Gene Ontology

Ставровская Елена

2015

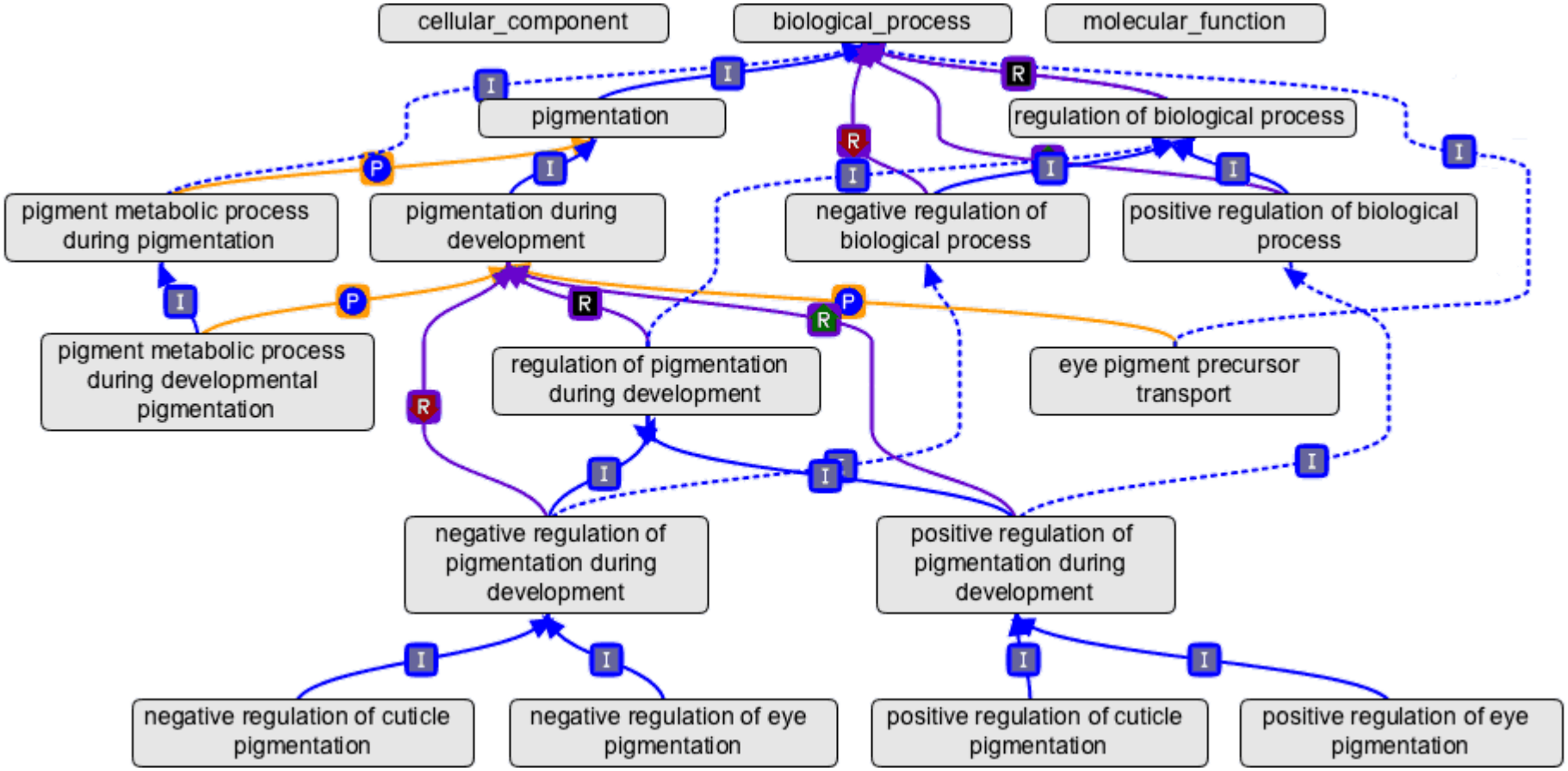
Gene Ontology или *GO*

- биоинформатический проект, посвященный созданию унифицированной терминологии для аннотации генов и генных продуктов всех биологических видов

Gene Ontology или *GO*

- **Молекулярные функции** (англ. molecular function) — специфическая активность генного продукта на молекулярном уровне, например, связывание углеводов илиАТФазная активность.
- **Биологические процессы** (англ. biological process) — сложные явления, необходимые для жизнедеятельности организмов и происходящие благодаря осуществлению последовательности молекулярных функций, например, митоз или биосинтез пуринов.
- **Клеточные компоненты** (англ. cellular component) — части клетки или внеклеточного пространства, где осуществляется функция генного продукта, например, ядроили рибосома.

Gene Ontology или GO



Статистически значимые Go-термы

- Мы хотим выявить те Go-термы, которые перепредставлены в одной выборке по сравнению с другой

Задача:

- Есть набор генов в геноме человека, рядом с которыми выявлен некоторый эффект (например, особое сочетание эпигенетических модификаций)
- Вопрос: что это за гены?
- В качестве референсной выборки (относительно которой считаем перепредставленность) рассматриваем весь геном

Нам понадобятся пакеты:

- org.Hs.eg.db – пакет с аннотацией генов в геноме человека
- GOFunction – пакет для вычисления перепредставленных термов GO

```
library(org.Hs.eg.db)
library(GOFunction)
```

Загружаем идентификаторы генов:

```
geneNames<-
   read.table("cg_no_fb_h3k27me3_h3k4me1.txt")
head(geneNames)
```

```
      V1
1 DUX4L8
2 DUX4L6
3 DUX4L5
4 DUX4L3
5 DUX4L2
6  CDC27
```

Нам нужен вектор идентификаторов,
а у нас таблица

```
geneNames<-geneNames[,1]
```

Проблема: для использования GoFunction требуется именовать все гены в Entrez gene ID

- Для этого нам и нужен org.Hs.eg.db!
- Что в нем есть?

```
> org.Hs.eg.db
```

Что там есть?

```
> org.Hs.eg.db
OrgDb object:
| DBSCHEMAVERSION: 2.1
| Db type: OrgDb
| Supporting package: AnnotationDbi
| DBSCHEMA: HUMAN_DB
| ORGANISM: Homo sapiens
| SPECIES: Human
| EGSOURCEDATE: 2015-Mar17
| EGSOURCENAME: Entrez Gene
| EGSOURCEURL: ftp://ftp.ncbi.nlm.nih.gov/gene/DATA
| CENTRALID: EG
| TAXID: 9606
| GOSOURCENAME: Gene Ontology
| GOSOURCEURL: ftp://ftp.geneontology.org/pub/go/godatabase/archive/latest-lite/
| GOSOURCEDATE: 20150314
| GOEGSOURCEDATE: 2015-Mar17
| GOEGSOURCENAME: Entrez Gene
| GOEGSOURCEURL: ftp://ftp.ncbi.nlm.nih.gov/gene/DATA
| KEGGSOURCENAME: KEGG GENOME
| KEGGSOURCEURL: ftp://ftp.genome.jp/pub/kegg/genomes
| KEGGSOURCEDATE: 2011-Mar15
| GPSOURCENAME: UCSC Genome Bioinformatics (Homo sapiens)
```

Что там есть?

Колонки базы данных, которые можно использовать в качестве ключей для поиска

```
> keytypes(org.Hs.eg.db)
 [1] "ENTREZID"     "PFAM"         "IPI"          "PROSITE"      "ACCNUM"
 [6] "ALIAS"        "ENZYME"       "MAP"          "PATH"         "PMID"
[11] "REFSEQ"       "SYMBOL"       "UNIGENE"      "ENSEMBL"      "ENSEMBLPROT"
[16] "ENSEMBLTRANS" "GENENAME"     "UNIPROT"      "GO"           "EVIDENCE"
[21] "ONTOLOGY"     "GOALL"        "EVIDENCEALL"  "ONTOLOGYALL"  "OMIM"
[26] "UCSCKG"
> columns(org.Hs.eg.db)
 [1] "ENTREZID"     "PFAM"         "IPI"          "PROSITE"      "ACCNUM"
 [6] "ALIAS"        "CHR"          "CHRLOC"       "CHRLOCEND"    "ENZYME"
[11] "MAP"          "PATH"         "PMID"         "REFSEQ"       "SYMBOL"
[16] "UNIGENE"      "ENSEMBL"      "ENSEMBLPROT"  "ENSEMBLTRANS" "GENENAME"
[21] "UNIPROT"      "GO"           "EVIDENCE"     "ONTOLOGY"     "GOALL"
[26] "EVIDENCEALL"  "ONTOLOGYALL"  "OMIM"         "UCSCKG"
```

Все доступные колонки базы данных

Переименовываем

```
geneIDs <- select(    org.Hs.eg.db,
keys=as.character(geneNames),
columns=c('ENTREZID'),    keytype='SYMBOL'  )
```

Entrez gene ID

наш тип идентификатора

```
> head(geneIDs)
  SYMBOL ENTREZID
1 DUX4L8    26583
2 DUX4L6   653544
3 DUX4L5   653545
4 DUX4L3   653548
5 DUX4L2   728410
6  CDC27      996
```

опять таблица, нужно вынуть столбец с ENTREZID:

```
geneIDs <- geneIDs[,2]
```

Добываем референсную выборку генов

```
refgeneIDs <-keys(org.Hs.eg.db,
   keytype="ENTREZID")
```

Считаем перепредставленные термы GO:

```
sigTerm <- GOFunction(geneIDs, refgeneIDs)
```

(может занять некоторое время)

Смотрим результат

```
> sigTerm
          goid                                                                name refnum interestnum       pvalue      adjustp
125 GO:0006024                                        glycosaminoglycan biosynthetic process    109          19 3.714290e-05 4.140598e-02
1   GO:0007155                                                          cell adhesion   1365         168 0.000000e+00 0.000000e+00
2   GO:0007156                homophilic cell adhesion via plasma membrane adhesion molecules    142          44 0.000000e+00 0.000000e+00
30  GO:0007167                               enzyme linked receptor protein signaling pathway   1092         120 3.242666e-10 1.506185e-06
90  GO:0007188 adenylate cyclase-modulating G-protein coupled receptor signaling pathway    139          24 4.392640e-06 6.801119e-03
104 GO:0007200   phospholipase C-activating G-protein coupled receptor signaling pathway     68          15 1.371819e-05 1.838065e-02
4   GO:0007399                                                  nervous system development   2027         234 0.000000e+00 0.000000e+00
38  GO:0007610                                                                behavior    649          80 2.355169e-09 8.636461e-06
126 GO:0008360                                                   regulation of cell shape    128          21 3.787107e-05 4.188267e-02
33  GO:0009888                                                       tissue development   1703         166 1.145247e-09 4.835957e-06
111 GO:0010769                regulation of cell morphogenesis involved in differentiation    267          35 2.179144e-05 2.729195e-02
56  GO:0022603                             regulation of anatomical structure morphogenesis    796          88 6.895652e-08 1.715871e-04
78  GO:0030029                                           actin filament-based process    565          65 9.653966e-07 1.724680e-03
18  GO:0030030                                           cell projection organization   1184         136 7.027712e-13 5.440499e-09
100 GO:0030198                                      extracellular matrix organization    382          46 1.167561e-05 1.626960e-02
67  GO:0031346                          positive regulation of cell projection organization    214          34 3.744842e-07 7.773909e-04
81  GO:0034329                                                cell junction assembly    215          33 1.231165e-06 2.118013e-03
106 GO:0035249                                    synaptic transmission, glutamatergic     77          16 1.598241e-05 2.082637e-02
```