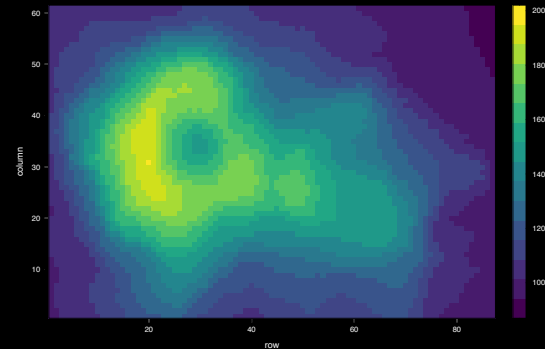


Evomics 2023

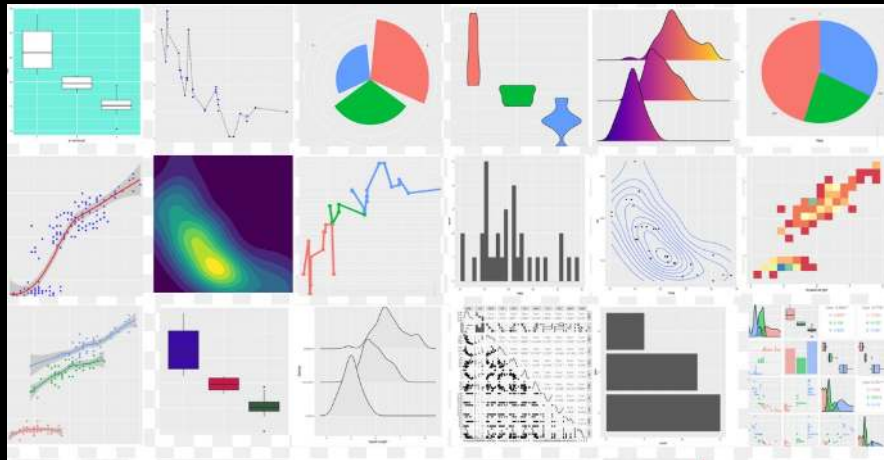
 & ggplot2



Special thanks to Dr Scott Handley and Washington University in St. Louis, Bernard Becker Medical Library, Research Computing for contributing their slides and teaching materials to this session.

Outline

- **What is R**
 - Why is it useful?
 - What can be done with R?
 - RStudio
- **Installing packages**
- **General data structures**



What is R?

A free software environment for statistical computing and graphics

```
R Console
R version 4.2.1 (2022-06-23) -- "Funny-Looking Kid"
Copyright (C) 2022 The R Foundation for Statistical Computing
Platform: x86_64-apple-darwin17.0 (64-bit)

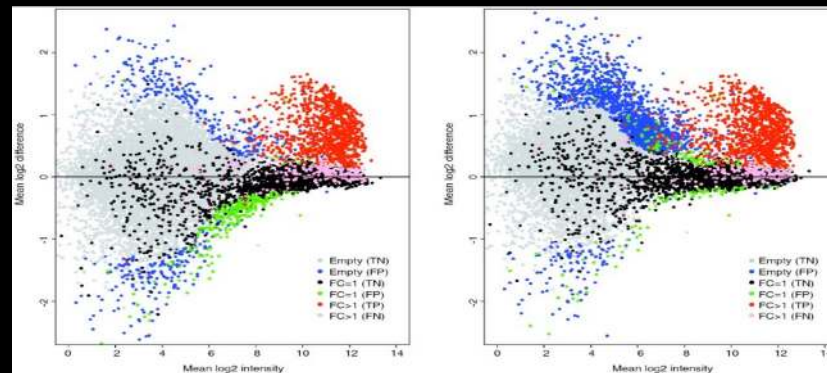
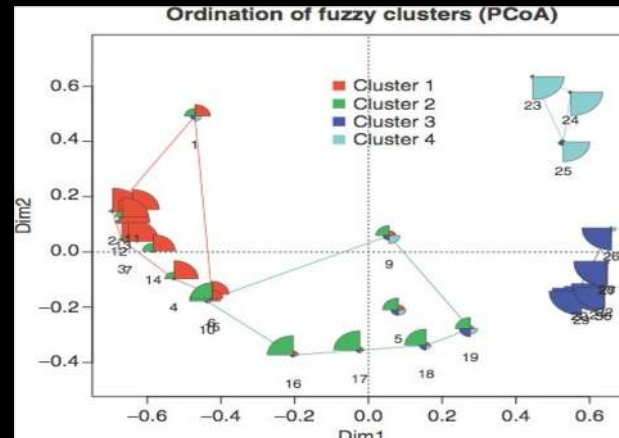
R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

[R.app GUI 1.79 (8095) x86_64-apple-darwin17.0]
> |
```



Why is R useful?

- **Open source**
- **Data management and manipulation**
 - Importing data in various formats (like text files, excel files, etc.)
 - Manipulating data (subsetting and filtering tables, merging, transposing, etc.)
- Cutting-edge **graphical data visualization**
- Support for rich **statistical simulation and modeling**
- Well established system of **packages and documentation**
- **Active development** and dedicated **community**



Why is R useful?

- **Open source**
- **Data management and manipulation**
 - Importing data in various formats (like text files, excel files, etc.)
 - Manipulating data (subsetting and filtering tables, merging, transposing, etc.)



Stay tuned for our lab session!

- **Cutting-edge graphical data visualization**



David Barnett (05/24)
and Jim Whiting (05/25)
sessions

- Support for rich **statistical simulation and modeling**
- Well established system of packages and documentation
- **Active development** and dedicated **community**



Next slides →

Where to learn more about R

- The R Project Homepage: <http://www.r-project.org>
- Quick R Homepage: <http://www.statmethods.net>
- Bioconductor: <http://www.bioconductor.org>
- An Introduction to R (long!): <http://cran.r-project.org/doc/manuals/R-intro.html>
- Guide to downloading and installing R and RStudio: <https://rstudio-education.github.io/hopr/starting.html>
- R Graph gallery (inspiration for R charts) <https://r-graph-gallery.com>
- Google - there are tons of tutorials, guides, demos, packages and more



R for Biologists

- **Bioconductor** (<http://bioconductor.org>)

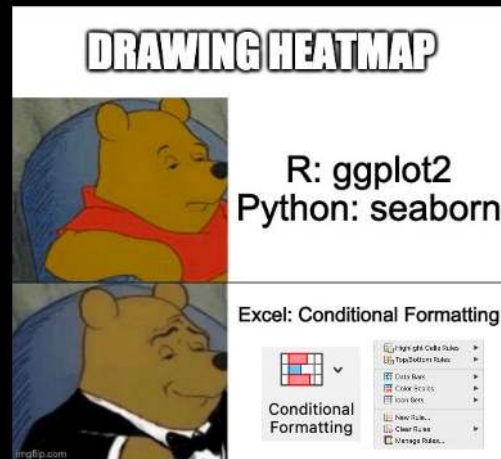
- 2,140 packages (05-27-2022):
 - Variant detection: coding changes, PolyPhen database
 - Annotation: pathway analysis, access GO, KEGG, NCBI and many others
 - High-throughput assays: flow cytometry, mass spec
 - Transcription factor binding detection

- **Phylogenetics and Evolution** (<http://cran.r-project.org/web/views/Phylogenetics.html>)

- Ancestral State Reconstruction
- Phylogenetic Inference
- Trait Evolution

- **Ecology** (<http://cran.r-project.org/web/views/Environmetrics.html>)

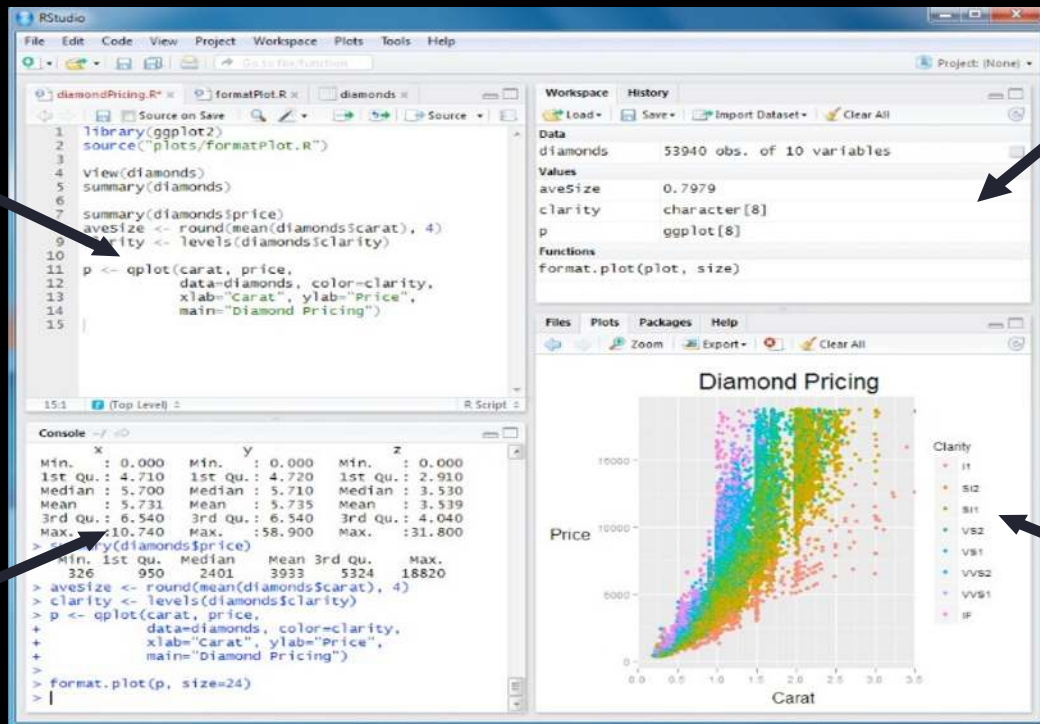
- Ordination
- Cluster Analysis
- Ecological Theory
- Population Dynamics
- Spatial Data Analysis



R Studio

Enter commands
here

Workspace data



View results here

Help and Plots
viewer

Handy R Studio tricks

The screenshot shows the R Studio interface with the following components:

- Source Editor:** Contains R code for creating a dataframe and a bar plot.
- Environment:** Shows the Global Environment with a variable `chris_df` containing 2 observations of 2 variables.
- Console:** Shows the R prompt and the command `?geom_bar` being entered.
- Help Viewer:** Displays the documentation for `geom_bar` from the `ggplot2` package.

```
1 # Import previously installed packages
2 library(ggplot2)
3 library(dplyr)
4
5 # Create a dataframe with some example data
6 list_of_objects <- c("potato", "lemon")
7 object_counts <- c(5, 10)
8
9 chris_df <- data.frame(list_of_objects, object_counts)
10
11 # Create a bar plot
12 ggplot(data=chris_df, aes(x=list_of_objects, y=object_counts)) +
13   geom_bar(stat="identity")
14
```

Console output:

```
>
> ?geom_bar
```

Help Viewer content:

geom_bar {ggplot2}

Bar charts

Description

There are two types of bar charts: `geom_bar()` and `geom_col()`. `geom_bar()` makes the height of the bar proportional to the number of cases in each group (or if the weight aesthetic is supplied, the sum of the weights). If you want the heights of the bars to represent values in the data, use `geom_col()` instead. `geom_bar()` uses `stat_count()` by default: it counts the number of cases at each x position. `geom_col()` uses `stat_identity()`: it leaves the data as is.

Usage

Can add a
“?” before a
function
and open
this help
information
viewer

R Markdown Cheat Sheet

learn more at rmarkdown.rstudio.com

rmarkdown 0.2.50 Updated: 8/14



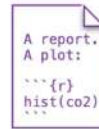
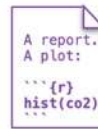
1. Workflow R Markdown is a format for writing reproducible, dynamic reports with R. Use it to embed R code and results into slideshows, pdfs, html documents, Word files and more. To make a report:

i. **Open** - Open a file that uses the .Rmd extension.

ii. **Write** - Write content with the easy to use R Markdown syntax

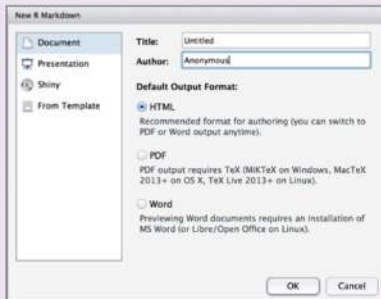
iii. **Embed** - Embed R code that creates output to include in the report

iv. **Render** - Replace R code with its output and transform the report into a slideshow, pdf, html or ms Word file.



2. Open File Start by saving a text file with the extension .Rmd, or open an RStudio Rmd template

- In the menu bar, click **File ► New File ► R Markdown...**
- A window will open. Select the class of output you would like to make with your .Rmd file
- Select the specific type of output to make with the radio buttons (you can change this later)
- Click OK



3. Markdown Next, write your report in plain text. Use markdown syntax to describe how to format text in the final report.

syntax

Plain text
End a line with two spaces to start a new paragraph.
italics and `_italics_`
bold and `__bold__`
^{superscript^2}
~~strikethrough~~
[\[link\]\(www.rstudio.com\)](http://www.rstudio.com)

Header 1

Header 2

Header 3

Header 4

Header 5

Header 6

endash: --
emdash: ---
ellipsis: ...
inline equation: $\$A = \pi \cdot r^2\$$
image: 

horizontal rule (or slide break):

becomes

Plain text
End a line with two spaces to start a new paragraph.
italics and *italics*
bold and **bold**
^{superscript²}
~~strikethrough~~
[link](http://www.rstudio.com)

Header 1

Header 2

Header 3

Header 4

Header 5

Header 6

endash: --
emdash: ---
ellipsis: ...

inline equation: $A = \pi \cdot r^2$



horizontal rule (or slide break):

4. Choose Output Write a YAML header that explains what type of document to build from your R Markdown file.

YAML

A YAML header is a set of key: value pairs at the start of your file. Begin and end the header

```
---  
title: "Untitled"  
author: "Anonymous"  
output: html_document  
---
```

This is the start of my

The RStudio
template writes
the YAML header

R installing packages

- **Bioconductor** (<http://bioconductor.org>)
 - BiocManager::install("Package Name")
- CRAN install (<https://cran.r-project.org/web/packages/>)
 - install.packages("Package Name")
- **GitHub install**
 - devtools::install_github("Package Name")
 - remotes::install_github("Package Name")

```
install.packages("light")
```



```
library("light")
```

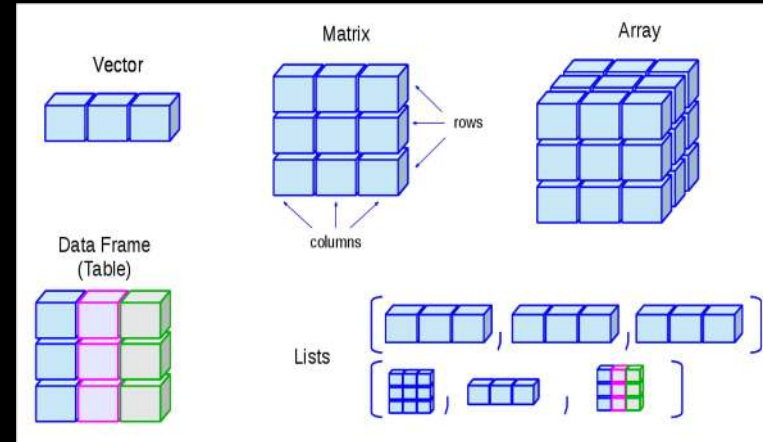


Images sourced from <https://www.wikihow.com/Change-a-Light-Bulb>



General data structures

- **Vector** - ordered collection of data
`vector_1 <- c(2, 3, 4, 10)`
`vector_2 <- c("potato", "lemonade", "avocado")`
- **Matrix** - 2D collection of vectors with same data type
- **Array** - multiple dimension collection of vectors
- **Dataframe** - matrix-like with multiple data types (like an excel table with text and numbers)
- **Lists** - ordered collection of objects but can contain other lists or vectors inside it



<http://venus.ifca.unican.es/Rintro/dataStruct.html>

But..

**which dataset should we use to do all
these things?**

Arabidopsis thaliana mutants *psbo1* and *psbo2*

WT



psbo1



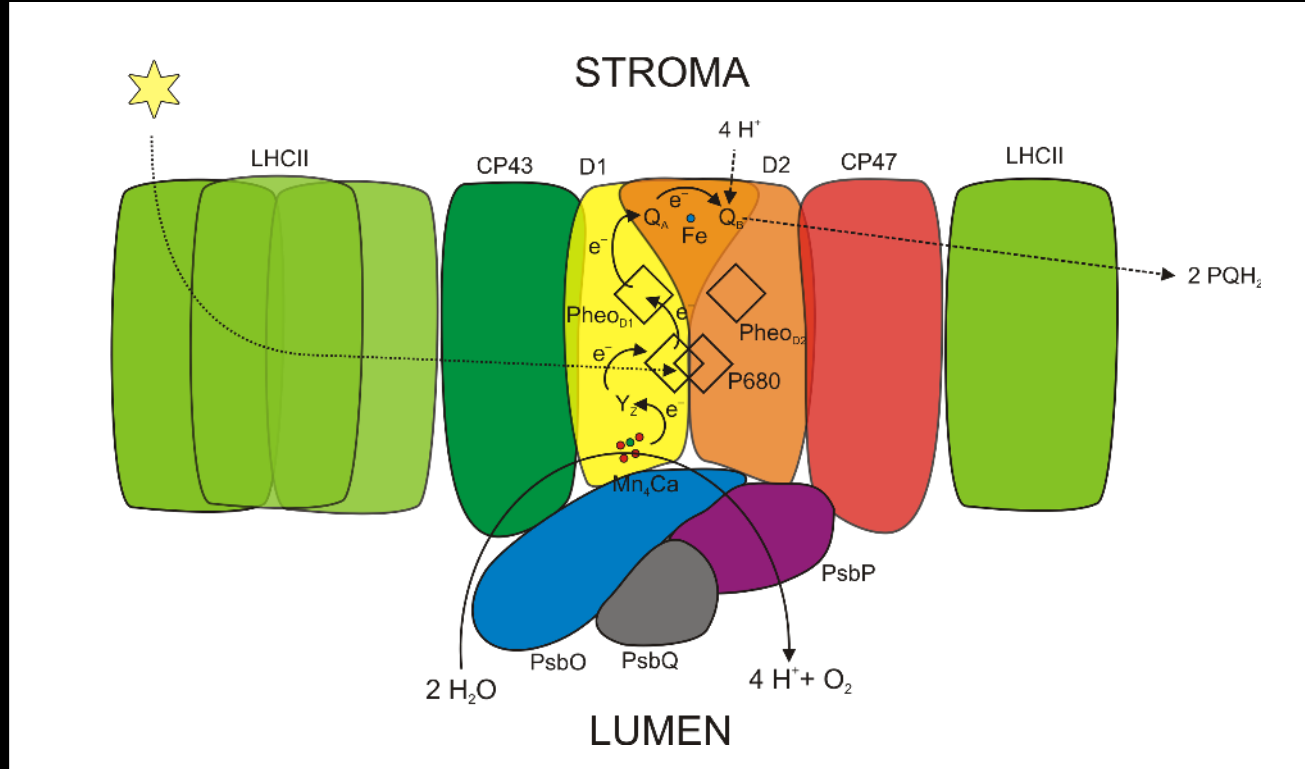
psbo2



PsbO protein

- Subunit of photosystem II
- Important for water splitting
- *Arabidopsis*: PsbO1 and PsbO2

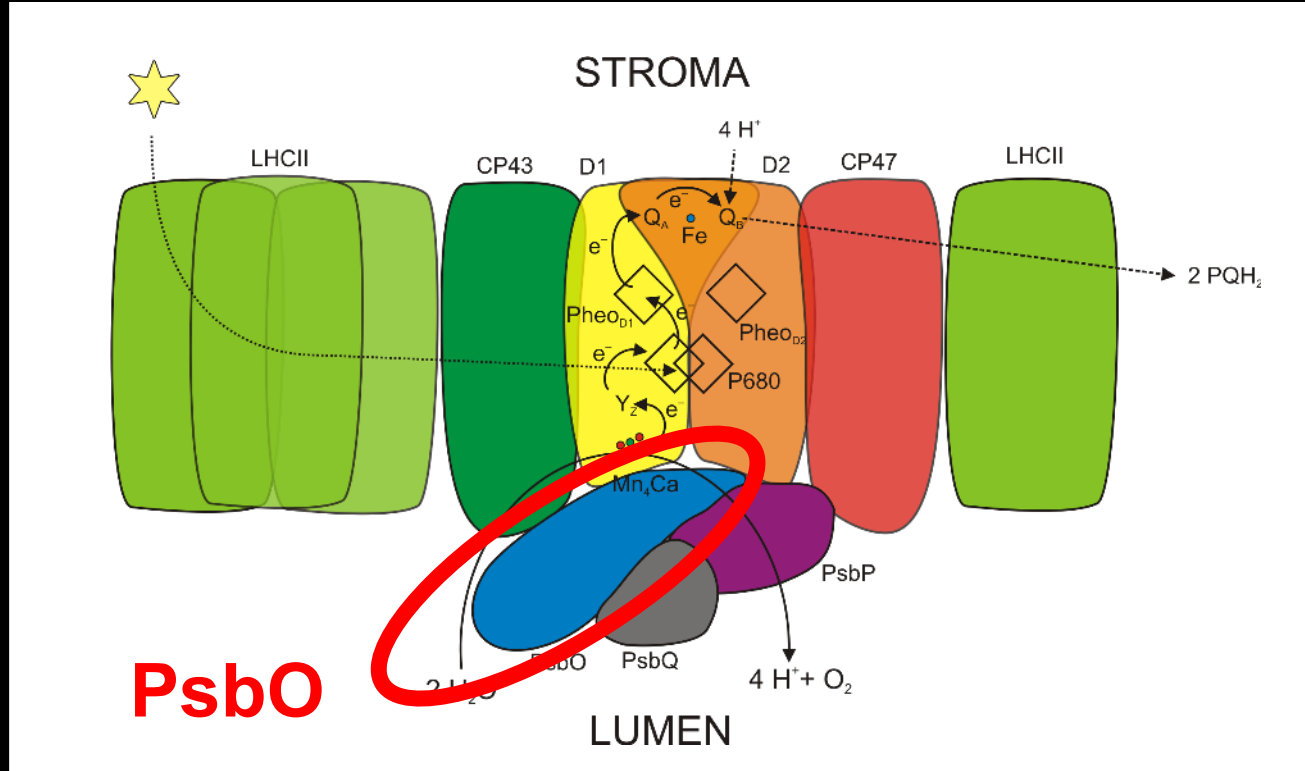
Photosystem II



PsbO protein

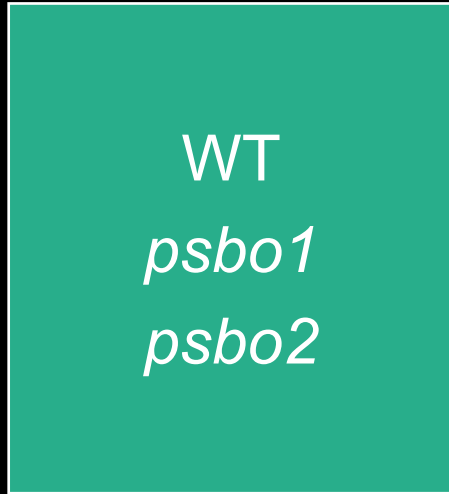
- Subunit of photosystem II
- Important for water splitting
- *Arabidopsis*: PsbO1 and PsbO2

Photosystem II

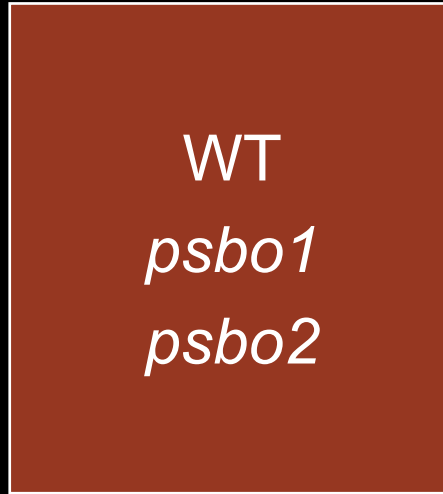


Experimental design

Control

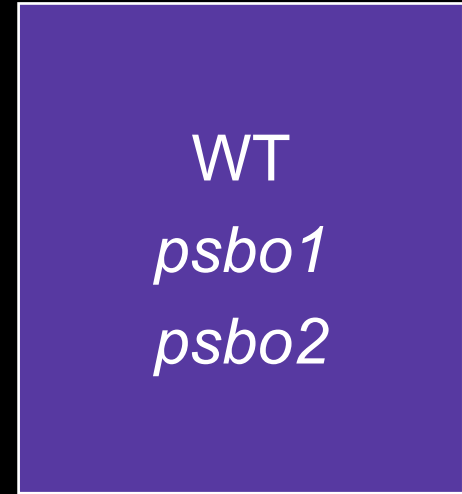


Drought



- water

Salt stress



+ NaCl

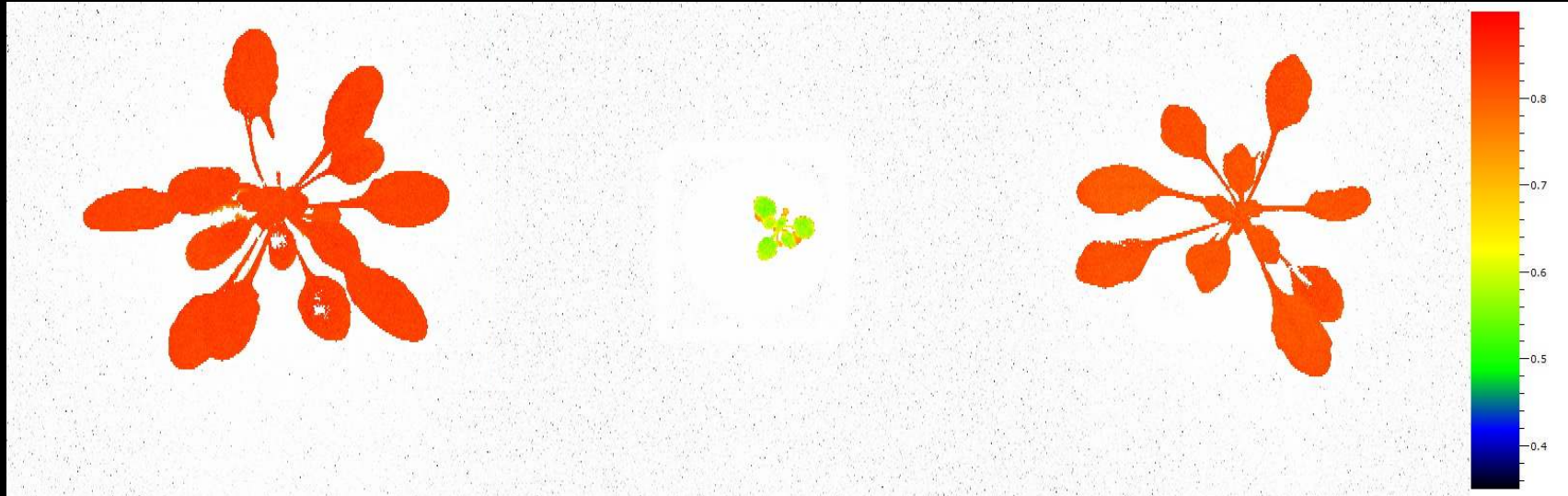
Measurement – chlorophyll fluorescence

- Leaf rosette area
- F_V/F_M (QY_max) – maximum quantum yield of photosystem II

WT

psbo1

psbo2



Yey!

Let's start the practical!

Open the Rstudio server by typing: `<IP>:8787`

