ook. When you execute code within the notebook, the results appear beneath the code.

Try executing this chunk by clicking the \*Run\* button within the chunk or by placing your cursor inside it and pressing \*Cmd+Shift+Enter\*.

```{r}

library(readr)

var\_meth\_S15 <- read\_csv("../var\_meth\_data/blink\_var\_meth/S15/S15\_link.csv",

col\_types=cols())

View(var\_meth\_S15)

```

Add a new chunk by clicking the \*Insert Chunk\* button on the toolbar or by pressing \*Cmd+Option+I\*.

When you save the notebook, an HTML file containing the code and output will be saved alongside it (click the \*Preview\* button or press \*Cmd+Shift+K\* to preview the HTML file).

The preview shows you a rendered HTML copy of the contents of the editor. Consequently, unlike \*Knit\*, \*Preview\* does not run any R code chunks. Instead, the output of the chunk when it was last run in the editor is displayed.

```{r}

library(dplyr)

var\_meth\_S15 %>%

filter(snp\_pos == 2905, meth\_pos == (3008-133+1)) %>%

group\_by(snp\_allele) %>%

summarise(meth = mean(meth\_state=="+"))

```

```{r}

#show methylation level for all snp positions and CpG positions (meth\_pos)

#mean methylation across all CpG sites associated with every snp position

#data table for all mice (importing)

#think about which graphs to generate

```

```{r}

library(dplyr)

var\_meth\_S15 %>%

filter(snp\_pos, meth\_pos) %>%

group\_by(snp\_allele) %>%

summarise(meth = mean(meth\_state=="+"))

```

```{r}

library(dplyr)

var\_meth\_S15 %>%

filter(snp\_pos, meth\_pos) %>%

group\_by(snp\_pos, snp\_allele) %>%

summarise(meth = mean(meth\_state=="+"))

```

```{r}

library(dplyr)

var\_meth\_S15 %>%

filter(snp\_pos, meth\_pos) %>%

group\_by(snp\_pos, snp\_allele, meth\_pos) %>%

summarise(meth = mean(meth\_state=="+"))

```

```{r}

common\_path = "../var\_meth\_data/blink\_var\_meth/"

sub\_dirs = list.files(common\_path)

sub\_dirs

read\_csv(sub\_dirs(pattern = "blink.csv"))

```

```{r}

common\_path = "../var\_meth\_data/blink\_var\_meth/"

sub\_dirs = list.files(common\_path)

lapply(X=sub\_dirs, FUN=read\_csv(../var\_meth\_data/blink\_var\_meth/(sub\_dirs)/(sub\_dirs)\_blink.csv))

```

```{r}

#Extract files from multiple folders, and sets end of file name “S1” etc (which is stored in “sub\_dirs”) as extra column/heading name

common\_path = "../var\_meth\_data/blink\_var\_meth/"

sub\_dirs = list.files(common\_path)

lapply(

X=sub\_dirs,

FUN=function(x){

read\_csv(

paste0(common\_path, x, '/', x, '\_link.csv'),

col\_types=cols()) %>%

mutate(sample=x)})

```

```{r}

#bind row - next step

bind\_rows(lapply(

X=sub\_dirs,

FUN=function(x){

read\_csv(

paste0(common\_path, x, '/', x, '\_link.csv'),

col\_types=cols()) %>%

mutate(sample=x)}))

#the below stores the bind\_rows(lapply(... output as "var\_meth\_table". then views it in a tab.

var\_meth\_table <- bind\_rows(lapply(

X=sub\_dirs,

FUN=function(x){

read\_csv(

paste0(common\_path, x, '/', x, '\_link.csv'),

col\_types=cols()) %>%

mutate(sample=x)}))

View(var\_meth\_table)

#use uppercase for "View".

```{r}

meth\_level <- var\_meth\_table %>%

filter(snp\_pos, meth\_pos) %>%

group\_by(snp\_pos, snp\_allele, meth\_pos) %>%

summarise(meth = mean(meth\_state=="+"))

#20.7.21

meth\_level\_CpG <- var\_meth\_table %>%

group\_by(sample, snp\_pos, snp\_allele, meth\_pos) %>%

summarise(meth = mean(meth\_state=="+"))

View(meth\_level\_CpG)

#sample, snp position and allele. average across all CpG sites #20.7.21

meth\_level\_snp <- meth\_level\_CpG %>%

group\_by(sample, snp\_pos, snp\_allele) %>%

summarise(meth = mean(meth))

View(meth\_level\_snp)

```{r}

#Extract files from multiple folders, and sets end of file name “S1” etc (which is stored in “sub\_dirs”) as extra column/heading name

common\_path2 = "../var\_meth\_data/blink\_var\_meth\_2/"

sub\_dirs = list.files(common\_path2)

lapply(

X=sub\_dirs,

FUN=function(x){

read\_csv(

paste0(common\_path2, x, '/', x, '\_snps.csv'),

col\_types=cols()) %>%

mutate(sample=x)})

```

```{r}

#bind row - next step

snps\_table <- bind\_rows(lapply(

X=sub\_dirs,

FUN=function(x){

read\_csv(

paste0(common\_path2, x, '/', x, '\_snps.csv'),

col\_types=cols()) %>%

mutate(sample=x)}))

View(snps\_table)

```

```{r}

#ref\_prop\_exp is WGS, ref\_prop\_obs is WGBS

cor(snps\_table$ref\_prop\_exp, snps\_table$ref\_prop\_obs)

```

```{r}

library(ggplot2)

ggplot(snps\_table) +

aes(x = ref\_prop\_obs, y = ref\_prop\_exp) +

labs(x="RAF in WGBS", y="RAF in WGS") +

geom\_point(colour = "#0c4c8a") +

theme\_minimal()

```

```{r}

```

```{r}

snps\_table <- bind\_rows(lapply(

X=sub\_dirs,

FUN=function(x){

read\_csv(

paste0(common\_path2, x, '/', x, '\_snps.csv'),

col\_types=cols()) %>%

mutate(sample=x, alt\_prop\_exp = 1 - ref\_prop\_exp,alt\_prop\_obs = 1 - ref\_prop\_obs)}))

View(snps\_table)

```

library(readr)

samples\_ID\_strain <- read\_csv("../var\_meth\_data/samples\_ID\_strain1.csv",

col\_types=cols())

View(samples\_ID\_strain)

```{r}

#ref\_prop\_exp is WGS, ref\_prop\_obs is WGBS

mutate(snps\_table, alt\_prop\_exp = 1 - snps\_table$ref\_prop\_exp)

mutate(snps\_table, alt\_prop\_obs = 1 - snps\_table$ref\_prop\_obs)

```

snps\_table\_strain <- inner\_join(samples\_ID\_strain, snps\_table, by = "sample")

View(snps\_table\_strain)

```{r}

library(readr)

samples\_ID\_strain\_mouseno <- read\_csv("../var\_meth\_data/samples\_ID\_strain\_mouseno.csv",

col\_types=cols())

View(samples\_ID\_strain\_mouseno)

library(dplyr)

snps\_table\_strain <- inner\_join(samples\_ID\_strain\_mouseno, snps\_table, by = "sample")

```

```{r}

#ref\_prop\_exp is WGS, ref\_prop\_obs is WGBS

mutate(snps\_table, alt\_prop\_exp = 1 - snps\_table$ref\_prop\_exp)

mutate(snps\_table, alt\_prop\_obs = 1 - snps\_table$ref\_prop\_obs)

library(ggplot2)

ggplot(snps\_table\_strain) +

aes(x = alt\_prop\_obs, y = alt\_prop\_exp) +

labs(x="AAF in WGBS", y="AAF in WGS") +

geom\_point(colour = "#0c4c8a") +

theme\_minimal()

```

```{r}

fig1 <- ggplot(snps\_table\_strain) +

aes(x = alt\_prop\_obs, y = alt\_prop\_exp) +

labs(x="AAF in WGBS", y="AAF in WGS") +

theme\_minimal()

fig1 + facet\_grid(vars(snps\_table\_strain$strain), vars(snps\_table\_strain$mouse\_number), scales = "free") +

geom\_point(aes(colour = factor(strain)), alpha = 0.5) +

scale\_y\_continuous(breaks=c(0, 0.5, 1), labels =c('', 0.5, 1)) +

scale\_x\_continuous(breaks =c(0, 0.5, 1), labels =c('', 0.5, 1))

fig1corr <- cor.test(snps\_table\_strain$alt\_prop\_obs, snps\_table\_strain$alt\_prop\_exp)

fig1corr

library(ggpubr)

fig1 <- ggplot(snps\_table\_strain) +

aes(x = alt\_prop\_obs, y = alt\_prop\_exp) +

labs(x="AAF in WGBS", y="AAF in WGS") +

theme\_minimal()

fig1 + facet\_grid(vars(snps\_table\_strain$strain), vars(snps\_table\_strain$mouse\_number), scales = "free") +

geom\_point(aes(colour = factor(strain)), alpha = 0.5) +

scale\_y\_continuous(breaks=c(0, 0.5, 1), labels =c('', 0.5, 1)) +

scale\_x\_continuous(breaks =c(0, 0.5, 1), labels =c('', 0.5, 1))

fig1complete <- fig1 + facet\_grid(vars(snps\_table\_strain$strain), vars(snps\_table\_strain$mouse\_number), scales = "free") +

geom\_point(aes(colour = factor(strain)), alpha = 0.5) +

scale\_y\_continuous(breaks=c(0, 0.5, 1), labels =c('', 0.5, 1)) +

scale\_x\_continuous(breaks =c(0, 0.5, 1), labels =c('', 0.5, 1))

```

```{r}

#ref: library(dplyr)

#library(dplyr)

#var\_meth\_S15 %>%

#filter(snp\_pos == 2905, meth\_pos == (3008-133+1)) %>%

#group\_by(snp\_allele) %>%

#summarise(meth = mean(meth\_state=="+"))

library(dplyr)

snps\_table\_strain %>%

filter (sample == "S1") %>%

cor.test(snps\_table\_strain$alt\_prop\_obs, snps\_table\_strain$alt\_prop\_exp)

#cannnot include cor.test in pipe

```

```{r}

fig1complete

```

```{r}

library(ggpubr)

fig1complete + stat\_cor(method = "pearson", size = 2)

```

```

```{r}

fig2 <- ggplot(snps\_table\_strain) +

aes(x = alt\_prop\_obs, y = alt\_prop\_exp) +

labs(x="AAF in WGBS", y="AAF in WGS") +

theme\_minimal()

```

```{r}

snps\_table\_strain\_meth <- inner\_join(snps\_table\_strain, meth\_level, by = "sample")

View(snps\_table\_strain\_meth)

```

```{r}

#20.7.21

meth\_level\_CpG <- var\_meth\_table %>%

group\_by(sample, snp\_pos, snp\_allele, meth\_pos) %>%

summarise(meth = mean(meth\_state=="+"))

View(meth\_level\_CpG)

#sample, snp position and allele. average across all CpG sites #20.7.21

meth\_level\_snp <- meth\_level\_CpG %>%

group\_by(sample, snp\_pos, snp\_allele) %>%

summarise(meth = mean(meth))

View(meth\_level\_snp)

meth\_level\_snp\_strain <- inner\_join(meth\_level\_snp, samples\_ID\_strain\_mouseno, by = "sample")

View(meth\_level\_snp\_strain)

```

```{r}

library(ggplot2)

fig2 <- ggplot(meth\_level\_snp\_strain) +

aes(x = factor(snp\_pos), y = meth) +

geom\_point(aes(colour = snp\_allele)) +

labs(x="SNP position in rDNA unit", y="Methylation level") +

theme\_minimal()

fig2 + facet\_grid(row = vars(strain), scales = "free")

```

for each mouse and snp, difference in methylation across all CpG sites, so only kept the snp\_positions for whcih there was a significant difference in methylation level

```{r}

library(dplyr)

var\_meth\_S15 %>%

filter(snp\_pos == 2905, meth\_pos == (3008-133+1)) %>%

group\_by(snp\_allele) %>%

summarise(meth = mean(meth\_state=="+"))

library(dplyr)

meth\_level\_snp\_strain %>%

group\_by(mouse\_number, snp\_pos, meth)

```

idea: maybe add another column, for methylation difference?

```{r}

meth\_level\_CpG\_strain <- inner\_join(meth\_level\_CpG, samples\_ID\_strain\_mouseno, by = "sample")

View(meth\_level\_CpG\_strain)

```

Fran's code:

```{r}

#adapted for my own table

meth\_level\_CpG\_strain %>%

group\_by(sample, strain, ID, snp\_pos) %>%

mutate(group=if\_else(snp\_allele==min(snp\_allele), 'm', 'M')) %>%

do(w = wilcox.test(meth ~ group, data=., exact=FALSE)) %>%

summarise(sample, strain, ID, snp\_pos, pval=w$p.value, .groups='drop') %>%

group\_by(sample) %>%

mutate(padj=p.adjust(pval, method = 'fdr')) %>%

group\_by(strain, snp\_pos) %>%

summarise(num\_sig=sum(padj<0.01), num\_samples=n()) %>%

filter(num\_samples>=3, num\_sig>num\_samples/2)

```

```{r}

#not adapted

strains\_wgbs\_raw %>%

group\_by(sample, snp\_pos, snp\_allele, meth\_pos) %>%

filter(n()>=10) %>%

group\_by(sample, strain, id, snp\_pos, meth\_pos) %>%

filter(n()>=50, n\_distinct(snp\_allele)>1) %>%

group\_by(sample, strain, id, snp\_pos, snp\_allele, meth\_pos) %>%

summarise(meth=sum(meth\_state=='+')/n(), .groups='drop')

```

need to edit var\_meth\_table via innerjoin to include strain and ID etc.

```{r}

var\_meth\_table\_strain <- inner\_join(var\_meth\_table, samples\_ID\_strain\_mouseno, by = "sample")

View(var\_meth\_table\_strain)

#did this wrong first time, so just use var\_meth\_table from now on

```

```{r}

var\_meth\_table %>%

group\_by(sample, snp\_pos, snp\_allele, meth\_pos) %>%

filter(n()>=10) %>%

group\_by(sample, strain, ID, snp\_pos, meth\_pos) %>%

filter(n()>=50, n\_distinct(snp\_allele)>1) %>%

group\_by(sample, strain, ID, snp\_pos, snp\_allele, meth\_pos) %>%

summarise(meth = mean(meth\_state=="+"), .groups='drop')

```

```{r}

packageVersion("tibble")

```

```{r}

var\_meth\_table\_filtered <- var\_meth\_table %>%

group\_by(sample, snp\_pos, snp\_allele, meth\_pos) %>%

filter(n()>=10) %>%

group\_by(sample, strain, ID, snp\_pos, meth\_pos) %>%

filter(n()>=50, n\_distinct(snp\_allele)>1) %>%

group\_by(sample, strain, ID, snp\_pos, snp\_allele, meth\_pos) %>%

summarise(meth = mean(meth\_state=="+"), .groups='drop')

View(var\_meth\_table\_filtered)

```

```{r}

var\_meth\_sig <- var\_meth\_table\_filtered %>%

group\_by(sample, strain, ID, snp\_pos) %>%

mutate(group=if\_else(snp\_allele==min(snp\_allele), 'm', 'M')) %>%

do(w = wilcox.test(meth ~ group, data=., exact=FALSE)) %>%

summarise(sample, strain, ID, snp\_pos, pval=w$p.value, .groups='drop') %>%

group\_by(sample) %>%

mutate(padj=p.adjust(pval, method = 'fdr')) %>%

group\_by(strain, snp\_pos) %>%

summarise(num\_sig=sum(padj<0.01), num\_samples=n()) %>%

filter(num\_samples>=3, num\_sig>num\_samples/2)

View(var\_meth\_sig)

```

take var\_meth\_sig and try different filters till you get all strains.

```{r}

var\_meth\_sig\_fil6 <- var\_meth\_sig %>% filter(num\_sig == 6)

View(var\_meth\_sig\_fil6)

```

```{r}

var\_meth\_sig\_fil456 <- var\_meth\_sig %>% filter(num\_sig >= 4)

View(var\_meth\_sig\_fil456)

var\_meth\_sig\_fil56 <- var\_meth\_sig %>% filter(num\_sig >= 5)

View(var\_meth\_sig\_fil56)

```

next step: get all the snp positions from this table and set as x axis for the methylation level graph.

```{r}

var\_meth\_fil\_meth <- var\_meth\_table\_filtered %>%

group\_by(strain, ID, snp\_pos, snp\_allele) %>%

summarise(meth = mean(meth))

View(var\_meth\_fil\_meth)

var\_meth\_fig2\_table <- inner\_join(var\_meth\_fil\_meth, var\_meth\_sig\_fil456, by = ('snp\_pos'))

View(var\_meth\_fig2\_table)

#problem wiith var\_meth\_fig2\_table - two columns for snp\_pos or strain

library(ggplot2)

fig2 <- ggplot(var\_meth\_sig\_fil456meth) +

aes(x = factor(snp\_pos), y = meth) +

geom\_point(aes(colour = snp\_allele)) +

labs(x="SNP position in rDNA unit", y="Methylation level") +

theme\_minimal()

fig2 + facet\_grid(row = vars(strain), scales = "free")

```

try to edit this by including group by ID - see what happens, save under different table name

```{r}

var\_meth\_sigID <- var\_meth\_table\_filtered %>%

group\_by(sample, strain, ID, snp\_pos) %>%

mutate(group=if\_else(snp\_allele==min(snp\_allele), 'm', 'M')) %>%

do(w = wilcox.test(meth ~ group, data=., exact=FALSE)) %>%

summarise(sample, strain, ID, snp\_pos, pval=w$p.value, .groups='drop') %>%

group\_by(sample) %>%

mutate(padj=p.adjust(pval, method = 'fdr')) %>%

group\_by(strain, ID, snp\_pos) %>%

summarise(num\_sig=sum(padj<0.01), num\_samples=n()) %>%

filter(num\_samples>=3, num\_sig>num\_samples/2)

View(var\_meth\_sigID)

#this returned no data

```

join by two columns

```{r}

var\_meth\_fig2\_table <- inner\_join(var\_meth\_fil\_meth, var\_meth\_sig, by = c('strain','snp\_pos'))

View(var\_meth\_fig2\_table)

```

now try the graph using var\_meth\_fig2\_table

```{r}

library(ggplot2)

fig2 <- ggplot(var\_meth\_fig2\_table) +

aes(x = factor(snp\_pos), y = meth) +

geom\_point(aes(colour = snp\_allele)) +

labs(x="SNP position in rDNA unit", y="Methylation level") +

theme\_minimal()

fig2 + facet\_grid(row = vars(strain), scales = "free")

```

rotate numbers

```{r}

library(ggplot2)

fig2 <- ggplot(var\_meth\_fig2\_table) +

aes(x = as.factor(snp\_pos), y = meth) +

geom\_point(aes(colour = snp\_allele)) +

labs(x="SNP position in rDNA unit", y="Methylation level") +

scale\_y\_continuous(breaks=c(0, 0.2, 0.4, 0.6, 0.8, 1), labels =c('', 0.2, 0.4, 0.6, 0.8, 1)) +

theme\_minimal() +

theme(axis.text.x=element\_text(angle=90))+

fig2 + facet\_grid(rows = vars(strain))

```

3008 is -1, 3009 is 1. if position is below or equal to 3008, subtract 3009, if it is above 3008 subtract 3008.

```{r}

library(dplyr)

var\_meth\_fig2\_adjtable <- var\_meth\_fig2\_table %>%

mutate(adj\_snp\_pos=if\_else(snp\_pos<= 3008, snp\_pos-3009, snp\_pos-3008))

View(var\_meth\_fig2\_adjtable)

#okay this worked

```

now re-do the graph

```{r}

library(ggplot2)

fig2 <- ggplot(var\_meth\_fig2\_adjtable) +

aes(x = as.factor(adj\_snp\_pos), y = meth) +

geom\_point(aes(colour = snp\_allele), size = 1, alpha = 0.5, position = position\_dodge(width = 0.9)) +

labs(x="SNP position in rDNA unit", y="Methylation level") +

scale\_y\_continuous(breaks=c(0, 0.2, 0.4, 0.6, 0.8, 1), labels =c('', 0.2, 0.4, 0.6, 0.8, 1)) +

theme\_minimal() +

theme(axis.text.x=element\_text(angle=90))

fig2 + facet\_grid(rows = vars(strain))

```

position = position\_dodge(width = 0.90