Gene Mutations Associated with HIV-1 Drug Resistance

Linsui Deng
denglinsui@ruc.edu.cn

Institute of Statistics & Big Data
Renmin University of China

June 15, 2021

Content

- Introduction
  - HIV Data Set
  - Multiple Hypothesis Testing
  - Multilayer Hypothesis Testing
- Multilayer Hypothesis Testing
  - Model
  - Algorithm
- Result and Discussion
  - False Discovery Proportion and Discovery Proprtion
  - The Mutative Frequencies versus Quality of Detection
  - Genes Associated with HIV-1 Drug Resistance
- Guidance for Future Clinic Expriments

1.0.0 BackGround

- 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.

1.0.1 Our Target

- Detect gene mutations associated with HIV-1 drug resistance.

1.0.2 Challenges

- Sample size v.s. the complete gene mutations;
  - Shrink the candidate set via expert information;
  - Obtaining data is expensive.
- Reliable simultaneous inference on various genes.
  - Not just variable selection!

Introduction

HIV Data Set

1.1.1 Source of HIV Data Set

- We use HIV Data Set described and analyzed by Rhee et al. (2006);
- The ground truth is provided by Rhee et al. (2005);
- Additional information is avaiable at PI, NRTI, NNRTI and THE WORLD HEALTH ORGANIZATION 2009 LIST OF MUTATIONS.

1.1.2 HIV-1 Drugs

$$\textbf{Drug Class}[1] \begin{cases} \text{PI:} & \text{APV ATV IDV LPV NFV RTV SQV} \\ \text{NRTI:} & \text{X3TC ABC AZT D4T DDI TDF} \\ \text{NNRTI:} & \text{DLV EFV NVP} \end{cases}$$

1.1.3 Genotype

- Position: $P1 \sim P99$ in PI and $P1 \sim P240$ in NRTI and NNRTI;
- Mutative Direction: On each position, there are several possible mutation directions, A, B, $\cdots$

[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$\cdots$)

1.2.1 Two Detection Cases

- Why?
  - The targets of genotype detection (1.1.3) vary;
- What?
  - **Case I:** Detect the mutative positions, e.g. $P1$, $P2$, $\cdots$.
  - **Case II:** Detect the mutative positions and the mutative directions simultaneously, e.g. $P1.A$, $P1.B$, $\cdots$.
- How? (Multiple Testing)
  1. Determine dectection case (e.g. Case I w.r.t APV);
  2. Obtain Hypothesises $\{H_i\}_{i\in\{P_1,P_2,\cdots,P_{99}\}}$;
     - $H_{P1} = 1 \Rightarrow$ the gene mutation on position $P1$ associated with HIV-1 drug resistence.
  3. Run selection procedure to test H_i simultaneously.

1.2.2 **Discovery Table**: Applying a selection procedure, we get Table 1

Table: Discovering Table of *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

1.2.3 Criteria for Selection

- False Discovery Proportion: $FDP = \frac{V}{R \vee 1}$.
- Discovery Proportion: $DP = \frac{S}{m_1}$.
- **Groupwise FDP**:
  - When the gene can be seperated into groups $A_1, A_2, \cdots, A_G$;
  - $H_{A_i} = 0 \Leftrightarrow \exists k \in A_i \ s.t. \ H_k = 0$;
  - Implement multiple testing on $\{H_{A_i}\}_{i=1,2,\ldots,G}$

1.3.1 Intuition and Idea

- Can we consider Case I and Case II together?
- Conduct multiple testing with several group information;

1.3.2 Possible Group Information

- The drug types within the same class;
- The genes within the same position.

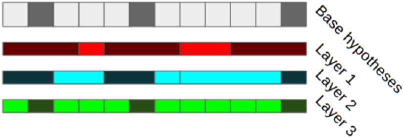

Figure: Multiple Group Information induces Multilayer Hypothesis Testing Problem

2.1.1 Multi-Task Model

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

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

where

- $\mathbf{Y} \in \mathbb{R}^{n\times r}$, $\mathbf{X} \in \mathbb{R}^{n\times p}$, $\mathbf{B} \in \mathbb{R}^{p\times r}$ and $\mathbf{E} \in \mathbb{R}^{n\times r}$;
- Y_{ij}: the response of isolates i with one drug type j;
- X_{ij}: the indicator of isolates i corresponding to gene mutative direction $j \in \{P1.A, P1.B, \cdots\}$;
- B_{ij}: 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**.

2.1.2 Single-Task Model with Group Info

Denote $y = \mathrm{vec}(\mathbf{Y})$, $\epsilon = \mathrm{vec}(\mathbf{E})$, $\beta = \mathrm{vec}(\mathbf{B})$, $\mathbb{X} = \mathbf{I}_r \otimes \mathbf{X}$, the model becomes[2]

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

Then, the group information is avaiable and FDP is defined groupwise.

- **Layer I**:
  - By gene mutative position;
  - Group partition: $\{A_{P1}, A_{P2}, \quad \cdots, A_{P99}\}$.
- **Layer II**:
  - By gene mutative position and direction;
  - Group partition: $\{A_{P1.A}, A_{P1.B}, \cdots, A_{P99.d}\}$.

[2] We remove the gene mutative direction whose frequencies is less than 3 and the possible duplicates.

2.1 Possible Algorithms:

- BH Procedure (Benjamini and Hochberg (1995));
- P-filter (Barber and Ramdas (2017));
- Knockoff (Barber and Candès (2015));
- Multilayer Knockoff (Katsevich and Sabatti (2019)).

Among these method,

- Multilayer Knockoff and p-filter could simultaneously control false discovery rate in different layers;
- p-value based procedure performs stablier than Knockoff method does.

Thus, we prefer **p-filter**.

Multilayer Hypothesis Testing

2.2.1 What does p-filter do?

- Control the groupwise FDR at α_k within the Layer k.

2.2.2 How to reject? (e.g. Two Layers)

- Given two dimension threshold (t_1, t_2), the rejection set is

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

2.2.3 How to control? (At Layer k)

- Construct p values for group g through Simes Test p_g^{Simes};
- Then expected false rejection is approximately bounded by $\sum_{g \in \mathcal{H}_0^k} \mathbf{1}\{p_g^{Simes} \leq t_k\} \preceq G_k \times t_k$;
- The upper estimator of FDP is $\widehat{FDP}_k = \frac{G_k \times t_k}{R(t_1, t_2) \vee 1}, m = 1, 2, \cdots$.

Result and Discussion

FDP and DP

We conducted these algorithms with p filter code, multilayer knockoff code and Knockoff Guide.

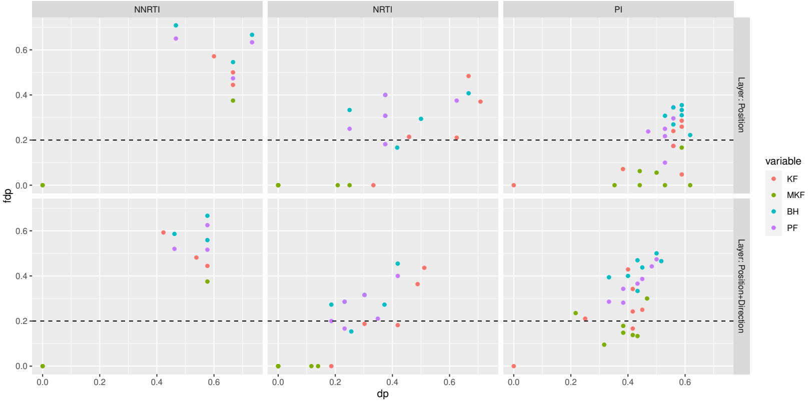

Figure: The FDP and DP for different drug types of the drug class PI, NRTI and NNRTI. KF is knockoff filter (Case I), MKF is multilayer knockoff filter (Case I and Case II), BH is Bejamini Hochberg procedure (Case I) and PF is p-filter (Case I and Case II)

Result and Discussion

FDP and DP

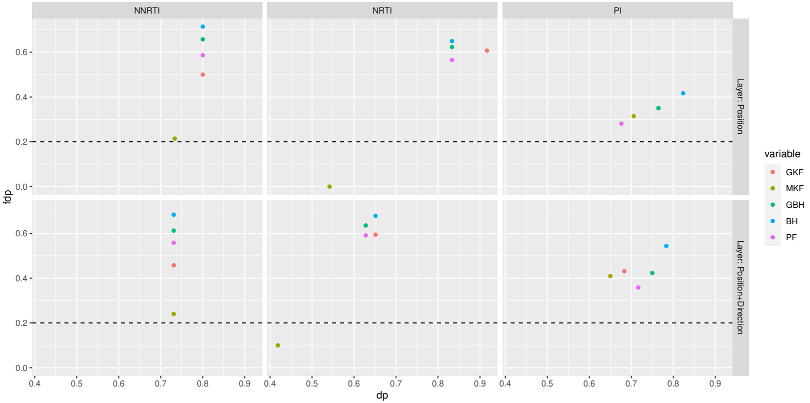

Figure: The FDP and DP for three drug classes PI, NRTI and NNRTI. GKF is group knockoff filter (Layer I), MKF is multilayer knockoff filter (Layer I and Layer II), BH is Bejamini Hochberg procedure (Case I), BH is groupwise Bejamini Hochberg procedure (Layer I) and PF is p-filter(Layer I and Layer II).

We also investigated whether low mutative frequencies leads to low accurancy in Case II. The selection procedure used in this section is p-filter.

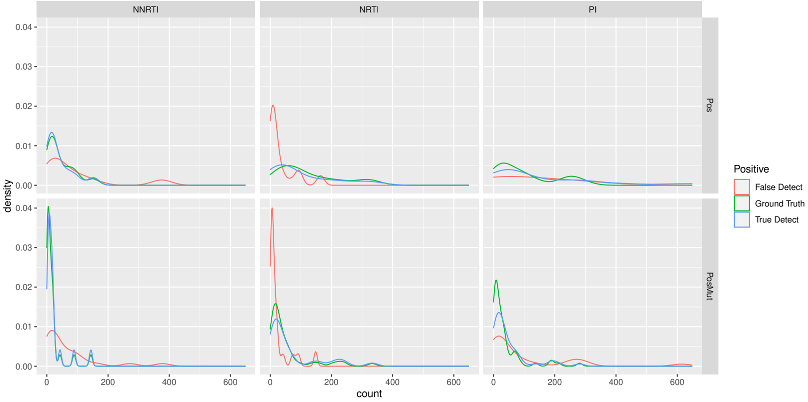

Figure: The empirical density of the position + mutative direction count for three drug classe.

Additional information is provided in PI, NRTI and NNRTI and THE WORLD HEALTH ORGANIZATION 2009 LIST OF MUTATIONS.

Table: New Discovered HIV-1 Drug Resistence related Gene Mutation

Drug Class	Neg Gene	Neg 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

Guidance for Future Clinic Expriments

Guidance for Future Clinic Expriments:

- We report some valuable gene mutations for three drug classes respectively;
- The new discoveried genes is worth considering in future clinic experiments;
- In the past, fewer experiments is conducted on NRTI and NNRTI, so the supporting information is less.
- However, the success in PI highly sustains more experiments on NRTI and NNRTI.

Thank You!

More Informations

More Informations could be found on

- Github Address: Github
- Visualization: Shiny
- Report: Report

Reference I

Rina Foygel Barber and Emmanuel J. Candès. Controlling the false discovery rate via knockoffs. *The Annals of Statistics*, 43(5): 2055–2085, 2015. ISSN 0090-5364.

Rina Foygel Barber and Aaditya Ramdas. The p-filter: multilayer false discovery rate control for grouped hypotheses. 79(4):1247–1268, 2017. ISSN 1369-7412.

Yoav Benjamini and Yosef Hochberg. 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, 1995. ISSN 00359246.

Ran Dai and Rina Barber. The knockoff filter for fdr control in group-sparse and multitask regression. In Maria Florina Balcan and Kilian Q. Weinberger, editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 1851–1859, 2016.

Reference II

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

S. Y. Rhee, W. J. Fessel, A. R. Zolopa, L. Hurley, T. Liu, J. Taylor, D. P. Nguyen, S. Slome, D. Klein, M. Horberg, J. Flamm, S. Follansbee, J. M. Schapiro, and R. W. Shafer. 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, 2005. ISSN 0022-1899 (Print) 0022-1899.

S. Y. Rhee, J. Taylor, G. Wadhera, A. Ben-Hur, D. L. Brutlag, and R. W. Shafer. Genotypic predictors of human immunodeficiency virus type 1 drug resistance. *Proc Natl Acad Sci U S A*, 103(46):17355–60, 2006. ISSN 0027-8424 (Print) 0027-8424.