breakpoints accuracy analysis for all simulated datasets

Qian Feng

11/12/2019

Section 1 ; rec proportions

Step 1: Load libraries

```
library(data.table)
library(stringr)
library(ggplot2)
library(ggpubr)
```

Step 2: Read results into R

```
simulation_30_70_output <- fread("/Users/fengqian/Downloads/UniMelb_shared-master/algorithm_simulation/p
                            , header=FALSE
                            , data.table = FALSE)
simulation_40_60_output <- fread("/Users/fengqian/Downloads/UniMelb_shared-master/algorithm_simulation/p
                            , header=FALSE
                            , data.table = FALSE)
simulation_50_50_output <- fread("/Users/fengqian/Downloads/UniMelb_shared-master/algorithm_simulation/p
                            , header=FALSE
                            , data.table = FALSE)
simulation_60_40_output <- fread("/Users/fengqian/Downloads/UniMelb_shared-master/algorithm_simulation/p
                            , header=FALSE
                            , data.table = FALSE)
simulation_70_30_output <- fread("/Users/fengqian/Downloads/UniMelb_shared-master/algorithm_simulation/p
                            , header=FALSE
                            , data.table = FALSE)
colnames(simulation_30_70_output ) <- c("chunk","target","db1","db2",   "rec","sv", "bkp_error","identi
colnames(simulation_40_60_output ) <- c("chunk","target","db1","db2",   "rec","sv", "bkp_error","identi
colnames(simulation_50_50_output ) <- c("chunk","target","db1","db2",   "rec","sv", "bkp_error","identi
colnames(simulation_60_40_output ) <- c("chunk","target","db1","db2",   "rec","sv", "bkp_error","identi
colnames(simulation_70_30_output ) <- c("chunk","target","db1","db2",   "rec","sv", "bkp_error","identi
```

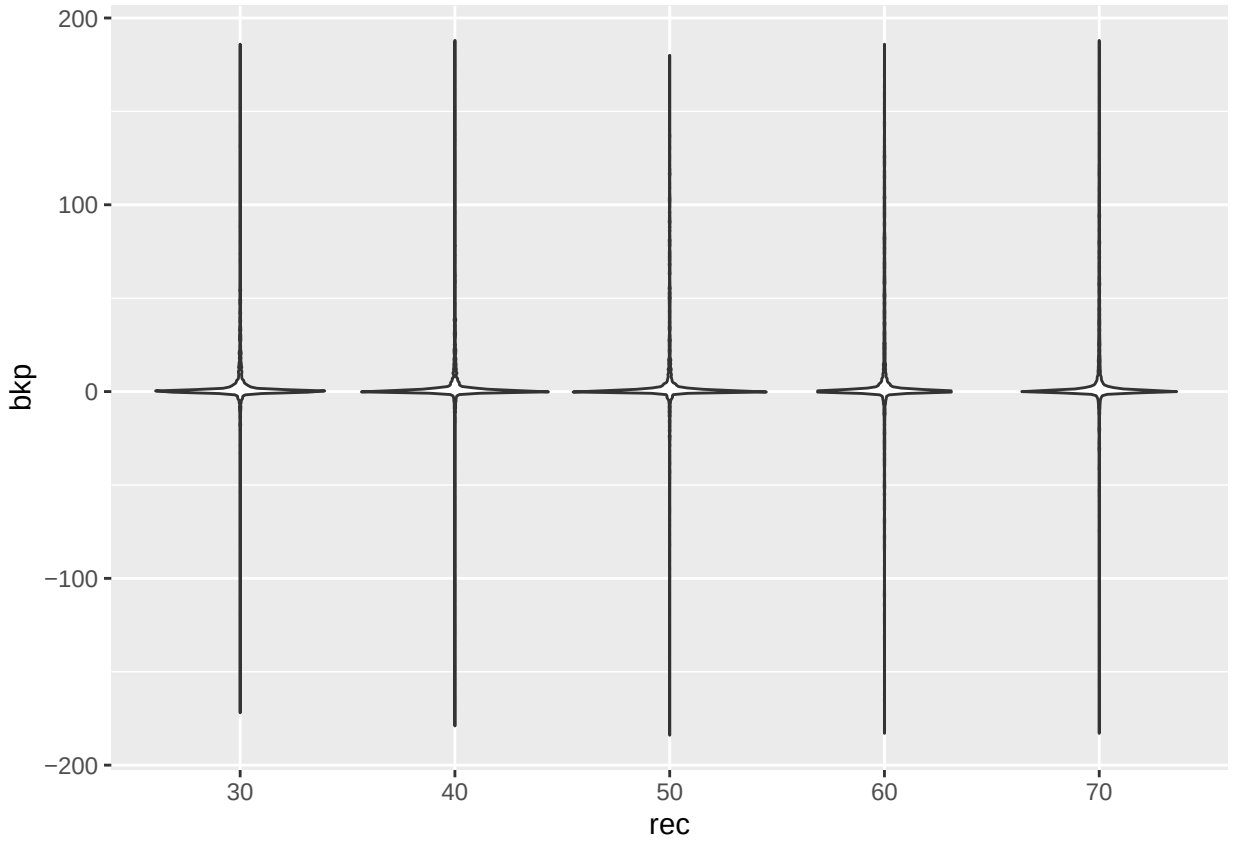
Step 3: organize the dataframe and draw the violin plot

```
bkp_data <- list()
bkp_data[[1]] = data.frame(rec = rep(30,length(simulation_30_70_output[,2])), bkp= simulation_30_70_outp
bkp_data[[2]] = data.frame(rec = rep(40,length(simulation_40_60_output[,2])), bkp= simulation_40_60_outp
bkp_data[[3]] = data.frame(rec = rep(50,length(simulation_50_50_output[,2])), bkp= simulation_50_50_outp
bkp_data[[4]] = data.frame(rec = rep(60,length(simulation_60_40_output[,2])), bkp= simulation_60_40_outp
bkp_data[[5]] = data.frame(rec = rep(70,length(simulation_70_30_output[,2])), bkp= simulation_70_30_outp
df=do.call(rbind, bkp_data)
df$rec<- as.factor(df$rec)
#head(df)
```

##draw a vioplots

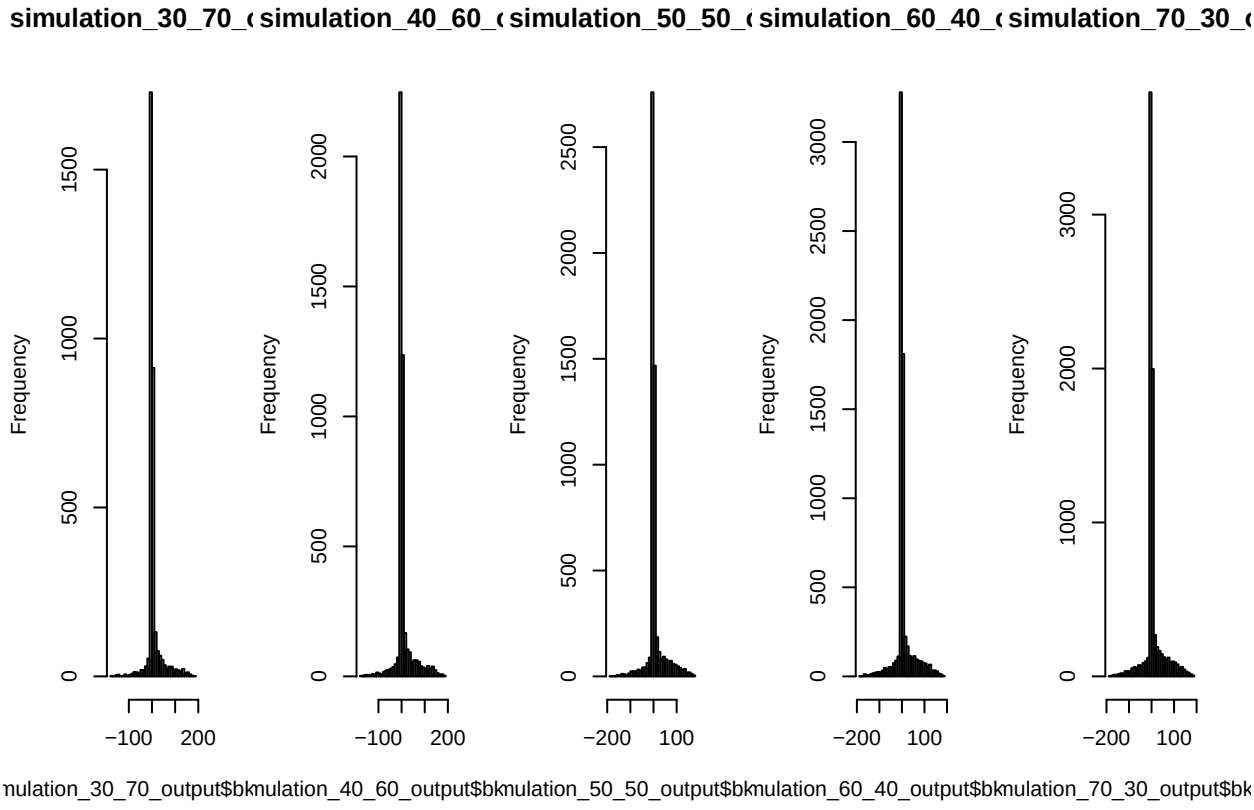
```
df_once <- df[df$group=="jump_once",]
p<-ggplot(df_once, aes(x=rec, y=bkp)) +
```

```
  geom_violin()
p
```



#hist

```
par(mfrow=c(1,5))
hist(simulation_30_70_output$bkp_error,breaks=50)
hist(simulation_40_60_output$bkp_error,breaks=50)
hist(simulation_50_50_output$bkp_error,breaks=50)
hist(simulation_60_40_output$bkp_error,breaks=50)
hist(simulation_70_30_output$bkp_error,breaks=50)
```

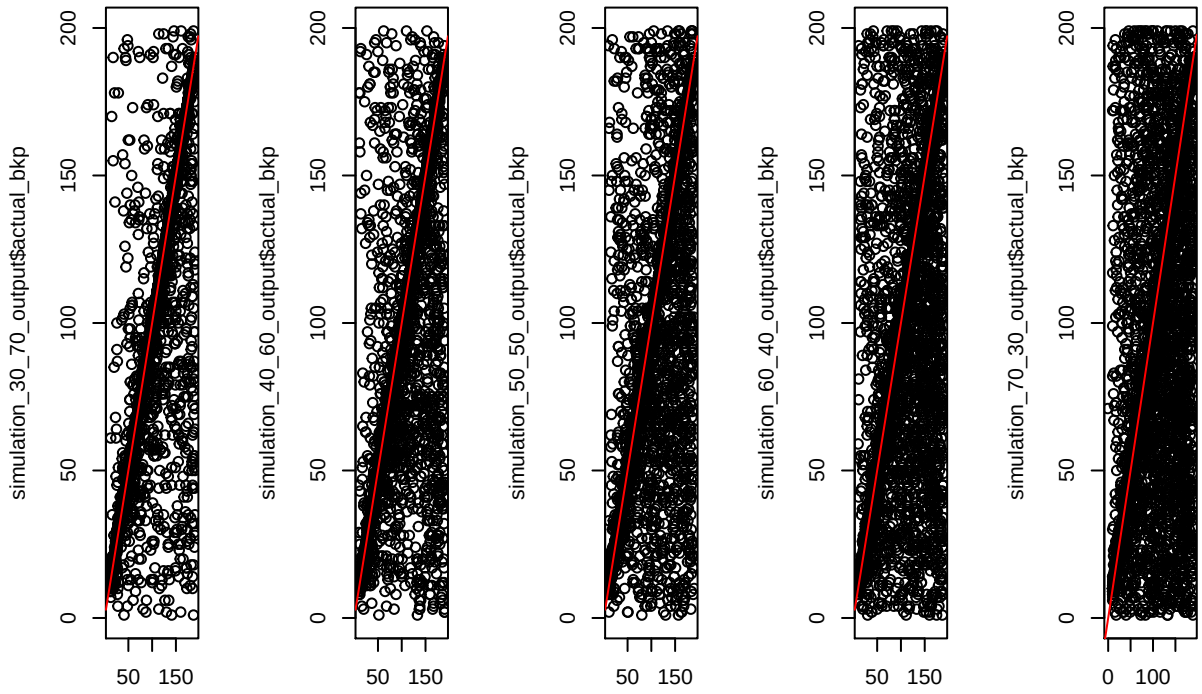


hist for comparing different replicates at only one setting

take the example of jump once + 30% rec proportion, conclusion is different replicates show different distributions, but they all have biggest frequency at position 0.

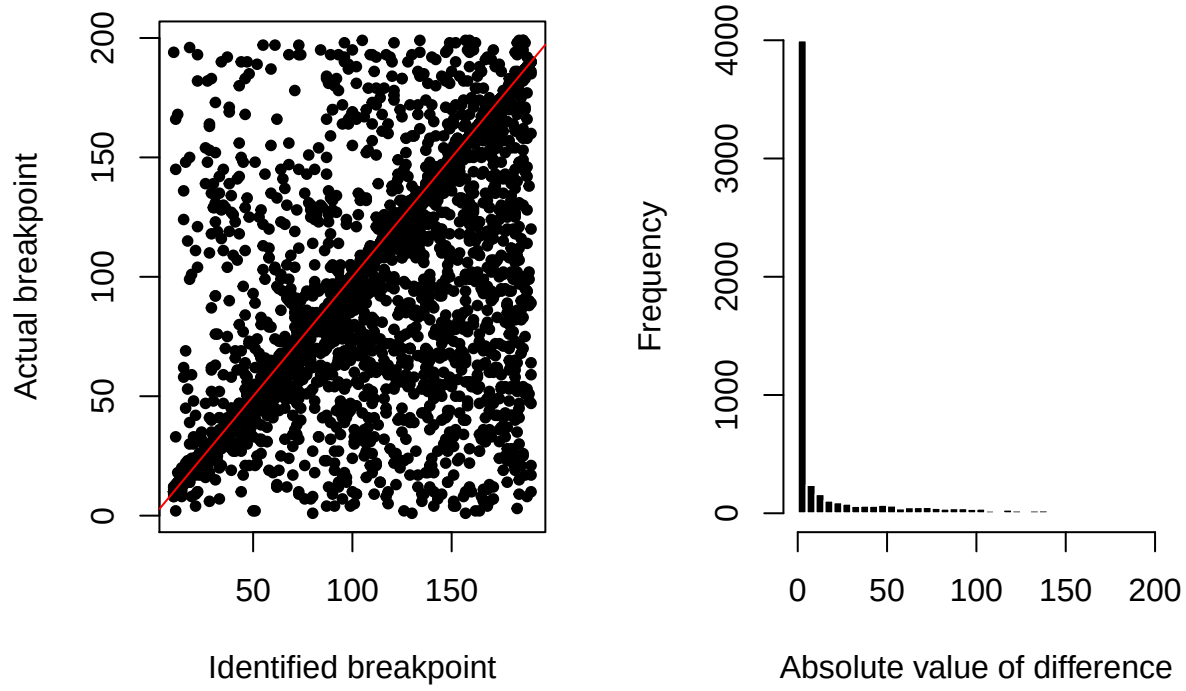
draw the actual_bkp vs identified_bkp plots

```
par(mfrow=c(1,5))
plot(simulation_30_70_output$identified_bkp,simulation_30_70_output$actual_bkp)
abline(0,1,col="red")
plot(simulation_40_60_output$identified_bkp,simulation_40_60_output$actual_bkp)
abline(0,1,col="red")
plot(simulation_50_50_output$identified_bkp,simulation_50_50_output$actual_bkp)
abline(0,1,col="red")
plot(simulation_60_40_output$identified_bkp,simulation_60_40_output$actual_bkp)
abline(0,1,col="red")
plot(simulation_70_30_output$identified_bkp,simulation_70_30_output$actual_bkp)
abline(0,1,col="red")
```



ılation_30_70_output$idenılation_40_60_output$idenılation_50_50_output$idenılation_60_40_output$idenılation_70_30_output$iden

```
#pdf(file="/Users/fengqian/Downloads/cabios-template/figures/bkp.pdf", height=4, width=8)
par(mfrow=c(1,2))
plot(simulation_50_50_output$identified_bkp,simulation_50_50_output$actual_bkp,pch=20,xlab="Identified
abline(0,1,col="red")
hist(abs(simulation_50_50_output$bkp_error),xlab="Absolute value of difference",breaks=seq(0,200,5),col=
```



```
#dev.off()
```

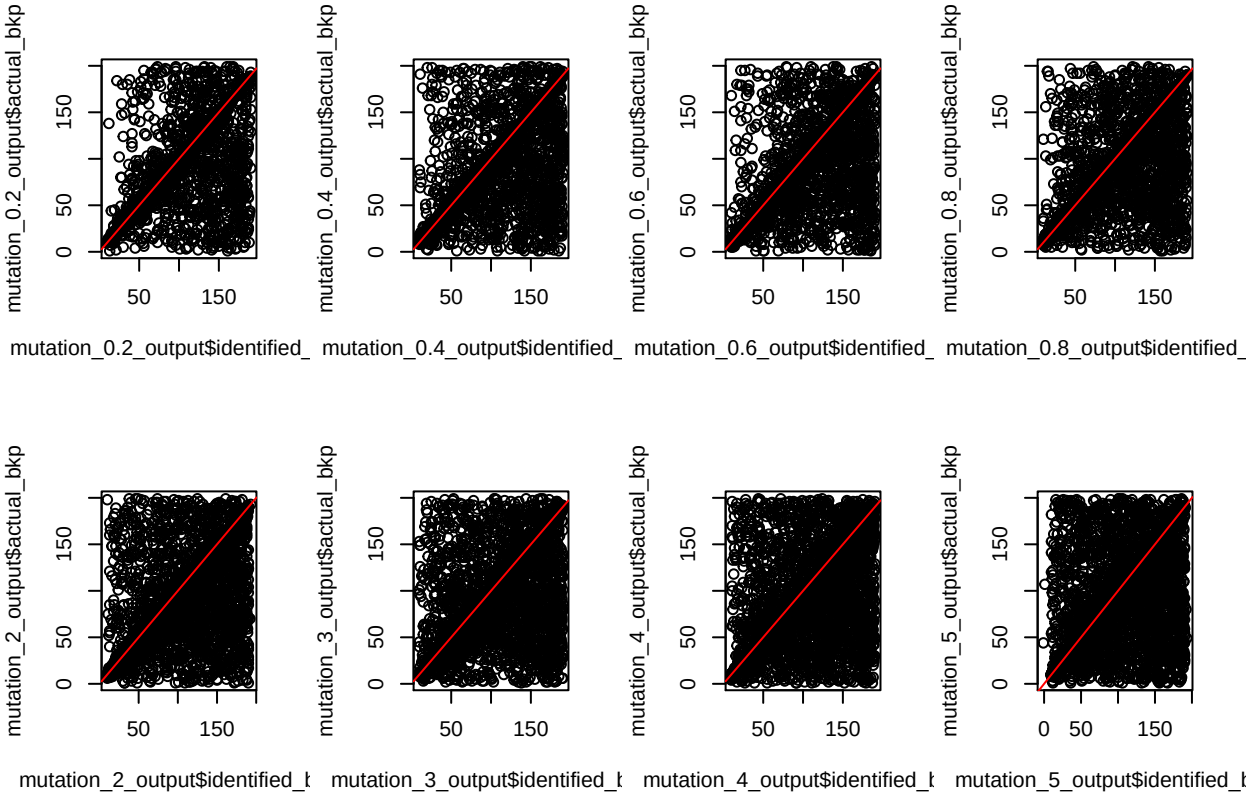
Section 2 : mutations

##read data into R

```
mutation_0.2_output <- fread("/Users/fengqian/Downloads/UniMelb_shared-master/algorithm_simulation/prote
                             , header=FALSE
                             , data.table = FALSE)
mutation_0.4_output <- fread("/Users/fengqian/Downloads/UniMelb_shared-master/algorithm_simulation/prote
                             , header=FALSE
                             , data.table = FALSE)
mutation_0.6_output <- fread("/Users/fengqian/Downloads/UniMelb_shared-master/algorithm_simulation/prote
                             , header=FALSE
                             , data.table = FALSE)
mutation_0.8_output <- fread("/Users/fengqian/Downloads/UniMelb_shared-master/algorithm_simulation/prote
                             , header=FALSE
                             , data.table = FALSE)
mutation_2_output <- fread("/Users/fengqian/Downloads/UniMelb_shared-master/algorithm_simulation/protein
                             , header=FALSE
                             , data.table = FALSE)
mutation_3_output <- fread("/Users/fengqian/Downloads/UniMelb_shared-master/algorithm_simulation/protein
                             , header=FALSE
                             , data.table = FALSE)
mutation_4_output <- fread("/Users/fengqian/Downloads/UniMelb_shared-master/algorithm_simulation/protein
                             , header=FALSE
                             , data.table = FALSE)
mutation_5_output <- fread("/Users/fengqian/Downloads/UniMelb_shared-master/algorithm_simulation/protein
                             , header=FALSE
                             , data.table = FALSE)
colnames(mutation_0.2_output ) <- c("chunk","target","db1","db2",   "rec","sv", "bkp_error","identified_
colnames(mutation_0.4_output ) <- c("chunk","target","db1","db2",   "rec","sv", "bkp_error","identified_
colnames(mutation_0.6_output ) <- c("chunk","target","db1","db2",   "rec","sv", "bkp_error","identified_
colnames(mutation_0.8_output ) <- c("chunk","target","db1","db2",   "rec","sv", "bkp_error","identified_
colnames(mutation_2_output ) <- c("chunk","target","db1","db2", "rec","sv", "bkp_error","identified_bkp"
colnames(mutation_3_output ) <- c("chunk","target","db1","db2", "rec","sv", "bkp_error","identified_bkp"
colnames(mutation_4_output ) <- c("chunk","target","db1","db2", "rec","sv", "bkp_error","identified_bkp"
colnames(mutation_5_output ) <- c("chunk","target","db1","db2", "rec","sv", "bkp_error","identified_bkp"
```

```
par(mfrow=c(2,4))
plot(mutation_0.2_output$identified_bkp,mutation_0.2_output$actual_bkp)
abline(0,1,col="red")
plot(mutation_0.4_output$identified_bkp,mutation_0.4_output$actual_bkp)
abline(0,1,col="red")
plot(mutation_0.6_output$identified_bkp,mutation_0.6_output$actual_bkp)
abline(0,1,col="red")
plot(mutation_0.8_output$identified_bkp,mutation_0.8_output$actual_bkp)
abline(0,1,col="red")
```

```
plot(mutation_2_output$identified_bkp,mutation_2_output$actual_bkp)
abline(0,1,col="red")
plot(mutation_3_output$identified_bkp,mutation_3_output$actual_bkp)
abline(0,1,col="red")
plot(mutation_4_output$identified_bkp,mutation_4_output$actual_bkp)
abline(0,1,col="red")
plot(mutation_5_output$identified_bkp,mutation_5_output$actual_bkp)
abline(0,1,col="red")
```

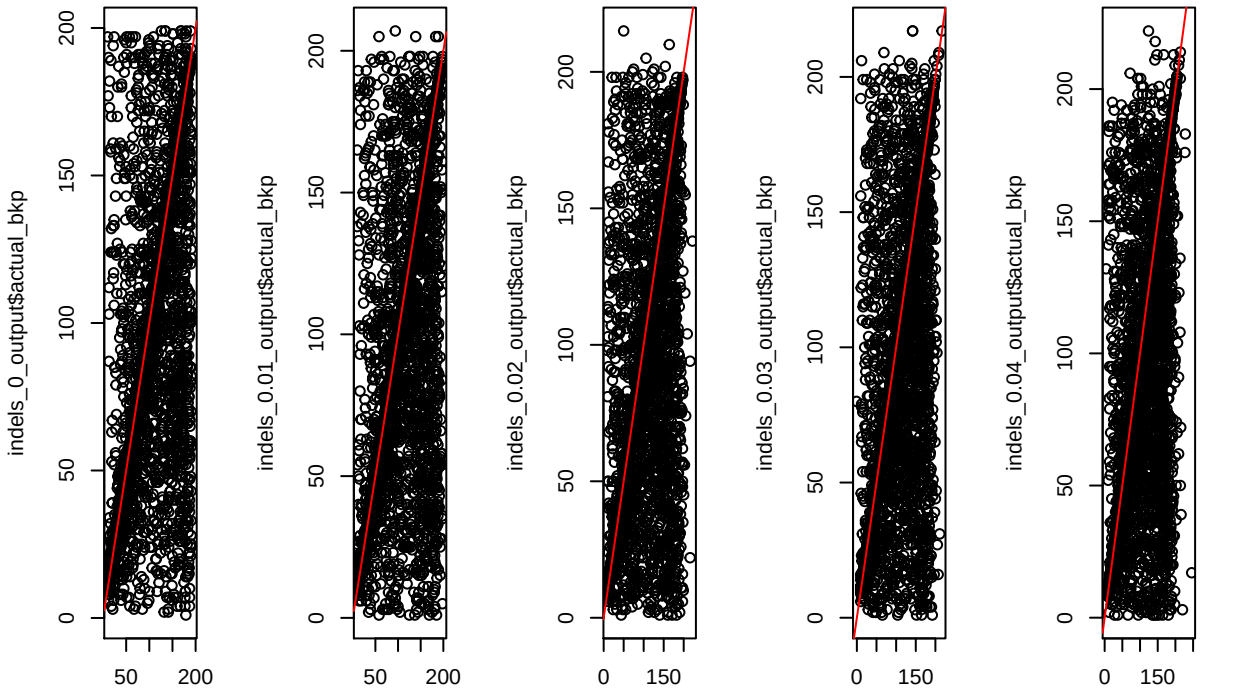


Section 3 : indels

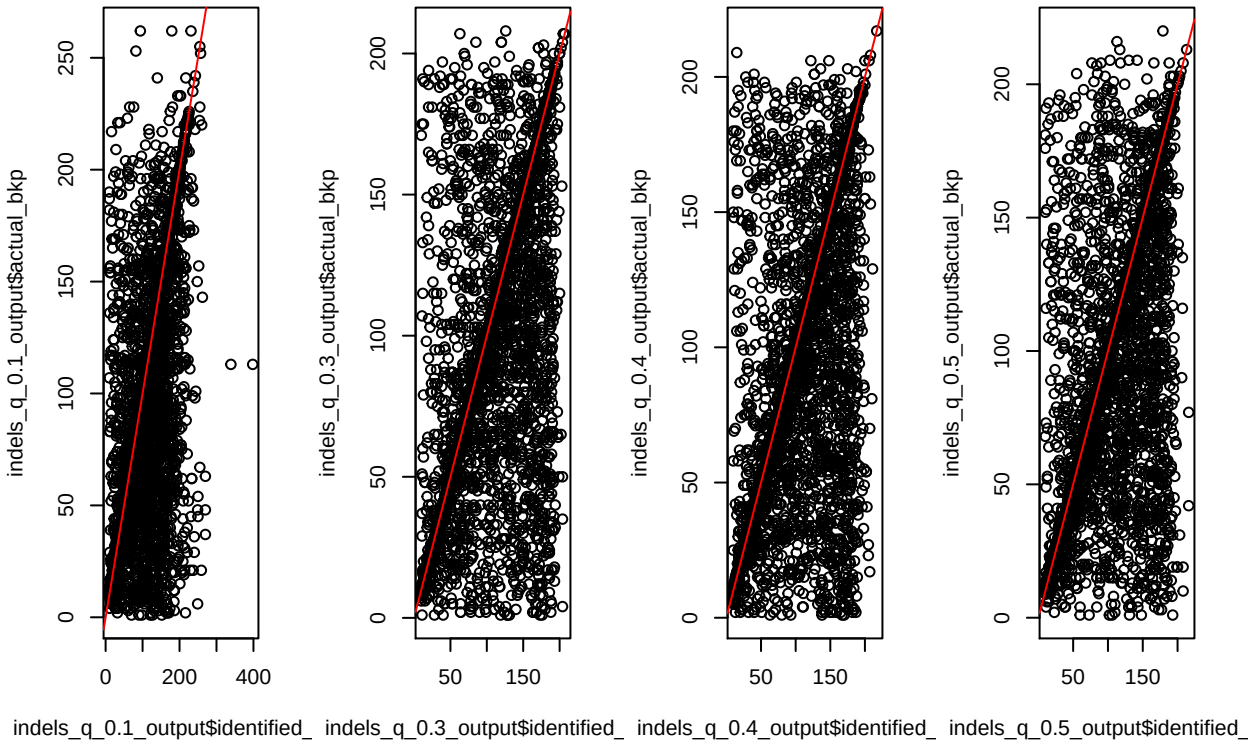
```
indels_0_output <- fread("/Users/fengqian/Downloads/UniMelb_shared-master/algorithm_simulation/protein/s
                         , header=FALSE
                         , data.table = FALSE)
indels_0.01_output <- fread("/Users/fengqian/Downloads/UniMelb_shared-master/algorithm_simulation/protei
                         , header=FALSE
                         , data.table = FALSE)
indels_0.02_output <- fread("/Users/fengqian/Downloads/UniMelb_shared-master/algorithm_simulation/protei
                         , header=FALSE
                         , data.table = FALSE)
indels_0.03_output <- fread("/Users/fengqian/Downloads/UniMelb_shared-master/algorithm_simulation/protei
                         , header=FALSE
                         , data.table = FALSE)
indels_0.04_output <- fread("/Users/fengqian/Downloads/UniMelb_shared-master/algorithm_simulation/protei
                         , header=FALSE
                         , data.table = FALSE)
```

```
indels_q_0.1_output <- fread("/Users/fengqian/Downloads/UniMelb_shared-master/algorithm_simulation/prot
                             , header=FALSE
                             , data.table = FALSE)
indels_q_0.3_output <- fread("/Users/fengqian/Downloads/UniMelb_shared-master/algorithm_simulation/prot
                             , header=FALSE
                             , data.table = FALSE)
indels_q_0.4_output <- fread("/Users/fengqian/Downloads/UniMelb_shared-master/algorithm_simulation/prot
                             , header=FALSE
                             , data.table = FALSE)
indels_q_0.5_output <- fread("/Users/fengqian/Downloads/UniMelb_shared-master/algorithm_simulation/prot
                             , header=FALSE
                             , data.table = FALSE)
colnames(indels_0_output ) <- c("chunk","target","db1","db2",   "rec","sv", "bkp_error","identified_bkp
colnames(indels_0.01_output ) <- c("chunk","target","db1","db2",    "rec","sv", "bkp_error","identified
colnames(indels_0.02_output ) <- c("chunk","target","db1","db2",    "rec","sv", "bkp_error","identified
colnames(indels_0.03_output ) <- c("chunk","target","db1","db2",    "rec","sv", "bkp_error","identified
colnames(indels_0.04_output ) <- c("chunk","target","db1","db2",    "rec","sv", "bkp_error","identified
colnames(indels_q_0.1_output ) <- c("chunk","target","db1","db2",   "rec","sv", "bkp_error","identified
colnames(indels_q_0.3_output ) <- c("chunk","target","db1","db2",   "rec","sv", "bkp_error","identified
colnames(indels_q_0.4_output ) <- c("chunk","target","db1","db2",   "rec","sv", "bkp_error","identified
colnames(indels_q_0.5_output ) <- c("chunk","target","db1","db2",   "rec","sv", "bkp_error","identified
```

```
par(mfrow=c(1,5))
plot(indels_0_output$identified_bkp,indels_0_output$actual_bkp)
abline(0,1,col="red")
plot(indels_0.01_output$identified_bkp,indels_0.01_output$actual_bkp)
abline(0,1,col="red")
plot(indels_0.02_output$identified_bkp,indels_0.02_output$actual_bkp)
abline(0,1,col="red")
plot(indels_0.03_output$identified_bkp,indels_0.03_output$actual_bkp)
abline(0,1,col="red")
plot(indels_0.04_output$identified_bkp,indels_0.04_output$actual_bkp)
abline(0,1,col="red")
```




```
par(mfrow=c(1,4))
plot(indels_q_0.1_output$identified_bkp,indels_q_0.1_output$actual_bkp)
abline(0,1,col="red")
plot(indels_q_0.3_output$identified_bkp,indels_q_0.3_output$actual_bkp)
abline(0,1,col="red")
plot(indels_q_0.4_output$identified_bkp,indels_q_0.4_output$actual_bkp)
abline(0,1,col="red")
plot(indels_q_0.5_output$identified_bkp,indels_q_0.5_output$actual_bkp)
abline(0,1,col="red")
```



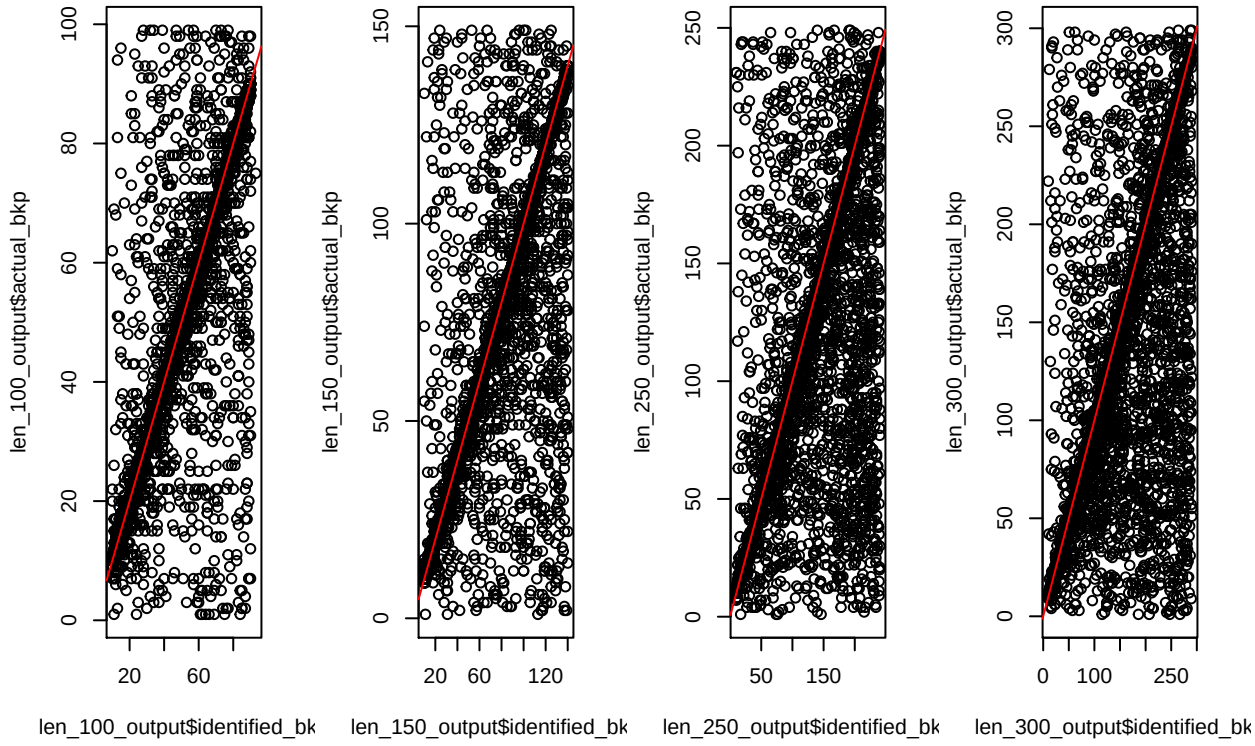
Section 4 : length

```
len_100_output <- fread("/Users/fengqian/Downloads/UniMelb_shared-master/algorithm_simulation/protein/s
                        , header=FALSE
                        , data.table = FALSE)
len_150_output <- fread("/Users/fengqian/Downloads/UniMelb_shared-master/algorithm_simulation/protein/s
                        , header=FALSE
                        , data.table = FALSE)
len_250_output <- fread("/Users/fengqian/Downloads/UniMelb_shared-master/algorithm_simulation/protein/s
                        , header=FALSE
                        , data.table = FALSE)
len_300_output <- fread("/Users/fengqian/Downloads/UniMelb_shared-master/algorithm_simulation/protein/s
                        , header=FALSE
                        , data.table = FALSE)

colnames(len_100_output ) <- c("chunk","target","db1","db2",    "rec","sv", "bkp_error","identified_bkp
colnames(len_150_output ) <- c("chunk","target","db1","db2",    "rec","sv", "bkp_error","identified_bkp
colnames(len_250_output ) <- c("chunk","target","db1","db2",    "rec","sv", "bkp_error","identified_bkp
colnames(len_300_output ) <- c("chunk","target","db1","db2",    "rec","sv", "bkp_error","identified_bkp
```

```
par(mfrow=c(1,4))
plot(len_100_output$identified_bkp,len_100_output$actual_bkp)
abline(0,1,col="red")
plot(len_150_output$identified_bkp,len_150_output$actual_bkp)
abline(0,1,col="red")
plot(len_250_output$identified_bkp,len_250_output$actual_bkp)
abline(0,1,col="red")
plot(len_300_output$identified_bkp,len_300_output$actual_bkp)
```

```
abline(0,1,col="red")
```




```
cor(len_100_output$identified_bkp,len_100_output$actual_bkp)
```

```
## [1] 0.8088791
```

```
cor(len_150_output$identified_bkp,len_150_output$actual_bkp)
```

```
## [1] 0.7520821
```

```
cor(simulation_50_50_output$identified_bkp,simulation_50_50_output$actual_bkp)
```

```
## [1] 0.7327853
```

```
cor(len_250_output$identified_bkp,len_250_output$actual_bkp)
```

```
## [1] 0.7063186
```

```
cor(len_300_output$identified_bkp,len_300_output$actual_bkp)
```

```
## [1] 0.7043425
```

Section 5 : AA models

```
DAYHOFF_output <- fread("/Users/fengqian/Downloads/UniMelb_shared-master/algorithm_simulation/protein/s
                        , header=FALSE
                        , data.table = FALSE)
JTT_output <- fread("/Users/fengqian/Downloads/UniMelb_shared-master/algorithm_simulation/protein/snake
                        , header=FALSE
                        , data.table = FALSE)
LG_output <- fread("/Users/fengqian/Downloads/UniMelb_shared-master/algorithm_simulation/protein/snake_
```

```
                              , header=FALSE
                              , data.table = FALSE)
AB_output <- fread("/Users/fengqian/Downloads/UniMelb_shared-master/algorithm_simulation/protein/snake_
                              , header=FALSE
                              , data.table = FALSE)
MTMAM_output <- fread("/Users/fengqian/Downloads/UniMelb_shared-master/algorithm_simulation/protein/snal
                              , header=FALSE
                              , data.table = FALSE)
colnames(DAYHOFF_output ) <- c("chunk","target","db1","db2",    "rec","sv", "bkp_error","identified_bkp
colnames(JTT_output ) <- c("chunk","target","db1","db2",    "rec","sv", "bkp_error","identified_bkp",
colnames(LG_output ) <- c("chunk","target","db1","db2", "rec","sv", "bkp_error","identified_bkp",    "ac
colnames(AB_output ) <- c("chunk","target","db1","db2", "rec","sv", "bkp_error","identified_bkp",    "ac
colnames(MTMAM_output ) <- c("chunk","target","db1","db2",  "rec","sv", "bkp_error","identified_bkp",
```

```
par(mfrow=c(1,5))
plot(DAYHOFF_output$identified_bkp,DAYHOFF_output$actual_bkp)
abline(0,1,col="red")
plot(JTT_output$identified_bkp,JTT_output$actual_bkp)
abline(0,1,col="red")
plot(LG_output$identified_bkp,LG_output$actual_bkp)
abline(0,1,col="red")
plot(AB_output$identified_bkp,AB_output$actual_bkp)
abline(0,1,col="red")
plot(MTMAM_output$identified_bkp,MTMAM_output$actual_bkp)
abline(0,1,col="red")
```