Gene Mutations Associated with HIV-1 Drug Resistance

Linsui Deng

denglinsui@ruc.edu.cn

Update: May 20, 2021

Contents

1 Introduction

Recently, understanding the genotype-phenotype correlation guides clinic treatment. Rhee et al. (2006) related HIV-1 protease and reverse transcriptase mutations to in vitro susptibility with 16 antiretroviral drugs. We want to go in deep with their data. Briefly 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.

Discovering all is equivalent to discovering none. Our target is not just making more discoveries but with theoretical false discoveries controlling. The common criteria of controlling false discoveries are false discovery rate Barber and Candès (2015); Benjamini and Hochberg (1995); Benjamini and Yekutieli (2001) and k-familywise error rate Holm (1979); Janson and Su (2016); Ren et al. (2020). We are going to explore the dataset provided by Rhee et al. (2006) and exploit more HIV-1 drug resistant phenotype related genotype. We will focus on selecting with false discovery rate control.

1.1 HIV-1 Dataset

Note that we are interested in the gene mutations associated with HIV-1 drug resistance. We use a data set described and analyzed by Rhee et al. (2006) and it's available at HIV Data Set. The ground truth is provided by Rhee et al. (2005).

The dataset consists of the genotype and the drug resistance of several drug types with respect to each genotype. There are three drug classes and each contains several specific drug types (7+6+3). Our data set contains the drug resistance of drug types and the ground truth is for drug classes.

Drug Class[1]	PI:	APV ATV IDV LPV NFV RTV SQV
	NRTI:	X3TC ABC AZT D4T DDI TDF
	NNRTI:	DLV EFV NVP

Genotypes were derived from the amino acid sequences of positions $1 \sim 99$ in PI class and $1 \sim 240$ in NRTI and NNRTI classes. On each position, there are several possible mutation directions. Isolates included viruses from the plasma of HIV-1-infected persons and laboratory viruses with drug-resistance mutations resulting from site directed mutagenesis or in vitro passage. Database for which both sequences and in vitro susceptibility results were available.

1.2 Multiple Hypothesis Testing

Since the target of genotype varies, detecting the drug-associated genotypes can be separated to two types:

- **Case I:** Detect the mutative positions, e.g. $P1$, $P2$, $\cdots$.
- **Case II:** Detect the mutative positions and the mutative directions simultaneously, e.g. $P1.A$, $P1.B$, $\cdots$.

For each case, if we assign an order on these genes, we can formalize our problem as multiple hypothesis problem. The null hypothesis is $\mathcal{H}_0$ where $i \in \mathcal{H}_1$ if i-th gene associates with HIV-I drug resistance for drug d and $j \in \mathcal{H}_0$ otherwise. Applying a rejection procedure based on data, we will have Table 1.

Benjamini and Hochberg (1995) proposed to control the false discovery rate FDR, which is defined as

$$FDR = \mathbb{E}\left(\frac{V}{R \vee 1}\right).$$

[1] To avoid ambiguity, we use drug class to denote macro-categories (PI, NRTI, NNRTI) and use drug type to denote the micro-categories (APV, ATV, IDV, LPV$\cdots$)

Table 1: Outcomes when testing m hypothesis

Hypothesis	$\mathcal{H}_0$	$\mathcal{H}_1$	
Reject	V	S	R
Fail to Reject	U	T	W
Total	m_0	m_1	m

also, the false discovery proportion FDP is defined as

$$FDP = \frac{V}{R \vee 1}.$$

Our target is rejecting more hypothesis while controlling the FDR at a target level α. However, in reality, since the expriment conducts only one time, we actually control FDP.

To measure the efficiency of the rejection, we define discovery rate as

$$DP = \frac{S}{m_1}.$$

If the gene can be seperated into groups $A_1, A_2, \cdots, A_G$, we can also define the groupwise FDR. The induced groupwise null hypothesis is $\mathcal{H}_0^{Group}$, where $g \in \mathcal{H}_0^{Group}$ if there exists $i \in A_g$ such that $i \in \mathcal{H}_0$, $g = 1, 2, \cdots, G$. Then the FDP, FDR and DP can be defined accordingly.

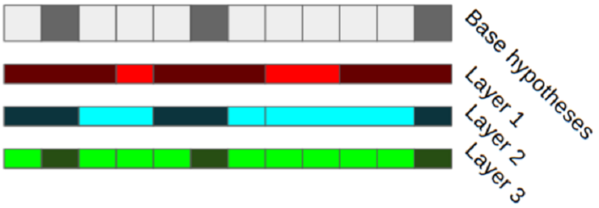

Figure 1: Multiple Group Information induces Multilayer Hypothesis Testing Problem

2 Multilayer Hypothesis Testing

Inspired by Dai and Barber (2016), we considered a multi-task problem. Upon fixing a drug class, like PI, the model is

$$\mathbf{Y} = \mathbf{X}\mathbf{B} + \mathbf{E}.$$

where the response $\mathbf{Y} \in \mathbb{R}^{n\times r}$, $\mathbf{X} \in \mathbb{R}^{n\times p}$, $\mathbf{B} \in \mathbb{R}^{p\times r}$ and $\mathbf{E} \in \mathbb{R}^{n\times r}$. The column of $\mathbf{Y}$ denotes the response of one particular drug type, the column of $\mathbf{X}$ is the indicator of a given gene mutative direction $(P1.A, P1.B, \cdots)$. B_{ij} denotes the underlying effect of drug resistance associated with gene i with respect to HIV drug type j .

Since the mechanisms of different drug types within a drug class are similar, we can assume B is **row-sparsed**. This assumption also coincides with the ground truth from Rhee et al. (2006).

Denote $y = \text{vec}(\mathbf{Y})$, $\epsilon = \text{vec}(\mathbf{E})$, $\beta = \text{vec}(\mathbf{B})$, $\mathbb{X} = \mathbf{I}_r \otimes \mathbf{X}$, the model becomes

$$y = \mathbb{X}\beta + \epsilon.$$

We remove the gene mutative direction whose frequencies are less than 5 and the possible duplicates. Then, the group information is avaiable and the false discovery rate is defined groupwise.

- **Layer I**: The group is separated by the gene mutative position and the induced partition is $\{A_{P1}, A_{P2}, \ \cdots, A_{P99}\}$;
- **Layer II**: The group is separated by the gene mutative position and gene mutative direction. The induced partition is

$$\{A_{P1.A}, A_{P1.B}, \cdots, A_{P99.d}\}.$$

2.1 Multilayer p-filter

Both Barber and Ramdas (2017) and Katsevich and Sabatti (2019) can deal multilayer hypothesis testing. Katsevich and Sabatti (2019) suffer from the instability discussed in Section 2.2.3, so we introduce the p-filter procedure.

P-filter aims to control the groupwise FDR at α_m within each layer. Consider a vector of p-value $\{p_i\}$ corresponding to hypotheiss $\mathcal{H}_0$ and we want to test with the groupwise FDR in Layer I and Layer II under control. Given two dimension threshold (t_1, t_2), the rejection set is

$$R(t_1, t_2) = \left\{ i : p_{g_1(i)1} > t_1, p_{g_2(i)2} > t_2, i \in A^m_{g_m(i)} \right\}.$$

For Layer m, the group partition is $A^m_1, \cdots, A^m_{G_m}$ and the induced hypothesis is $\mathcal{H}^m_0$. For $g \in \{1, 2, \cdots, G_m\}$, we can construct p value for group g through Simes Test p^{Simes}_g. Then expected false rejection is approximately bounded by

$$\begin{aligned}\sum_{g \in \mathcal{H}^m_0} \mathbf{1}\{p^{Simes}_g \le t_m\} \approx \sum_{g \in \mathcal{H}^m_0} \Pr\{p^{Simes}_g \le t_m\} \\ \le \#\{g : g \in \mathcal{H}^m_0\} \times t_m \\ \le G_m \times t_m.\end{aligned}$$

Hence, the upper estimator of FDP is

$$\widehat{FDP}_m = \frac{G_m \times t_m}{R(t_1, t_2) \vee 1}, m = 1, 2, \cdots.$$

The (t_1, t_2) is found by recursively coordinate updation. The general p-filter algorithm is given in Algorithm 1.

2.2 Comparison with Existing Method

Barber and Candès (2015), Janson and Su (2016) and Dai and Barber (2016) have examined this dataset. Janson and Su (2016) considered control the gene mutative direction (Case II: $P1.A, P1.B$) with FWER control and we won't discuss it here.

2.2.1 Inconsistency between Gene Detection and Interpretation

Barber and Candès (2015) and Dai and Barber (2016) applied the knockoff filter on this dataset. They implemented their procedure on the gene mutative

Algorithm 1 Multilayer p-filter

```
Input: a vector of $p$-values $P \in [0,1]^n$; target FDR levels $\alpha_1, \ldots, \alpha_M$; partition $m$ given by $A_1^m, \ldots, A_{G_m}^m \subseteq [n]$ for $m = 1, \ldots, M$
Initialize: thresholds $t_1 = \alpha_1, \ldots, t_M = \alpha_M$
  repeat
    for $m = 1$ to $M$ do
      Define $\hat{R}_m(\cdot)$ as
          $$\hat{R}_m\left(t_1, \ldots, t_M\right) = \left\{g \in [G_m] : \hat{R}\left(t_1, \ldots, t_M\right) \cap A_g^m \neq \emptyset\right\}$$
      Let
          $$t_m \leftarrow \max\left\{T \in [0, t_m] : \frac{G_m T}{1 \vee \left|\hat{R}_m\left(t_1, \ldots, t_{m-1}, T, t_{m+1}, \ldots, t_M\right)\right|} \leq \alpha_m\right\}$$
    end for
  until the thresholds $t_1, \ldots, t_M$ are all unchanged in the last round
Output: adaptive thresholds $\hat{t}_1 = t_1, \ldots, \hat{t}_M = t_M$.
```

direction (Case II: $P1.A, P1.B, \cdots$). However, they presented their results with respect to the gene mutative position (Case I: $P1, P2, \cdots$). The result seems satisfactory but the explanation is questionable. We want to reveal the result in Case I.

2.2.2 Inconsistency between Ground Truth and available Data

Recall that Rhee et al. (2005) provieded the ground truth for three major drug classes, PI, NRTI and NNRTI. The inference was performed on the minor drug types in Barber and Candès (2015). This approach contradicts the form of ground truth. Dai and Barber (2016) transformed the problem into a multi-task problem and unified different drug types within one drug class. Therefore, we follow their framework.

2.2.3 Instability of Knockoff Filter

Katsevich and Sabatti (2019) extended knockoff filter to multilayer hypothesis testing, but we won't interpret the result based on multilayer knockoff. The reason is the discovery drastically changes along with k, the threshod of gene mutation frequencies for maintaining. When I chose $k = 3$, the total discoveries are 27 (22 true and 5 false) while it becomes 0 when $k = 5$. This phenomenon suggests knockoff filter is unstable sometimes.

3 Result and Discussion

We conducted p-filter procedure with the p filter code and the multilayer knockoff with the multilayer knockoff code. Part of the data extraction and analysis refers to Knockoff Guide. The group knockoff filter and group BH procedure are special cases for multilayer knockoff and p-filter respectively.

3.1 False Discovery Proportion and Discovery Proprtion

Figure 2 models different drug types separately while Figure 3 treats different drug types within one drug class as a whole. In terms of FDP control, MKF performs well but it sacrifices the discovery proportion. From Figure 2, MKF is also unstable because there is no discovery in some cases.

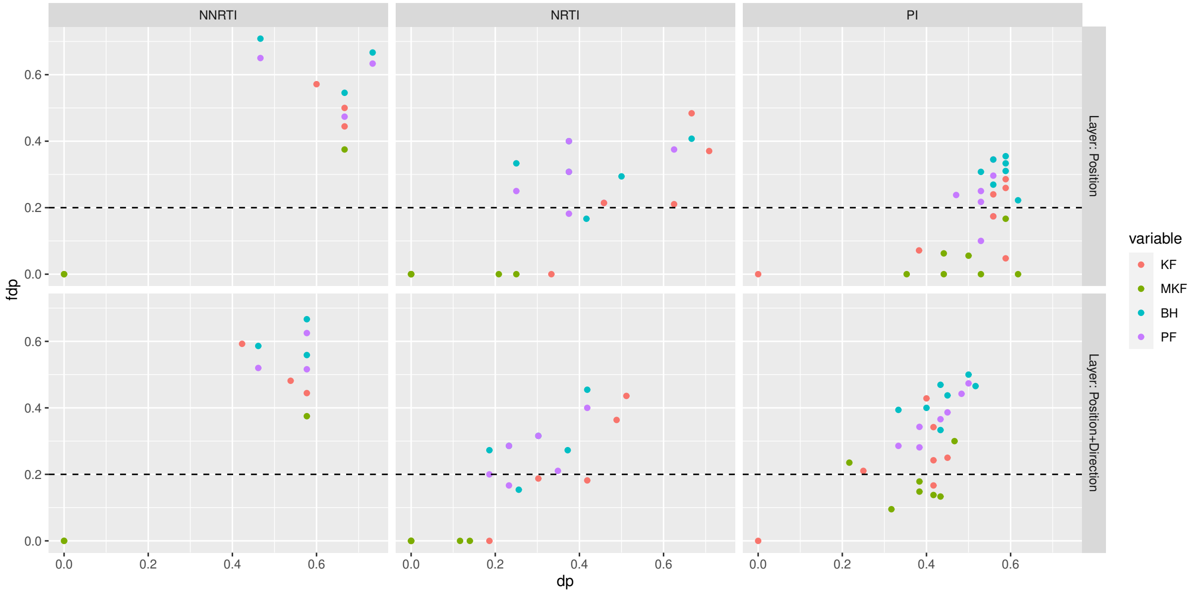

Figure 2: The FDP and DP for different drug types of the drug class PI, NRTI and NNRTI. KF is knockoff filter (Case I), MKF is multilayer knockoff filter (Case I and Case II), BH is Benjamini Hochberg procedure (Case I) and PF is p-filter (Case I and Case II). The gene mutative direction whose frequencies below 5 is removed. The FDP target is 20% in all cases.

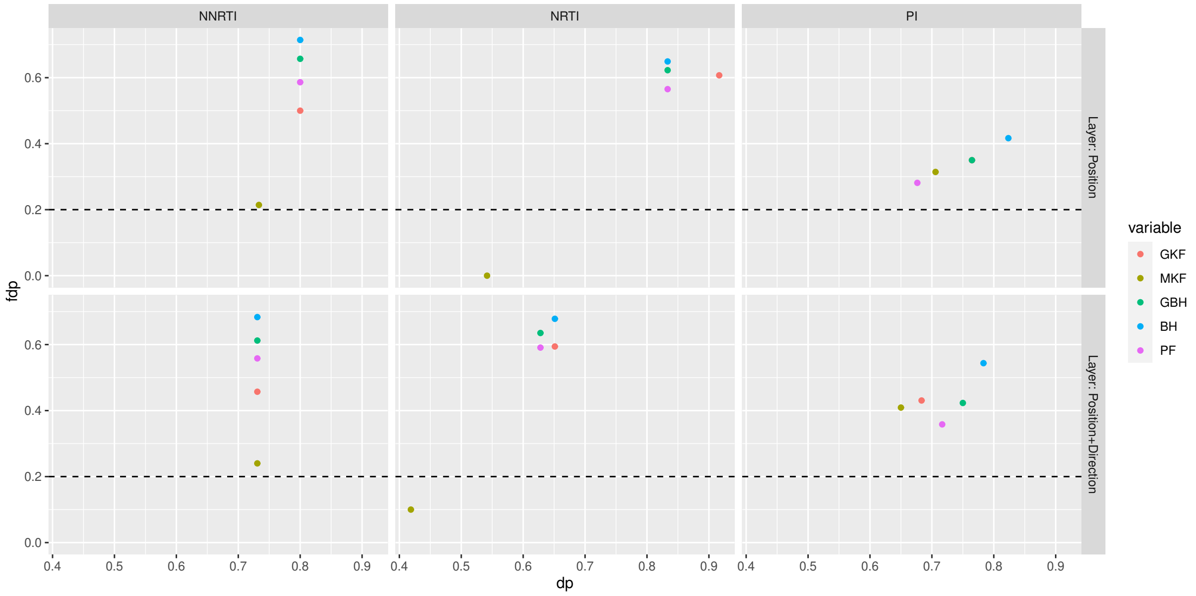

Figure 3: The FDP and DP for three drug classes PI, NRTI and NNRTI. GKF is group knockoff filter (Layer I), MKF is multilayer knockoff filter (Layer I and Layer II), BH is Benjamini Hochberg procedure (Case I), BH is groupwise Benjamini Hochberg procedure (Layer I) and PF is p-filter(Layer I and Layer II). The gene mutative direction whose frequencies below 5 is removed. The FDP target is 20% in all cases and layers.

3.2 The Mutative Frequencies versus Quality of Detection

We also investigated whether low mutative frequencies lead to low accuracy in Case II. The selection procedure used in this section is p-filter. Figure 4 shows the emprical density of the mutation count. The mutations are divided into false discoveries, true discoveries and ground truth. From Figure 4, the mutation count concentrates on low frequencies and this is because low frequent mutations are prevalent.

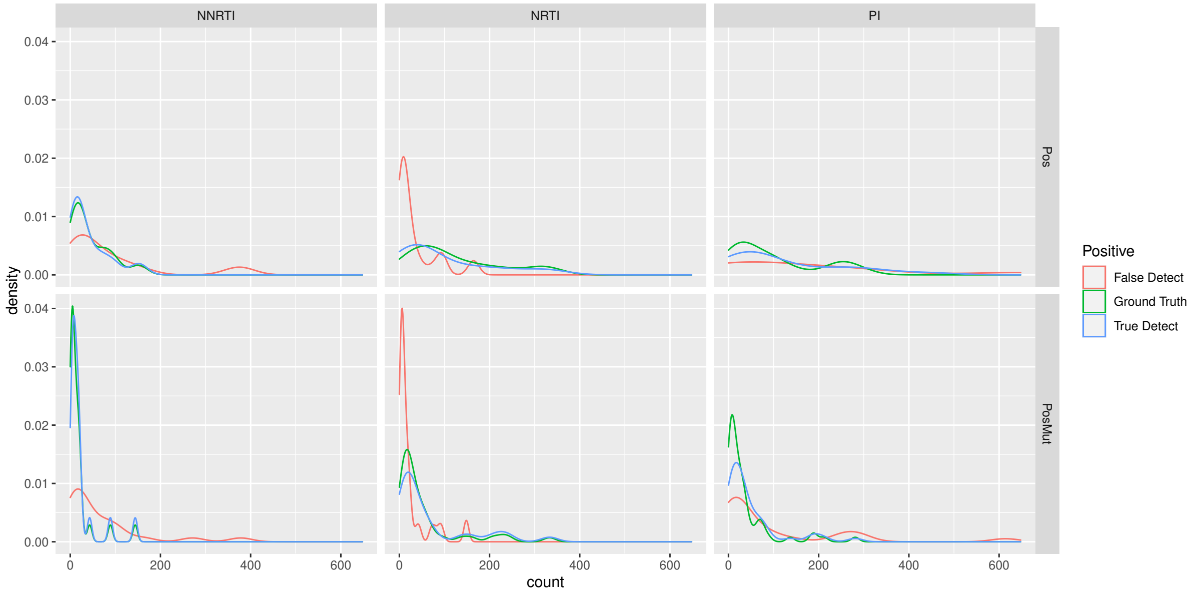

Figure 4: The empirical density of the position + mutative direction count for three drug classes. We use p-filter as our selection procedure. From the figure, the density of true detection nearly matches that of ground truth because they are highly correalted. No specific pattern indicates the low frequencies would cause false discoveries. Indeed, most of gene mutations with low frequencies are removed in the selection procedure.

3.3 Genes Associated with HIV-1 Drug Resistance

From Figure 2 and Figure 3, it seems the false discovery rate is not controlled, but does only the gene mutations suggested by Rhee et al. (2005) is valid? Another question is whether our selection excavates more related genes.

Table 2: New Discovered HIV-1 Drug Resistance related Gene Mutation

Drug Class	Negative Mutative Gene	Negative Mutative Gene Position
PI	P10.I P10.L P10.V P20.R P36.I P36.L P37.S P63.H P63.P P64.I P64.V P67.Y P71.T P71.V P82.I P91.S P93.L	36 37 63 64 91 93
NRTI	P103.N P118.V P121.H P135.T P142.V P162.Y P180.V P181.C P181.V P203.D P215.D P227.L P35.I P35.R P40.F P4.S P70.G P83.K	103 118 121 135 142 162 180 181 227 35 40 4 83
NNRTI	P101.H P101.Q P135.T P138.A P139.R P179.D P179.E P184.V P215.Y P219.N P49.R P74.V P98.G	135 139 179 184 215 219 49 74 98

As discussed in Section 3.1, we remove MKF. We consider unified approach, i.e., integrating different drug types within one drug class. We take the intersection of gene mutations selected by GKF, GBH, BH and PF and focus on "false" discoveries.

Geno Clinical Review collects abundant literature about HIV-1 drug resistance to PI, NRTI and NNRTI respectively. THE WORLD HEALTH ORGANIZATION 2009 LIST OF MUTATIONS also provides more information. When additional support on these websites is available, we mark our new discoveries red.

From Table 2, in PI class, our new discoveries mostly have sustainability, while in NRTI class, no much evidence justifies our finding. However, our finding intuitively points out some gene mutations worth researching.

References

Barber, R. F. and Candès, E. J. (2015). Controlling the false discovery rate via knockoffs. *The Annals of Statistics*, 43(5):2055–2085.

Barber, R. F. and Ramdas, A. (2017). The p-filter: multilayer false discovery rate control for grouped hypotheses. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(4):1247–1268.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300.

Benjamini, Y. and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics*, 29(4):1165–1188.

Dai, R. and Barber, R. (2016). The knockoff filter for fdr control in group-sparse and multitask regression. In Balcan, M. F. and Weinberger, K. Q., editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 1851–1859, New York, New York, USA. PMLR.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70.

Janson, L. and Su, W. (2016). Familywise error rate control via knockoffs. *Electron. J. Statist.*, 10(1):960–975.

Katsevich, E. and Sabatti, C. (2019). Multilayer knockoff filter: Controlled variable selection at multiple resolutions. *The Annals of Applied Statistics*, 13(1):1–33, 33.

Ren, Z., Wei, Y., and Candès, E. (2020). Derandomizing knockoffs. *Arxiv*.

Rhee, S. Y., Fessel, W. J., Zolopa, A. R., Hurley, L., Liu, T., Taylor, J., Nguyen, D. P., Slome, S., Klein, D., Horberg, M., Flamm, J., Follansbee, S., Schapiro, J. M., and Shafer, R. W. (2005). Hiv-1 protease and reverse-transcriptase mutations: correlations with antiretroviral therapy in subtype b isolates and implications for drug-resistance surveillance. *J Infect Dis*, 192(3):456–65.

Rhee, S. Y., Taylor, J., Wadhera, G., Ben-Hur, A., Brutlag, D. L., and Shafer, R. W. (2006). Genotypic predictors of human immunodeficiency virus type 1 drug resistance. *Proc Natl Acad Sci U S A*, 103(46):17355–60.