGDSpider2-cli.jar -inputfile test.vcf -inputformat VCF -outputfile test.bayenv -outputformat BAYENV -spid VCF_BAYENV.spid
```

注：上述命令不加 `-spid` 参数运行，可自动生成配置文件模板 `template_VCF_BAYENV.spid`，需依据实际数据情况修改该模板中的内容，而后加 `-spid` 参数再次运行上述命令。

3. 准备环境变量文件

在环境变量文件中，依据 SNP 数据集中的群体顺序，列出参与环境关联分析的环境变量的值。每个环境变量必须进行标准化，即减去平均值后除以参与分析群体的标准差。每个环境变量占一行，数值之间以 Tab 键分隔。具体文件格式参考 BAYENV2 软件提供的示例文件 `PCs.env` (https://bitbucket.org/tguenther/bayenv2_public/src/master/example_input.zip)。

注：常见环境变量一般可从公共数据库和网站等获取，例如可获取气候变量的常用公共资源包括 IRI/LDEO Climate Data Library (<http://iridl.ldeo.columbia.edu/>)、WorldClim Database (<https://www.worldclim.org/>) 等。

4. 估算协方差矩阵

输入一组中性位点，估算群体间的协方差矩阵，即估算群体间等位基因频率变化的零模型，用于下一步检测环境相关性中排除掉等位基因频率变异与群体结构、环境或表型变异间假的关联。一般可将 **BayeScan** 或其他软件和方法鉴定到的受选择位点 (如 F_{ST} 离群 SNPs) 从 SNP 数据集中剔除，再进行 SNP 数据集缺失数据、连锁不平衡的过滤。过滤后最终保留几千个 SNPs，协方差矩阵的估算即可收敛。

协方差矩阵的估算通常需要用不同的随机种子 (random seed) 独立运行三次以上，以确保矩阵估算已经收敛。在下一步环境关联分析中，可使用多次协方差矩阵估算结果的平均值。

```
$ java -jar PGDSpider2-cli.jar -inputfile neutral_loci.vcf -inputformat VCF -outputfile neutral_loci.bayenv -outputformat BAYENV -spid VCF_BAYENV.spid
$ bayenv2 -i neutral_loci.bayenv -p 17 -k 50000 -r 3240 > neutral_loci.matrix
```

5. 检测环境相关性

使用 **BAYENV2** 软件自带的脚本 `calc_bf.sh` (https://bitbucket.org/tguenther/bayenv2_public/src/master/calc_bf.sh)，估算每个 SNP 对每个环境变量的贝叶斯因子 (Bayes factor, BF)。输出文件中贝叶斯因子的大小，可代表环境变量与等位基因频率变异之间相关的支持度；贝叶斯因子的值越大，说明该 SNP 与特定环境变量之间的关联越强。

检测环境相关性同样需要用不同的随机种子独立运行三次及以上，取多次运行结果的平均值，或多次分析均鉴定到的环境关联 SNPs。

```
$ calc_bf.sh test.bayenv ENVIRONFILE neutral_loci.matrix.mean 17 50000 2
```

注: `calc_bf.sh` 后的参数分别为 **BAYENV** 格式输入文件、环境变量文件、协方差矩阵文件、群体数目、MCMC 链迭代数目、环境变量数目。为保证脚本正常运行，上述参数应严格按照以上顺序添加至脚本后。

6. 结果分析及绘图

生成文件 `bf_environs.ENVIRONFILE` 第一列为 SNP ID，后续几列为该 SNP 与各个环境变量间的贝叶斯因子。该文件可用于绘制 SNP 等位基因频率变异与环境变量间相关联的曼哈顿图或频率分布直方图。下文中的图 1 和图 2 是用 Guo *et al.* (2016a) 中 17 个波罗的海大西洋鲱鱼群体对盐度的环境关联分析的结果生成的

(<http://datadryad.org/review?doi=doi:10.5061/dryad.d8b93>), 供读者参考。

- 1) 绘制曼哈顿图，以各个 SNP ID 为横坐标，以贝叶斯因子以 10 为底的对数为纵坐标，以 $\log_{10}[\text{Bayes factor (BF)}] > 1.5$ 作为判定 SNP 是否与该环境变量强相关的阈值 (Jeffreys, 1961) (图 1)。

加载 R 包 ggplot2:

```
> library(ggplot2)
```

读取 BAYENV2 输出文件:

```
> bf <- read.table("bf_envIRON.ENVIROnFILE")
```

```
> colnames(bf) <- c("SNP_ID","Salinity")
```

生成一系列 SNP 序号:

```
> bf$NUM <- c(1:nrow(bf))
```

绘图:

```
> man_plot <- ggplot(bf,aes(x=NUM,y=log10(Salinity)))+
  geom_point(size=2)+
  geom_hline(aes(yintercept=1.5),size=1.5,linetype="dashed")+
  xlab("SNPs")+
  ylab(expression(log[10](BF)))+
  ggtitle("Salinity")+
  theme_bw()+
  theme(
    axis.text=element_text(size=30),
    axis.title=element_text(size=35),
    plot.title=element_text(size=35,hjust=0.5),
    panel.border=element_rect(size=1.5),
    panel.grid=element_blank()
  )
```

保存图像:

```
> ggsave("man_plot.pdf",plot=man_plot,width=28,height=6)
```

图 1. SNP 等位基因频率变异与环境变量相关联的曼哈顿图 虚线表示 $\log_{10}[\text{Bayes factor (BF)}] = 1.5$ 的阈值。

- 2) 绘制 SNP 等位基因频率变异与环境变量相关联的频率分布直方图, 从该图可以看出贝叶斯因子值的分布 (图 2)。

加载 R 包 scales:

```
> library(scales)
```

绘图:

```
> freq_plot <- ggplot(bf,aes(x=log10(Salinity)))+
  geom_histogram(bins=100,aes(y=after_stat(count/sum(count)*100)),
    color="#BABABA",fill="#BABABA")+
  xlab(expression(log[10](BF)))+
  ylab("Frequency (percentage)")+
  ggtitle("Salinity")+
  theme_classic()+
  scale_x_continuous(limits=c(-1,1.5),oob=squish)+
  theme(axis.text=element_text(size=20),
    axis.title=element_text(size=20),
    plot.title=element_text(size=30,hjust=0.5),
    panel.grid=element_blank()
  )
```

保存图像:

```
> ggsave("freq_plot.pdf",plot=freq_plot,width=8,height=8)
```

图 2. SNP 等位基因频率变异与环境变量间相关联的频率分布直方图

以 $\log_{10}[\text{Bayes factor (BF)}] > 1.5$ 为标准，可统计与特定环境变量强相关的 SNP 数目，这些 SNPs 通常被认为与群体的本地适应相关。在实践中，环境关联分析可与 F_{ST} 离群值分析（详见本手册另一篇文章（金铃和郭宝成, 2021））等结合进行。

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. 金铃, 郭宝成. (2021). 群体基因组水平的适应性位点检测. *Bio-101*.
2. 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.

3. Coop, G., Witonsky, D., Di Rienzo, A. and Pritchard, J. K. (2010). [Using environmental correlations to identify loci underlying local adaptation](#). *Genetics* 185(4): 1411-1423.
4. Fraik, A. K., Margres, M. J., Epstein, B., Barbosa, S., Jones, M., Hendricks, S., Schönfeld, B., Stahlke, A. R., Veillet, A., Hamede, R., McCallum, H., Lopez-Contreras, E., Kallinen, S. J., Hohenlohe, P. A., Kelley, J. L. and Storfer, A. (2020). [Disease swamps molecular signatures of genetic-environmental associations to abiotic factors in Tasmanian devil \(*Sarcophilus harrisii*\) populations](#). *Evolution* 74(7): 1392-1408.
5. Gunther, T. and Coop, G. (2013). [Robust identification of local adaptation from allele frequencies](#). *Genetics* 195(1): 205-220.
6. 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.
7. 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.
8. 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.
9. Jeffreys, H. (1961). [Theory of probability](#). Oxford University Press. London. ISBN: 0-19-850368-7.
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. Magalhaes, I. S., Whiting, J. R., D'Agostino, D., Hohenlohe, P. A., Mahmud, M., Bell, M. A., Skulason, S. and MacColl, A. D. C. (2021). [Intercontinental genomic parallelism in multiple three-spined stickleback adaptive radiations](#). *Nat Ecol Evol* 5(2): 251-261.
12. Pardo-Diaz, C., Salazar, C. and Jiggins, C. D. (2015). [Towards the identification of the loci of adaptive evolution](#). *Methods Ecol Evol* 6(4): 445-464.
13. Rellstab, C., Gugerli, F., Eckert, A. J., Hancock, A. M. and Holderegger, R. (2015). [A practical guide to environmental association analysis in landscape genomics](#). *Mol Ecol* 24(17): 4348-4370.

附表 1. 可进行环境关联分析的方法和软件概览 其中一些方法可用其他软件或 R 包实现。(译自 Rellstab *et al.*, 2015 Table 1)

方法	参考文献	关联类型	采样策略	是否考虑了中性遗传结构	是否考虑了空间自相关	个体或群体数据	是否有混样数据模式	是否校正样本大小	软件/R 包
分类 (Categories)		分类型	分类型	可能	可能	均可	可能	可能	多种统计方法
空间分析方法 (Spatial analysis method, SAM)	Joost <i>et al.</i> (2007)	逻辑斯蒂型	梯度型/分散型	可能 (在 SAMβADA 中)	可能 (在 SAMβADA 中)	个体	否	否	SAM (Joost <i>et al.</i> , 2008), SAMβADA (Stucki <i>et al.</i> , 2017)
多重逻辑回归 (Multiple logistic regression)		逻辑斯蒂型	梯度型/分散型	可能	可能	个体	否	否	R (R development Core Team 2011)
广义估计方程 (Generalized estimating equations, GEEs)	Carl and Kuhn (2007)	逻辑斯蒂型	梯度型/分散型	否	是	个体	否	否	Geepack (Yan and Fine, 2004)
Partial Mantel test	Smouse <i>et al.</i> (1986)	线性/秩线性	梯度型/分散型	是	可能	均可	否	否	ECODIST (Geoslee and Urban, 2007); VEGAN (Oksanen <i>et al.</i> , 2013)

多重线性回归/一般线性模型 (Multiple linear regression/General linear models)		线性	梯度型/分散型	可能	可能	均可	否	否	R (R Development Core Team 2011); TASSEL (Bradbury <i>et al.</i> , 2007)
典范相关分析 (Canonical correlation analysis, CCA)	Legendre and Legendre (2012)	线性	梯度型/分散型	可能	可能	均可	否	否	VEGAN (Oksanen <i>et al.</i> , 2013)
(偏)冗余分析 ((Partial) redundancy analysis, RDA)	Legendre and Legendre (2012)	线性	梯度型/分散型	可能	可能	均可	否	否	VEGAN (Oksanen <i>et al.</i> , 2013)
BAYENV	Coop <i>et al.</i> (2010)	线性/秩线性	梯度型/分散型	是	否	群体	是 (在 BAYENV2 中)	是	BAYENV (Coop <i>et al.</i> , 2010); BAYENV2 (Günther and Coop, 2013)

空间广义线性混合模型 (Spatial generalized linear mixed model, SGLMM)	Guillot <i>et al.</i> (2014)	线性	梯度型/分散型	是	是	均可	否	是	Ginland (Guillot <i>et al.</i> , 2014)
潜在因子混合模型 (Latent factor mixed models, LFMMs)	Frichot <i>et al.</i> (2013)	线性	梯度型/分散型	是	否	均可	否	否	LFMM (Frichot <i>et al.</i> , 2013); LEA (Frichot and Francois, 2015)
GWAS 混合模型 (GWAS mixed models)		线性	梯度型/分散型	是	否	个体	否	否	EMMA (Kang <i>et al.</i> , 2008); TASSEL (Bradbury <i>et al.</i> , 2007); LME4 (Bates <i>et al.</i> , 2014)
基于 F_{ST} 的方法	de Villemereuil and Gaggiotti (2015)	基于分化的	梯度型/分散型	是	否	均可	否	是	BAYESCENV (de Villemereuil and Gaggiotti, 2015)