

使用abagen包提取AHBA基因表达数据

Alex / 2023-06-26 / free_learner@163.com / learning-archive.org

一、背景

得益于Allen Human Brain Atlas (AHBA)基因数据库的发布，研究者可以将磁共振脑成像数据与基因数据关联起来。同磁共振数据一样，在进行统计分析之前，基因数据需要进行一系列的处理，这里介绍使用Python环境下abagen包提取AHBA基因表达数据的基本方法。所有内容来自于[官方文档](#)和相应的论文：

Markello, R. D., Arnatkeviciute, A., Poline, J. B., Fulcher, B. D., Fornito, A., & Misic, B. (2021). Standardizing workflows in imaging transcriptomics with the abagen toolbox. *elife*, 10, e72129.

Arnatkeviciute, A., Fulcher, B. D., & Fornito, A. (2019). A practical guide to linking brain-wide gene expression and neuroimaging data. *Neuroimage*, 189, 353-367.

二、AHBA基因数据库

AHBA数据库包含6名捐献者，其中2名捐献者采集了双侧大脑的样本，4名捐献者采集了左侧大脑的样本。总共包含3702个样本，分布在皮层、皮层下、小脑和脑干区域，并提供了每个样本的坐标。每个样本通过微阵列（microarray）来测量基因表达水平，微阵列上有很多探针（probe），每个探针对应于一种基因，总共涉及20000多个基因。需要注意的是，一个微阵列中多个探针可能测量的是同一个基因，也就是说探针是重复的。所以，AHBA数据库的基本结构可以理解为 样本*探针 的矩阵，样本来自于不同捐献者的不同脑区，探针对应不同的或者相同的基因。由于我对于基因数据了解甚少，这里的描述可能理解有误，请参考前面的文献资料。

三、使用abagen

1. 安装abagen

Python版本要求3.6+，我这里测试的Python版本是3.7.16，abagen版本是0.1.3。

```
pip install abagen
```

2. 使用命令行提取数据

```
abagen --output_file=/home/alex/expression.csv --data_dir=/home/alex/ahba \  
--atlas_info=/home/alex/myatlas_info.csv /home/alex/myatlas.nii.gz
```

这里 `expression.csv` 是最终提取的数据，结构为 脑区*基因，每行表示一个脑区，用数字标签来标识，每列表示一个基因；`--data_dir` 选项表示AHBA数据的路径，如果第一次运行会自动下载数据，需要花较长时间；`--atlas_info` 表示分区模板的信息（如下图所示），具体地，包含3列，第一列是脑区数字标签，第二列是每个脑区位于左半球还是右半球等，第三列是每个脑区属于什么结构，比如皮层或者皮层下等。最后一个参数是分区模板，我这里测试的是体积空间的分区模板（要求位于MNI152空间），也可以提供皮层空间分区模板（要求位于fsaverage5空间）。

id	hemisphere	structure
1	L	cortex
2	R	cortex
3	L	cortex
4	R	cortex
5	L	cortex
6	R	cortex

如果分区模板中某些脑区较小，可能会存在空值的情况，也就是没有匹配的基因数据。这个时候可以考虑设置 `--missing` 选项，另外，根据使用文档中的建议，如果使用 `--missing` 选项，可以加上 `--norm_all`，比如：

```
abagen --output_file=expression.csv --data_dir=/home/alex/ahba \  
--atlas_info=/home/alex/myatlas_info.csv \  
--missing=centroids --norm_all /home/alex/myatlas.nii.gz
```

3. 使用脚本提取数据

```
import abagen  
expression, report= abagen.get_expression_data('/home/alex/myatlas.nii.gz',  
                                              atlas_info='/home/alex/myatlas_info.csv',  
                                              missing='centroids', norm_matched=False,  
                                              return_report=True,  
                                              data_dir='/home/alex/ahba')  
expression.to_csv('/home/alex/expression.csv', header=True, index=True)
```

相比于命令行的方式，脚本的方式可以输出一个处理流程的报告（如下图所示）。此外，除了提取每个脑区的基因表达数据，也可以以样本为单位提取数据，具体情况请参考官方使用文档。

Regional microarray expression data were obtained from 6 post-mortem brains (1 female, ages 24.0--57.0, 42.50 +/- 13.38) provided by the Allen Human Brain Atlas (AHBA, <https://human.brain-map.org>; [H2012N]). Data were processed with the abagen toolbox (version 0.1.3; <https://github.com/rmarkello/abagen>) using a 246-region volumetric atlas in MNI space.

First, microarray probes were reannotated using data provided by [A2019N]; probes not matched to a valid Entrez ID were discarded. Next, probes were filtered based on their expression intensity relative to background noise [Q2002N], such that probes with intensity less than the background in $\geq 50.00\%$ of samples across donors were discarded, yielding 31,569 probes. When multiple probes indexed the expression of the same gene, we selected and used the probe with the most consistent pattern of regional variation across donors (i.e., differential stability; [H2015N]), calculated with: