Notes on using Leaf Proteomics Project analysis tools in R

1. The first step in using any of these tools is to load the R Project file in R Studio. Click on D14.RProj in your file browser.
2. Scripts are found in the scripts directory
3. The most important script file is transformations.R
   1. This scripts aggregates several subscripts to import and process ion area data
   2. It produces 4 output dataframes:
      1. climate\_locs – includes all environmental and trait data associated with leaf samples (*Table 1*)
      2. protein\_samples\_D14 – contains amounts per leaf area (mg/mm2) for each of the 2581 identified proteins, for each sample. Also lists Mercator/MAPMAN identifier codes for each protein.
      3. protein\_D14 – contains protein amounts per area (mg/mm2) aggregated by protein functional groups. Functional groups included are shown in Table 2.
      4. protein\_D14 – contains fractional protein amounts (per leaf area amounts divided by total protein amount\*) aggregated by protein functional groups. Functional groups included are shown in Table 2.
         * Total protein amount is calculated by summing the per leaf area amount of all identified proteins (and thus is an approximation of actual total leaf protein per area)
   3. LiCOR data, chlorophyll measurements, d13C data, leaf NP and soil NP are not included in climate\_locs by default (as data are not available for all samples, and rows with missing data are excluded from subsequent analyses). These data can be included by telling the script to include them, by running any of the following commands before running the script:
      1. include\_photosynthesis = TRUE
      2. include\_chlorophyll = TRUE
      3. include\_d13C = TRUE
      4. include\_leaf\_N = TRUE
      5. include\_leaf\_P = TRUE
      6. include\_soil\_N = TRUE
      7. include\_soil\_P = TRUE
4. Protein amount data can be combined with environmental and trait data using prep\_data.R or prep\_data\_mg\_per\_mm2.R, which will use fractional or per leaf area protein amounts, respectively.
   1. The resulting dataframe is called data, and additionally contains unique identifiers for each species\*site combination, allowing data to be aggregated by species\*site. This dataframe is used in subsequent plotting functions.
5. Plots of protein amounts aggregated by species\*site are generated using the aggplot\_save\_combined\_3 function.
   1. This function will be available if you have already run transformations.R or can be loaded by running: source(functions.R)
   2. This function has a number of arguments (options) which can be set which dictate what data is plotted and how the figure will look. Comments on options for each argument can be found in the comments at the top of the function code (type aggplot\_save\_combined\_3() in the console to print the function code).
   3. For examples on how to use this function, please see the aggplot\_save\_combined\_3.R script
6. Aggregated data outputs are currently generated as need (e.g. the code to create the means dataframe in heatmaps\_extra.R, see Supp 1.)
   1. We should discuss what more useable solutions might be regarding this

[1] "sample" "Latitude" "Longitude"

[4] "leaf\_age" "irradiance" "gap"

[7] "LMA\_g\_per\_m2" "LWC\_percent" "species"

[10] "site\_revised" "tavg" "dnrg"

[13] "isot" "tsea" "tmax"

[16] "tmin" "trng" "twet"

[19] "tdry" "twrm" "tcld"

[22] "prec" "pwmt" "pdmt"

[25] "psea" "pwet" "pdry"

[28] "pwrm" "pcld" "soilN"

[31] "soilP" "soilC" "alpha"

[34] "site\_id.x" "tavg\_recent\_year" "tmax\_recent\_year"

[37] "prec\_recent\_year" "VPD\_recent\_year" "site\_id.y"

[40] "tavg\_recent\_month" "tmax\_recent\_month" "prec\_recent\_month"

[43] "VPD\_recent\_month" "leaf\_rad" "species\_confirmed"

[46] "date" "gap\_mean" "gap\_SE"

[49] "leafrad\_mean" "leafrad\_SE" "LMA\_mean"

[52] "LMA\_SE"

*Table 1.* Columns in climate\_locs dataframe

[1] "sample"

[2] "total\_protein"

[3] "cell\_wall"

[4] "DNA"

[5] "glycolysis"

[6] "hormone\_metabolism"

[7] "lipid\_metabolism"

[8] "mitochondrial\_electron\_transport\_ATP\_synthesis"

[9] "protein"

[10] "protein.folding"

[11] "redox"

[12] "RNA"

[13] "secondary\_metabolism"

[14] "signalling"

[15] "stress"

[16] "TCA\_org\_transformation"

[17] "CHO\_metabolism"

[18] "glutathione\_S\_transferases"

[19] "LHC\_I"

[20] "LHC\_II"

[21] "light\_reactions"

[22] "ATP\_synthase\_chloroplastic"

[23] "cytochrome\_b6f"

[24] "NADH\_DH"

[25] "other\_electron\_carrier"

[26] "photosystem\_I"

[27] "photosystem\_II"

[28] "polyamine\_metabolism"

[29] "calvin\_cycle"

[30] "photorespiration"

[31] "rubisco\_large\_subunit"

[32] "rubisco\_small\_subunit"

[33] "stress.abiotic"

[34] "stress.abiotic.heat"

[35] "stress.biotic"

[36] "PSII\_min\_LHCII"

[37] "PSI\_min\_LHCI"

[38] "Photosystems"

[39] "electron\_transport\_minATPsynth"

[40] "Rubisco"

[41] "electron\_transport"

[42] "LHC"

[43] "Photosystems\_min\_LHC"

[44] "LHCI\_per\_PSI"

[45] "LHCII\_per\_PSII"

[46] "LHC\_per\_PS"

*Table 2.* Columns in protein\_D14 and protein\_stand\_D14 dataframes

Supp 1. – example code for generating output data aggregated by species\*site ID

source('scripts/transformations\_photomax.R')

source('scripts/prep\_data.R')

data <- merge(data, rbact\_, by = 'sample')

data$rbact\_sum <- data$rbact\_sum / data$total\_protein

data <- merge(data, iso\_, by = 'sample')

data$isoprene\_synthase <- data$isoprene\_synthase / data$total\_protein

means <- data %>% dplyr::group\_by(ID) %>% dplyr::summarise(calvin\_cycle\_mean = mean(calvin\_cycle, na.rm=TRUE),

rubisco\_mean = mean(Rubisco, na.rm=TRUE),

photosystems\_mean = mean(Photosystems, na.rm=TRUE),

ATP\_synthase\_chloroplastic\_mean = mean(ATP\_synthase\_chloroplastic, na.rm=TRUE),

electron\_transport\_mean = mean(electron\_transport\_minATPsynth, na.rm=TRUE),

photorespiration\_mean = mean(photorespiration,na.rm=TRUE),

protein\_mean = mean(protein, na.rm=TRUE),

heatstress\_mean = mean(stress.abiotic.heat, na.rm=TRUE),

rbact\_mean = mean(rbact\_sum, na.rm=TRUE),

isoprene\_synthase\_mean = mean(isoprene\_synthase, na.rm=TRUE))

data <- full\_join(data, means)

data <- distinct(data, calvin\_cycle\_mean, rubisco\_mean, photosystems\_mean, ATP\_synthase\_chloroplastic\_mean, electron\_transport\_mean, photorespiration\_mean, protein\_mean,

heatstress\_mean, rbact\_mean, isoprene\_synthase\_mean, .keep\_all=TRUE)