Readme (Becker et al.)

<u>Software Dependencies</u>
RStudio 2023.06.1+524 "Mountain Hydrangea" Release (547dcf861cac0253a8abb52c135e44e02ba407a1, 2023-07-07) for windows
Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) RStudio/2023.06.1+524 Chrome/110.0.5481.208 Electron/23.3.0 Safari/537.36

Windows 64-bit, v10.0.22631 Build 22631

Microsoft® Excel® for Microsoft 365 MSO (Version 2407 Build 16.0.17830.20056) 64-bit

This software has also been tested on:
RStudio 2021.09.2+382 "Ghost Orchid" Release (fc9e217980ee9320126e33cdf334d4f4e105dc4f, 2022-01-04) for macOS
Mozilla/5.0 (Macintosh; Intel Mac OS X 10_14_6) AppleWebKit/537.36 (KHTML, like Gecko) QtWebEngine/5.12.10 Chrome/69.0.3497.128 Safari/537.36

MacOS v10.14.6 ("Mojave")

No non-standard hardware are required.

<u>Installation</u> (10 min)

Navigate to "https://dailies.rstudio.com/version/2023.06.1+524.pro1/" and download the correct installer (for Windows 10+: "https://s3.amazonaws.com/rstudio-ide-build/electron/windows/RStudio-2023.06.1-524.exe")

- If these links are no longer active, versions can also be found at https://cran.r-project.org/.

Go through the prompts to download the software to your machine. Open Rstudio when it is done.

Run the following script to install the required packages:

install.packages(c("pbapply","stringr","ggplot2","reshape2","ggrepel","RColorBrewer","readxl","tidyverse","dplyr","purr","broom","msigdbr")

Volcano plot demo (R markdown file)

Take the imputed and $\log_2$ transformed data from Perseus, saved as a .csv file. It has nine columns and looks like this:

	A	B	C	D	E	F	G	H	I	J
1	ProteinID	Gene	Descriptio	control_1 l	control_2 l	control_3 l	treatment_	treatment_	treatment_3	Intensity
2	contaminant_GR78_P	owl|Q0586		26.4023	26.0612	26.2288	27.2474	27.503	27.8726	
3	contaminant_INT-STD	BSA		34.2082	36.0645	35.3108	37.2463	34.0636	34.1459	
4	contaminant_KERATIN	contamina		26.2269	28.4168	28.5231	27.1569	27.6684	27.4281	

Place the data, as a .csv, in a new folder "data" nested within a folder you will use for the analysis. In this example, the shown path is:
"C:\Users\andbp\Documents\R_data\Readme_example\data"

Open Data File S1. In line 18, change the working directory path to your folder.
(C:\Users\andbp\Documents\R_data\Readme_example)

Define the fnames function by running "names<-list.files.()" and then "fnames" in the console.

In line 19, set the number within [c(#)] to the number corresponding to the "data" folder after fnames has been run (will be '1' if there are no other files in the folder)

Change line 26 to specify the name of the file after the "fnames" argument. In this example, it would be:
"files = lapply(paste(fnames,"Example_data_volcano.csv", sep= "/"), read.csv, sep = ",")"

Run the rest of the code chunk (through line 37).
- If this does not run, check that your input data file is formatted the same way with the correct column headers (ProteinID, Gene, Description)

Run the next code chunk (line 42) to merge the data.

Run the next code chunk (lines 47-49). A dataframe named "df" should populate in the R environment.
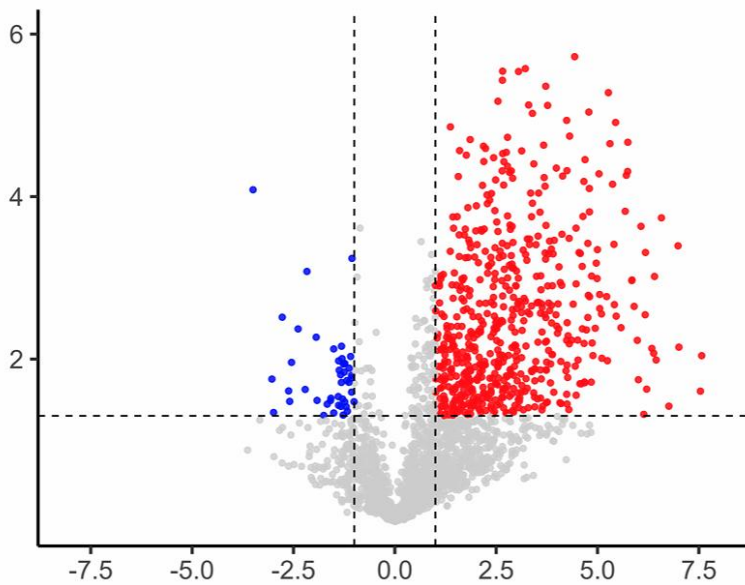- Check the dataframe to ensure the data are displaying properly (will look analogous to the .csv file in Excel)

Run the next code chunk (lines 54-79).

Run the next code chunk (lines 84-89).

Run the next code chunk (lines 94-108).

Run the next code chunk (lines 113-113) to generate the volcano plot. It should display in the R markdown file and should look like this:

To save the volcano plot or results file, run line 138 or 139, respectively.

Lines 142-170 have ways to label the "results2" dataframe with different lists of proteins to either color specific data points on the plots differently or label specific data points with black circles on the plot.

Spearman correlation and multiple linear regression analysis (R markdown file)

Take the imputed and $\log_2$ transformed data from Perseus for both the labeling and expression protein intensities of each replicate, saved as .csv files. They have nine columns and look like this:

Expression data:

	A	B	C	D	E	F	G	H	I
1	ProteinID	Gene	Descriptic	parental_1 Intensity	parental_2 Intensity	parental_3 Intensity	SRB1_1 Intensity	SRB1_2 Intensity	SRB1_3 Intensity
2	contaminant_GR78_I	contamina		33.4565	33.4772	33.3932	32.9565	33.338	32.891
3	contaminant_GR78_{	contamina		31.5381	27.7942	31.8916	31.87	30.0789	32.8117
4	contaminant_GR78_'	contamina		28.7261	25.1626	24.3567	23.7152	23.6963	23.739

Labeling data:

	A	B	C	D	E	F	G	H	I	J
1	ProteinID	Gene	Descriptio	control_1 l	control_2 l	control_3 l	treatment_	treatment_	treatment_3	Intensity
2	contaminant_GR78_P	owl|Q0586	26.4023	26.0612	26.2288	27.2474	27.503	27.8726		
3	contaminant_INT-STD	BSA	34.2082	36.0645	35.3108	37.2463	34.0636	34.1459		
4	contaminant_KERATIN	contamina	26.2269	28.4168	28.5231	27.1569	27.6684	27.4281		

Note the column headers: they begin with "parental" or "SRB1" for the expression data or "control" or "treatment" for the labeling data.

Place the .csv files in a new folder In this example, the shown path is: "C:\Users\andbp\Documents\R_data\Readme_example\data"

Save Data File S2 into the same folder and open it. In line 17, change the working directory path to your folder.
(C:\Users\andbp\Documents\R_data\Readme_example)

In line 18, specify the name of the file that contains the protein intensities for the labeling data.

In line 19, specify the name of the file that contains the protein intensities for the expression data.

Run the chunk (lines 6–20)

In lines 27 and 39, check that the values in the "cols = matches()" arguments match the column headers in your data.

In line 109, change the values in the msigdbr() argument to match the desired database (GO Molecular Function in this example) to be used for the Spearman correlation.

Run the chunk (lines 24–125). A preview of the results of the Spearman correlation will appear.

Run the next chunk (lines 128–193) to visualize the results of the multiple linear regression and the Spearman correlation analyses.