

main

August 31, 2021

1 Examine tissue specific genes for correlation with gene expression or cell type proportion

```
[1]: library(dplyr)
```

Attaching package: 'dplyr'

The following objects are masked from 'package:stats':

filter, lag

The following objects are masked from 'package:base':

intersect, setdiff, setequal, union

1.1 Functions

```
[2]: get_tpm <- function(){
  cc_file = paste0("/ceph/projects/v4_phase3_paper/inputs/counts/",
                  "text_files_counts/tpm/_m/caudate/gene/log2tpm.csv")
  dd_file = paste0("/ceph/projects/v4_phase3_paper/inputs/counts/",
                  "text_files_counts/tpm/_m/dlpfc/gene/log2tpm.csv")
  hh_file = paste0("/ceph/projects/v4_phase3_paper/inputs/counts/",
                  "text_files_counts/tpm/_m/hippocampus/gene/log2tpm.csv")
  cc = data.table::fread(cc_file) %>% tibble::column_to_rownames("names") %>%
    t %>% as.data.frame %>% tibble::rownames_to_column("RNum")
  dd = data.table::fread(dd_file) %>% tibble::column_to_rownames("names") %>%
    t %>% as.data.frame %>% tibble::rownames_to_column("RNum")
  hh = data.table::fread(hh_file) %>% tibble::column_to_rownames("names") %>%
    t %>% as.data.frame %>% tibble::rownames_to_column("RNum")
  return(bind_rows(cc, hh, dd))
}
memTPM <- memoise::memoise(get_tpm)
```

```

get_pheno <- function(){
  filename = "/ceph/projects/v4_phase3_paper/inputs/phenotypes/_m/
  ↪merged_phenotypes.csv"
  df = data.table::fread(filename) %>%
    filter(Age > 13, Race %in% c("AA", "EA"), Dx %in% c("CTL", "SZ"))
  return(df)
}
memPHENO <- memoise::memoise(get_pheno)

```

2 Extract tissue specific eGenes

```

[3]: eFeature = data.table::fread("../_m/genes/significant_geneSNP_pairs_3tissues.
  ↪tsv") %>%
  filter(N_Regions_Shared == 1) %>% select(-N_Regions_Shared)
eFeature %>% head(2)

```

	gene_id	variant_id	Caudate	DLPFC	Hippocampus
	<chr>	<chr>	<int>	<int>	<int>
A data.table: 2 × 5	ENSG00000002746.14	chr7:43035480:T:C	1	0	0
	ENSG00000003096.13	chrX:118122592:G:A	1	0	0

2.1 Prepare data

```

[4]: df = memPHENO() %>% inner_join(memTPM(), by="RNum")
df %>% dim

```

1. 1173 2. 49603

2.2 Linear model for expression and brain region

```

[5]: pvals = c(); genes = c()
for(gene_id in eFeature$gene_id){
  model = paste(paste0(gene_id, "~ Region"), "Dx + Sex + Age + mitoRate +_
  ↪rRNA_rate",
               "overallMapRate + RIN + ERCCsumLogErr + totalAssignedGene +_
  ↪snpPC1",
               "snpPC2 + snpPC3", sep=" + ")
  fitted = anova(lm(model, data=df))
  #fit_lm = aov(lm(model, data=df))
  pvals = c(pvals, fitted["Region", "Pr(>F)"])
  genes = c(genes, gene_id)
}
pval_df = data.frame("gene_id"=genes, "p_values"=pvals)
print(sum(pvals > 0.05))
pval_df %>% head(2)

```

[1] 1

		gene_id	p_values
		<chr>	<dbl>
A data.frame: 2 × 2	1	ENSG00000002746.14	0
	2	ENSG00000003096.13	0

2.3 Comparison of expression

```
[6]: dt = df %>% select(Region, all_of(eFeature$gene_id)) %>%
  aggregate(. ~ Region, ., mean) %>%
  mutate(Region = gsub("HIPPO", "Hippocampus", Region)) %>%
  tibble::column_to_rownames("Region") %>%
  t %>% as.data.frame %>% tibble::rownames_to_column("gene_id") %>%
  inner_join(eFeature, by="gene_id", suffix=c("_Expression", "_eQTL")) %>%
  select(-c("variant_id")) %>% inner_join(pval_df, by="gene_id")
tt = dt %>% select(ends_with("Expression"))
dt = dt %>% mutate("Max Expression"=gsub("_Expression", "",
  ↳colnames(tt)[apply(tt, 1, which.max)]),
  "Min Expression"=gsub("_Expression", "",
  ↳colnames(tt)[apply(tt, 1, which.min)]),
  "Mean Expression"=rowMeans(tt),
  "Ratio (DLPFC / Caudate)" = DLPFC_Expression/
  ↳Caudate_Expression,
  "Ratio (Hippocampus / Caudate)" = Hippocampus_Expression/
  ↳Caudate_Expression,
  "Ratio (Hippocampus / DLPFC)" = Hippocampus_Expression/
  ↳DLPFC_Expression)
dt %>% data.table::fwrite("eQTL_regionSpecific_summary.tsv", sep='\t')
dt %>% head(2)
```

		gene_id	Caudate_Expression	DLPFC_Expression	Hippocampus_Expression
		<chr>	<dbl>	<dbl>	<dbl>
A data.frame: 2 × 14	1	ENSG00000002746.14	3.911108	4.002739	1.966804
	2	ENSG00000003096.13	5.132414	1.538676	1.724228

```
[7]: sum(dt$`Ratio (DLPFC / Caudate)` > 0.9)
```

284

```
[8]: sum(dt$`Ratio (Hippocampus / Caudate)` > 0.9)
```

86

```
[9]: nochange = sum(dt$`Ratio (DLPFC / Caudate)` > 0.9) + sum(dt$`Ratio (Hippocampus_
  ↳ / Caudate)` > 0.9)
print(nochange)
nochange / dim(eFeature)[1]
```

[1] 370

0.185185185185185

```
[10]: sum(dt$`Ratio (Hippocampus / DLPFC)` > 0.9)
```

325

```
[11]: ## Low expression genes
sum(dt$`Mean Expression` < 1)
sum(dt$`Mean Expression` < 1) / dim(eFeature)[1]
```

238

0.119119119119119

```
[12]: sum(dt$Caudate_eQTL == 1 & dt$`Max Expression` == "Caudate")
sum(dt$Caudate_eQTL == 1 & dt$`Max Expression` == "Caudate" &
      (dt$`Ratio (DLPFC / Caudate)` < 0.9 | dt$`Ratio (Hippocampus / Caudate)` <
      ↪0.9))
sum(dt$Caudate_eQTL == 1 & dt$`Max Expression` == "Caudate" &
      (dt$`Ratio (DLPFC / Caudate)` < 0.9 | dt$`Ratio (Hippocampus / Caudate)` <
      ↪0.9)) / dim(eFeature)[1]
```

1741

1738

0.86986986986987

```
[13]: sum(dt$DLPFC_eQTL == 1 & dt$`Max Expression` == "DLPFC")
sum(dt$Hippocampus_eQTL == 1 & dt$`Max Expression` == "Hippocampus")
```

0

0

```
[14]: sum(dt$Caudate_eQTL == 1 & dt$`Min Expression` == "Caudate")
sum(dt$DLPFC_eQTL == 1 & dt$`Min Expression` == "DLPFC")
sum(dt$Hippocampus_eQTL == 1 & dt$`Min Expression` == "Hippocampus")
```

52

5

0

```
[15]: sum(eFeature$Caudate == 1)
sum(eFeature$DLPFC == 1)
sum(eFeature$Hippocampus == 1)
```

1932

66

0

2.3.1 Summary

- All specific genes are caudate, and caudate has the highest expression!

2.4 Reproducibility information

```
[16]: Sys.time()  
proc.time()  
options(width=120)  
sessioninfo::session_info()
```

```
[1] "2021-08-31 07:05:00 EDT"
```

```
      user  system elapsed  
144.475   10.422  143.881
```

Session info

```
setting  value
```

```
version  R version 4.0.3 (2020-10-10)
```

```
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     2021-08-31
```

Packages

package	* version	date	lib	source
assertthat	0.2.1	2019-03-21	[1]	CRAN (R 4.0.2)
base64enc	0.1-3	2015-07-28	[1]	CRAN (R 4.0.2)
cachem	1.0.5	2021-05-15	[1]	CRAN (R 4.0.3)
cli	3.0.0	2021-06-30	[1]	CRAN (R 4.0.3)
crayon	1.4.1	2021-02-08	[1]	CRAN (R 4.0.3)
data.table	1.14.0	2021-02-21	[1]	CRAN (R 4.0.3)
DBI	1.1.1	2021-01-15	[1]	CRAN (R 4.0.2)
digest	0.6.27	2020-10-24	[1]	CRAN (R 4.0.2)
dplyr	* 1.0.7	2021-06-18	[1]	CRAN (R 4.0.3)
ellipsis	0.3.2	2021-04-29	[1]	CRAN (R 4.0.3)
evaluate	0.14	2019-05-28	[1]	CRAN (R 4.0.2)
fansi	0.5.0	2021-05-25	[1]	CRAN (R 4.0.3)
fastmap	1.1.0	2021-01-25	[1]	CRAN (R 4.0.2)
generics	0.1.0	2020-10-31	[1]	CRAN (R 4.0.2)
glue	1.4.2	2020-08-27	[1]	CRAN (R 4.0.2)
htmltools	0.5.1.1	2021-01-22	[1]	CRAN (R 4.0.2)
IRdisplay	1.0	2021-01-20	[1]	CRAN (R 4.0.2)

IRkernel	1.2	2021-05-11	[1]	CRAN	(R 4.0.3)
jsonlite	1.7.2	2020-12-09	[1]	CRAN	(R 4.0.2)
lifecycle	1.0.0	2021-02-15	[1]	CRAN	(R 4.0.3)
magrittr	2.0.1	2020-11-17	[1]	CRAN	(R 4.0.2)
memoise	2.0.0	2021-01-26	[1]	CRAN	(R 4.0.2)
pbdZMQ	0.3-5	2021-02-10	[1]	CRAN	(R 4.0.3)
pillar	1.6.1	2021-05-16	[1]	CRAN	(R 4.0.3)
pkgconfig	2.0.3	2019-09-22	[1]	CRAN	(R 4.0.2)
purrr	0.3.4	2020-04-17	[1]	CRAN	(R 4.0.2)
R6	2.5.0	2020-10-28	[1]	CRAN	(R 4.0.2)
repr	1.1.3	2021-01-21	[1]	CRAN	(R 4.0.2)
rlang	0.4.11	2021-04-30	[1]	CRAN	(R 4.0.3)
sessioninfo	1.1.1	2018-11-05	[1]	CRAN	(R 4.0.2)
tibble	3.1.2	2021-05-16	[1]	CRAN	(R 4.0.3)
tidyselect	1.1.1	2021-04-30	[1]	CRAN	(R 4.0.3)
utf8	1.2.1	2021-03-12	[1]	CRAN	(R 4.0.3)
uuid	0.1-4	2020-02-26	[1]	CRAN	(R 4.0.2)
vctrs	0.3.8	2021-04-29	[1]	CRAN	(R 4.0.3)
withr	2.4.2	2021-04-18	[1]	CRAN	(R 4.0.3)

[1] /home/jbenja13/R/x86_64-pc-linux-gnu-library/4.0

[2] /usr/lib/R/library