main

August 26, 2022

1 Examine sample make-up

[1]:

```
suppressMessages({library(SummarizedExperiment)
                  library(tidyverse)
                  library(ggpubr)})
```

1.1 Samples after quality control

[2]:

```
save_ggplots <- function(p, fn, w, h){
    for(ext in c('.pdf', '.svg')){
        ggsave(paste0(fn, ext), plot=p, width=w, height=h)
    }
}
```

1.1.1 Load Caudate data

[3]:

```
# Load counts and phenotype R variable
load("../../input/counts/_m/caudate_brainseq_phase3_hg38_rseGene_merged_n464.
 ↪rda")
### Subset and recode
keepIndex = which((rse_gene$Dx %in% c('Control', "Schizo")) &
                  rse_gene$Race %in% c('CAUC', 'AA'))
rse_gene = rse_gene[, keepIndex]
### Extract phenotypes
pheno_C <- colData(rse_gene) %>% as.data.frame
```

1.1.2 Load DLPFC data

[4]:

```
# Load counts and phenotype R variable
load("../../input/counts/_m/
 ↪dlpfc_ribozero_brainseq_phase2_hg38_rseGene_merged_n453.rda")
### Subset and recode
keepIndex = which((rse_gene$Dx %in% c('Control', "Schizo")) &
                  rse_gene$Race %in% c('CAUC', 'AA'))
rse_gene = rse_gene[, keepIndex]
### Extract phenotypes
pheno_D <- colData(rse_gene) %>% as.data.frame
```

1.1.3 Load Hippocampus data

[5]:

```
# Load counts and phenotype R variable
load("../../input/counts/_m/hippo_brainseq_phase2_hg38_rseGene_merged_n447.rda")
### Subset and recode
keepIndex = which((rse_gene$Dx %in% c('Control', "Schizo")) &
                  rse_gene$Race %in% c('CAUC', 'AA'))
rse_gene = rse_gene[, keepIndex]
### Extract phenotypes
pheno_H <- colData(rse_gene) %>% as.data.frame
```

1.1.4 Load DG data

[6]:

```
# Load counts and phenotype R variable
load("../../input/counts/_m/astellas_dg_hg38_rseGene_n263.rda")
### Subset and recode
keepIndex = which((rse_gene$Dx %in% c('Control', "Schizo")) &
                  rse_gene$Race %in% c('CAUC', 'AA'))
rse_gene = rse_gene[, keepIndex]
### Extract phenotypes
pheno_dg <- colData(rse_gene) %>% as.data.frame
```

1.1.5 Merge data

[7]:

```
allCols <- intersect(intersect(intersect(colnames(pheno_C), colnames(pheno_D)),
                               colnames(pheno_H)),
                     colnames(pheno_dg))
pheno = rbind(pheno_C[, allCols], pheno_D[, allCols],
              pheno_H[, allCols], pheno_dg[, allCols]) %>%
    filter(Age > 17) %>% mutate(Race=gsub("CAUC", "EA", Race))
```

1.2 STRUCTURE analysis

[8]:

```
ancestry = data.table::fread("../../input/ancestry_structure/structure.
↪out_ancestry_proportion_raceDemo_compare")
ancestry %>% head(2)
```

A data.table: 2 × 4

id	Afr	Eur	group
<chr>	<dbl>	<dbl>	<chr>
Br2374	0.007	0.993	CAUC
Br1857	0.001	0.999	CAUC

[9]:

```
ancestry %>% mutate_if(is.character, as.factor) %>%
    group_by(group) %>% summarize(AA=mean(Afr), EA=mean(Eur))
```

A tibble: 2 × 3

group	AA	EA
<fct>	<dbl>	<dbl>
AA	0.782219451	0.2177805
CAUC	0.007510536	0.9924895

[10]:

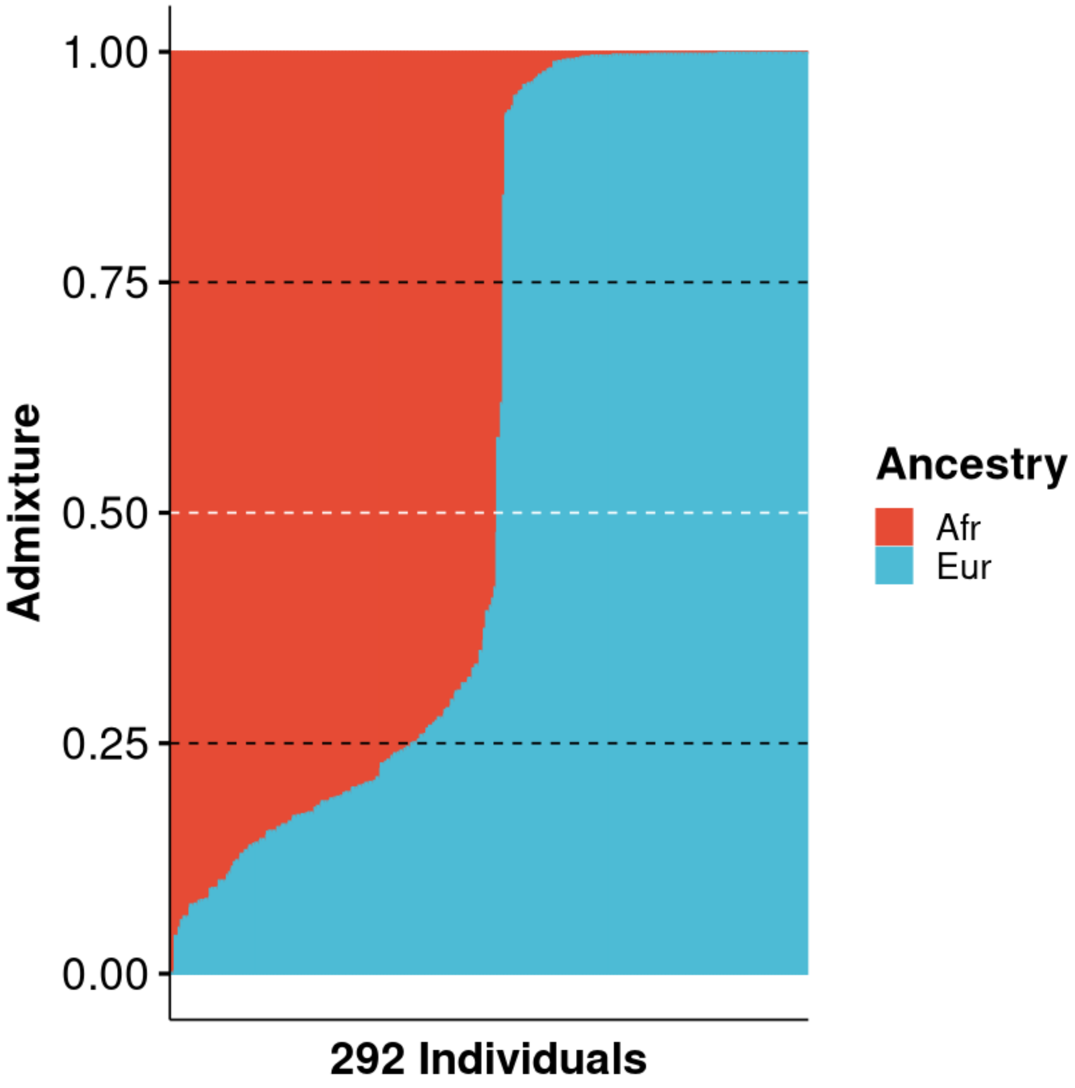
```
ancestry %>% inner_join(pheno, by=c("id"="BrNum")) %>%
    filter(Age > 17, Dx == "Control") %>% select(group, Afr, Eur) %>%
    mutate_if(is.character, as.factor) %>% distinct %>%
    group_by(group) %>%
    summarize(AA_mean=mean(Afr), AA_sd=sd(Afr), AA_max=max(Afr),␣
  ↪AA_min=min(Afr),
              EA_mean=mean(Eur), EA_sd=sd(Eur), EA_max=max(Eur),␣
  ↪EA_min=min(Eur))
```

A tibble: 2 × 9

group	AA_mean	AA_sd	AA_max	AA_min	EA_mean	EA_sd	EA_max
<fct>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
AA	0.78962609	0.10611682	0.999	0.381	0.2103739	0.10611682	0.619
CAUC	0.03087879	0.02997578	0.156	0.001	0.9691212	0.02997578	0.999

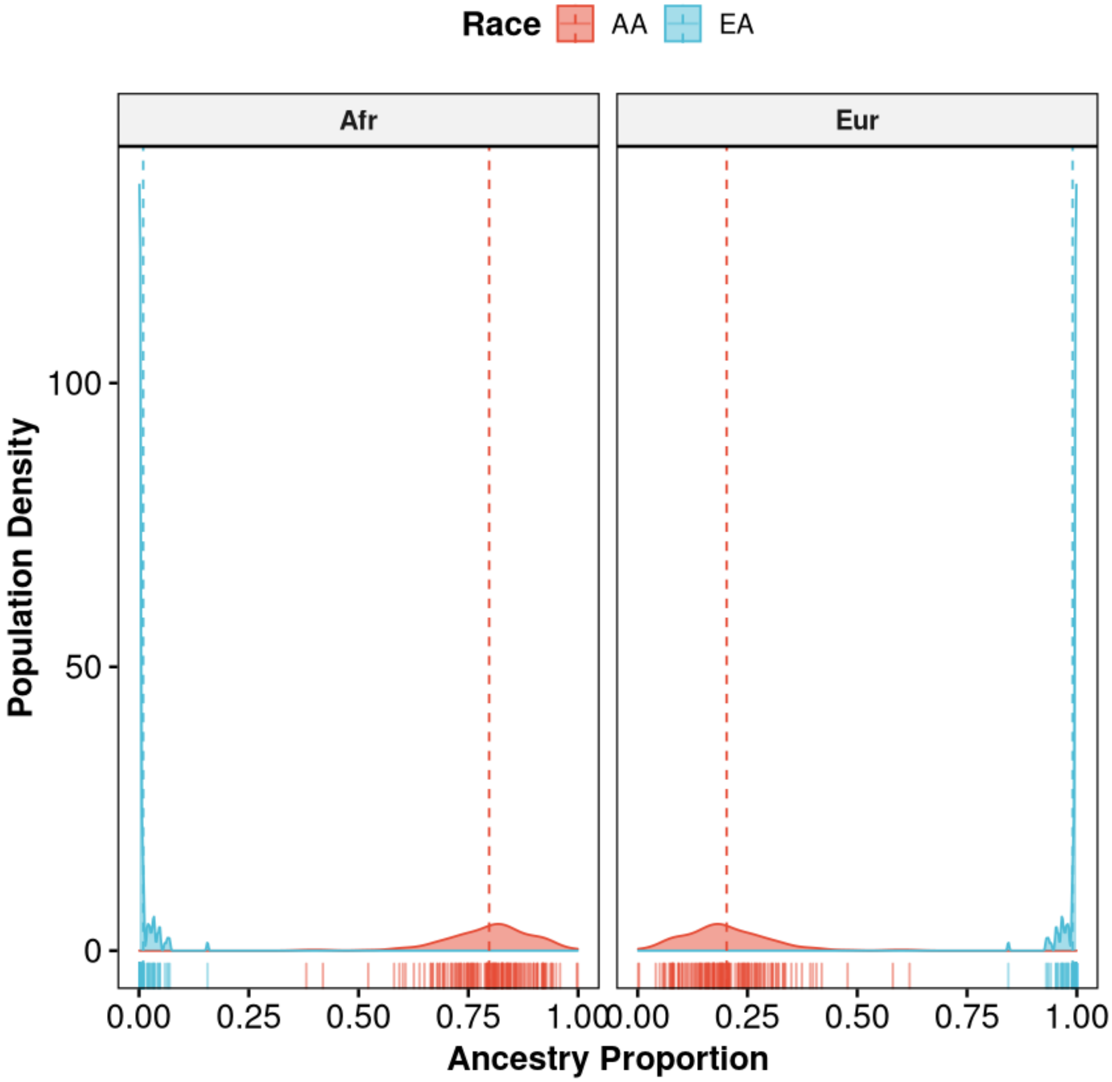
[11]:

```
brp = ancestry %>% inner_join(pheno, by=c("id"="BrNum")) %>%
    filter(Age > 17, Dx == "Control") %>% select(id, Race, Afr, Eur) %>%
    mutate_if(is.character, as.factor) %>% distinct %>%
    pivot_longer(-c("Race", "id"), names_to="Ancestry", values_to="Proportion")␣
  ↪%>%
    mutate_if(is.character, as.factor) %>% group_by(Ancestry) %>%
    mutate(ID = fct_reorder(id, desc(Proportion))) %>%
    ggbarplot(x="ID", y="Proportion", fill = "Ancestry", color="Ancestry",
              palette="npg", ylab="Admixture", xlab="292 Individuals",
              ggtheme=theme_pubr(base_size=20), legend="right") +
    geom_hline(yintercept=0.5, linetype="dashed", color="white") +
    geom_hline(yintercept=0.75, linetype="dashed", color="black") +
    geom_hline(yintercept=0.25, linetype="dashed", color="black") +
    font("xy.title", face="bold") + font("legend.title", face="bold") +
    rremove("x.text") + rremove("x.ticks")
save_ggplots(brp, "ancestry_structure_barplot", 12, 5)
brp
```



[12]:

```
bxp = ancestry %>% inner_join(pheno, by=c("id"="BrNum")) %>%
    filter(Age > 17, Dx == "Control") %>% select(id, Race, Afr, Eur) %>%
    mutate_if(is.character, as.factor) %>% distinct %>%
    pivot_longer(-c("Race", "id"), names_to="Ancestry", values_to="Proportion")␣
 ↪%>%
    ggdensity(x="Proportion", color="Race", fill="Race", facet.by="Ancestry",
              ncol=2, rug=TRUE, add="mean", palette="npg", ylab="Population␣
 ↪Density",
              xlab="Ancestry Proportion", panel.labs.font=list(face='bold'),
              ggtheme=theme_pubr(base_size=15, border=TRUE)) +
    font("xy.title", face="bold") + font("legend.title", face="bold")
save_ggplots(bxp, "ancestry_structure_distribution", 10, 5)
bxp
```



1.3 eQTL analysis

[13]:
```
pheno %>% dim
```

1. 1291 2. 21

[14]:
```
print(paste("There are", unique(pheno$BrNum) %>% length, "unique BrNum."))
```

```
[1] "There are 485 unique BrNum."
```

[15]:
```
pheno %>% select(BrNum, Region) %>% distinct %>%
    mutate_if(is.character, as.factor) %>%
    group_by(Region) %>% count()
```

A grouped_df: 4 × 2

Region	n
<fct>	<int>
Caudate	394
DentateGyrus	161
DLPFC	360
HIPPO	376

```
[16]: pheno %>% select(BrNum, Race) %>% distinct %>%
          mutate_if(is.character, as.factor) %>%
          group_by(Race) %>% count()
```

A grouped_df: 2 × 2

Race	n
<fct>	<int>
AA	249
EA	236

```
[17]: pheno %>% select(BrNum, Race, Region) %>% distinct %>%
          mutate_if(is.character, as.factor) %>%
          group_by(Region, Race) %>% count()
```

A grouped_df: 8 × 3

Region	Race	n
<fct>	<fct>	<int>
Caudate	AA	205
Caudate	EA	189
DentateGyrus	AA	78
DentateGyrus	EA	83
DLPFC	AA	200
DLPFC	EA	160
HIPPO	AA	207
HIPPO	EA	169

```
[18]: pheno %>% select(BrNum, Sex, Region) %>% distinct %>%
          mutate_if(is.character, as.factor) %>%
          group_by(Region, Sex) %>% count()
```

A grouped_df: 8 × 3

Region	Sex	n
<fct>	<fct>	<int>
Caudate	F	121
Caudate	M	273
DentateGyrus	F	48
DentateGyrus	M	113
DLPFC	F	114
DLPFC	M	246
HIPPO	F	121
HIPPO	M	255

```
[19]: pheno %>% group_by(Region) %>%
        summarise_at(vars(c("Age")), list(mean = mean, sd = sd))
```

A tibble: 4 × 3

Region	mean	sd
<chr>	<dbl>	<dbl>
Caudate	49.65508	15.58123
DentateGyrus	50.06770	15.43849
DLPFC	47.36772	15.36858
HIPPO	47.03652	15.28105

```
[20]: pheno %>% group_by(Region, Race) %>%
        summarise_at(vars(c("Age")), list(mean = mean, sd = sd))
```

A grouped_df: 8 × 4

Region	Race	mean	sd
<chr>	<chr>	<dbl>	<dbl>
Caudate	AA	48.98595	14.31824
Caudate	EA	50.38085	16.85304
DentateGyrus	AA	50.18423	15.53374
DentateGyrus	EA	49.95819	15.44210
DLPFC	AA	47.63338	14.77009
DLPFC	EA	47.03565	16.12621
HIPPO	AA	47.26860	14.84346
HIPPO	EA	46.75225	15.84035

```
[21]: pheno %>% filter(RIN != "NA") %>% mutate("RIN"=as.numeric(unlist(RIN))) %>%
          group_by(Region) %>% summarise_at(vars(c("RIN")), list(mean = mean, sd =␣
       ↪sd))
```

A tibble: 4 × 3

Region	mean	sd
<chr>	<dbl>	<dbl>
Caudate	7.860152	0.8665752
DentateGyrus	5.208403	1.1871187
DLPFC	7.667222	0.9209920
HIPPO	7.598138	1.0308426

```
[22]: pheno %>% filter(RIN != "NA") %>% mutate("RIN"=as.numeric(unlist(RIN))) %>%
          group_by(Region, Race) %>% summarise_at(vars(c("RIN")), list(mean = mean,␣
       ↪sd = sd))
```

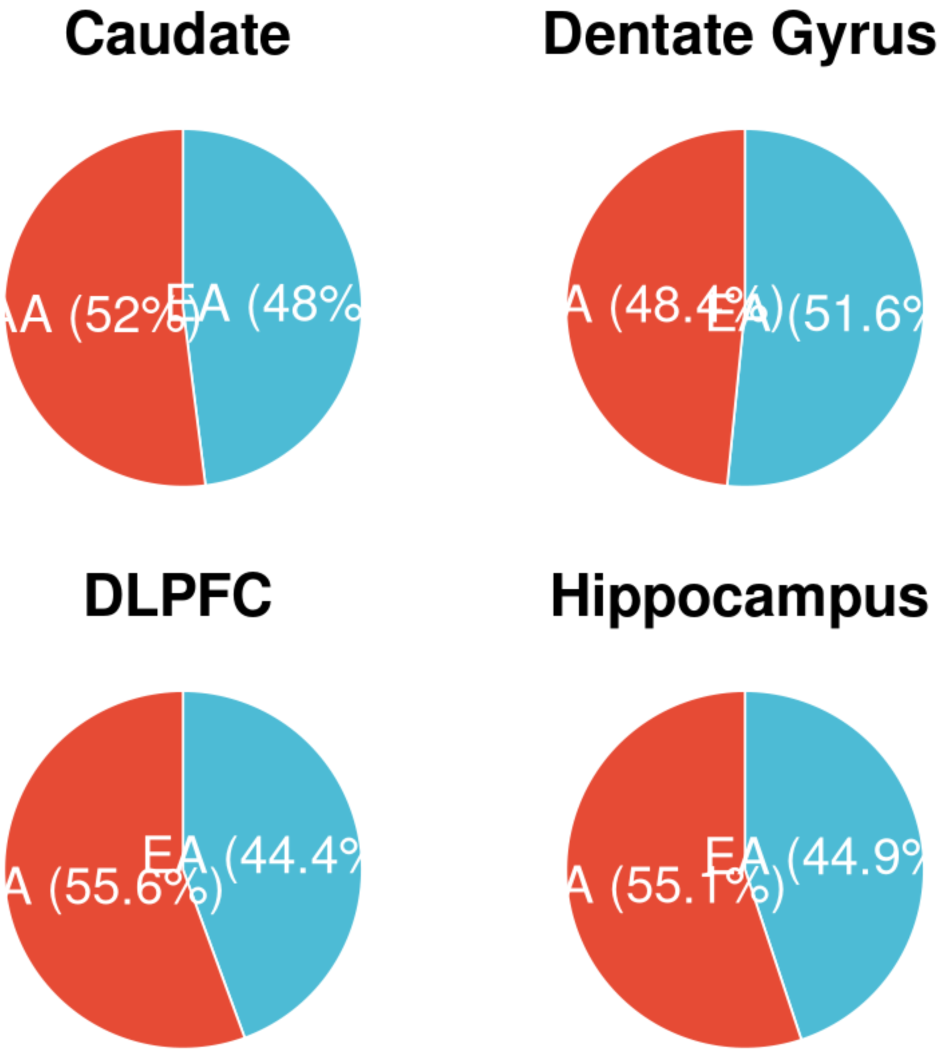
A grouped_df: 8 × 4

Region	Race	mean	sd
<chr>	<chr>	<dbl>	<dbl>
Caudate	AA	7.860976	0.8435098
Caudate	EA	7.859259	0.8931664
DentateGyrus	AA	5.206349	1.2062837
DentateGyrus	EA	5.210714	1.1760765
DLPFC	AA	7.661500	0.9452169
DLPFC	EA	7.674375	0.8926849
HIPPO	AA	7.582126	1.0549556
HIPPO	EA	7.617751	1.0032885

1.3.1 Pie chart

[23]:
```
plot_pie <- function(tissue){
    pie = pheno %>% mutate_if(is.character, as.factor) %>% group_by(Region,␣
  ↪Race) %>%
        count %>% as.data.frame %>% group_by(Region) %>%
        transmute(Race, Percent = round(n/sum(n)*100, 1)) %>%
        mutate(Labels=paste0(Race, " (", Percent, "%)")) %>% filter(Region ==␣
  ↪tissue) %>%
        ggpie("Percent", label="Labels", fill="Race", color="white",␣
  ↪palette="npg",
              lab.pos="in", lab.font=c(8, "bold", "white"),
              ggtheme=theme_pubr(base_size=20, legend="none"))
    return(pie)
}
```

[24]:
```
## Get and annotate plot
cc_pie = annotate_figure(plot_pie("Caudate"),
                         top = text_grob("Caudate", face = "bold", size = 26))
gg_pie = annotate_figure(plot_pie("DentateGyrus"),
                         top = text_grob("Dentate Gyrus", face = "bold", size =␣
  ↪26))
dd_pie = annotate_figure(plot_pie("DLPFC"),
                         top = text_grob("DLPFC", face = "bold", size = 26))
hh_pie = annotate_figure(plot_pie("HIPPO"),
                         top = text_grob("Hippocampus", face = "bold", size =␣
  ↪26))
## Arrange figure
figure <- ggarrange(cc_pie, gg_pie, dd_pie, hh_pie, ncol = 2, nrow = 2)
save_ggplots(figure, "ancestry_piecharts", 10, 10)
figure
```



1.4 Adult individuals for expression related analysis

[25]:
```
pheno = pheno %>% filter(Age > 17, Dx == "Control", Race == "AA")
pheno %>% dim
```

1. 425 2. 21

[26]:
```
print(paste("There are", unique(pheno$BrNum) %>% length, "unique BrNum."))
```

```
[1] "There are 151 unique BrNum."
```

[27]:

```
pheno %>% select(BrNum, Region) %>% distinct %>%
    mutate_if(is.character, as.factor) %>%
    group_by(Region) %>% count()
```

A grouped_df: 4 × 2

Region	n
<fct>	<int>
Caudate	122
DentateGyrus	47
DLPFC	123
HIPPO	133

[28]:

```
pheno %>% select(BrNum, Race) %>% distinct %>%
    mutate_if(is.character, as.factor) %>%
    group_by(Race) %>% count()
```

A grouped_df: 1 × 2

Race	n
<fct>	<int>
AA	151

[29]:

```
pheno %>% select(BrNum, Race, Region) %>% distinct %>%
    mutate_if(is.character, as.factor) %>%
    group_by(Region, Race) %>% count()
```

A grouped_df: 4 × 3

Region	Race	n
<fct>	<fct>	<int>
Caudate	AA	122
DentateGyrus	AA	47
DLPFC	AA	123
HIPPO	AA	133

[30]:

```
pheno %>% select(BrNum, Sex, Region) %>% distinct %>%
    mutate_if(is.character, as.factor) %>%
    group_by(Region, Sex) %>% count()
```

A grouped_df: 8 × 3

Region	Sex	n
<fct>	<fct>	<int>
Caudate	F	50
Caudate	M	72
DentateGyrus	F	16
DentateGyrus	M	31
DLPFC	F	48
DLPFC	M	75
HIPPO	F	53
HIPPO	M	80

[31]:

```
pheno %>% group_by(Region) %>%
  summarise_at(vars(c("Age")), list(mean = mean, sd = sd))
```

A tibble: 4 × 3

Region	mean	sd
<chr>	<dbl>	<dbl>
Caudate	45.63770	14.72979
DentateGyrus	45.85043	16.32827
DLPFC	44.12511	14.97092
HIPPO	43.30015	14.73609

```
[32]: pheno %>% group_by(Region, Race) %>%
        summarise_at(vars(c("Age")), list(mean = mean, sd = sd))
```

A grouped_df: 4 × 4

Region	Race	mean	sd
<chr>	<chr>	<dbl>	<dbl>
Caudate	AA	45.63770	14.72979
DentateGyrus	AA	45.85043	16.32827
DLPFC	AA	44.12511	14.97092
HIPPO	AA	43.30015	14.73609

```
[33]: pheno %>% filter(RIN != "NA") %>% mutate("RIN"=as.numeric(unlist(RIN))) %>%
          group_by(Region) %>% summarise_at(vars(c("RIN")), list(mean = mean, sd =␣
        ↪sd))
```

A tibble: 4 × 3

Region	mean	sd
<chr>	<dbl>	<dbl>
Caudate	7.829508	0.7993477
DentateGyrus	5.447368	1.2173824
DLPFC	7.696748	0.8851169
HIPPO	7.715038	0.9754173

```
[34]: pheno %>% filter(RIN != "NA") %>% mutate("RIN"=as.numeric(unlist(RIN))) %>%
          group_by(Region, Race) %>% summarise_at(vars(c("RIN")), list(mean = mean,␣
        ↪sd = sd))
```

A grouped_df: 4 × 4

Region	Race	mean	sd
<chr>	<chr>	<dbl>	<dbl>
Caudate	AA	7.829508	0.7993477
DentateGyrus	AA	5.447368	1.2173824
DLPFC	AA	7.696748	0.8851169
HIPPO	AA	7.715038	0.9754173

1.5 Reproducibility Information

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

```
[1] "2022-08-26 12:09:00 EDT"
```

```
    user   system  elapsed 
  31.643    1.333 1452.757 
```

$platform **$version** ’R version 4.2.1 (2022-06-23)’

$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 ’2022-08-26’

$pandoc ’2.18 @ /usr/bin/pandoc’

$packages A packages_info: 93 × 11

	package <chr>	ondiskversion <chr>	loade <chr
abind	abind	1.4.5	1.4-5
assertthat	assertthat	0.2.1	0.2.1
backports	backports	1.4.1	1.4.1
base64enc	base64enc	0.1.3	0.1-3
Biobase	Biobase	2.56.0	2.56.
BiocGenerics	BiocGenerics	0.42.0	0.42.
bitops	bitops	1.0.7	1.0-7
broom	broom	1.0.0	1.0.0
car	car	3.1.0	3.1-0
carData	carData	3.0.5	3.0-5
cellranger	cellranger	1.1.0	1.1.0
cli	cli	3.3.0	3.3.0
colorspace	colorspace	2.0.3	2.0-3
cowplot	cowplot	1.1.1	1.1.1
crayon	crayon	1.5.1	1.5.1
data.table	data.table	1.14.2	1.14.
DBI	DBI	1.1.3	1.1.3
dbplyr	dbplyr	2.2.1	2.2.1
DelayedArray	DelayedArray	0.22.0	0.22.
digest	digest	0.6.29	0.6.2
dplyr	dplyr	1.0.9	1.0.9
ellipsis	ellipsis	0.3.2	0.3.2
evaluate	evaluate	0.16	0.16
fansi	fansi	1.0.3	1.0.3
farver	farver	2.1.1	2.1.1
fastmap	fastmap	1.1.0	1.1.0
forcats	forcats	0.5.2	0.5.2
fs	fs	1.5.2	1.5.2
gargle	gargle	1.2.0	1.2.0
generics	generics	0.1.3	0.1.3
purrr	purrr	0.3.4	0.3.4
R6	R6	2.5.1	2.5.1
RCurl	RCurl	1.98.1.8	1.98-
readr	readr	2.1.2	2.1.2
readxl	readxl	1.4.1	1.4.1
repr	repr	1.1.4	1.1.4
reprex	reprex	2.0.2	2.0.2
rlang	rlang	1.0.4	1.0.4
rstatix	rstatix	0.7.0	0.7.0
rvest	rvest	1.0.3	1.0.3
S4Vectors	S4Vectors	0.34.0	0.34.
scales	scales	1.2.1	1.2.1
sessioninfo	sessioninfo	1.2.2	1.2.2
stringi	stringi	1.7.8	1.7.8
stringr	stringr	1.4.1	1.4.1
SummarizedExperiment	SummarizedExperiment	1.26.1	1.26.
svglite	svglite	2.1.0	2.1.0
systemfonts	systemfonts	1.0.4	1.0.4
tibble	tibble	3.1.8	3.1.8
tidyr	tidyr	1.2.0	1.2.0