Data integration

Andrej Blejec

16. 1. 2023

Contents

1 Purpose of this document

2 File management system

Files are managed under the pISA-tree framework. To take advantage of pISA-tree metadata information, we will use package `pisar`

```
library(pisar)
pisa <- pisa()
```

Final document is in subdirectory `reports`:

```
basename(outputFile)
```

```
## [1] ".pdf"
```

2.1 Info from pISA structure

Input/data directory

```
.inroot
```

```
## [1] "../../../input"
```

Results directory

```
.oroot
```

```
## [1] "../../../output/knitr-pISA-Main-render-knitr-pISA-Main-HTML-C\\Users\\ablejec\\AppData\\Roaming
```

```
## project:  _p_Omics
```

```
## Investigation:  _I_Omics
```

```
## Study:          _S_multiOmics
```

```
## Assay:          _A_multiOmics-integration-R
```

3 Data

It is advisable to first read phenodata and featuredata, followed by actual data input. This enables smart selection of samples, based on the sample selection column with the assay name.

3.1 Phenodata

```
(pfn <- getMeta(.ameta
    , "Phenodata"))
```

```
## [1] "../../phenodata_20221001.txt"
```

```
dir(file.path(.aroot,dirname(pfn)), pattern = basename(pfn))
```

```
## [1] "phenodata_20221001.txt"
```

```
phdata <- read.table(file.path(.aroot,pfn)
  , header = TRUE
  ,  sep = "\t"
  , stringsAsFactors = FALSE
  , row.names=1
  )
dim(phdata)
```

```
## [1] 32 15
```

```
names(phdata)
```

```
##  [1] "SampleID"                      "Treatment"
##  [3] "Harvest"                       "SamplingDay"
##  [5] "DaysOfStressH"                 "PlantNo"
##  [7] "Sample.type"                   "Date"
##  [9] "Heat.Recovery.Days"            "TreatmentxDatexPlant"
## [11] "TreatmentxSamplingDay"         "TreatmentxSamplingDayxPlantNo"
## [13] "Transcriptomics"               "Metabolomics"
## [15] "Hormonomics"
```

Check and use the sample selection column, if present.

```
.aname
```

```
## [1] "_A_multiOmics-integration-R"
```

```
selectId <- substr(gsub("-",".",.aname),2,nchar(.aname))
selectId <- .vzorci
selectId
```

```
## [1] NA
```

```
if(selectId %in% names(phdata)) pdata <- phdata[!is.na(phdata[,selectId]),] else
pdata <- phdata
dim(pdata)
```

```
## [1] 32 15
```

Selected samples overview

```
table(pdata$Treatment, pdata$SamplingDay)
```

```
##
##     1 7 8 14
##   C 4 4 4  4
##   H 4 4 4  4
```

```
summary(pdata)
```

```
##    SampleID          Treatment            Harvest      SamplingDay
##  Length:32          Length:32          Min.   :1.00   Min.   : 1.0
##  Class :character   Class :character   1st Qu.:1.75   1st Qu.: 5.5
##  Mode  :character   Mode  :character   Median :2.50   Median : 7.5
##                                        Mean   :2.50   Mean   : 7.5
##                                        3rd Qu.:3.25   3rd Qu.: 9.5
##                                        Max.   :4.00   Max.   :14.0
##  DaysOfStressH      PlantNo      Sample.type
##  Min.   : 0.00   Min.   : 7.00   Length:32
##  1st Qu.: 0.00   1st Qu.:10.75   Class :character
##  Median : 0.50   Median :14.50   Mode  :character
##  Mean   : 3.75   Mean   :14.50
##  3rd Qu.: 7.25   3rd Qu.:18.25
##  Max.   :14.00   Max.   :22.00
##      Date           Heat.Recovery.Days TreatmentxDatexPlant
##  Length:32          Length:32          Length:32
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
##
##
##
##  TreatmentxSamplingDay TreatmentxSamplingDayxPlantNo Transcriptomics
##  Length:32             Length:32                     Min.   :1
##  Class :character      Class :character              1st Qu.:1
##  Mode  :character      Mode  :character              Median :1
##                                                      Mean   :1
##                                                      3rd Qu.:1
##                                                      Max.   :1
##   Metabolomics  Hormonomics
##  Min.   :1     Min.   :1
##  1st Qu.:1     1st Qu.:1
##  Median :1     Median :1
##  Mean   :1     Mean   :1
##  3rd Qu.:1     3rd Qu.:1
##  Max.   :1     Max.   :1
```

```
apply(pdata,2,function(x) summary(as.factor(x)))
```

```
## $SampleID
```

```
## AD001 AD002 AD003 AD004 AD005 AD006 AD007 AD008 AD013 AD014 AD015
##     1     1     1     1     1     1     1     1     1     1     1
## AD016 AD017 AD018 AD019 AD020 AD025 AD026 AD027 AD028 AD037 AD038
##     1     1     1     1     1     1     1     1     1     1     1
## AD039 AD040 AD045 AD046 AD047 AD048 AD057 AD058 AD059 AD060
##     1     1     1     1     1     1     1     1     1     1
##
## $Treatment
##  C  H
## 16 16
##
## $Harvest
## 1 2 3 4
## 8 8 8 8
##
## $SamplingDay
##  1  7  8 14
##  8  8  8  8
##
## $DaysOfStressH
##  0  1  7  8 14
## 16  4  4  4  4
##
## $PlantNo
##  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21 22
##  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2
##
## $Sample.type
## adult leaf
##         32
##
## $Date
## 04/11/2020 10/11/2020 11/11/2020 17/11/2020
##          8          8          8          8
##
## $Heat.Recovery.Days
##  0_0  1_0 14_0  7_0  8_0
##   16    4    4    4    4
##
## $TreatmentxDatexPlant
## C_2020-11-04_10  C_2020-11-04_7  C_2020-11-04_8  C_2020-11-04_9
##               1               1               1               1
## C_2020-11-10_11 C_2020-11-10_12 C_2020-11-10_13 C_2020-11-10_14
##               1               1               1               1
## C_2020-11-11_15 C_2020-11-11_16 C_2020-11-11_17 C_2020-11-11_18
##               1               1               1               1
## C_2020-11-17_19 C_2020-11-17_20 C_2020-11-17_21 C_2020-11-17_22
##               1               1               1               1
## H_2020-11-04_10  H_2020-11-04_7  H_2020-11-04_8  H_2020-11-04_9
##               1               1               1               1
## H_2020-11-10_11 H_2020-11-10_12 H_2020-11-10_13 H_2020-11-10_14
##               1               1               1               1
## H_2020-11-11_15 H_2020-11-11_16 H_2020-11-11_17 H_2020-11-11_18
##               1               1               1               1
```

```
## H_2020-11-17_19 H_2020-11-17_20 H_2020-11-17_21 H_2020-11-17_22
##               1               1               1               1
##
## $TreatmentxSamplingDay
##  C_1 C_14  C_7  C_8  H_1 H_14  H_7  H_8
##    4    4    4    4    4    4    4    4
##
## $TreatmentxSamplingDayxPlantNo
##  C_1_10   C_1_7   C_1_8   C_1_9 C_14_19 C_14_20 C_14_21 C_14_22
##       1       1       1       1       1       1       1       1
##  C_7_11  C_7_12  C_7_13  C_7_14  C_8_15  C_8_16  C_8_17  C_8_18
##       1       1       1       1       1       1       1       1
##  H_1_10   H_1_7   H_1_8   H_1_9 H_14_19 H_14_20 H_14_21 H_14_22
##       1       1       1       1       1       1       1       1
##  H_7_11  H_7_12  H_7_13  H_7_14  H_8_15  H_8_16  H_8_17  H_8_18
##       1       1       1       1       1       1       1       1
##
## $Transcriptomics
##  1
## 32
##
## $Metabolomics
##  1
## 32
##
## $Hormonomics
##  1
## 32
```

3.2 Featuredata

```
(ffn <- getMeta(.ameta,"Featuredata"))
```

```
## [1] ""
```

```
if(ffn != "")
fdata <- read.table(file.path(.iroot,ffn)
    , sep = "\t"
    , header = TRUE
    , na.strings = c("", "-")
    , stringsAsFactors = FALSE
    , row.names = 1
    ) else fdata <- NULL
head(fdata)
```

```
## NULL
```

First few columns, if present.

```
fdata[,1:2]
```

```
## NULL
```

3.3 Process data files

4 Metadata files

4.1 Project

Key	Value
project:	_p_Omics
Title:	Omics data analysis
Description:	Omics data analysis protocol
pISA projects path:	C:/Users/majaz/Desktop/pISA-Projects
Local pISA-tree organisation:	National Institute of Biology (NIB)
pISA project creation date:	2020-06-01
pISA project creator:	Kristina Gruden
Project funding code:	H2020-SFS-2019-2
Project coordinator:	Markus Teige
Project partners:	UNIVIE, UBO, UU, FAU, HUTTON, UDUR, WUR, UP, CRAG, NIB, HZPC, Meijer, SOL
Project start date:	2020-06-01
Project end date:	2024-06-01
Principal investigator:	Kristina Gruden
License:	CC BY 4.0
Sharing permission:	Public
Upload to FAIRDOMHub:	Yes

4.2 Investigation

Key	Value
Investigation:	_I_Omics
Title:	Omics
Description:	Minimal reproducible example for multi-Omics integration pipeline Investigation
Phenodata:	./phenodata_20221001.txt
pISA Investigation creation date:	2022-10-01
pISA Investigation creator:	MZ
Principal investigator:	Kristina Gruden
License:	CC BY 4.0
Sharing permission:	Public
Upload to FAIRDOMHub:	Yes

4.3 Study

Key	Value
Study:	_S_multiOmics
Title:	multiOmics
Description:	Minimal reproducible example for multi-Omics integration pipeline Study
Raw Data:	
pISA Study creation date:	2022-10-01
pISA Study creator:	MZ
Principal investigator:	Kristina Gruden
License:	CC BY 4.0
Sharing permission:	Public

Key	Value
Upload to FAIRDOMHub:	Yes

4.4 Assay

Key	Value
Assay:	_A_multiOmics-integration-R
Assay Class:	DRY
Assay Type:	R
Title:	multiOmics-integration
Description:	Minimal reproducible example for multi-Omics integration pipeline Assay
pISA Assay creation date:	2022-10-01
pISA Assay creator:	MZ
Analyst:	AB, MZ
Phenodata:	../../phenodata_20221001.txt
Featuredata:	
Data hormonomics:	./input/data_hormonomics.txt
Data metabolomics:	./input/data_metabolomics.txt
Data qPCR:	./input/data_qPCR.txt
Principal investigator:	Kristina Gruden
License:	MIT
Sharing permission:	Public
Upload to FAIRDOMHub:	Yes

5 SessionInfo

Windows 10 x64 (build 19044) R version 4.0.2 (2020-06-22) Platform: x86_64-w64-mingw32/x64 (64-bit)
Running under: Windows 10 x64 (build 19044)

Matrix products: default

locale: [1] LC_COLLATE=Slovenian_Slovenia.1250 [2] LC_CTYPE=Slovenian_Slovenia.1250
[3] LC_MONETARY=Slovenian_Slovenia.1250 [4] LC_NUMERIC=C
[5] LC_TIME=Slovenian_Slovenia.1250
system code page: 1252

attached base packages: [1] stats graphics utils datasets grDevices methods
[7] base

other attached packages: [1] xtable_1.8-4 amisc_0.1.0 Hmisc_4.6-0
[4] ggplot2_3.3.5 Formula_1.2-4 survival_3.2-7
[7] lattice_0.20-41 pisar_0.1.0.9000 knitr_1.30
[10] rmarkdown_2.6

loaded via a namespace (and not attached): [1] tidyselect_1.1.0 xfun_0.19 purrr_0.3.4
[4] splines_4.0.2 colorspace_1.4-1 vctrs_0.3.6
[7] generics_0.1.0 htmltools_0.5.2 yaml_2.2.1
[10] base64enc_0.1-3 rlang_0.4.10 pillar_1.4.7
[13] foreign_0.8-80 glue_1.4.2 withr_2.3.0
[16] RColorBrewer_1.1-2 jpeg_0.1-8.1 lifecycle_0.2.0
[19] stringr_1.4.0 munsell_0.5.0 gtable_0.3.0
[22] htmlwidgets_1.5.3 evaluate_0.14 latticeExtra_0.6-29 [25] fastmap_1.1.0 htmlTable_2.1.0 scales_1.1.1
[28] backports_1.2.0 checkmate_2.0.0 gridExtra_2.3

[31] png_0.1-7 digest_0.6.27 stringi_1.5.3
[34] dplyr_1.0.2 grid_4.0.2 tools_4.0.2
[37] magrittr_2.0.1 tibble_3.0.4 cluster_2.1.0
[40] crayon_1.3.4 pkgconfig_2.0.3 ellipsis_0.3.1
[43] Matrix_1.2-18 data.table_1.13.2 rstudioapi_0.13
[46] R6_2.5.0 rpart_4.1-15 nnet_7.3-14
[49] compiler_4.0.2
Project path:D:/ DEJAVNOSTI/ OMIKE/ pISA-projects/ multiOmics-integration/ _p_Omics/ _I_Omics/ _S_multiOmics/ _A_multiOmics-integration-R/ other/ 03_Step5/ Main file :`../doc/knitr-pISA-Main.Rmd`Project file: [link]