main

February 13, 2023

1 Correlate variables to remove redundant metrics

[1]:

```r
suppressPackageStartupMessages({
    library(here)
    library(dplyr)
    library(SummarizedExperiment)
})
```

1.1 Functions

[2]:

```r
load_phenotypes <- function(region){
    pheno_file <- here("input/phenotypes/_m/phenotypes.csv")
    pheno = data.table::fread(pheno_file) |> filter(Region == region) |>
        mutate_if(is.list, ~sapply(., sum)) |>
        mutate_if(is.numeric, scales::rescale)
    return(pheno)
}

check_dup <- function(df){
    sample    <- df |> select_if(is.numeric)
    sample    <- Filter(function(x) sd(x) != 0, sample)
    variables <- names(sample)
    return(cytominer::correlation_threshold(variables, sample, cutoff=0.95))
}

check_corr <- function(df){
    sample <- df |> select_if(is.numeric)
    sample <- Filter(function(x) sd(x) != 0, sample)
    dt     <- sample |> corrr::correlate() |>
        corrr::stretch() |> tidyr::drop_na() |>
        filter(abs(r) > 0.95) |>
        distinct(r, .keep_all=TRUE)
    varX <- distinct(dt, x)$x
    varX <- varX[-which(varX %in% intersect(varX, distinct(dt, y)$y))]
    vars <- unique(c(distinct(dt, x)$x, distinct(dt, y)$y))
    return(setdiff(vars, varX))
}
```

```r
remove_variables <- function(pheno_df){
    if(length(check_corr(pheno_df)) != 0){
        pheno_df <- pheno_df |> select(-check_corr(pheno_df))
    }
    return(pheno_df)
}
```

1.2 Main

1.2.1 Load phenotypes

[3]:
```r
caudate <- load_phenotypes("Caudate")
dlpfc   <- load_phenotypes("DLPFC")
hippo   <- load_phenotypes("HIPPO")
```

[4]:
```r
caudate |> dim()
dlpfc |> dim()
hippo |> dim()
```

1. 418 2. 57

1. 388 2. 57

1. 408 2. 57

1.2.2 Drop correlated

[5]:
```r
check_corr(caudate)
check_corr(dlpfc)
check_corr(hippo)
```

```
Correlation computed with
• Method: 'pearson'
• Missing treated using: 'pairwise.complete.obs'
```

1. ’Exonic_Rate’ 2. ’gene_Assigned’ 3. ’Intronic_Rate’ 4. ’End_2_Sense_Rate’ 5. ’Mapped_Unique_Reads’ 6. ’NonGlobin_Reads’

```
Correlation computed with
• Method: 'pearson'
• Missing treated using: 'pairwise.complete.obs'
```

1. ’Exonic_Rate’ 2. ’Low_Quality_Reads’ 3. ’gene_Assigned’ 4. ’Ambiguous_Reads’ 5. ’Intronic_Reads’ 6. ’NonGlobin_Reads’ 7. ’Intronic_Rate’ 8. ’End_2_Sense_Rate’ 9. ’numReads’ 10. ’Mapped_Unique_Reads’ 11. ’MedianAvgTxCov’

```
Correlation computed with
• Method: 'pearson'
• Missing treated using: 'pairwise.complete.obs'
```

1. ’Exonic_Rate’ 2. ’gene_Assigned’ 3. ’Intronic_Rate’ 4. ’mitoRate’ 5. ’End_2_Sense_Rate’ 6. ’Low_Quality_Reads’ 7. ’Mapped_Unique_Reads’ 8. ’NonGlobin_Reads’ 9. ’Intronic_Reads’

[6]:

```
caudate <- remove_variables(caudate)
dlpfc   <- remove_variables(dlpfc)
hippo   <- remove_variables(hippo)
```

```
Correlation computed with
• Method: 'pearson'
• Missing treated using: 'pairwise.complete.obs'
Correlation computed with
• Method: 'pearson'
• Missing treated using: 'pairwise.complete.obs'
Correlation computed with
• Method: 'pearson'
• Missing treated using: 'pairwise.complete.obs'
Correlation computed with
• Method: 'pearson'
• Missing treated using: 'pairwise.complete.obs'
Correlation computed with
• Method: 'pearson'
• Missing treated using: 'pairwise.complete.obs'
Correlation computed with
• Method: 'pearson'
• Missing treated using: 'pairwise.complete.obs'
```

[7]:

```
caudate |> dim()
dlpfc |> dim()
hippo |> dim()
```

1. 418 2. 51

1. 388 2. 46

1. 408 2. 48

1.2.3 Commone variables

[8]:

```
vars <- intersect(colnames(caudate),intersect(colnames(dlpfc), colnames(hippo)))
vars
```

1. ’SAMPLE_ID’ 2. ’RNum’ 3. ’Region’ 4. ’Dataset’ 5. ’Protocol’ 6. ’RIN’ 7. ’BrNum’ 8. ’Dx’ 9. ’Race’ 10. ’Sex’ 11. ’Age’ 12. ’PMI’ 13. ’MoD’ 14. ’Mapping_Rate’ 15. ’Base_Mismatch’ 16. ’ExprProfEff’ 17. ’Intergenic_Rate’ 18. ’totalAssignedGene’ 19. ’Ambiguous_Alignment_Rate’ 20. ’rRNA_rate’ 21. ’End_1_Sense_Rate’ 22. ’Chimeric_Alignment_Rate’ 23. ’Low_Mapping_Quality’ 24. ’Genes_Detected’ 25. ’Mean3Bias’ 26. ’totalMapped’ 27. ’Intergenic_Reads’ 28. ’Read_Length’ 29. ’Mito_mapped’ 30. ’globinRate’ 31. ’IID’ 32. ’SOL’ 33. ’snpPC1’ 34. ’snpPC2’ 35. ’snpPC3’ 36. ’snpPC4’ 37. ’snpPC5’ 38. ’snpPC6’ 39. ’snpPC7’ 40. ’snpPC8’ 41. ’snpPC9’ 42. ’snpPC10’ 43. ’New_Dx’ 44. ’antipsychotics’ 45. ’lifetime_antipsych’

[9]:

```r
length(vars)
```

45

[10]:

```r
data.frame("Variables"=vars) |>
    data.table::fwrite("shared_variables.tsv", sep='\t')
```

1.3 Reproducibility

[11]:

```r
Sys.time()
proc.time()
options(width = 120)
sessioninfo::session_info()$platform
sessioninfo::session_info()$packages
```

```
[1] "2023-02-13 18:45:00 EST"
```

```
   user  system elapsed
  5.183   0.195   7.075
```

$version ’R version 4.2.2 (2022-10-31)’

$os ’Arch Linux’

$system ’x86_64, linux-gnu’

$ui ’X11’

$language ’(EN)’

$collate ’en_US.UTF-8’

$ctype ’en_US.UTF-8’

$tz ’America/New_York’

$date ’2023-02-13’

$pandoc ’3.0.1 @ /usr/bin/pandoc’

A packages_info: 57 × 11

	package	ondiskversion	loadedversion	p
	<chr>	<chr>	<chr>	<
base64enc	base64enc	0.1.3	0.1-3	/
Biobase	Biobase	2.58.0	2.58.0	/
BiocGenerics	BiocGenerics	0.44.0	0.44.0	/
bitops	bitops	1.0.7	1.0-7	/
cli	cli	3.6.0	3.6.0	/
colorspace	colorspace	2.1.0	2.1-0	/
corrr	corrr	0.4.4	0.4.4	/
crayon	crayon	1.5.2	1.5.2	/
data.table	data.table	1.14.6	1.14.6	/
DelayedArray	DelayedArray	0.24.0	0.24.0	/
digest	digest	0.6.31	0.6.31	/
dplyr	dplyr	1.1.0	1.1.0	/
evaluate	evaluate	0.20	0.20	/
fansi	fansi	1.0.4	1.0.4	/
fastmap	fastmap	1.1.0	1.1.0	/
generics	generics	0.1.3	0.1.3	/
GenomeInfoDb	GenomeInfoDb	1.34.9	1.34.9	/
GenomeInfoDbData	GenomeInfoDbData	1.2.9	1.2.9	/
GenomicRanges	GenomicRanges	1.50.2	1.50.2	/
ggplot2	ggplot2	3.4.1	3.4.1	/
glue	glue	1.6.2	1.6.2	/
gtable	gtable	0.3.1	0.3.1	/
here	here	1.0.1	1.0.1	/
htmltools	htmltools	0.5.4	0.5.4	/
IRanges	IRanges	2.32.0	2.32.0	/
IRdisplay	IRdisplay	1.1	1.1	/
IRkernel	IRkernel	1.3.2	1.3.2	/
jsonlite	jsonlite	1.8.4	1.8.4	/
lattice	lattice	0.20.45	0.20-45	/
lifecycle	lifecycle	1.0.3	1.0.3	/
magrittr	magrittr	2.0.3	2.0.3	/
Matrix	Matrix	1.5.3	1.5-3	/
MatrixGenerics	MatrixGenerics	1.10.0	1.10.0	/
matrixStats	matrixStats	0.63.0	0.63.0	/
munsell	munsell	0.5.0	0.5.0	/
pbdZMQ	pbdZMQ	0.3.9	0.3-9	/
pillar	pillar	1.8.1	1.8.1	/
pkgconfig	pkgconfig	2.0.3	2.0.3	/
purrr	purrr	1.0.1	1.0.1	/
R6	R6	2.5.1	2.5.1	/
RCurl	RCurl	1.98.1.10	1.98-1.10	/
repr	repr	1.1.6	1.1.6	/
rlang	rlang	1.0.6	1.0.6	/
rprojroot	rprojroot	2.0.3	2.0.3	/
S4Vectors	S4Vectors	0.36.1	0.36.1	/
scales	scales	1.2.1	1.2.1	/
sessioninfo	sessioninfo	1.2.2	1.2.2	/
SummarizedExperiment	SummarizedExperiment	1.28.0	1.28.0	/
tibble	tibble	3.1.8	3.1.8	/
tidyr	tidyr	1.3.0	1.3.0	/
tidyselect	tidyselect	1.2.0	1.2.0	/