基因组水平的遗传负荷分析

Genomic Genetic Load Analysis

胡靖扬, 匡卫民, 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

*通讯邮箱: yuli@ynu.edu.cn

引用格式: 胡靖扬, 匡卫民, 吴宏, 于黎. (2021). 基因组水平的遗传负荷分析. Bio-101 e1010600.

Doi: 10.21769/BioProtoc.2001010600.

How to cite: Hu, J. Y., Kuang, W. M., Wu, H. and Yu, L. (2021). Genomic Genetic Load Analysis. Bio-101 e1010600. Doi: 10.21769/BioProtoc.1010600. (in Chinese)

摘要: 有害突变会破坏基因功能从而降低个体对当前环境的生存适应能力产生遗传负荷 (genetic load)。基因组水平的遗传负荷是评估生物群体生存潜力的重要指标。有害突变主要包括有害的错义突变 (deleterious missense mutations) 以及造成基因功能缺失 (loss of function) 的突变。常用的 SnpEff 软件能够对群体中每个个体编码区的突变类型进行预测。该方法通过统计纯合等位基因和杂合等位基因突变数量, 评估目标群体的遗传负荷。此外, 通过基因功能注释, 确定产生遗传负荷相关基因的功能, 进一步评估遗传负荷对当前生物群体造成的潜在影响。

关键词: 遗传负荷, 有害突变, 有害错义突变, 功能缺失突变, SnpEff

软件运行环境及信息

1. 本文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 (<http://linux.vbird.org>)
2. SnpEff v.4.3t (<http://snpeff.sourceforge.net/SnpEff.html>)
3. Vcftools v0.1.16 (https://vcftools.github.io/man_latest.html)
4. Grantham_score_calculator (https://github.com/ashutoshkpandey/Annotation/blob/master/Grantham_score_calculator.py)
5. JAVA v.1.8 (<https://www.java.com/>)

6. Python v.2.7 (<https://www.python.org/>)

实验步骤

1. 获取数据

本实验方案需要获取相关研究物种的参考基因组和注释信息，同时还需要获取该物种的基因组 SNP 数据来开展遗传负载分析。首先，读者需要从存储原始数据的 Fastq 文件中获取存储变异位点信息 SNP 的 VCF 文件。详细流程可参考“GATK4.0 和全基因组数据分析实践” <https://www.plob.org/article/11630.html>。

2. 下载 SnpEff 软件 (Cingolani *et al.* 2012)

#下载软件的最新版本

```
wget
```

```
https://snpeff.blob.core.windows.net/versions/snpeff_latest_core.zip
```

利用 unzip 将软件直接解压到新建的 snpEff 文件夹中就可以使用。SnpEff 软件是基于 JAVA 开发的软件，所以解压之后可以直接运行 `java -jar snpEff.jar`，以确认你的系统里面是否安装有配套的 JAVA 1.8 及其以上的版本，另外还可以查看具体的使用方法。

3. 构建研究物种的参考基因组数据库

SnpEff 软件可以直接使用已经构建好的数据库进行数据分析。通过 `databases` 命令，查找可供下载的数据库，以 human 为例，运行以下命令：

```
java -jar snpEff.jar databases | grep -i human
```

可以相应获得 GRCh37 和 GRCh38 两个不同版本的数据库信息，根据实验需要进行下载使用。

但是基因组数据和注释信息经常有更新，新测序的物种没有现成的数据库供使用，因此常常需要自己构建参考物种(reference)的数据库。首先将参考物种的基因组数

据和基因组注释信息分别更名为 **sequences.fa** 和 **genes.gff**，存放在 **/snpEff/data/ref_1/**下面，然后在 **snpEff.config** 配置文件中添加参考物种基因组和版本信息(分别用 **ref_1.genome** 和 **ref_1** 命名)并指定数据库的绝对路径：

```
#genome, version
ref_1.genome : ref_1
data.dir = /snpEff/data/
```

最后运行 **java -jar snpEff.jar build -gff3 -v ref_1**，进行数据库构建，构建成功之后会在 **/snpEff/data/ref_1/**下面生成 **snpEffectPredictor.bin** 文件，表示数据库构建成功。

4. 注释不同个体的 SNP 信息

4.1 定义祖先等位基因组型

将与外群相同的等位基因定义为祖先等位基因，如果研究类群没有合适的外群，可以考虑将所有个体中主要的等位基因型(即超过 50%的个体都是纯合的等位基因)定义为祖先型，也可以考虑整合以上两个条件。将能够确定祖先等位基因的位点利用 **vcftools** (Danecek *et al.* 2011)中的**--positions** 进行提取，获得总 VCF 文件 **all.vcf**。

4.2 将每个个体的 VCF 从总 VCF 文件中分离出来

```
vcftools --vcf all.vcf --indv individual01 --recode --out
individual01
```

4.3 注释每个个体的 VCF 文件

```
java -jar snpEff.jar ref_1 individual01.recode.vcf
>individual01.ann
```

4.4 将每个注释结果划分为纯合(0/0 或者 1/1)和杂合(0/1)

```
grep "0/0" individual01.ann > individual01.ann.hom
grep "1/1" individual01.ann >> individual01.ann.hom
```

```
grep "0/1" individual01.ann > individual01.ann.het
```

5. 统计注释结果

5.1 SnpEff 软件将突变类型划分为同义突变(synonymous_variant)， 错义突变(missense_variant)和功能缺失突变(loss of function_variant; LOF)。因此，可以将注释结果中的纯合突变和杂合突变分别划分到不同的突变类型，并进行统计。

5.2 功能缺失突变主要包括翻译提前终止(stop_gained)，翻译起始缺失(start_lost)，剪切受体变异(splice_acceptor_variant)，剪切供体变异(splice_donor_variant)等，会直接造成基因原有功能的丢失，影响基因的表达，因此是最直接的遗传负荷。

5.3 错义突变是否是有害突变遗传负荷，可以通过 **Grantham Score (Grantham. 1974)**，即氨基酸变化的物理/化学性质改变的量化指标进行判断。通常 **Grantham Score** 大于或等于 150 被认为是有害突变遗传负荷。将每个错义突变位点利用 Python 程序 **Grantham_score_calculator.py**，计算其 **Grantham Score**，并统计大于 150 的位点，定义为错义有害突变遗传负荷。

6. 计算不同突变类型的遗传负荷

统计和计算纯合等位基因突变数量占有所有纯合和杂合突变等位基因数量的比率，来评估群体的遗传负荷。其中纯合等位基因数量是纯合位点数的两倍，杂合等位基因数量与杂合位点数相等。

7. 基因注释分析

对高频且纯合等位基因中功能缺失突变相关基因(LOF 基因)进行功能注释。将 LOF 基因与人或小鼠进行同源基因确定，获得相关基因列表。打开 **DAVID** 数据库 (<https://david.ncifcrf.gov/>)，选择 **Start Analysis**，将基因列表导入，根据基因列表选择相应的基因类别(Select Identifier)，如果使用的是 **ENSEMBL** 的基因编号则选择 **ENSEMBL_GENE_ID**，List Type 选择 **Gene List**，最后选择 **Submit List**。结果出来之后，对应查看相应的 **GO** 和 **KEGG** 功能富集分析结果。

方法运用

该流程已经成功运用在濒危动物穿山甲的遗传负荷检测中 (Hu *et al.* 2020)。研究结果发现穿山甲的遗传负荷显著增加，其中 LOF 基因主要富集到与代谢和疾病相关的通路。

参考文献

1. Cingolani, P., Platts, A., Wang le, L., Coon, M., Nguyen, T., Wang, L., Land, S. J., Lu, X. and Ruden, D. M. (2012). [A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3.](#) *Fly (Austin)* 6(2): 80-92.
2. 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. and Genomes Project Analysis, G. (2011). [The variant call format and VCFtools.](#) *Bioinformatics* 27(15): 2156-2158.
3. Grantham, R. (1974). [Amino acid difference formula to help explain protein evolution.](#) *Science* 185(4154): 862-864.
4. Hu, J.Y., Hao, Z.Q., Frantz, L., Wu, S.F., Chen W, Jiang, Y.F., Wu, H., Kuang, W.M., Li, H.P., Zhang, Y.P., Yu, L. (2020). [Genomic consequences of population decline in critically endangered pangolins and their demographic histories.](#) *Natl Sci Rev* 7:798-814.