Correlations between ADI-R Subscore A and Individual Specific DEGs

1. Introduction

Velmeshev, Dmitry, et al. (2019)에서 ASD 환자들의 single-cell profiles와 대조군의 profiles를 비교한 결과 L2/3, L5/6-CC을 포함한 여러 cell type에서 환자들의 ADI-R scores와 상관관계를 보였다. 또한 ADI-R scores와 가장 상관관계가 높은 differentially expressed genes (DEGs)는 transcriptional dysregulation의 정도가 높은 DEGs과는 관련이 적었다. 따라서, 유전자 발현량의 Fold change(ASD versus control) 수치만을 보기보다는 환자들에게 공통으로 발현되는 DEGs를 고려하여 ADI-R score의 상관관계를 분석하고자 한다.

논문에서 Combined ADI-R score을 산출할 때 각 점수의 rank를 구하여 합친다라고 간략하게만 나와 있고 B sub-score인 verbal과 nonverbal에 대한 처리가 나와 있지 않다. 그래서 이 Portfolio에서는 ADI-R score 중에서 variation이 score 중에서 가장 높은 A score(reciprocal social interaction의 이상 정도)와 cell type마다 갖고 있는 Individual specific DEGs의 수에 Fold change을 가중치로 부여한 값을 비교한다. 환자들의 ADI-R A score는 위 논문의 Data S1에서 확인할 수 있었고, Individual specific DEGs는 Data S4에서 5명 이상의 환자로부터 유의적으로 발현량 차이가 있는 유전자를 이용하였다.

2. Data Pre-processing

우선, excel 파일을 불러오고 데이터를 처리할 수 있는 Package를 불러온다.

```
# Load packages
library(tidyverse)
library(readxl)
```

이제 사용할 데이터를 읽어보자.

```
# Read excel files
S1_1 <- read_excel("aav8130_Data_S1.xlsx", sheet = 1)
S4_1 <- read_excel("aav8130_Data_S4.xls", sheet = 1)
S4_4 <- read_excel("aav8130_Data_S4.xls", sheet = 4)
```

데이터가 잘 읽혔는지 확인해 본다.

```
colnames(S1_1)
```

```
##  [1] "Patient ID"      "Diagnosis"       "Age"             "Sex"
##  [5] "PMI"             "Other diagnoses" "Medications"     "ADI-R"
##  [9] "...9"            "...10"           "...11"           "...12"
## [13] "Cause of death"
```

```
colnames(S4_1)
```

```
##  [1] "Cell type"
##  [2] "gene ID"
##  [3] "Gene name"
##  [4] "Gene biotype"
##  [5] "Fold change"
##  [6] "Sample fold change"
##  [7] "q value"
##  [8] "correlation (bulk mRNA/bulkized nuclear RNA)"
##  [9] "Epilepsy DEG"
## [10] "gene group"
## [11] "SFARI gene"
## [12] "Satterstrom"
## [13] "Sanders"
## [14] "cell type-specific expression"
```

```
colnames(S4_4)
```

```
## [1] "Cell type"
## [2] "Gene ID"
## [3] "Gene name"
## [4] "Direction of change (ASD/Control)"
## [5] "# of ASD patients with significant change"
## [6] "ASD patients with DE signal"
```

S1_1의 데이터를 살펴보면 S1_1의 열 이름 중에 ADI-R scores가 누락되어 1행이 존재하는 것을 알 수 있다. 따라서 1행을 삭제해 주고 열 이름을 다시 지정해주어야 한다.

```
S1_1 <- S1_1[-1,]
colnames(S1_1)[8:12] <- c('A','Bv','Bnv','C','D')
```

대조군 집단과 일부 환자들의 A score가 결측치로 되어 있기 때문에 filter 함수를 통해 이를 제외한다. 또한 Patient ID는 범주형 변수로 활용하기 위해 character로, A score는 양적 변수로 활용하기 위해 numeric으로 class를 바꾸어 준다. 사용할 정보는 이 두 개의 정보 밖에 없으므로 select 함수로 간단하게 해준다.

```
S1_1 <- S1_1 %>% filter(!(A=='NA')) %>%
  mutate('Patient ID' = as.character(`Patient ID`), A = as.numeric(A)) %>%
  select(`Patient ID`, A)
S1_1
```

```
## # A tibble: 12 x 2
##    `Patient ID`     A
##    <chr>        <dbl>
##  1 5278            22
##  2 5144            28
##  3 5403            30
##  4 5419            24
##  5 4899            22
##  6 5978            13
##  7 6033            26
##  8 5864            18
##  9 5939            29
## 10 5565            27
## 11 5294            17
## 12 4849            22
```

S4_4의 데이터의 경우 우리가 사용할 patient ID 정보가 `ASD patients with DE signal`에 한 문자열로 묶여 있다. 환자마다 각각의 유전자가 DEGs인지 확인하기 위해서 `str_count` 함수를 이용하여 `ASD patients with DE signal`에서 추출된 patient ID들 중에 알고 싶은 환자 ID가 포함되어 있는지 확인하면 된다. 해당 환자 ID가 포함되어 있다면 1이, 포함되어 있지 않다면 0이 결과값이 된다. A score를 갖고 있는 모든 환자에 대해서 조사하려면 `sapply` 함수를 사용한다. 이를 통해 생성된 data.frame을 `ind_DEGs`에 지정한다.

```
ind_DEGs <- sapply(S1_1$`Patient ID`,
                   function(ID) str_count(S4_4$`ASD patients with DE signal`, ID))
head(ind_DEGs)
```

```
##      5278 5144 5403 5419 4899 5978 6033 5864 5939 5565 5294 4849
## [1,]    0    0    0    0    0    1    1    1    1    1    0    1
## [2,]    0    0    1    0    0    0    1    0    0    1    0    1
## [3,]    0    0    1    0    1    1    0    1    1    1    0    1
## [4,]    0    0    0    0    0    0    1    1    0    1    0    1
## [5,]    0    1    0    0    0    1    1    0    1    1    0    1
## [6,]    0    1    0    0    0    0    0    1    1    1    0    1
```

`cbind` 함수를 이용하여 `ind_DEGs`를 S4_4에 붙여보자.

```
S4_4 <- cbind(S4_4, ind_DEGs)
```

이제 환자마다 cell type에 따른 indivisual specific DEGs의 수를 찾아본다. `group_by` 함수로 `Cell type`에 따라 grouping하고 각 환자에 대해서 `ind_DEGs` 행들의 합을 구한다. 이 때 단지 DEGs의 수만으로 A score와의 관계를 설명하는 것은 무리가 있다. 따라서 gene마다 얼마나 영향을 주는지를 고려하여 각 `Cell type`의 genes의 Fold Change의 절대값을 가중치로 두고 합한다. `summarize_at` 함수를 통해 이 값을 한꺼번에 구할 수 있다. 이 결과를 `effect`에 지정한다.

```
effect <- S4_4 %>% merge(S4_1 %>% select(`Cell type`, `Gene name`, `Fold change`)) %>%
  mutate("Fold change" = abs(`Fold change`)) %>%
  mutate_at(vars(`5278`:`4849`), function(x) x*.$`Fold change`) %>%
  group_by(`Cell type`) %>% summarise_at(vars(`5278`:`4849`), sum)
effect
```

```
## # A tibble: 13 x 13
##    `Cell type` `5278` `5144` `5403` `5419` `4899` `5978` `6033` `5864` `5939`
##    <chr>        <dbl>  <dbl>  <dbl>  <dbl>  <dbl>  <dbl>  <dbl>  <dbl>  <dbl>
##  1 AST-FB       0      0      0      0      0      0      0      0.463  0
##  2 AST-PP       0.804  0.791  2.18   0.462  0.452  3.84   1.96   2.74   1.90
##  3 IN-PV        0      0      0.289  0      0.183  0.748  0.728  0.684  1.17
##  4 IN-SST       0.150  0.705  0.288  0.150  0      1.23   0.945  0.288  1.23
##  5 IN-SV2C      0      0      0      0      0      0      0      0.357  0.357
##  6 IN-VIP       1.58   2.38   1.08   1.93   0.237  5.85   1.87   1.97   5.64
##  7 L2/3         1.97   3.78   3.76   2.77   4.35   5.71   6.19   4.38   6.29
##  8 L4           2.02   4.21   6.52   3.72   3.11   5.64   4.37   3.57   7.15
##  9 L5/6         0      0      0      0      0      0.552  0.552  0.552  0.344
## 10 L5/6-CC      0.468  0.269  0.862  0      1.40   1.57   2.07   0.827  1.67
## 11 Neu-NRGN-II  0      0.800  0      0.800  1.19   1.48   1.09   1.09   1.45
## 12 Oligodendr~  1.64   1.64   1.40   1.64   0.416  0.240  0      0.747  0.479
## 13 OPC          0.387  0.387  0      0      0      0      0      0      0
## # ... with 3 more variables: `5565` <dbl>, `5294` <dbl>, `4849` <dbl>
```

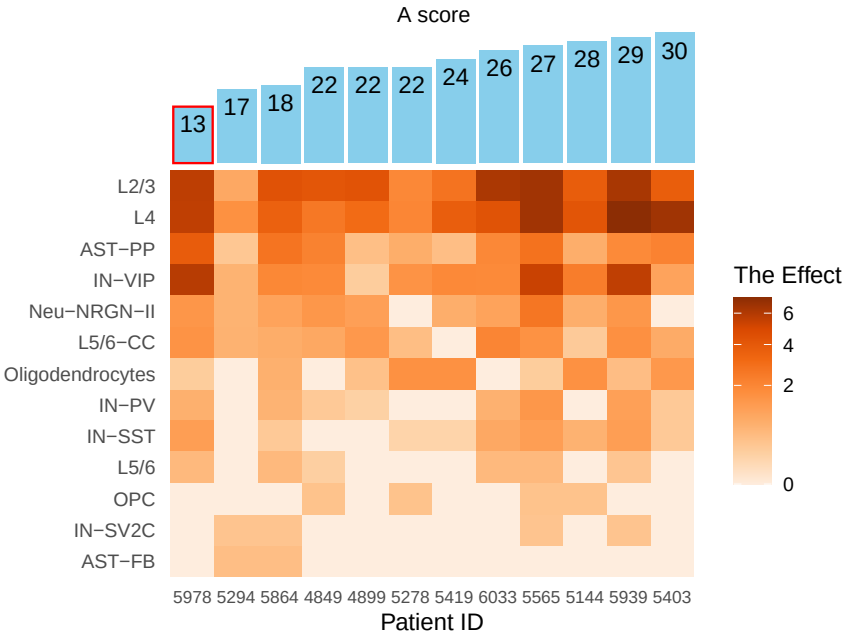
환자 A score와 effect를 비교하기 위해서 우선 gather 함수를 이용하여 effect를 long format으로 바꿔준다. 그리고 A score 정보를 가진 S1_1를 merge 함수를 통해 합쳐준다. 이 data.frame을 편의를 위해 df라고 지정한다.

```
# transform data.frame from wide into long format
df <- effect %>% gather(`Patient ID`, effect, -`Cell type`) %>% merge(S1_1)
```

3. Data Visualization

분석을 하기 전에 우선 데이터가 어떻게 분포되어 있는지 확인해 보자.

```
library(RColorBrewer)
library(cowplot)
tile <- df %>% ggplot(aes(reorder(`Patient ID`, A, median),
                          reorder(`Cell type`, effect, median), fill = effect)) +
  geom_tile() +
  labs(x = "Patient ID", y = NULL, fill = "The Effect") +
  scale_fill_distiller(trans = "sqrt", type = "seq", palette = 7, direction = 1) +
  theme_minimal() +
  theme(panel.grid = element_blank(),axis.text.x = element_text(size = 8))
xplot <- df %>% group_by(`Patient ID`) %>%
  summarise(A = mean(A)) %>%
  ggplot(aes(reorder(`Patient ID`, A), A, label = A)) +
  geom_bar(stat = "identity", fill = "skyblue", show.legend = F,
           color = ifelse(df %>% group_by(`Patient ID`) %>%
                            summarise(A = mean(A)) %>% .$A == 13, "red", NA)) +
  geom_text(vjust = 1.5, size = 4, show.legend = F) +
  ggtitle("A score") + theme_void() +
  theme(plot.title = element_text(size = 10, hjust = 0.5))
ggdraw() +
  draw_plot(xplot, .14, .68, width = .428, height = .25) +
  draw_plot(tile, 0, 0 , width = .7, height = .7)
```



ggplot를 이용하여 `Cell type`에 따른 A score와 DEGs의 `effect` 값을 비교했을 때, `Cell type`마다 DEGs의 차이가 크고 `effect` 값에 0이 많은 `Cell type`이 많아서 상관분석에 불리하다. 또한 빨간 막대로 표시된 환자는 나머지 환자들과 `effect`에서 완전히 다른 양상을 보였다. `Cell type`별로 Pearson correlation analysis를 구해보면 다음과 같다.

```
df %>% group_by(`Cell type`) %>%
  summarize(r = cor(A, effect, method = "pearson"),
            p.value = cor.test(A, effect, method = "pearson")$p.value) %>%
  arrange(p.value)
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
## # A tibble: 13 x 3
##    `Cell type`            r p.value
##    <chr>              <dbl>   <dbl>
##  1 AST-FB           -0.507   0.0924
##  2 L4                0.494   0.103
##  3 Oligodendrocytes  0.353   0.260
##  4 L2/3              0.264   0.407
##  5 OPC               0.224   0.484
##  6 L5/6             -0.202   0.528
##  7 IN-SST            0.199   0.535
##  8 AST-PP           -0.195   0.545
##  9 IN-PV             0.168   0.601
## 10 IN-SV2C          -0.0590  0.856
## 11 Neu-NRGN-II      -0.0459  0.887
## 12 L5/6-CC           0.0315  0.923
## 13 IN-VIP            0.0102  0.975
```

상관계수가 $p < .05$을 만족시키는 `Cell type`은 존재하지 않았다. 그 이유로 연구에 사용된 ASD 환자 샘플 수가 적어서 빨간 막대로 표시된 다른 양상을 보이는 환자에 의해 분석의 신뢰도가 낮아진 것으로 보인다. 따라서 이

환자를 제외하여 분석을 해봤다. 또한 effect의 0 값이 대부분인 Cell type의 경우 잘못된 상관관계가 나타날 수 있어서 이 세포들을 제외하고 다시 분석을 진행해보았다.

```
corr <- df %>% group_by(`Cell type`) %>% filter(A > 13 & median(effect) > 0) %>%
  summarize(r = cor(A, effect, method = "pearson"),
            p.value = cor.test(A, effect, method = "pearson")$p.value) %>%
  arrange(p.value)
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

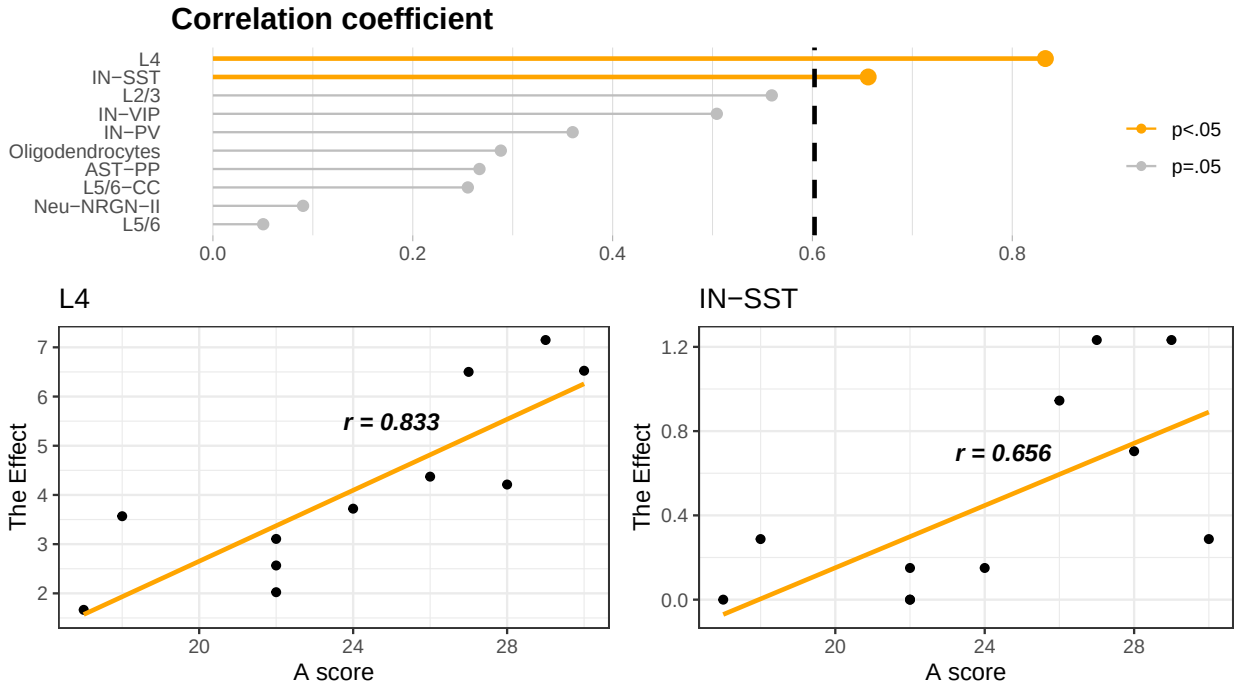
```
corr
```

```
## # A tibble: 10 x 3
##    `Cell type`           r p.value
##    <chr>             <dbl>   <dbl>
##  1 L4                0.833  0.00145
##  2 IN-SST            0.656  0.0285
##  3 L2/3              0.559  0.0736
##  4 IN-VIP            0.504  0.114
##  5 IN-PV             0.360  0.277
##  6 Oligodendrocytes 0.288  0.390
##  7 AST-PP            0.267  0.428
##  8 L5/6-CC           0.255  0.449
##  9 Neu-NRGN-II       0.0902 0.792
## 10 L5/6              0.0503 0.883
```

분석 결과 $p < .05$을 만족하는 Cell type은 L4, IN-SST이었다. 상관관계 분석 결과와 $p < .05$를 만족하는 Cell type의 A score와 분포를 ggplot을 이용해 나타었다.

```
p1 <- df %>% filter(A > 13 & `Cell type` == 'L4') %>%
  ggplot(aes(A, effect)) + geom_point() +
  geom_smooth(method = 'lm', se = F, col = "orange") + ggtitle('L4') +
  geom_text(data = data.frame(x = 25, y = 5.5, text = "r = 0.833"),
            aes(x, y, label = text), size = 4, fontface = "bold.italic") +
  xlab("A score") + ylab("The Effect") + theme_bw()
p2 <- df %>% filter(A > 13 & `Cell type` == 'IN-SST') %>%
  ggplot(aes(A, effect)) + geom_point() +
  geom_smooth(method = 'lm', se = F, col = "orange") + ggtitle('IN-SST') +
  geom_text(data = data.frame(x = 24.5, y = 0.7, text = "r = 0.656"),
            aes(x, y, label = text), size = 4, fontface = "bold.italic") +
  xlab("A score") + ylab("The Effect") + theme_bw()
p3 <- corr %>% mutate(`Cell type` = reorder(`Cell type`, r)) %>%
  ggplot(aes(col = ifelse(corr$`Cell type` %in% c("L4","AST-FB", "IN-SST"),
                          "p<.05", "p .05"))) +
  geom_point(aes(r, `Cell type`),
             size = ifelse(corr$`Cell type` %in% c("L4","AST-FB", "IN-SST"),
                                    3, 2)) +
  geom_segment(aes(0, `Cell type`, xend = r, yend = `Cell type`),
               size = ifelse(corr$`Cell type` %in% c("L4","AST-FB", "IN-SST"), 1, 0.5)) +
  geom_vline(aes(xintercept = 0.6021), lty = 2, size = 1) +
  scale_color_manual(values = c("orange", "grey")) +
  theme_light() +
  theme(panel.grid.major.y = element_blank(),
```

```
    panel.border = element_blank(),
    axis.ticks.y = element_blank(),
    legend.title = element_blank(),
    title = element_text(face = "bold", size = 12)) +
  labs(title = "Correlation coefficient", x = NULL, y = NULL)
plot_grid(p3, plot_grid(p1, p2, nrow = 1), nrow = 2, rel_heights = c(1, 1.5))
```



4. Discussion

이 분석에서의 한계은 각각의 Patient와 Control 간의 Fold change 정보를 이용할 수 없어서, 전체 환자들과 Control에 대한 Fold change를 이용했다는 점이다. 또한 A score 점수가 없거나 이질적인 환자들을 배제했다. Sample size가 작다는 것도 이 분석의 결정적인 한계이다.

하지만 이 분석을 통해 Excitatory neurons(L4) 뿐만 아니라 GABAergic interneurons(IN-SST)에서도 환자 5명 이상에서 공통으로 나타나는 DEGs를 고려했을 때 Excitatory neurons보다 작음에도 불구하고 Clinical severity(reciprocal social interaction의 이상)와 상관있을 수 있다는 결과를 얻었다. 이와 관련하여 IN-SST에서 발현되는 유전자에 대해서 더 알아보면 좋을 것 같다.