PHPA.R user instruction

- PHPA.R is a program written in R to convert MS signal intensities to the corresponding QC loading amounts according to the selection of using peak height or peak area.
- The PHPA.R script is freely available for non-commercial use.
- The instructions are given below:

1) Download and install R studio following the instruction in Rstudio website: https://rstudio.com/
2) Download the PHPA.R and save it in a folder
3) Within the same folder, prepare four input files
   A sample metabolite-intensity table from peak height measurement (**file 1**)
   A sample metabolite-intensity table from peak area measurement (**file 2**)
   A QC intensity table from peak height measurement (**file 3**)
   A QC intensity table from peak area measurement (**file 4**)
   in the following formats:
   Sample metabolite-intensity table containing all real samples (file 1 & 2, prepared in .csv format)
   Column 1: alignment ID.
   Column 2: retention time.
   Column 3: m/z value.
   Column 4 to the last column: MS signal intensities of real samples.
   The example dataset format is as follows:

Alignment ID	Average Rt(min)	Average Mz	1442_0	1874_0	1896_0	1958_0	2019_0	2035_0	3770_0	1442_29	1874_29	1896_29	1958_29	2019_29	2035_29	3770_29
0	3.163	67.05682	263	478	312	1717	0	118	856	698	1246	2276	924	571	524	180
2	8.021	69.03413	0	0	0	888	1920	0	972	0	0	518	76	134	752	196
3	12.119	69.03471	0	0	0	823	2596	0	1164	76	0	344	0	0	686	231
4	10.364	71.01347	2562	2358	1706	1975	2950	2034	1477	1836	1992	1572	2041	1822	1581	2452
5	10.688	71.01355	1906	2608	2160	2135	1620	2193	1777	1526	1471	1442	2345	1405	1355	2514
6	3.995	71.01392	0	223	91	183	6427	78	203	78	108	97	80	80	0	127
8	22.432	72.96916	504	340	182	263	1507	274	499	746	545	407	336	544	674	352

   QC intensity tables (files 3 & 4, prepared in .csv format)
   Column 1: alignment ID. Note: The IDs in the QC intensity table should be the same as the IDs in the sample metabolite-intensity table. Both IDs should be ordered in the same sequence.
   Column 2: retention time.
   Column 3: m/z value.
   Column 4 and after: intensities of serial diluted QC samples. The sequence should be from low concentration to high concentration. The analytical replicates should be displayed consecutively.
   The example dataset format is as follows:

Alignment ID	Average Rt(min)	Average Mz	QC-0.5uL-1	QC-0.5uL-2	QC-0.5uL-3	QC-1.5uL-1	QC-1.5uL-2	QC-1.5uL-3	QC-2.5uL-1	QC-2.5uL-2	QC-2.5uL-3	QC-3.5uL-1	QC-3.5uL-2	QC-3.5uL-3	QC-5uL-1	QC-5uL-2	QC-5uL-3
0	3.16	67.0568	120	148	0	510	496	407	694	625	600	667	846	833	973	1334	1477
2	8.02	69.0341	75	0	0	182	158	188	398	378	127	432	632	366	657	614	613
3	12.12	69.0347	90	0	0	312	97	248	348	581	513	524	590	589	846	784	762
4	10.36	71.0135	355	470	401	1156	1132	1152	1828	1706	1901	1998	1894	2157	2620	2348	2758
5	10.69	71.0136	1225	1022	775	1902	1750	1598	1808	1637	1655	1918	1720	1976	1995	1699	1742
6	4.00	71.0139	151	82	88	331	446	392	788	610	695	653	835	644	776	991	939
8	22.43	72.9692	0	102	94	492	428	304	604	728	535	782	659	836	1023	1113	1190

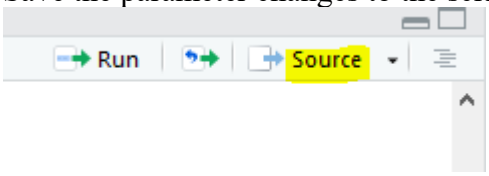
4) Open the PHPA.R script (see below) in Rstudio (https://rstudio.com/) and change the parameters therein (see Table 1 for the explanation of these parameters)

```
#Parameter setting
 #File input
Calibration_datapath = ""
Peak_area_FileName = "Leukemia_Peak_area.csv"
Peak_height_FileName = "Leukemia_Peak_height.csv"
QC_PA_FileName = "QC_Peak_area.csv"
QC_PH_FileName = "QC_Peak_height.csv"
 #QC information
QC_conc = c(0.1,0.3,0.5,0.7,1)
rep_QC = 3
 #Calibration settings
QC_zero_num = 1
R2_threshold = 0.8
k_threshold = 0
```

Table 1. Instruction of parameters in QC_cal.R script that can be tuned as needed.

Parameter	Function
data.path	Assign the data path for the folder that contains PHPA.R and other 4 required csv files for intensity correction
Peak_height_FileName	Set the file name for **File 1**
Peak_area_FileName	Set the file name for **File 2**
QC_PH_FileName	Set the file name for **File 3**
QC_PA_FileName	Set the file name for **File 4**
QC_conc	A set of numbers representing the relative concentrations of serial diluted QC samples
rep_QC	Number of analytical replicates for each serial diluted QC sample
QC_zero_num	Set the threshold for the number of zero intensity data points in the QC intensity table for a given metabolic feature. If the number of zero intensity data points is larger than the threshold, the QC calibration curve cannot be established. The default threshold is 2, which means the maximum allowed zero intensity QC data points is 2.
R2_threshold	Set the R^2 threshold to categorize metabolic features based on their regression coefficients in the QC samples. $R^2 \geqslant 0.8$ is the default threshold for considering a high-quality linear regression.
k_threshold	Set the k value (the slope of the linear calibration curve) threshold to categorize metabolic features based on their slopes in the linear regression.

5) Save the parameter changes to the script. Click “source” on top right to run the script.



6) After running the script, there will be two output files.
File 1: The converted intensity table for all real biological samples. Default name: Calibrated intensity table.csv
In this file, every metabolic feature will be noted according to their performance in linear regression process (**Table 2**).

Table 2. Labels for metabolic features and their definitions.

Label	Definition
Good regression	The established QC-based calibration curve meets the requirement of slope (k_threshold) and square pf corelated coefficient (R2_threshold)
Poor regression	The established QC-based calibration curve DOESN’T meet the requirement of slope (k_threshold) and square pf corelated coefficient (R2_threshold), and these features won’t be calibrated.
Insufficient QC data points	For a given metabolic feature, if the number of QC data points is smaller than the defined threshold, QC-based calibration curve cannot be established and signal calibration cannot be accomplished, thus “Insufficient QC data points” will be assigned to that feature. These features won’t be calibrated

File 2: A table describing whether peak height or peak area-based measurement was used to generate the converted intensity for every metabolic feature in each sample. Default name: decision table.csv