**基于****PulseDIA技术的搜库结果预处理使用说明书**

**软件名称：** **PulseDIA expert**

**版本号：version 1.0**

**一、功能简介**

PulseDIA expert是一款专门用于DIA-NN, Spectronaut和OpenSWATH这三种流行的搜库软件对pulseDIA质谱数据的搜库结果进行自动化处理的程序。

对于pulseDIA质谱数据的搜库，一般分不同part分别进行搜库，最后再将同一样本的不同part的定量结果整合在一起。用户仅需提供各个搜库软件直接导出的蛋白鉴定信息表，设置想要的参数，即可快速得到不同part的定量整合结果。

**二、使用说明**

PulseDIA expert主要是用来对DIA-NN搜库结果进行protein与peptide矩阵信息的提取，然后再将不同part文件的按peptide矩阵或protein矩阵格式进行整合，一般取同一样本不同part的intensity的平均值或最大值作为该样本的定量结果。对于OpenSWATH和Spectronaut的搜库结果，PulseDIA expert仅提供不同part的peptide矩阵信息的合并，如果想导出合并后的protein矩阵，建议再使用已发表的ProteomeExpert's网站上的pep2prot工具将peptide矩阵转化成protein矩阵。

pulse\_dia\_combine.R程序适用于Windows、Linux、Mac OS等操作系统。下述使用说明及案例演示均基于Windows操作系统。

**2.1 环境变量配置**

      在软件使用之前，需要先确保R软件正确安装在自己的Windows电脑端，然后再进行环境变量的配置。配置环境变量的步骤如下：

       1）打开本地电脑的"高级系统设置"；

       2）点击系统属性对话框中的环境变量选项；

       3）将R软件的安装路径复制到用户环境变量或系统环境变量中的Path路径下；

       4）在cmd命令提示符或powershell中输入R，回车，如果提示R相关的版本信息，表明配置成功。

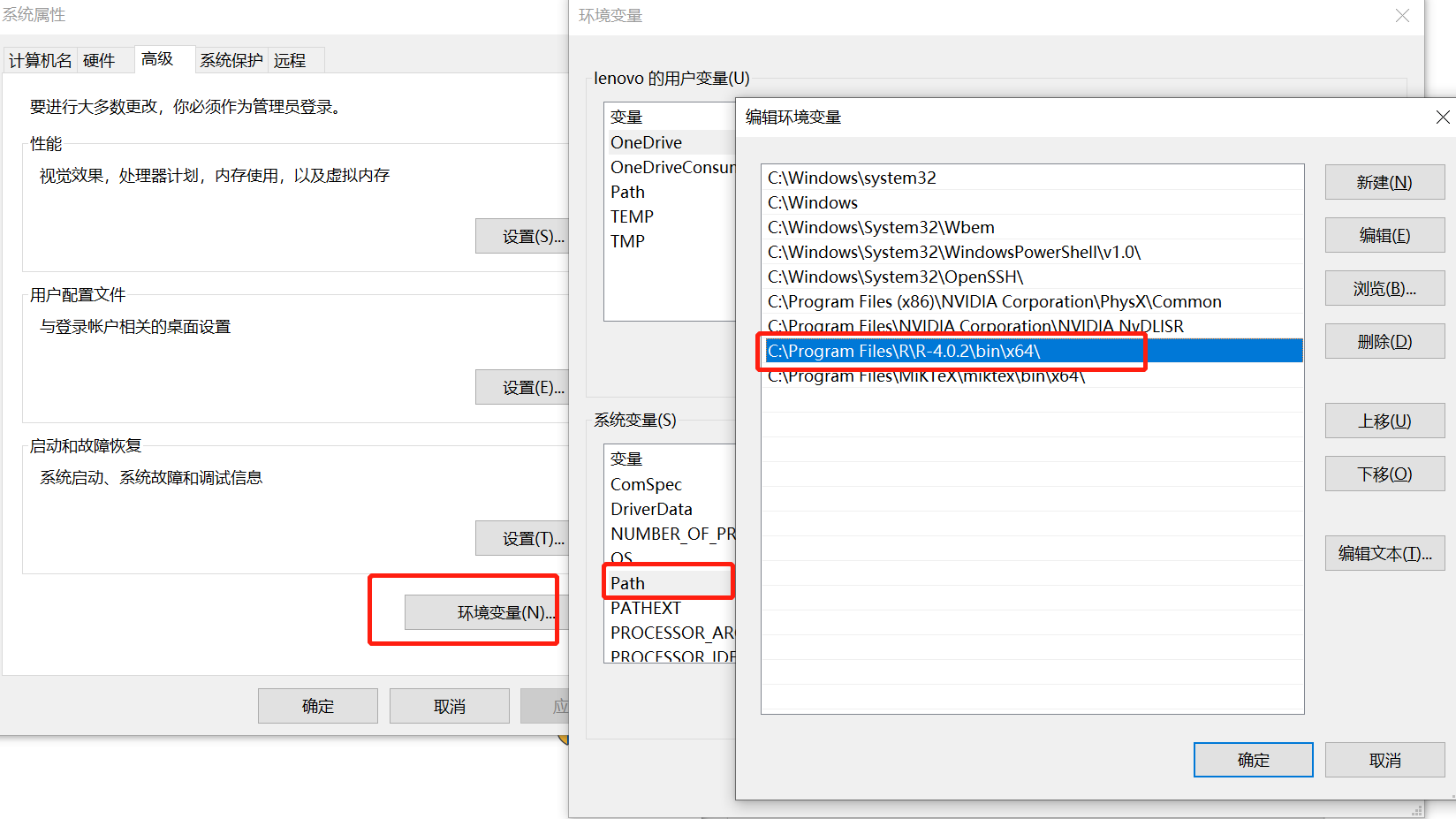


图1. 环境变量配置

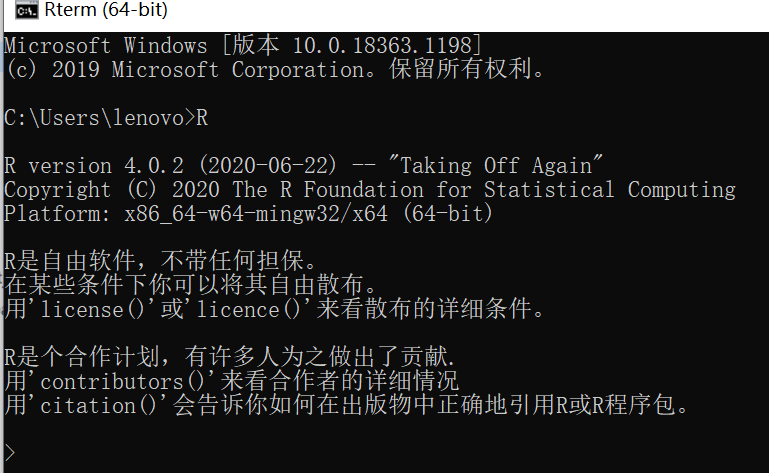


 图2. R的环境变量配置成功

**2.2 参数说明**

 程序运行之前， 可以输入命令Rscript .\pulse\_dia\_combine.R -h查看帮助文档信息：

-i|--input 必选参数，输入文件的路径

-o|--out 必选参数, 结果文件输出路径及文件名

-m|--combine\_method 可选参数,合并同一样本不同part文件的intensity方式,可选值: mean或max, 默认为mean

-s|--DIA\_method 可选参数, 搜库软件类型，可选值：DIANN, Spectronaut , OpenSWATH, 默认为DIANN

-x|--outmatrix 可选参数，输出矩阵的类型，可选值：peptide or protein, 默认为protein矩阵

-f|--fdr 可选参数, 提取DIANN 搜库结果时的FDR阈值, 默认为0.01

-p|--spn.pep.col 可选参数, 参数类型为integer。表示用Spectronaut搜库时导出的结果文件中peptide所在列数, 默认为第2列。

-t|--spn.prot.col 可选参数, 参数类型为integer。表示用Spectronaut搜库时导出的结果文件中protein所在列数, 默认为第1列

-a|--spn.sample.col.start 可选参数, 参数类型为integer。表示用Spectronaut搜库时导出的结果文件中定量信息起始列数, 默认为第3列开始

-b|--pulse.bool 可选参数, 是否需要做PulseDIA 定量结果的合并, 默认为T。若选择F,只能把不同表格的结果汇总成一张表

-h|--help 显示帮助文档

**2.3 程序运行**

打开cmd命令提示符或Windows PowerShell，输入指令：Rscript -i  TMT矩阵信息表 -f 样本信息表，其他参数根据分析需求自行设定。

输入指令：Rscript .\pulse\_dia\_combine.R -h可查看帮助文档。

注意：脚本程序、输入文件均需使用绝对路径。

**三、 案例演示**

**3.1 输入文件格式要求**

在程序运行前，须确保输入文件满足特定的格式要求。对三个软件搜库结果格式的统一要求为：搜库结果中的质谱文件须用"\_part1"、"\_part2"、"\_part3"等区分同一样本不同pulse的DIA文件, 且每个文件名中"\_part"仅允许出现一次，同一样本的文件名中除"\_part1"、"\_part2"、"\_part3"等不同外，其他部分的命名须保持一致。如sample1的3个pulseDIA文件的命名分别为："B20191210\_pulseDIA\_3\_part1.mzXML","B20191210\_pulseDIA\_3\_part2.mzXML","B20191210\_pulseDIA\_3\_part3.mzXML"。

下述案例以每个软件为例，对各自的格式要求作出说明。

3.1.1 DIA-NN输入文件格式

输入文件仅支持tsv格式，不同part的结果单独存放在一个tsv文件中，才可以合并不同part的结果。

每个tsv文件至少须包含以下列名的内容：

File.Name: 文件名，字符型

Precursor.Quantity: Precursor定量信息，数值型

Protein.Ids: 蛋白名，字符型

Modified.Sequence: 修饰肽段序列，字符型

Q.Value：Precursor的FDR, 浮点型

Protein.Q.Value: Protein的FDR, 浮点型

|  |
| --- |

命令运行示例：Rscript .\pulse\_dia\_combine.R -i .\test\_data -o .\diann\protein\_matrix

.\test\_data路径中存放了不同part的tsv格式的结果文件，\diann路径下存放了输出的结果文件protein\_matrix文件。

3.1.2 Spectronaut输入文件格式

输入文件仅支持从Spectronaut的report模块中导出的.xls格式，不同part的结果单独存放在一个xls文件中，才可以合并不同part的结果。

每个文件的格式形如：

|  |  |  |  |
| --- | --- | --- | --- |
| prot | peptide | B20191210\_pulseDIA\_3\_part1.htrms | B20191211\_pulseDIA\_3\_part1.htrms |
| 1/SMED30022992.p1 | 96\_GLPVTEMTK\_2 | 15046200 | 7461500 |
| 1/SMED30028142.p1 | 176\_GLQYENHDK\_2 | 42993000 | 14706400 |
| 1/SMED30004937.p1 | 188\_GLSEDIPPLDK\_2 | 40821100 | 32995600 |
| 1/SMED30026107.p1 | 334\_GLVDEVEQVR\_2 | 1273020 | 57906800 |
| 1/SMED30004348.p1 | 410\_GLWIAPTQK\_2 |  | 47134000 |
| 1/SMED30023391.p1 | 567\_GMINQGELTNLR\_2 | 68960400 | 54431600 |

pulse\_dia\_combine.R适用于合并Spectronaut搜库结果中不同part的peptide矩阵信息，不推荐使用该程序直接合并不同part的protein信息。

运行命令示例：Rscript .\pulse\_dia\_combine.R -i .\test\_data -o spn,peptide\_matrix.txt -s Spectronaut -x peptide

.\test\_data路径中存放了不同part的.xls格式的结果文件，默认输出路径下存放了输出的结果文件spn,peptide\_matrix.txt 。

3.1.3 Openswath输入文件格式

输入文件仅支持txt格式，不同part的结果可以存放在一个或多个txt文件中。

每个文件的格式形如：

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| peptide | prot | B20191210\_pulseDIA\_3\_part1.mzXML | B20191210\_pulseDIA\_3\_part2.mzXML | B20191210\_pulseDIA\_3\_part3.mzXML |
| 96\_GLPVTEMTK\_2 | 1/SMED30022992.p1 | 15046200 | 7461500 | 4564580 |
| 176\_GLQYENHDK\_2 | 1/SMED30028142.p1 | 42993000 | 14706400 | 16416900 |
| 188\_GLSEDIPPLDK\_2 | 1/SMED30004937.p1 | 40821100 | 32995600 | 32567900 |
| 334\_GLVDEVEQVR\_2 | 1/SMED30026107.p1 | 1273020 | 57906800 | 40609200 |
| 410\_GLWIAPTQK\_2 | 1/SMED30004348.p1 |  | 47134000 |  |
| 567\_GMINQGELTNLR\_2 | 1/SMED30023391.p1 | 68960400 | 54431600 | 55860300 |

pulse\_dia\_combine.R适用于合并Openswath搜库结果中不同part的peptide矩阵信息，不推荐使用该程序直接合并不同part的protein信息。

运行命令示例：Rscript .\pulse\_dia\_combine.R -i .\test\_data -o .\os\peptide\_matrix.txt -s openswath -x peptide

.\test\_data路径中存放了txt格式的结果文件，默认输出路径下存放了输出的结果文件os.peptide\_matrix.txt。