HIV_Data

Denglinsui

2021/3/16

Introduction

Recently, understanding the genotype-phenotype correlation guiding clinic treatment. Rhee et al. (2006) related HIV-1 protease and reverse transcriptase mutations to in vitro susptibility to 16 antiretroviral drugs. We want to further research with respect to their data. Breifly speaking, we are interested in the gene mutations resistance to each drugs. This task is challenging because the sample size is not large enough compared to the complete gene mutations. A possible solution is taking advantage of expert information and considering a small collection. However, we wish to discover more potential interesting mutations and to suggest candidates for the further experiment.

Discoverying all is equivalent to discoverying none. Our target is not just making more discoveries but with theoretical false discoveries controlling. The common criterions of controlling false discoveries are false discovery rate (Benjamini and Hochberg 1995; Benjamini and Yekutieli 2001; Barber and Candès 2015) and k-familywise error rate(Holm 1979; Janson and Su 2016; Ren, Wei, and Candès, n.d.). We are going to explore the dataset provided by Rhee et al. (2006) and to exploit more HIV-1 drug resistant phenotype related genotype.

To summary, our interested research problem are:

- Finding the genes related to HIV-1 drug resistance;
- Catching and checking the potential interactive effect among the genes.
- Studying the correctness of discoveries versus the frequencies of them.

Some discussions about question are below:

- The first question could be satisfactory answered (Benjamini and Hochberg 1995; Barber and Candès 2015; Holm 1979; Janson and Su 2016). However, some new techniques grew up(Candès et al. 2018; Sesia2018; Ren, Wei, and Candès, n.d.) and we can try these innovative methods.
- The second question includes two parts: (1) Which interactive features should use consider? Are founded interactive features significant?
- The intuition of last question is from instability of result obtained from small sample size. This question will not be answered rigorously.

Explore the HIV-1 Drug Resistance Dataset

The dataset is available at HIV DRUG RESISTANCE DATABASE.

Antiretroviral Drug

There are four major types of antiretroviral drugs: eight protease inhibitors(PIs), seven nucleoside and one nucleotide reverse transicriptase inhibitors(NRTIs), three nonnucleoside reverse transcriptase inhibitors(NNRTIs) amd one fusion inhibitor. The first three drug classes are included in this dataset.

DrugClass	Drug
PI	APV ATV IDV LPV NFV RTV SQV

DrugClass	Drug
NRTI	X3TC ABC AZT D4T DDI TDF
NNRTI	DLV EFV NVP

Data Description

We demostrate the data structure with PIs as an example.

To measure the discovery accurancy, a panel of active gene mutation obtained from larger dataset is given below.

```
knitr::kable(t(res_PI$tsm_df))
```

```
## Warning in kable_pipe(x = structure(c("Position", "Mutations", "10", "F R", :
## The table should have a header (column names)
```

Position	10	11	20	23	24	30	32	33	34	35	43	46	47	48	50	53	54	55	58	66	67	71	73	74	76	79	82	84	85	88
Mutations	F I R	I	I T V	I	I	N	I	F	Q	G	T	I L V	A V	M V	L V	L Y	A L M S T V	R	E	F	F	I	A C S T	A P S	V	A	A F S T	A C V	V	D S T

The structure of our dataset is: The first three rows are id of a record. Row `APV` to row `SQV` are the drug names of PIs. The remaining rows are the gene information.

```
str(res_PI$gene_df[,1:15])
```

```
## 'data.frame':    848 obs. of  15 variables:
##  $ IsolateName: chr  "CA10676" "CA37880" "CA9984" "CA17003" ...
##  $ PseudoName : chr  "CA622" "CA622" "CA624" "CA628" ...
##  $ MedlineID  : int  10839657 15995959 11897594 15995959 10839657 15995959 10839657 15273130 NA 1083
##  $ APV        : num  2.3 76 2.8 6.5 8.3 82 2.3 0.5 33 14.9 ...
##  $ ATV        : num  NA NA NA 9.2 NA 75 NA NA 38 NA ...
##  $ IDV        : num  32.7 131 12 2.1 100 400 12 0.5 19 22.3 ...
##  $ LPV        : num  NA 200 NA 5.3 NA 400 NA 0.5 90 NA ...
##  $ NFV        : num  23.4 50 100 5 161.1 ...
##  $ RTV        : num  51.6 200 41 36 170.2 ...
##  $ SQV        : num  37.8 156 145.6 13 100 ...
##  $ P1         : chr  "-" "-" "-" "-" ...
##  $ P2         : chr  "-" "-" "-" "-" ...
##  $ P3         : chr  "-" "-" "-" "-" ...
##  $ P4         : chr  "-" "-" "-" "-" ...
##  $ P5         : chr  "-" "-" "-" "-" ...
```

The missing value of the responses are:

```
res_PI$gene_df %>%
  select(c("APV","ATV","IDV","LPV","NFV","RTV","SQV")) %>%
  is.na() %>%
  colSums()
```

```
## APV ATV IDV LPV NFV RTV SQV
##  80 519  21 331   4  53  22
```

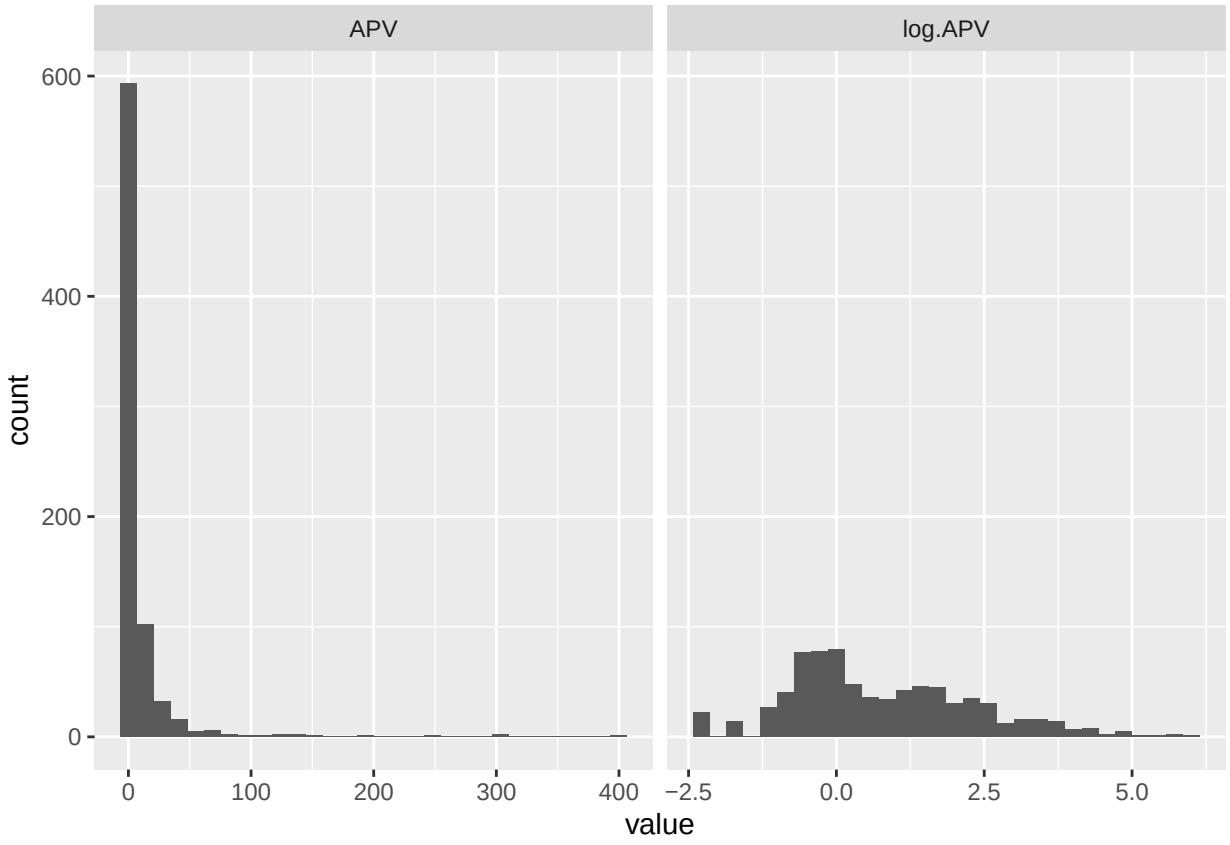
The number of gene position is

```
ncol(res_PI$gene_df)-which(names(res_PI$gene_df)=="P1")
```

```
## [1] 98
```

It better to use the logarithm of the response in the analysis:

```
data <-
  res_PI$gene_df%>% filter(!is.na(APV)) %>%
  select(APV, IsolateName) %>%
  mutate(log.APV = log(APV))
data <- melt(data, id = "IsolateName")

ggplot(data = data, aes(x=value))+
  geom_histogram(bins = 30) +
  facet_grid('~variable',scales="free",)
```

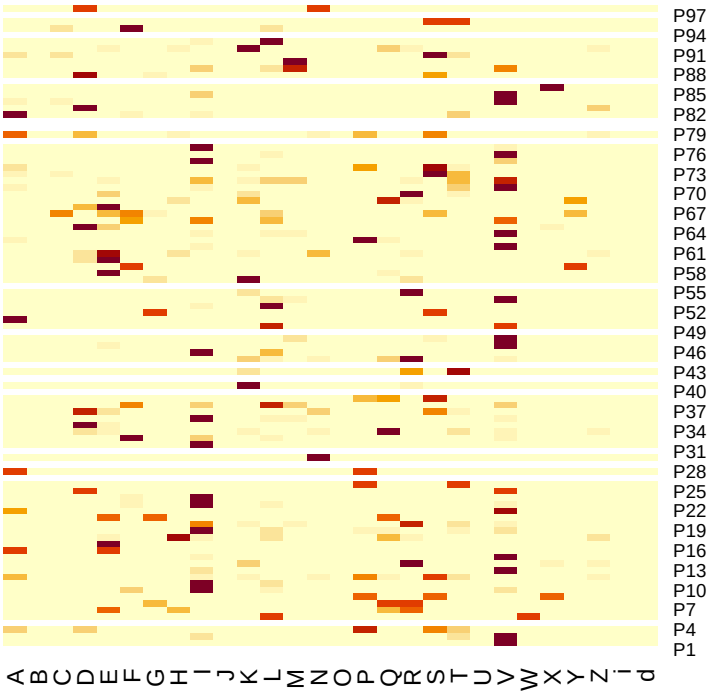


Gene Mutations

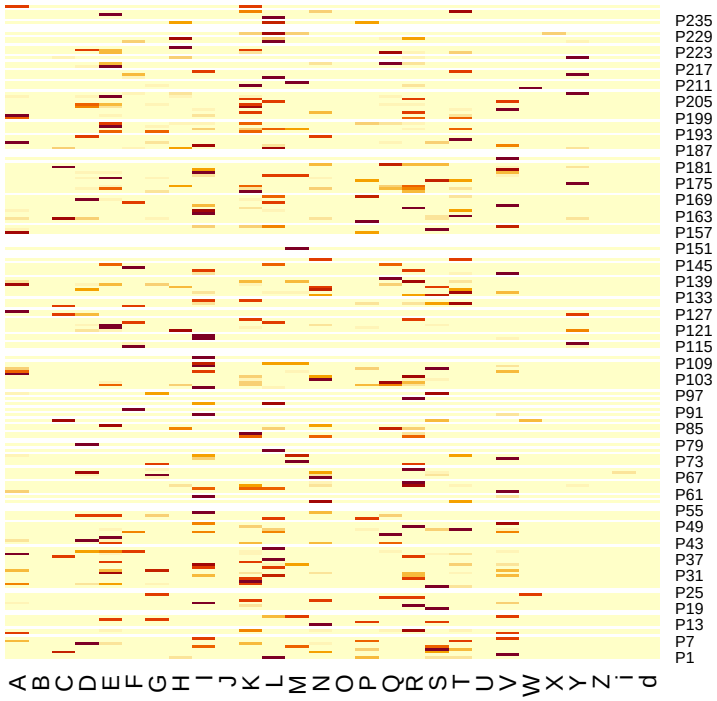
Now, let us scrutinize the gene mutations. The complete gene sets varies from different drug class and we illustrate the rough distribution of the times of gene mutation for the three drug class. The `y-axis` represents the position of gene, the `x-axis` denotes the type of mutations and the color means the frequencies of the x mutation happening on position y.

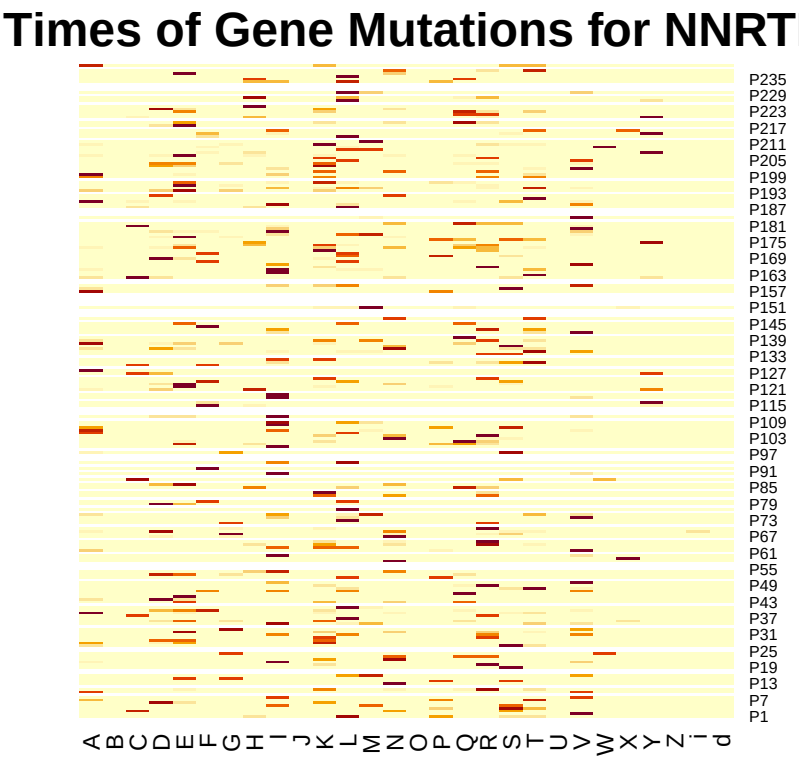
Compared to the gene mutations with low frequencies, we prefer those with high frequencies due to the value of further exploration and the stability of the result.

Times of Gene Mutations for PI



Times of Gene Mutations for NRTI





Preliminary Exploration

Knockoff and BH procedure

We preparatorily apply FDR controlling procedure, Knockoff and BH procedure, to this dataset. This step is the standard step in Barber and Candès (2015) together with the method of Candès et al. (2018).

To begin with, we have to flatten the matrix of the gene mutations associated with paticular genes.

```
res <- res_PI
X <- Flatten_X(res)
```

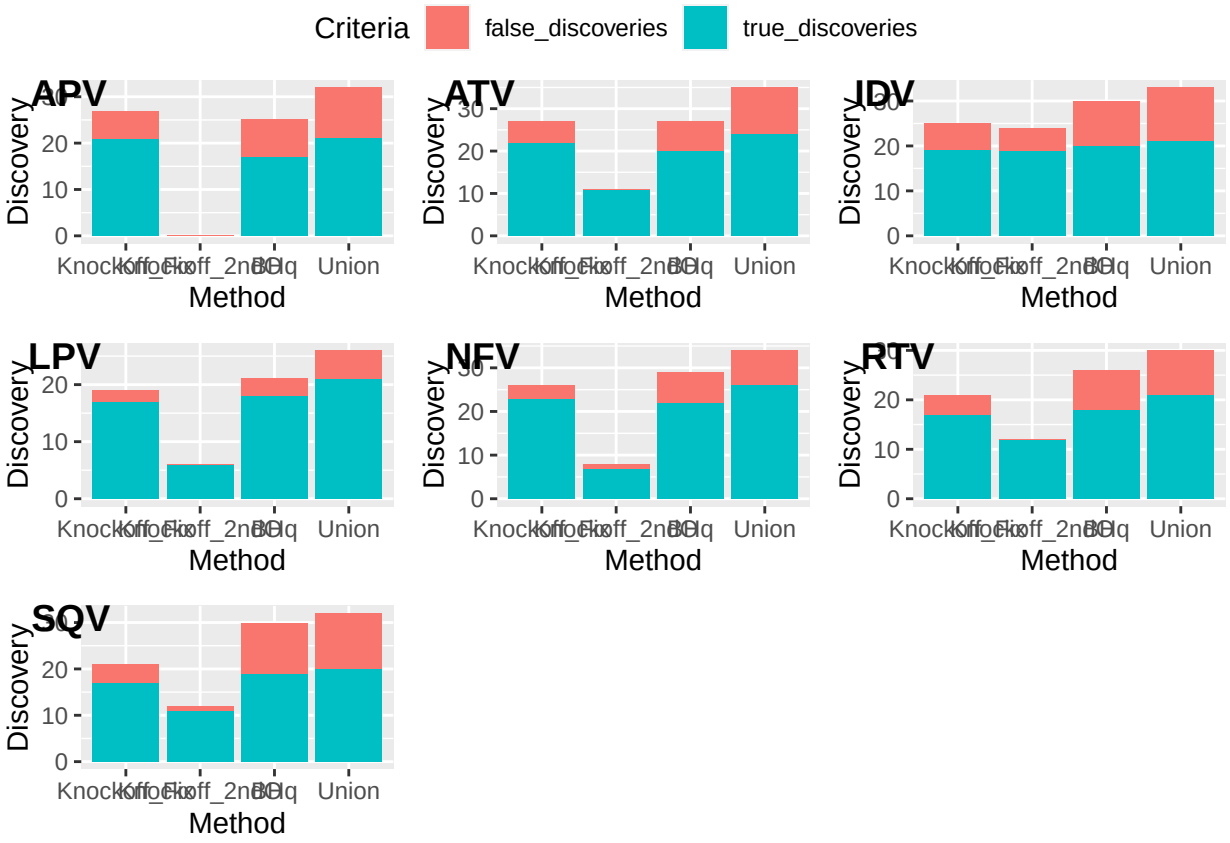
After that, we apply Knockoff and BHq procedure to select features.

```
fdr = 0.20

# Define Datasets for PIs
gene_df <- res$gene_df
pos_start <- which(names(gene_df) == 'P1')
tsm_df <- res$tsm_df
Y = gene_df[,4:(pos_start-1)]

# Run Knockoff and BHq algorithm
results = lapply(Y, function(y) knockoff_and_bhq(X, y, fdr))
```

Finally, we show the discoveries of each drug as following:

```
p.total <- ggarrange(pic[[1]], pic[[2]], pic[[3]], pic[[4]], pic[[5]],
                     pic[[6]], pic[[7]],
```

```
                    labels = c("APV","ATV","IDV","LPV","NFV","RTV","SQV"),
                    ncol = 3, nrow = 3,
                    common.legend = TRUE)
p.total
```



Aprior Algorithm

Besides the effect of single mutation, we are also interested in the across effect of several mutations. As discussed above, we have to explore the gene combinations with high frequencies. We can apply `Aprior Algorithm` to detect the association rule, especially the support set. The effect of three drug classes need to be analyzed seperately as before.

we apply `Aprior Algorithm` to the data mutation separately. Here we only use PI drugs as an example.

```
MyTrans<-as(X[,colSums(X)!=0],"transactions")
MyRules<-apriori(data=MyTrans,
                 parameter=list(support=0.1,
                                confidence = 0,
                                minlen=2,
                                target="rules"))
```

```
## Apriori
##
## Parameter specification:
##  confidence minval smax arem  aval originalSupport maxtime support minlen
##           0    0.1    1 none FALSE            TRUE       5     0.1      2
##  maxlen target  ext
```

```
##      10  rules TRUE
##
## Algorithmic control:
##  filter tree heap memopt load sort verbose
##     0.1 TRUE TRUE  FALSE TRUE    2    TRUE
##
## Absolute minimum support count: 84
##
## set item appearances ...[0 item(s)] done [0.00s].
## set transactions ...[364 item(s), 848 transaction(s)] done [0.00s].
## sorting and recoding items ... [20 item(s)] done [0.00s].
## creating transaction tree ... done [0.00s].
## checking subsets of size 1 2 3 4 done [0.00s].
## writing ... [177 rule(s)] done [0.00s].
## creating S4 object  ... done [0.00s].
```

```
MyRules.sorted<-sort(MyRules,by=c("support"),decreasing=TRUE)
inspect(head(MyRules.sorted))
```

```
##     lhs        rhs     support   confidence coverage  lift     count
## [1] {P93.L} => {P63.P} 0.2983491 0.8939929  0.3337264 1.234700 253
## [2] {P63.P} => {P93.L} 0.2983491 0.4120521  0.7240566 1.234700 253
## [3] {P90.M} => {P63.P} 0.2912736 0.8790036  0.3313679 1.213998 247
## [4] {P63.P} => {P90.M} 0.2912736 0.4022801  0.7240566 1.213998 247
## [5] {P10.I} => {P63.P} 0.2712264 0.7823129  0.3466981 1.080458 230
## [6] {P63.P} => {P10.I} 0.2712264 0.3745928  0.7240566 1.080458 230
```

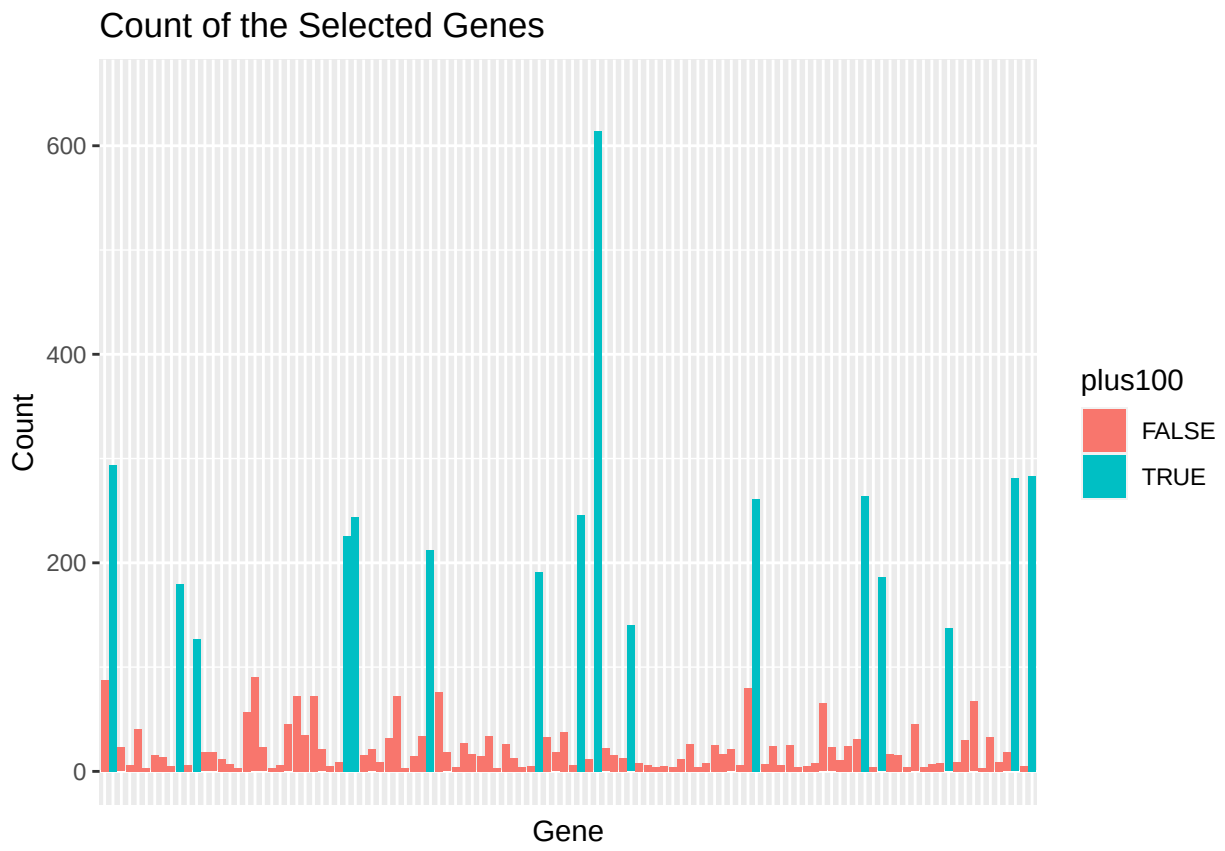
It should be noted here we do not hope to analyze the association rule among the gene mutations but to find highly frequent itemsets. It provides guidance of choosing the gene mutation set to analyzing the interactive effects.

Significant Features / Frequencies

In this part, we are going to investigate the frequencies of siginificant features and the relationship between the itemsets found by Aprior algoritm.

The frequencies of siginicant genes are following:

```
ggplot(data = X.select.plot, aes(x=Gene,y=Count, fill = plus100))+
  geom_bar(stat="identity") +
  labs(title = "Count of the Selected Genes") +
  ylim(0,650) +
  theme(axis.text.x = element_blank()) +
  theme(axis.ticks.x = element_blank())
```



We calculate the times of detected gene mutation in the frequent itemsets. This result will help us shrink the possible interation. However, only detected gene mutations with high frequency will be considered.

```
Union_Apriori <- unique(c(unlist(MyRules@lhs@itemInfo)[MyRules@lhs@data@i],
                          unlist(MyRules@rhs@itemInfo)[MyRules@rhs@data@i]))
Intersect <- intersect(Union_Set,Union_Apriori)
Intersect.ind <- which(unlist(MyRules@lhs@itemInfo) %in% Intersect)
App_Total <- sum(MyRules@lhs@data@i %in% Intersect.ind) +
  sum(MyRules@rhs@data@i %in% Intersect.ind)
print(sprintf("There %d times of detected gene mutation in the frequent itemsets.",App_Total))
```

```
## [1] "There 81 times of detected gene mutation in the frequent itemsets."
```

Future Plan

My future plan is: * Determining the potential interactive genes mutations set; * Checking whether the interaction is significant with both multiple testing procedure and finite sample analysis; * Reasearching the relationship between the validity of discovering procedure with the frequencies of the gene mutations.

Reference

Barber, Rina Foygel, and Emmanuel J. Candès. 2015. "Controlling the False Discovery Rate via Knockoffs." Journal Article. *The Annals of Statistics* 43 (5): 2055–85. https://doi.org/10.1214/15-aos1337.

Benjamini, Yoav, and Yosef Hochberg. 1995. "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing." Journal Article. *Journal of the Royal Statistical Society. Series B (Methodological)* 57 (1): 289–300. http://www.jstor.org/stable/2346101.

Benjamini, Yoav, and Daniel Yekutieli. 2001. "The Control of the False Discovery Rate in Multiple Testing Under Dependency." Journal Article. *The Annals of Statistics* 29 (4): 1165–88. http://www.jstor.org/stable/2674075.

Candès, Emmanuel, Yingying Fan, Lucas Janson, and Jinchi Lv. 2018. "Panning for Gold: 'Model-X' Knockoffs for High Dimensional Controlled Variable Selection." Journal Article 80 (3): 551–77. https://doi.org/10.1111/rssb.12265.

Holm, Sture. 1979. "A Simple Sequentially Rejective Multiple Test Procedure." Journal Article. *Scandinavian Journal of Statistics* 6 (2): 65–70. http://www.jstor.org/stable/4615733.

Janson, Lucas, and Weijie Su. 2016. "Familywise Error Rate Control via Knockoffs." Journal Article. *Electron. J. Statist.* 10 (1): 960–75. https://doi.org/10.1214/16-EJS1129.

Ren, Zhimei, Yuting Wei, and Emmanuel Candès. n.d. "Derandomizing Knockoffs." Conference Proceedings. In.

Rhee, S. Y., J. Taylor, G. Wadhera, A. Ben-Hur, D. L. Brutlag, and R. W. Shafer. 2006. "Genotypic Predictors of Human Immunodeficiency Virus Type 1 Drug Resistance." Journal Article. *Proc Natl Acad Sci U S A* 103 (46): 17355–60. https://doi.org/10.1073/pnas.0607274103.