基因组水平的近交系数分析

Genomic Inbreeding Coefficient Analysis

匡卫民，胡靖扬，吴宏，于黎*

云南大学生命科学学院，省部共建云南生物资源保护与利用国家重点实验室，昆明 650091

*通讯作者邮箱：yuli@ynu.edu.cn

引用格式：匡卫民，胡靖扬，吴宏，于黎. (2021). 基因组水平的近交系数分析. Bio-101 e1010603. Doi: 10.21769/BioProtoc. 1010603.

How to cite: Kuang, W. M., Hu, J. Y., Wu, H. and Yu, L. (2021). Genomic Inbreeding Coefficient Analysis. Bio-101 e1010603. Doi: 10.21769/BioProtoc.1010603. (in Chinese)

摘要：在基因组水平上，连续性纯合片段（runs of homozygosity；ROHs）是指染色体的某一段区域内存在的连续纯合状态现象。ROHs 的累积长度可以用来估计个体的近交系数。ROHs 是在双亲传递祖先单倍型相同的两个拷贝的过程中产生的，长单倍型片段来源于最近共同祖先，反映较为近期的近交事件，而短单倍型片段来源于亲缘关系较远的共同祖先，反映较早期的近交事件。通过计算 ROHs 片段长度总和占整个基因组长度的比例来代表近交系数（F_{ROH}）。目前，PLINK 软件广泛应用在 ROHs 的检测分析中，该方法是在个体水平的基因组上通过设定杂合基因型和允许缺失基因型的数量来计算连续性纯合基因型间隔的片段长度。

关键词：基因组，近交系数，ROHs，PLINK

软件运行环境及信息

1. 文本中所有软件的运行及数据处理和分析均在 Linux 操作系统环境上进行（相关学习教程请参考：http://linux.vbird.org）
2. PLINK v1.90 (www.cog-genomics.org/plink/1.9；Purcell *et al.*, 2007)
3. VCFtools v0.1.16 (https://vcftools.github.io/man_latest.html；Danecek *et al.*, 2011)
4. R v.4.0.2 (https://cran.r-project.org/src/base/R-4)

实验步骤

一、准备 VCF 文件

本实验教程需要读者提前准备好记录全基因组的单核苷酸多态性位点的VCF 文件。VCF 文本文件格式请参考和阅读 https://github.com/samtools/hts-specs，VCF 文件命名为 Chr1.vcf。

注：为了准确评估近交程度和基因组上 ROHs 的分布情况，建议读者选择拼装到染色体水平的基因组作为参考序列。GATK 的变异检测和过滤流程请参考和阅读 GATK 官方说明书（https://gatk.broadinstitute.org/hc/en-us#variant-discovery-ovw）或其他实验流程，本实验不叙述这部分的步骤。

二、使用 VCFtools 进行格式转换

利用 VCFtools 软件将 VCF 文件转换为 ped 和 map 文件格式。

vcftools --vcf Chr1.vcf --plink --out Chr1

运行结束后会生成两个文件：Chr1.ped 和 Chr1.map。

plink --noweb --file Chr1 –make-bed --out Chr1

运行结束后会生成三个文件：Chr1.bed, Chr1.bim, Chr1.fam。

注：输入文件和输出文件均不需要文件的后缀。

三、计算 ROHs

PLINK 软件检测基因组中的纯合片段是基于基因型计数的方法，即通过设定杂合子最大数量和允许缺失基因型的数量，在基因组上鉴定连续纯合基因型间隔的长度。

在 Linux 环境中，输入 plink --homozyg --help，如下：

```
PLINK v1.90p 64-bit (21 Jan 2020)          www.cog-genomics.org/plink/1.9/
(C) 2005-2020 Shaun Purcell, Christopher Chang   GNU General Public License v3

  plink <input flag(s)...> [command flag(s)...] [other flag(s)...]
  plink --help [flag name(s)...]

Commands include --make-bed, --recode, --flip-scan, --merge-list,
--write-snplist, --list-duplicate-vars, --freqx, --missing, --test-mishap,
--hardy, --mendel, --ibc, --impute-sex, --indep-pairphase, --r2, --show-tags,
--blocks, --distance, --genome, --homozyg, --make-rel, --make-grm-gz,
--rel-cutoff, --cluster, --pca, --neighbour, --ibs-test, --regress-distance,
--model, --bd, --gxe, --logistic, --dosage, --lasso, --test-missing,
--make-perm-pheno, --unrelated-heritability, --tdt, --dfam, --qfam, --tucc,
--annotate, --clump, --gene-report, --meta-analysis, --epistasis,
--fast-epistasis, and --score.
```

图 1. PLINK v1.90 运行参数示例。

通常设置--homozyg-snp 、--homozyg-kb、--homozyg-window-het、--homozyg-window-missing 四个参数即可，其余为默认参数设置。

单独计算每条染色体的 ROHs，命令如下：

plink --file tmp.ped --homozyg --homozyg-window-snp 20 --homozyg-kb 10 --allow-no-sex --noweb --homozyg-window-het 1 --homozyg-window-missing 20 –out Chr1

输出文件：Chr1.hom；Chr1.hom.indiv；Chr1.hom. summary

将每个个体的所有染色体的.hom 结果文件合并到一个文件内：

cat Chr*.hom > AllChr.hom

四、统计 ROHs 和计算近交系数

参考 (McQuillan *et al.*, 2008) 的近交系数 (F_{ROH}) 计算方法，计算公式为 $F_{ROH} = L_{ROH}/L_{auto}$，其中 L_{ROH} 表示每个个体 ROHs 的总长度，L_{auto} 表示 SNPs 覆盖到的常染色体总长度。根据读者的需求，按照对不同长度的 ROH 进行分别统计，如统计 ROH > 2.5 Mb 的总长度：

awk '$9>2500{sum+=$9}END{print sum}' AllChr.hom > AllChr_GT2.5Mb.hom

ROH 长度的选择根据读者需要进行设置，一般按 ROHs 不同的长度分为短 ROHs (< 200 kb)，中等 ROHs (200 kb-2.5 Mb)，超长 ROHs (> 2.5 Mb)。ROHs 的丰度和长

短分布可以为种群进化历史提供额外的信息。短 ROHs 可能是古老的近交事件遗留的结果，因为重组事件会将 ROH 打断成碎片，超长 ROHs 能直接反映最近的近交事件。然而，中等 ROHs 通常被忽略掉，因为很难辨别是背景相关的结果还是古老的种群瓶颈的结果 (Díez-del-Molino *et al.,* 2018)。另外，可以通过设定 ROH 固定的长度估计最近的近交事件发生的大致时间，比如 2.5 Mb 的 ROH 大约发生在 20 个世代以内(g = 100/2*ROH_{length}，其中 g 为世代数，ROH_{length} 代表检测的 ROH 遗传距离长度，ROH 遗传距离近似物理距离) (Van Der Valk *et al.,* 2019)。

另外，在以上单条染色体计算流程的基础上，我们提供了对基因组 22 条染色体批量计算的 shell 脚本供读者参考，如下：

```
#! /bin/bash
for i in {1..22}  ##批量对 22 条染色体进行计算
do vcftools --vcf Chr$i.vcf --plink --out Chr$i
plink --noweb --file Chr$i --make-bed --out Chr$i
plink --file Chr$i.ped --homozyg --homozyg-window-snp 20 --homozyg-kb 10 --allow-no-sex --noweb --homozyg-window-het 1 --homozyg-window-missing 20 --out Chr$i
done
cat Chr*.hom > AllChr.hom  ##合并 22 条染色体的 ROHs 结果
awk '$9>2500{sum+=$9}END{print sum}' AllChr.hom > AllChr_GT2.5Mb.hom
## ROH>2.5Mb 的总长度
```

五、呈现和解析 ROHs 结果

ROHs 统计和呈现方式多种多样，读者可根据自己的需求进行结果展示。下面列举两种呈现方式供读者参考 (图 2 和图 4)。

1. 不同长度的 ROH 计算的 F_{ROH} 值的统计和描述，示例如下：

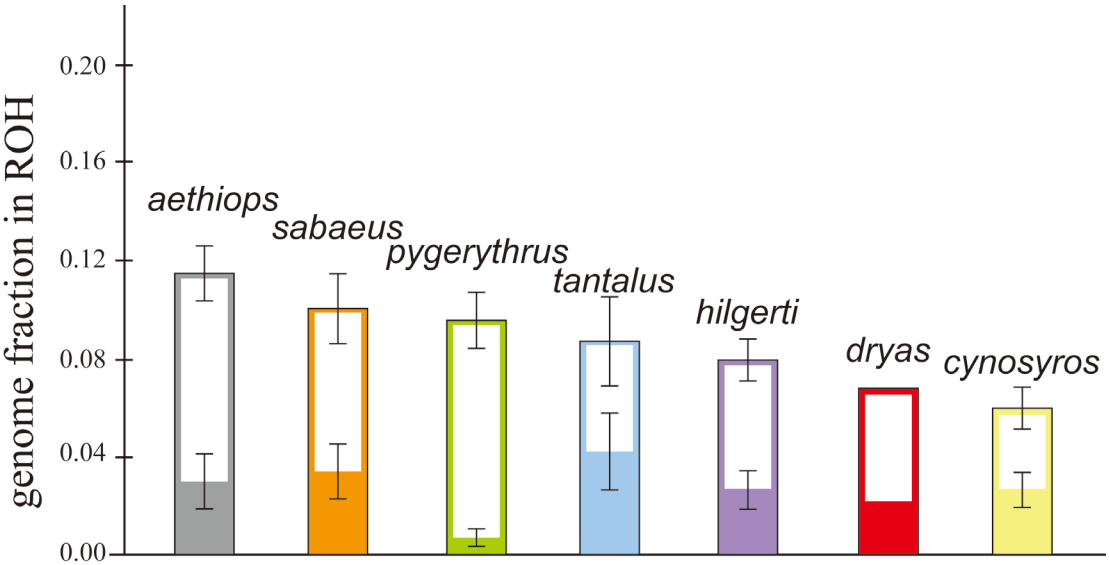

图 2. F_{ROH} > 100 kb (中空条形图) 和 F_{ROH} > 2.5 Mb (非中空条形图) (图片来源于 Van Der Valk *et al.*, 2020 中的 Figure 4B)

2. 每条染色体上 ROHs 的分布密度和集中区域。我们以 (Kuang *et al.*, 2020) 发表的怒江金丝猴 (*Rhinopithecus strykeri*) 数据为例，进行绘图。首先，提前准备好以下格式文件 (test.bed)，第一列为样品编号，第二列为染色体编号，第三列为 SNP1 起始位置，第四列为 SNP2 终止位置，第五列为 ROH 的长度 (kb)。

```
R.strykeri-01   CM017351.1     3807643 5282306 1474.664
R.strykeri-01   CM017351.1     6308741 7373516 1064.776
R.strykeri-01   CM017351.1     7763826 9232327 1468.502
R.strykeri-01   CM017351.1     12590063        13602052        1011.990
R.strykeri-01   CM017351.1     17668002        19625723        1957.722
R.strykeri-01   CM017351.1     25153493        26602426        1448.934
R.strykeri-01   CM017351.1     28622294        29803411        1181.118
R.strykeri-01   CM017351.1     30796982        32028220        1231.239
R.strykeri-01   CM017351.1     40333943        41385374        1051.432
R.strykeri-01   CM017351.1     45028692        46649663        1620.972
R.strykeri-02   CM017351.1     9860350 10923980        1063.631
R.strykeri-02   CM017351.1     24976105        26188455        1212.351
R.strykeri-02   CM017351.1     37436853        38626796        1189.944
R.strykeri-02   CM017351.1     38626893        39889405        1262.513
R.strykeri-02   CM017351.1     87623418        88660653        1037.236
R.strykeri-02   CM017352.1     68503769        69590607        1086.839
R.strykeri-02   CM017352.1     72739748        74382749        1643.002
R.strykeri-02   CM017352.1     83350802        84387353        1036.552
R.strykeri-02   CM017352.1     85339984        86631759        1291.776
R.strykeri-02   CM017352.1     90029727        91585062        1555.336
```

图 3. 绘制每条染色体上 ROHs 分布密度图的输入文件格式。

然后，在 R 中进行绘图，R 代码如下：

Library (CMplot) ##提前安装 CMplot 包,安装命令为 install.packages ("CMplot")

ROH<-read.table ("test.bed",header=TRUE)

df1<-data.frame (ROH)

CMplot (df1,type="p", plot.type="d",bin.size=1e3,chr.den.col=c ("yellow", "red"), file="pdf", memo="", dpi=300, bin.range=c (1,10), file.output=TRUE, verbose=TRUE, width=9,height=6)

图 4. 怒江金丝猴物种的 ROHs 在染色体上的分布情况 (以川金丝猴基因组为参考序列)。不同颜色代表在染色体上覆盖到同一 ROH 上的个体数目。

参考文献：

1. Danecek, P., Auton, A., Abecasis, G., Albers, C.A., Banks, E., Depristo, M.A., Handsaker, R. E., Lunter, G., Marth, G.T., Sherry, S.T., McVean, G., Durbin, R. 2011. The variant call format and VCFtools. *Bioinformatics.* 27(15):2156-2158.
2. Díez-del-Molino, D., Sánchez-Barreiro, F., Barnes, I., Gilbert, M.T.P., and Dalén, L. 2018. Quantifying temporal genomic erosion in endangered species. *Trends Ecol Evol.* 33(3): 176-185.
3. Kuang, W.M., Hu, J.Y., Wu, H., Fen. X.T., Dai, Q.Y., Fu, Q.M., Xiao, W., Frantz, L., Roos, C., Nadler, T., Irwin, D.M., Zhou, L.C, Yang, X., Yu L. 2020. Genetic Diversity,

Inbreeding Level, and Genetic Load in Endangered Snub-Nosed Monkeys (*Rhinopithecus*). *Frontiers in Genetics*. 11(1574):1574-1583.

4. McQuillan, R., Leutenegger, A.L., Abdel-Rahman, R., Franklin, C.S., Pericic, M., Barac-Lauc, L., SmolejNarancic, N., Janicijevic, B., Polasek, O., Tenesa, A. 2008. Runs of homozygosity in European populations. *Am. J. Hum. Genet*. 83(3):359–372.
5. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A., Bender, D., Maller, J.L., Sklar, P., de Bakker, P.I.W., Daly, M.J., Sham, P.C. 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet.* 81(3):559-575.
6. Robinson, JA., Räikkönen, J., Vucetich, L.M., Vucetlch, J.A., Peterson, RO., Lohmueller, K.E., Wayne, R.K. 2019. Genomic signatures of extensive inbreeding in Isle Royale wolves, a population on the threshold of extinction. *Sci Adv.* 5(5): eaau0757.
7. Van Der Valk, T., Diez-Del-Molino, D., Marques-Bonet, T., Guschanski, K., and Dalen, L. 2019. Historical genomes reveal the genomic consequences of recent population decline in eastern gorillas. *Current biology.* 29(1): 165-170 e166
8. Van Der Valk, T., Gonda, C.M., Silegowa, H., Almanza, S., Sifuentes-Romero, I., Hart, T.B., Detwiler, K.M., Guschanski, K. 2020. The genome of the endangered dryas monkey provides new insights into the evolutionary history of the vervets. *Mol Biol Evol.* 37(1):183-194.

