Test de Fisher unilateral

Guillermo Ayala Gallego

5/16/23

Table of contents

Análisis de sobre representación

- Hasta este momento hemos obtenido una ordenación de los genes.
- Esta lista la hemos estudiado pretendiendo que cuando mayor sea la expresión diferencial del gen este aparezca antes en la lista.
- De este modo el primer gen en la lista es el **marginalmente** tiene una mayor expresión diferencial y así sucesivamente.
- De hecho, el p-valor no es más que una medida de esa diferenciación.
- Una expresión muy utilizada en la literatura es que que hay una asociación entre el gen (su expresión) y el fenotipo.
- Cuando fijamos una tasa de error (FDR) lo que hacemos es fijar un punto de corte.
- Los genes que están antes del punto de corte se consideran significativos y los que siguen no.
- ¿Cómo interpretamos este grupo de genes?

Test de Fisher

- Sea G es el conjunto de genes considerado (posiblemente los contemplados en nuestro estudio).

- S_0 ($\subset G$) el conjunto de genes que nuestro estudio ha indicado como significativo.
- S_1 un conjunto de genes predefinido (misma función, misma localización).

	S_1	S_1^c	Total
S_0	n_{11}	n_{12}	$n_{1\cdot}$
S_0^c	n_{21}	n_{22}	$n_{2\cdot}$
Total	$n_{\cdot 1}$	$n_{\cdot 2}$	N

- Bajo la hipótesis de que no hay ningún tipo de asociación entre fila y columna, entonces la probabilidad de **lo observado** es

$$P(N_{11} = n_{11}) = \frac{\binom{n_{1\cdot}}{n_{11}}\binom{n_{2\cdot}}{n_{\cdot 1}-n_{11}}}{\binom{n}{n_{\cdot 1}}}.$$

- Supongamos que estamos contrastando la posible sobrerepresentación.
- **Bajo la hipótesis de independencia** (condicionada a las marginales), **rechazaríamos** la hipótesis nula de independencia para un valor mayor o igual al observado.
- El p-valor sería

$$p = P(N_{11} \geq n_{11}) = \sum_{t=n_{11}}^{\min\{n_{1\cdot},n_{\cdot 1}\}} = \frac{\binom{n_{1\cdot}}{t}\binom{n_{2\cdot}}{n_{\cdot 1}-t}}{\binom{N}{n_{\cdot 1}}}.$$

- Un ejemplo

	S_1	S_1^c	Total
S_0	30	40	70
S_0^c	120	156	276
Total	150	196	346

Podemos utilizar el **test exacto de Fisher direccional** (o unilateral o de una cola).

```
conteos = matrix(c(30,120,40,156),ncol=2)
fisher.test(conteos,alternative = "greater")
```

```
    Fisher's Exact Test for Count Data

data:  conteos
p-value = 0.589
alternative hypothesis: true odds ratio is greater than 1
95 percent confidence interval:
 0.6018802       Inf
sample estimates:
odds ratio
 0.9750673
```

Colecciones de grupos de genes

1. Cargamos paquete y datos

```
pacman::p_load(EnrichmentBrowser,Biobase,SummarizedExperiment)
data(gse21942,package="tamidata")
se21942 = makeSummarizedExperimentFromExpressionSet(gse21942)
```

Podemos ver que los genes vienen identificados por los correspondientes al fabricante AffyID o PROBEID.

```
head(rowData(se21942))
```

```
DataFrame with 6 rows and 7 columns
              PROBEID    ENTREZID         ENSEMBL      SYMBOL          GO
          <character> <character>     <character> <character> <character>
1007_s_at   1007_s_at         780 ENSG00000204580        DDR1  GO:0001558
1053_at       1053_at        5982 ENSG00000049541        RFC2  GO:0000722
117_at         117_at        3310 ENSG00000173110       HSPA6  GO:0005524
121_at         121_at        7849 ENSG00000125618        PAX8  GO:0000978
1255_g_at   1255_g_at        2978 ENSG00000048545      GUCA1A  GO:0000272
1294_at       1294_at        7318 ENSG00000182179        UBA7  GO:0004839
            EVIDENCE    ONTOLOGY
          <character> <character>
1007_s_at         IEA          BP
1053_at           TAS          BP
117_at            IEA          MF
121_at            IEA          MF
1255_g_at         IEA          BP
1294_at           IBA          MF
```

Y los metadatos o variables fenotípicas.

```
head(colData(se21942))
```

```
DataFrame with 6 rows and 40 columns
              Source.Name Comment..Sample_description.
              <character>                   <character>
GSM545846.CEL GSM545846 1       Gene expression data..
GSM545845.CEL GSM545845 1       Gene expression data..
GSM545844.CEL GSM545844 1       Gene expression data..
```

```
GSM545843.CEL GSM545843 1       Gene expression data..
GSM545842.CEL GSM545842 1       Gene expression data..
GSM545841.CEL GSM545841 1       Gene expression data..
              Comment..Sample_source_name. Characteristics.Organism.
                                 <character>              <character>
GSM545846.CEL       peripheral blood mon..             Homo sapiens
GSM545845.CEL       peripheral blood mon..             Homo sapiens
GSM545844.CEL       peripheral blood mon..             Homo sapiens
GSM545843.CEL       peripheral blood mon..             Homo sapiens
GSM545842.CEL       peripheral blood mon..             Homo sapiens
GSM545841.CEL       peripheral blood mon..             Homo sapiens
              Term.Source.REF  Term.Accession.Number Characteristics.cell.type.
                  <character>            <character>                <character>
GSM545846.CEL             EFO http://purl.org/obo/..     peripheral blood mon..
GSM545845.CEL             EFO http://purl.org/obo/..     peripheral blood mon..
GSM545844.CEL             EFO http://purl.org/obo/..     peripheral blood mon..
GSM545843.CEL             EFO http://purl.org/obo/..     peripheral blood mon..
GSM545842.CEL             EFO http://purl.org/obo/..     peripheral blood mon..
GSM545841.CEL             EFO http://purl.org/obo/..     peripheral blood mon..
              Characteristics.disease.state. Term.Source.REF.1
                                  <character>       <character>
GSM545846.CEL             multiple sclerosis               EFO
GSM545845.CEL             multiple sclerosis               EFO
GSM545844.CEL             multiple sclerosis               EFO
GSM545843.CEL             multiple sclerosis               EFO
GSM545842.CEL             multiple sclerosis               EFO
GSM545841.CEL             multiple sclerosis               EFO
              Term.Accession.Number.1 Protocol.REF Term.Source.REF.2
                            <character>  <character>         <logical>
GSM545846.CEL             EFO_0003885 P-GSE21942-2                NA
GSM545845.CEL             EFO_0003885 P-GSE21942-2                NA
GSM545844.CEL             EFO_0003885 P-GSE21942-2                NA
GSM545843.CEL             EFO_0003885 P-GSE21942-2                NA
GSM545842.CEL             EFO_0003885 P-GSE21942-2                NA
GSM545841.CEL             EFO_0003885 P-GSE21942-2                NA
              Protocol.REF.1 Term.Source.REF.3        Extract.Name
                 <character>         <logical>         <character>
GSM545846.CEL   P-GSE21942-3                NA GSM545846 extract 1
GSM545845.CEL   P-GSE21942-3                NA GSM545845 extract 1
GSM545844.CEL   P-GSE21942-3                NA GSM545844 extract 1
GSM545843.CEL   P-GSE21942-3                NA GSM545843 extract 1
GSM545842.CEL   P-GSE21942-3                NA GSM545842 extract 1
GSM545841.CEL   P-GSE21942-3                NA GSM545841 extract 1
```

```
              Material.Type Protocol.REF.2 Term.Source.REF.4
                <character>    <character>         <logical>
GSM545846.CEL     total RNA   P-GSE21942-4                NA
GSM545845.CEL     total RNA   P-GSE21942-4                NA
GSM545844.CEL     total RNA   P-GSE21942-4                NA
GSM545843.CEL     total RNA   P-GSE21942-4                NA
GSM545842.CEL     total RNA   P-GSE21942-4                NA
GSM545841.CEL     total RNA   P-GSE21942-4                NA
              Labeled.Extract.Name       Label Protocol.REF.3 Term.Source.REF.5
                       <character> <character>    <character>         <logical>
GSM545846.CEL       GSM545846 LE 1      biotin   P-GSE21942-5                NA
GSM545845.CEL       GSM545845 LE 1      biotin   P-GSE21942-5                NA
GSM545844.CEL       GSM545844 LE 1      biotin   P-GSE21942-5                NA
GSM545843.CEL       GSM545843 LE 1      biotin   P-GSE21942-5                NA
GSM545842.CEL       GSM545842 LE 1      biotin   P-GSE21942-5                NA
GSM545841.CEL       GSM545841 LE 1      biotin   P-GSE21942-5                NA
              Hybridization.Name Array.Design.REF Term.Source.REF.6
                     <character>      <character>         <logical>
GSM545846.CEL          GSM545846        A-AFFY-44                NA
GSM545845.CEL          GSM545845        A-AFFY-44                NA
GSM545844.CEL          GSM545844        A-AFFY-44                NA
GSM545843.CEL          GSM545843        A-AFFY-44                NA
GSM545842.CEL          GSM545842        A-AFFY-44                NA
GSM545841.CEL          GSM545841        A-AFFY-44                NA
              Protocol.REF.4 Term.Source.REF.7 Protocol.REF.5 Term.Source.REF.8
                 <character>         <logical>    <character>         <logical>
GSM545846.CEL   P-GSE21942-6                NA   P-GSE21942-7                NA
GSM545845.CEL   P-GSE21942-6                NA   P-GSE21942-7                NA
GSM545844.CEL   P-GSE21942-6                NA   P-GSE21942-7                NA
GSM545843.CEL   P-GSE21942-6                NA   P-GSE21942-7                NA
GSM545842.CEL   P-GSE21942-6                NA   P-GSE21942-7                NA
GSM545841.CEL   P-GSE21942-6                NA   P-GSE21942-7                NA
              Array.Data.File Comment..ArrayExpress.FTP.file. Protocol.REF.6
                  <character>                      <character>    <character>
GSM545846.CEL   GSM545846.CEL          ftp://ftp.ebi.ac.uk/..   P-GSE21942-1
GSM545845.CEL   GSM545845.CEL          ftp://ftp.ebi.ac.uk/..   P-GSE21942-1
GSM545844.CEL   GSM545844.CEL          ftp://ftp.ebi.ac.uk/..   P-GSE21942-1
GSM545843.CEL   GSM545843.CEL          ftp://ftp.ebi.ac.uk/..   P-GSE21942-1
GSM545842.CEL   GSM545842.CEL          ftp://ftp.ebi.ac.uk/..   P-GSE21942-1
GSM545841.CEL   GSM545841.CEL          ftp://ftp.ebi.ac.uk/..   P-GSE21942-1
              Term.Source.REF.9     Normalization.Name Derived.Array.Data.File
                      <logical>            <character>             <character>
GSM545846.CEL                NA GSM545846_sample_tab..  GSM545846_sample_tab..
```

```
GSM545845.CEL                NA GSM545845_sample_tab..  GSM545845_sample_tab..
GSM545844.CEL                NA GSM545844_sample_tab..  GSM545844_sample_tab..
GSM545843.CEL                NA GSM545843_sample_tab..  GSM545843_sample_tab..
GSM545842.CEL                NA GSM545842_sample_tab..  GSM545842_sample_tab..
GSM545841.CEL                NA GSM545841_sample_tab..  GSM545841_sample_tab..
             Comment..Derived.ArrayExpress.FTP.file.
                                        <character>
GSM545846.CEL                  ftp://ftp.ebi.ac.uk/..
GSM545845.CEL                  ftp://ftp.ebi.ac.uk/..
GSM545844.CEL                  ftp://ftp.ebi.ac.uk/..
GSM545843.CEL                  ftp://ftp.ebi.ac.uk/..
GSM545842.CEL                  ftp://ftp.ebi.ac.uk/..
GSM545841.CEL                  ftp://ftp.ebi.ac.uk/..
             FactorValue..DISEASE.STATE. Term.Source.REF.10
                                <factor>        <character>
GSM545846.CEL          multiple sclerosis                EFO
GSM545845.CEL          multiple sclerosis                EFO
GSM545844.CEL          multiple sclerosis                EFO
GSM545843.CEL          multiple sclerosis                EFO
GSM545842.CEL          multiple sclerosis                EFO
GSM545841.CEL          multiple sclerosis                EFO
             Term.Accession.Number.2     GROUP
                         <character> <numeric>
GSM545846.CEL             EFO_0003885         1
GSM545845.CEL             EFO_0003885         1
GSM545844.CEL             EFO_0003885         1
GSM545843.CEL             EFO_0003885         1
GSM545842.CEL             EFO_0003885         1
GSM545841.CEL             EFO_0003885         1
```

2. Preparamos el **ExpressionSet**.

```
se21942=probe2gene(se21942)
```

¿Qué hemos hecho?

Modificar los identificadores de los genes sustituyéndolos por los ENTREZID.

```
head(rowData(se21942))
```

```
DataFrame with 6 rows and 0 columns
```

Introducir una variable fenotípica **GROUP** con valores 0 y 1.

3. Análisis de expresión diferencial utilizando el procedimiento **Limma**.

```
se21942 = deAna(expr = se21942) ## t-test moderados
```

Podemos ver los resultados en **fData**.

```
head(rowData(se21942))
```

```
DataFrame with 6 rows and 4 columns
              FC limma.STAT       PVAL  ADJ.PVAL
       <numeric>  <numeric>  <numeric> <numeric>
780   0.18679961  2.8245963 0.00820538 0.0439322
5982 -0.20334770 -3.5661800 0.00119963 0.0104985
3310 -0.00376277 -0.0326465 0.97416583 0.9890259
7849  0.01222596  0.3640900 0.71826277 0.8519014
2978  0.05324267  1.3586019 0.18407863 0.3832102
7318  0.01396915  0.1890567 0.85128092 0.9258587
```

4. Grupos de genes con Gene Ontology.

```
hsaGO =  getGenesets("hsa",onto="BP")
```

¿Cómo son estos grupos? ¿A qué se refieren?

```
head(names(hsaGO))
```

```
[1] "GO:0000002_mitochondrial_genome_maintenance"
[2] "GO:0000003_reproduction"
[3] "GO:0000012_single_strand_break_repair"
[4] "GO:0000017_alpha-glucoside_transport"
[5] "GO:0000018_regulation_of_DNA_recombination"
[6] "GO:0000019_regulation_of_mitotic_recombination"
```

¿Qué genes, con código ENTREZID, forman el grupo 3?

```
hsaGO[[3]]
```

```
 [1] "1161"      "2074"      "3981"      "7141"      "7515"      "23411"
 [7] "54840"     "54840"     "55775"     "55775"     "55775"     "200558"
[13] "100133315"
```

```
names(hsaGO[3])
```

```
[1] "GO:0000012_single_strand_break_repair"
```

O bien con

```
hsaGO$"GO:0000012_single_strand_break_repair"
```

```
 [1] "1161"      "2074"      "3981"      "7141"      "7515"      "23411"
 [7] "54840"     "54840"     "55775"     "55775"     "55775"     "200558"
[13] "100133315"
```

5. Sobre representación con GO

Realizamos un análisis de sobre representación con el test de Fisher unilateral.

```
se21942.oraGO= sbea(method="ora", se=se21942, gs=hsaGO,
                    perm=0, alpha=0.05)
```

```
gsRanking(se21942.oraGO)
```

```
DataFrame with 376 rows and 4 columns
                 GENE.SET  NR.GENES NR.SIG.GENES      PVAL
               <character> <numeric>    <numeric> <numeric>
1   GO:0006120_mitochond..        41           24  4.45e-07
2   GO:0006974_cellular_..       244           85  2.69e-06
3   GO:0006886_intracell..       228           78  1.44e-05
4   GO:0006888_endoplasm..       108           43  2.22e-05
5   GO:0006396_RNA_proce..       100           40  3.73e-05
...                    ...       ...          ...       ...
372 GO:2001241_positive_..        10            5    0.0481
373 GO:0008654_phospholi..        23            9    0.0482
374 GO:0006478_peptidyl-..         2            2    0.0485
375 GO:0008089_anterogra..        30           11    0.0487
376 GO:0000165_MAPK_casc..       106           31    0.0497
```

Análisis de sobre-representación de tamidata::gse21942

1. Análisis de expresión diferencial

- Utiliza el procedimiento del test de la t moderado de **Limma**.

```
pacman::p_load(tami)
gse21942_deo = dema(x=gse21942,
                    y="FactorValue..DISEASE.STATE.",
                    test = rowtmod,correction = "BH",
                    fdr = .0001,foutput = "gse21942")
```

```
browseURL(glimpse(gse21942_deo))
```

```
Warning: replacing previous import 'utils::findMatches' by
'S4Vectors::findMatches' when loading 'AnnotationForge'
```

```
Registered S3 method overwritten by 'GGally':
  method from
  +.gg   ggplot2
```

2. Utilizamos un test de Fisher unilateral para realizar un análisis de sobre-representación.

```
gse21942.tidy = tidy(gse21942_deo)
sel0 = which(gse21942.tidy[,"adjp"] < .05)
set0 = gse21942.tidy[sel0,"ENTREZID"]
gse21942_gsao = ora(set0,gsc=hsaGO)
head(gse21942_gsao)
```

```
                                                    rawp       OR     OR_low
GO:0000002_mitochondrial_genome_maintenance    0.61520723 1.102853 0.05151511
GO:0000003_reproduction                        1.00000000 0.000000 0.00000000
GO:0000012_single_strand_break_repair          1.00000000 0.000000 0.00000000
GO:0000017_alpha-glucoside_transport           1.00000000 0.000000 0.00000000
GO:0000018_regulation_of_DNA_recombination     0.08272247 5.518878 0.73972582
GO:0000019_regulation_of_mitotic_recombination 1.00000000 0.000000 0.00000000
                                                    OR_up
GO:0000002_mitochondrial_genome_maintenance       Inf
GO:0000003_reproduction                           Inf
GO:0000012_single_strand_break_repair             Inf
```

```
GO:0000017_alpha-glucoside_transport             Inf
GO:0000018_regulation_of_DNA_recombination       Inf
GO:0000019_regulation_of_mitotic_recombination   Inf
```