ISSN 1982-5390
Outubro / 2020

DOCUMENTOS

165

Tutorial para análise funcional a partir de estudos de associação genômica ampla e transcriptômicos utilizando o banco de dados MeSH (Medical Subject Headings) no programa R

Embrapa

ISSN 1982-5390
Outubro/2020

Empresa Brasileira de Pesquisa Agropecuária
Embrapa Pecuária Sul
Ministério da Agricultura, Pecuária e Abastecimento

DOCUMENTOS 165

Tutorial para análise funcional a partir de estudos de associação genômica ampla e transcriptômicos utilizando o banco de dados MeSH (Medical Subject Headings) no programa R

Bruna Pena Sollero
Priscila Grynberg

Embrapa Pecuária Sul
Bagé/RS
2020

Exemplares desta publicação podem ser adquiridos na:

Embrapa Pecuária Sul
Rodovia BR-153, Km 632,9
Vila Industrial, Zona Rural, C. Postal 242
CEP 96401-970, Bagé, RS
Fone: +55 (53) 3240-4650
Fax: +55 (53) 3240-4651 www.embrapa.br
www.embrapa.br/fale-conosco/sac

Supervisão editorial
Lisiane Brisolara

Revisão de texto
Felipe Rosa

Normalização bibliográfica
Graciela Olivella Oliveira

Tratamento das ilustrações
Daniela Garcia Collares

Projeto gráfico da coleção
Carlos Eduardo Felice Barbeiro

Editoração eletrônica
Daniela Garcia Collares

Ilustração da capa
Bruna Pena Sollero

1ª edição
Publicação digitalizada (2020)

Dados Internacionais de Catalogação na Publicação (CIP)
Embrapa Pecuária Sul

Sollero, Bruna Pena
Tutorial para análise funcional a partir de estudos de associação genômica ampla e transcriptômicos utilizando o banco de dados MeSH (Medical Subject Headings) no programa R / Bruna Pena Sollero, Priscila Grynberg. — Bagé: Embrapa Pecuária Sul, 2020.
PDF (43 p.).— (Documentos / Embrapa Pecuária Sul, ISSN 1982-5390 ; 165)

1. Genótipo. 2. Programa de computador. I. Grynberg, Priscila. II. Título. III. Série.

CDD 576.53

Graciela Olivella Oliveira (CRB 10/1434)

Autores

Bruna Pena Sollero

Zootecnista, doutora em Melhoramento animal, pesquisadora da Embrapa Pecuária Sul, Bagé, RS

Priscila Grynberg

Bióloga, doutora em Bioinformática, pesquisadora da Embrapa Recursos Genéticos e Biotecnologia, Brasília, DF

Apresentação

As publicações técnicas da Série Embrapa são importantes veículos de informação, destinada a produtores, técnicos, empresários do agronegócio, pesquisadores, estudantes e público em geral interessados nas tecnologias desenvolvidas pela Empresa e seus colaboradores.

A Embrapa Pecuária Sul utiliza este veículo para comunicar suas tecnologias produzidas, recomendações, práticas agrícolas e resultados de pesquisa e desenvolvimento, direcionando ao público interessado informações ligadas à produção de forrageiras e pastagens, bovinocultura de corte e leite e ovinocultura dos campos sulbrasileiros. É com satisfação que oferecemos mais esta obra, destacando recente trabalho desenvolvido pela Embrapa, em benefício à sustentabilidade da pecuária nacional.

Nos últimos anos, esforços multidisciplinares vêm associando diversos métodos computacionais, estatísticos e matemáticos para organizar, processar e, principalmente, analisar, por meio da lente biológica, dados em escala genômica levando em consideração as particularidades da espécie, do experimento e do contexto.

Este documento tem como objetivo apresentar um protocolo baseado na linguagem R para análises de genômica funcional, a fim de explorar as funcionalidades de genes identificados em estudos genômicos ou transcritômicos. As linhas de comando apresentadas podem ser aplicadas para buscar o sentido biológico dos dados através das análises funcionais como, por exemplo, determinar se há enriquecimento de processos biológicos, vias metabólicas ou interações entre genes a partir de uma lista de genes identificados com relevância no fenômeno biológico ou fenótipo em questão.

Ao longo da obra, por meio de exemplos didaticamente construídos, serão apresentados recursos para se acessar repositórios de dados (NCBI e Ensembl, entre outros) que permitem aos pesquisadores a obtenção e análise de dados genéticos de uma maneira altamente personalizada.

Esperamos, por intermédio desta publicação, promover o compartilhamento de conhecimentos sobre aplicação de análises funcionais no âmbito dos genomas e dos transcritomas sob a luz da bioinformática, permitindo a interpretação do(s) fenômeno(s) biológico(s) neles existentes.

Daniel Portella Montardo
Chefe Geral

Sumário

Introdução

Dados biológicos advindos do conhecimento genômico e/ou transcritômico são relativamente complexos em função da quantidade excessiva de informações. Os resultados que se obtêm de experimentos de GWAS (Genome-Wide Association Analysis), RNA-Seq, proteômica, epigenômica, dentre outros, geralmente são uma lista de marcadores ou genes (com um valor de expressão relativa) significativos, associados a uma análise estatística. A interpretação destes dados possibilita compreender os resultados sob a luz da biologia. Diversas abordagens podem ser aplicadas para buscar o sentido biológico dos dados, como determinar se há enriquecimento de funções biológicas, vias metabólicas ou interações entre genes a partir de uma lista de genes identificados com relevância em determinado fenômeno biológico ou fenótipo em questão.

Mais frequentemente, as categorias funcionais de uma lista de genes de interesse são identificadas usando esquemas de classificação como o do consórcio Gene Ontology (GO) – KEEG, MSigDB, GeneSigDB e Ingenuity Pathway Analysis (IPA) são outros exemplos. A descoberta de que a lista de genes regulados contém uma super-representação de uma ou mais funções biológicas pode identificar o mecanismo subjacente à diferença entre as condições, dando à análise de dados um 'fator de plausibilidade'. Essa evidência é associada a um valor estatístico que resume a probabilidade de a super-representação não ocorrer por acaso. A análise de enriquecimento funcional é, portanto, uma estratégia amplamente usada para legitimar os estudos ômicos, mas é frequentemente afetada por percepções de revisores e da aceitabilidade estatística das análises em questão. Fato é que, quanto mais complexos os dados, maior o número de possibilidades para extrair informações. Neste contexto, é pertinente lembrar que essas possibilidades surgem frente às diferentes abordagens analíticas e também devido a vieses experimentais, anteriores à análise funcional. A heterogeneidade entre essas abordagens e, em particular, a heterogeneidade nos resultados que elas produzem é desafiadora para os interpretadores, e consequentemente interfere nos estudos sequenciais (Duroux et al., 2020). Além disso, nem todos os genes podem ser detectados com a mesma confiabilidade, na medida em que

alguns genes nunca são identificados como sendo ‘regulados’ (o sinal nunca muda). Mas, esse é um viés de detecção que pode refletir aspectos da tecnologia de investigação transcriptômica ou da sequência do SNP/gene alvo amostrado numa lâmina ou chip sob investigação.

De forma geral, estudos de classificação funcional são representativos quando existe significância, os termos apresentam-se em consonância com o sentido biológico, e quando é possível identificar as vias do metabolismo e das proteínas que se sabe atuarem na expressão ou variação de uma característica estudada. Afinal, não é necessário realizar uma análise de enriquecimento funcional de um tecido específico simplesmente para ser informado de que se está estudando esse tecido (Timmons et al., 2015), bem como não deve-se esperar que todas as classificações funcionais de genes apresentadas serão vinculadas à característica ou doença investigada. Cabe aqui, inclusive, enfatizar a importância e necessidade de anotações genômicas em quantidade e qualidade para refinar progressivamente análises funcionais.

Segundo Levin et al. (2018) existem mais de 23.000 ferramentas associadas à bioinformática desenvolvidas desde 1990, e mais de 1.500 bancos de dados biológicos das mais diversas categorias, de acordo com Fernández-Suárez et al. (2014), disponíveis na Internet. Esses “navegadores do genoma” (e/ou recursos de anotações) também oferecem recursos para pesquisa avançada, como o Table Browser (UCSC) e o BioMart (Ensembl), que permitem aos pesquisadores a obtenção e análise de dados de uma maneira altamente personalizada. São a partir desses downloads de dados (sequências, variantes, genes, proteínas etc.) personalizados que os bioinformatas analisam predições computacionais em laboratório (in silico).

Considerando essas premissas, este tutorial tem o objetivo de apresentar um pipeline que pode ser usado em análises de genômica funcional, utilizando a espécie animal como exemplo, acessando repositórios como o NCBI (National Center for Biotechnology Information), Biomart, Santa Cruz (UCSC) Genome Browser e aplicando pacotes do programa R vinculados ao Bioconductor[1].

[1] Disponível em: http://www.bioconductor.org/help/workflows

O Bioconductor oferece ferramentas para a análise e compreensão de dados genômicos a partir de uma extensa lista de recursos para mapeamento entre sondas e microarranjos, marcadores SNPs e genes, além de vias metabólicas, ontologia gênica e anotações. Estas operações são facilitadas via recursos de anotações como o MeSH (Medical Subject Headings)

(Tsuyuzaki et al., 2015) que, por meio da interface R, pode acessar os repositórios. Portanto, MeSH será a principal ferramenta abordada neste tutorial, uma vez representando mais uma opção viável para explorar as funcionalidades de genes identificados em estudos genômicos ou transcriptômicos.

Pré-requisitos para o acompanhamento deste pipeline: noções de programação em R e de conceitos da biologia molecular e genética.

Acessando recursos de anotações

2.1 Map2NCBI

Permite ao usuário obter uma lista de características (features) de interesse, como gene, pseudogene, RNA, CDS e/ou UTR diretamente da base de dados NCBI para qualquer espécie disponível, através de download via ftp.

A ferramenta permite filtrar e salvar informações para uso futuro. Por meio da função “GetGeneList”, é possível fazer o download de determinada versão do genoma da espécie e filtrar os dados conforme especificado pelo usuário. Após essa função, nenhum outro acesso ao NCBI ou internet é necessário.

• Função: GetGeneList (Species,latest = TRUE, savefiles = TRUE, destfile)

Obs.: a versão da montagem do genoma deve sempre corresponder àquela informada no arquivo do mapa de marcadores.

• Para mais informações:

- Package ‘Map2NCBI’[2]
- (Atenção: Este pacote vem sendo constantemente atualizado)
- Recomenda-se a versão 3.6.1 do R e RStudio 1.2

```
# Instalando e carregando o pacote no R:
install.packages (“Map2NCBI”)
require (“Map2NCBI”)
#Representando um exemplo com pesquisa no genoma bovino
bostaurus<-GetGeneList(“Bos_taurus”,latest = TRUE, savefiles =TRUE,
       destfile=”/usuario/documentos”)
```

O argumento “latest” tem, como padrão, verdadeiro/TRUE. Esse termo indica se a versão da montagem do genoma (assembly) deve ser a mais recente criada para tal espécie a fim de se obter as características genômicas. Se definido como FALSE, o usuário será solicitado a identificar a versão requerida. O parâmetro “savefiles” indica se o usuário quer salvar o resultado (TRUE) ou não (FALSE). Se “savefiles = TRUE”, o local de destino deverá ser especificado (“destfile”) ou o comando não será executado.

```
# Opções de versões do genoma:
# Digite no console a opção e aguarde...
```

A versão do genoma pode ser encontrada assim que o comando “GetGeneList” começa a rodar:

[2] Disponível em: https://cran.r-project.org/web/packages/Map2NCBI/Map2NCBI.pdf

```
The only assembly information in this file is:
1 GCF_002263795.1

# Salve o arquivo (caso prefira salvar com nome e extensão específicos):
#.csv file:
write.table(bostaurus, file = “btaurus_versao_x.csv”, append =FALSE, quote = TRUE,
       sep= “ “, dec = “.”, row.names =TRUE,col.names = TRUE)
```

A função “MapMarkers” pode ser utilizada quando quiser mapear marcadores SNPs aos genes (ou outro recurso como RNA, UTR etc.) mais próximos de acordo com o mapa da versão do genoma baixado pela ferramenta.

Dica 1-básica: Para iniciar as análises, indique ao R em qual diretório você está trabalhando, utilizando o comando setwd(). Se estiver trabalhando com o RStudio, basta clicar em Session -> Set Working Directory -> Choose Directory. Todos os arquivos necessários para a análise serão salvos ou carregados a partir do seu diretório de trabalho indicado. Exemplo: setwd(“C:/meudrive/meudiretorio/”).

```
# Leitura lista de SNPs a serem mapeados aos genes
list_snps<-read.csv(“list.csv”, header=T, sep=””)
# Exemplo de formato do arquivo “list.csv”: 1 coluna com cabeçalho e nome dos
marcadores em cada linha
      SNP
      ARS-BFGL-NGS-86145
      Hapmap35887-SCAFFOLD5242_24725
      Hapmap44154-BTA-87046
```

```
rownames(list_snps)<-list_snps[,1]
# Leitura do mapa de SNPs
# Exemplo de formato do arquivo “map.txt”: 3 colunas com cabeçalho e info dos
marcadores em cada linha
       ID BTA_chr BTA_pos
       ARS-BFGL-NGS-86145 2 134355559
       Hapmap35887-SCAFFOLD5242_24725 2 123247319
       Hapmap44154-BTA-87046 5 14953446

map<-read.table(“map.txt”, sep=””, stringsAsFactors=FALSE, header=TRUE)
map<-map[,1:3]
names(map)<-c(“Marker”, “chromosome”, “position”)
snp_map<-map[map[,”Marker”] %in% rownames(list_snps),]
# Extração dos genes mapeados de acordo com os marcadores listados
# Função MapMarkers:
listgenes<-MapMarkers(bostaurus, snp_map, nAut=29, other = FALSE, savefiles = TRUE,
       destfile=”usuario/documentos”)
#Para o argumento “nAut”, insira o número de autossomos da espécie
# Salve o arquivo com as informações dos genes mapeados
write.table(listgenes, file = “listofgenesmaped.csv”, append = FALSE, quote = TRUE,
       sep = “ “, row.names = TRUE, col.names = TRUE)
```

O arquivo de saída contém várias colunas com informações de genes mapeados, quando houver. Por meio dessa função de mapeamento, apenas um gene é indicado para cada marcador, esteja este localizado dentro de um gene ou <= 2.500/ > 2.500 e <= 5.000 / > 5.000 e <= 2.500 / > 25.000 / > 1Mb pares de bases (pb) antes ou após um gene. Na coluna “Inside” estão essas informações da localização (distância) de cada marcador relativa a um gene. Por exemplo: “Yes, Inside_Gene” ou “Nearest_feature_is_>_25,000_bp_After_Feature” etc.

Caso seja interessante obter todos os possíveis genes encontrados a partir das anotações recuperadas de determinada versão do genoma, pode-se utilizar um loop preparado no R, por exemplo, para mapeá-los com base em uma lista de marcadores determinando a distância a ser considerada a partir de cada marcador (por exemplo 100.000 pb para ambos os lados). Esta distância se baseia no grau de desequilíbrio de ligação existente em tal população estudada, na densidade de marcadores disponíveis e/ou no tamanho da "janela cromossômica", ou QTL, considerada em estudos de associação (Tang et al., 2019). Um exemplo desta aplicação será apresentado abaixo, na sessão 3.1.

2.2 biomaRT

O pacote biomaRT também permite acessar grandes quantidades de dados de maneira uniforme (servidor Ensembl Biomart). Exemplos de bancos de dados acessíveis com esta ferramenta são Ensembl, Uniprot e HapMap. Esses principais bancos de dados dão aos usuários do biomaRt acesso direto a um conjunto diversificado de dados e permitem uma ampla variedade de consultas on-line utilizando o programa R.

- Exemplo de extração de dados genômicos da espécie bovina via biomaRT/ Ensembl.

```
# Instalando e carregando o pacote no R:

if (!requireNamespace("BiocManager", quietly = TRUE))

      install.packages("BiocManager")

BiocManager::install("biomaRt")

library(biomaRt)
```

```
#Verificando opções com o Ensembl Gene mart:
listEnsembl()
#Verificando opções de espécies disponíveis. O número 20 indica quantas espécies você
quer que o comando imprima:
head(listDatasets(useMart("ensembl")), 20)
# Acessando as informações: Exemplo genoma bovino
ensembl <- useMart("ENSEMBL_MART_ENSEMBL", dataset = "btaurus_gene_ensembl")
#Filtros:
head(listFilters(ensembl), 20)
myFilter <- "chromosome_name"
#Valores: Neste exemplo, os números dos cromossomos autossomais a serem incluídos na
pesquisa
myValues <- c(1:29)
# Atributos (impressão até a 30ª linha):
head(listAttributes(ensembl),30)
myAttributes <- c("ensembl_gene_id","chromosome_name", "start_position",
      "end_position", "external_gene_name")
## Consultar e baixar o genoma com base nos filtros, valores e atributos:
result<- getBM(attributes = myAttributes, filters = myFilter,values = myValues,
      mart = ensembl)
```

O arquivo "result" recupera as informações listadas em atributos, ou "attributes" para todo "ensemble_gene_id" (gene) encontrado entre os 29 cromossomos autossomais de bovinos.

Obs.: A documentação do biomaRT especifica todas as possibilidades de recuperação de dados genômicos. Assim, diferentes tipos de pesquisas de dados genômicos da espécie bovina podem ser realizadas via biomaRT/Ensembl. Seguem alguns exemplos:

a. Teste uma busca: quais são os “marts” disponíveis no Ensembl?

```
listEnsembl()
```

b. Teste uma busca para extração de informações dos SNPs para a versão ARS-UCD1.2 do genoma bovino:

```
variation <- useEnsembl(biomart="snps")
listDatasets(variation)
#Aparecerá na tela uma série de opções (dentre elas, "btaurus_snp" com a versão do
genoma: ARS-UCD1.2)
> listDatasets(variation)
                  dataset
1            btaurus_snp
2       btaurus_structvar
3             chircus_snp

variation <- useEnsembl(biomart="snps", dataset="btaurus_snp")
listFilters(variation)
listAttributes(variation)
```

c. Ou ainda, teste a extração de informações sobre um marcador SNP, especificando atributos e filtros:

```
rs43710845<-getBM(attributes=c("refsnp_id", "refsnp_source", "chr_name", "allele",
      "ensembl_gene_stable_id", "ensembl_transcript_stable_id"),
       filters = "snp_filter", values = "rs43710845", mart = variation)
```

MeSH

O MeSH (Medical Subject Headings) é uma coleção de vocabulário abrangente da área de ciências da vida, contendo mais de 25.000 anotações clínicas e biológicas para 70 espécies, incluindo categorias além daquelas tradicionalmente representadas no GO (Gene Ontology).

Os termos MeSH são designados por artigos indexados no PubMed pela "National Library of Medicine" e podem ser diretamente mapeados a genes para o desenvolvimento de anotações gênicas. Atualmente tem-se 19 categorias e o tamanho do seu vocabulário é aproximadamente o dobro do tamanho do GO. As categorias biológicas incluídas se enquadram em: "Fenômenos e Processos", "Processos Químicos e Drogas", "Anatomia", além de "Doenças"; o que fortalece as interpretações (médica/biológica) dos dados com a diversidade de temas abrangentes. A aplicação desta ferramenta vem sendo utilizada em diversos estudos com animais domésticos e com espécie vegetal (Morota et al., 2015, 2016; Beissinger; Morota, 2017; Mota et al., 2018; Oliveira Júnior et al., 2019; Campos et al., 2019).

A abordagem dos estudos de associação genômica ampla (GWAS) utiliza genótipos e fenótipos para predizer regiões do genoma responsáveis pela ocorrência de um caráter (Sollero et al., 2017). A partir da obtenção dos efeitos dos SNPs que flanqueiam as "janelas" cromossômicas, pode-se então aplicar ferramentas de análise funcional para identificar genes participantes de vias regulatórias, processos biológicos e propor novas hipóteses para mecanismos de ações gênicas atuantes. Assim, com base numa lista de SNPs identificados como mais informativos para predizer um fenótipo em questão,

o primeiro passo é mapear estes marcadores de volta ao genoma de referência para identificar os genes localizados dentro ou próximos aos marcadores.

Nesta seção do tutorial, será apresentado o exemplo de um pipeline para análises via MeSH com dados já publicados de um estudo de associação genômica ampla para identificar variantes genéticas que controlam características relacionadas à qualidade da carne bovina (Xia et al., 2016). Este trabalho destacou 20 SNPs considerados significativamente associados à coloração da gordura e da carne, à gordura intramuscular, área de olho de lombo e força de cisalhamento.

Esta ferramenta pode ser utilizada para outras espécies e fenótipos/condições.

• Requisitos mínimos (downloads):

- BiocManager[3]

- biomaRt[4]

- org.Bt.eg.db[5]

- MeSH.db[6]

- MeSH.Bta.eg.db[7]

Observação: Todos estes pacotes são constantemente atualizados. É importante atualizá-los antes de rodar as análises.

---

[3] Disponível em: https://www.bioconductor.org/install/
[4] Disponível em: https://bioconductor.org/packages/release/bioc/html/biomaRt.html/
[5] Disponível em: https://bioconductor.org/packages/release/data/annotation/html/org.Bt.eg.db.html
[6] Disponível em: https://bioconductor.org/packages/release/data/annotation/html/MeSH.db.html
[7] Disponível em: https://bioconductor.org/packages/release/data/annotation/html/MeSH.Bta.eg.db.html

Notas:

- Bioconductor packages: biomaRt, ensembldb, org.Bt.eg.db etc.
- Recurso de Anotação: OrgDb, GO.db, KEGG.db, biomaRt etc.
- gene2pubmed: dados advindos do pubmed (outros: gendoo e RBBH)
- keys: são os ID utilizados.
- keytypes: são os tipos de ID utilizados- “ENTREZ ID”, “SYMBOL” etc.

O pipeline de análise de enriquecimento funcional (ORA- Overepresentation Analysis) via MeSH, a partir de dados genômicos, apresenta as seguintes etapas principais: identificação de uma lista de genes (entrezID) de interesse, acesso à anotação do genoma, que será o “background”, e teste com significância estatística para identificar genes, vias e/ou funções metabólicas que estão enriquecidos.

3.1 Preparo da lista de genes de interesse

- Leitura dos dados (input)

Dica 1: Acesse o banco de dados contendo todo o genoma da espécie e baixe o arquivo (como apresentado na seção 2) de acordo com a versão do genoma utilizando por Xia et al. (2016): a versão UMD3.1[8].

Dica 2: Nos casos de estudos pós-GWA, onde se tem uma lista de marcadores (SNPs) identificados como mais informativos para determinada característica, faz-se a leitura da lista de SNPs e do mapa para montar um arquivo de identificação/“endereço” dos marcadores (como apresentado na seção 2.1).

[8] Disponível em: https://bovinegenome.elsiklab.missouri.edu/node/61[UMD3.1]

De acordo com o exemplo (Xia et al., 2016), pode-se montar a lista de SNPs com informações de cromossomo e posição no genoma (arquivo dat):

chr	pos
13	32478877
13	32483878
13	32486720
13	32493419
13	32506077
15	32598283
9	27847649
9	27855750
9	27844027
3	18828779
3	18832469
18	54363951
7	58022778
13	43934808
6	45645474
7	58022778
9	38244421
13	67984463
13	77522320
16	15470163

Com base na lista de SNPs (dat) e dados do genoma completo (bostaurus –seção 2.1), escolhe-se a amplitude de varredura (-/+ 100.000 pb = 100kb; no exemplo) para se obter a lista de todos os genes (list) mapeados e seus IDs (ENTREZID).

```
# Mapeando SNPs aos genes anotados
# bostaurus: nome do arquivo de anotações do genoma bovino, versão UMD3.1
# *dat: lista de marcadores
chr<-numeric()
bp<-numeric()
ENTREZID<-numeric()
m<-1
for(k in 1:nrow(dat))
  {
       sub.gene<-subset(bostaurus, chromosome == dat[k,1])
       for(j in 1:nrow(sub.gene))
       {
       if((sub.gene$start[j] - 100000) <= (dat[k,2]) &
          (sub.gene$end[j] + 100000) >= (dat[k,2]))
       {
          chr[m]<-dat[k,1]
                   bp[m]<-dat[k,2]
                   ENTREZID[m]<-sub.gene$ENTREZID[j]
                   m<-m+1
        }
    }
}
dat2<-data.frame(chr, bp, ENTREZID)

# Verifica e remove possíveis entrez.id duplicados
table(duplicated(dat2$ENTREZID))
list<-dat2[!duplicated(dat2$ENTREZID),]
# Verifica e remove genes sem informação de ENTREZID
table(is.na(list$ENTREZID))
list<-list[!is.na(list$ENTREZID),]
```

A lista total de genes (bostaurus), de onde se extraiu esta lista “seleta” de marcadores mais informativos, deve ser salva como “background”, ou referência, para as análises subsequentes. Importante ressaltar que para fins de descobertas funcionais (ontologia) a partir de SNPs (estudos genômicos), esta lista “de fundo” pode compreender toda a anotação da espécie existente.

```
# Prepara a lista de genes selecionados para a análise funcional
sel <- unlist(list$ENTREZID)
class(sel)
# Prepara a lista de genes universais para a análise funcional
univ <- unlist(listgenes$ENTREZID))
class(univ)
```

Lembrete: Para estudos de expressão gênica (seção 3.3), pode-se utilizar como lista universal (Background) a base de arranjos, no caso de microarranjos, ou mesmo todos os transcritos identificados, considerando análise de sequenciamento de RNA (Yu, 2018).

3.2 ORA MeSH – Dados genômicos

Para finalmente extrair significado biológico dessas listas de genes (neste caso, mapeados a partir de uma lista de SNPs), a análise de “over-representation” ou enriquecimento será realizada com base na Ontologia MeSH. Esta análise determina quais termos biológicos foram significativamente enriquecidos (e quais genes compõem cada termo biológico) entre os genes listados (selecionados) com base no genoma de referência (background). O grau de

enriquecimento é calculado como uma probabilidade que indica quais termos específicos foram detectados com mais frequência do que o esperado ao acaso (Morota et al., 2015). O teste hipergeométrico (ou teste exato de Fisher) é amplamente usado para calcular essas probabilidades. Os termos identificados como significativos são extraídos para a formação de redes, compreensão de vias gênicas ou metabólicas e/ou formulação de novas hipóteses.

Dentre as categorias utilizadas no MeSH, no contexto animal, cinco delas são mais importantes: A (Anatomia), B (Organismos), C (Doença), D (Drogas e Químicos) e G (Fenômeno e Processo).

A primeira etapa da análise ORA-MeSH é carregar os pacotes necessários. O pacote MeSH.db contém a relação entre MeSH IDs e MeSH terms. O MeSH.xxx.eg.db e o org.xx.eg.db são pacotes de anotação que contêm a correspondência entre MeSH IDs e Entrez Gene IDs. No caso da espécie bovina, será utilizado MeSH.bta.eg.db e org.Bt.eg.db, mas pode-se citar outros exemplos, como:

MeSH_Mmu_eg_db: tabela de referência MeSH para *Mus musculus*
MeSH_Osa_eg_db: tabela de referência MeSH para *Oryza sativa*
MeSH_Zma_eg_db: tabela de referência MeSH para *Zea mays*
MeSH_Ssc_eg_db: tabela de referência MeSH para *Sus scrofa*
MeSH_Hsa_eg_db: tabela de referência MeSH para *Homo sapiens*

A correspondência entre Entrez Gene ID and MeSH IDs é gerada principalmente por três métodos: usando mineração textual (Gendoo), curadoria manual pelo NCBI (gene2-pubmed) e similaridade de sequência usando a pesquisa BLASTP (RBBH) (reciprocal BLAST best Hit). Gene2pubmed é utilizado em 15 “major” e 100 “minor” organismos, e faz a correspondência entre os IDs Entrez Gene e os NLM PubMed manualmente pelas equipes de curadoria do NCBI.

Curiosidade: para construir pacotes do tipo org.MeSH.XXX.db, focou-se nos organismos que satisfizeram três requisitos: 1) uso em pelo menos uma entre cinco ferramentas genômicas; 2) posse de um Entrez Gene ID, em vez de um Ensembl Gene ID; e 3) dados publicados/disponíveis em pelo menos 100 artigos. Finalmente, 120 organismos foram selecionados para estruturar a base de dados MeSH.

```
# Para instalação dos pacotes:
BiocManager :: install (“nome do pacote”)
# Para carregar os pacotes:
library(org.Bt.eg.db)
library(MeSH.db)
library(MeSH.Bta.eg.db)
```

Para referenciar a lista de genes universais e selecionados de acordo com a base de dados no “org.Bt.eg.db”, é necessário, primeiramente, acessar essa base:

```
# Acesso a base “org.Bt.eg.db”
key.symbol <- keys(org.Bt.eg.db, keytype = c(“SYMBOL”))
entrezUniverse = select(org.Bt.eg.db, as.character(key.symbol),columns = c(“ENTREZID”,
      “ENSEMBL”),keytype = “SYMBOL”)
head(entrezUniverse)

# Remove possíveis ENTREZID/SYMBOL duplicados
entrezUniverse2 <- entrezUniverse[!duplicated(entrezUniverse[,2]),]
entrezUniverse3 <- entrezUniverse2[!duplicated(entrezUniverse2[,1]),]
```

```
 # Genes Universais
genes.back = data.frame(univ)
colnames(genes.back) <- “ENTREZID”
### Relacionando a lista de genes em comum com o banco de dados org.Bt.eg.db e sua
lista universal (background genes):
table(genes.back$ENTREZID %in% entrezUniverse3$ENTREZID)
geneID.back <- merge(genes.back, entrezUniverse3, by =”ENTREZID”)
table(duplicated(geneID.back[,1]))
geneID2.back <- geneID.back[!duplicated(geneID.back[,1]),]

#Genes Selecionados
genes.sel = data.frame(sel)
colnames(genes.sel) <- “ENTREZID”
### Relacionando a lista de genes em comum com o banco de dados
org.Bt.eg.db e sua lista de genes selecionados:
table(genes.sel$ENTREZID %in% entrezUniverse3$ENTREZID)
geneID.sel  <- merge(genes.sel, entrezUniverse3, by =”ENTREZID”)
geneID2.sel <- geneID.sel[!duplicated(geneID.sig[,1]),]
ns = length(geneID2.sel[,1])
nt = length(geneID2.back[,1])
cat(paste(“Genes selecionados:”, ns, “ e Genes universais:”, nt -ns), “\n”)

univ= as.numeric(geneID.back[,1])

sel= as.numeric(geneID.sig[,1])
```

Nesta etapa, dois pacotes do R podem ser utilizados para realizar a análise de “over-representation” ou enriquecimento: meshr ou meshes.

O pacote meshr (Morota et al., 2015; Tsuyuzaki et al., 2015) aceita genes selecionados e universais como entrada e retorna os termos MeSH mais representativos e/ou significativos. É usado em conjunto com o pacote MeSH.db e um dos pacotes de anotação (por exemplo, MeSH.Bta.eg.db). Neste caso, é necessário gerar uma instância de parâmetro especificando os objetos contendo os genes “selecionados” e “universais”, o nome do pacote de anotação, a categoria MeSH, o banco de dados de correspondência entre Gene IDs e MeSH IDs, limiar de p-valor e a escolha de um método de correção para testes múltiplos, se for o caso.

A função de resumo (“*summary*”) retorna um objeto “data.frame” com informações sobre MeSH ID, valor P, termo MeSH, Entrez Gene ID e PubMed ID.

```
#ORA – meshr – Categoria D

library(meshr)

meshParams <- new("MeSHHyperGParams", geneIds = geneID2.sel[,1],

              universeGeneIds = geneID2.back[,1], annotation =

              "MeSH.Bta.eg.db", category = "D", database = "gene2pubmed",

               pvalueCutoff = 0.05, pAdjust = "none")


# Teste hipergeométrico

meshR <- meshHyperGTest(meshParams)

summary(meshR)
```

Para fins de interpretação, apresenta-se a seguir resultados das análises de enriquecimento geradas pelo pacote meshes (Yu, 2018), que é mais atualizado e, além de apresentar ferramentas de visualização para ajudar na interpretação dos resultados, permite também utilizar outra função (GSEA) quando for o caso de avaliação de dados de expressão gênica em larga escala. A função “enrichMeSH” também implementa o teste hipergeométrico para investigar associações de termos MeSH com genes “selecionados” e “universais”.

```
#ORA – meshes - Categoria A

library(meshes)

x <- enrichMeSH(sel, univ, MeSHDb = “MeSH.Bta.eg.db”, database=”gene2pubmed”, category

= “A”)

#Para traduzir entrezID em símbolo

x <- DOSE::setReadable(x, “org.Bt.eg.db”, “ENTREZID”)

#ORA – Categoria D

X2 <- enrichMeSH(sel, univ, MeSHDb = “MeSH.Bta.eg.db”, database=”gene2pubmed”, category

= “D”)

#Para traduzir entrezID em símbolo

X2 <- DOSE::setReadable(x2, “org.Bt.eg.db”, “ENTREZID”)
```

A partir do objeto “x” gerado (tabela 1), interpreta-se que, considerando um nível de significância alfa de 10%, o teste hipergeométrico para a categoria A (anatomia) identificou sete (7) termos de ontologia MeSH (Descrição) significativos dentro da lista selecionada comparado ao total de genes (universais); apontando quais genes se enquadraram em cada termo (geneID).

Tabela 1. Descrição de todos os termos MeSH (categoria A) e parâmetros gerados pelo teste hipergeométrico associados ao caracter.

ID	Descrição	Gene-Ratio	BgRatio	p-valor	p.adjust	q-valor	geneID
D007422	Intestines	1/19	27/25408	0,02000	0,10532	0,07675	S100A10
D010920	Placenta	2/19	344/25408	0,02685	0,10532	0,07675	S100A10/S100A11
D004622	Embryo, Mammalian	2/19	366/25408	0,03010	0,10532	0,07675	S100A10/S100A11
D007908	Lens, Crystalline	1/19	44/25408	0,03240	0,10532	0,07675	AKR1C4
D009504	Neutrophils	1/19	81/25408	0,05888	0,15310	0,11157	LBP
D050151	Subcutaneous Fat	1/19	107/25408	0,07707	0,15597	0,11366	LBP
D004717	Endometrium	1/19	117/25408	0,08398	0,15597	0,11366	AKR1C4
D001011	Aorta	1/19	166/25408	0,11713	0,18130	0,13212	S100A10
D019556	COS Cells	1/19	190/25408	0,13295	0,18130	0,13212	AKR1C4
D008168	Lung	1/19	200/25408	0,13946	0,18130	0,13212	AKR1C4
D004847	Epithelial Cells	1/19	260/25408	0,17757	0,20508	0,14945	LBP
D009865	Oocytes	1/19	279/25408	0,18930	0,20508	0,14945	SLC1A5
D008099	Liver	1/19	313/25408	0,20990	0,20990	0,15296	AKR1C4

Termos significativos (p-valor<0,05 ou p-valor<0,10).

A coluna GeneRatio da tabela 1 representa a razão entre o número de genes identificados na lista de genes selecionados representados dentro do termo e o número de todos os genes selecionados identificados com anotação nesta categoria. BgRatio representa a razão entre o número total de genes anotados e relacionados a determinado termo e todos os genes da lista universal (associados a todos os termos dentro de uma categoria).

O p-valor representa o nível de significância de cada termo representado pelos genes. O p-adjustado (p-adjust) ou q-valor é a correção do p-valor para testes de múltiplas hipóteses. Estes pacotes que utilizam o MeSH (meshr e meshes) possuem recursos para executar, ou não, a correção de teste múltiplo, permitindo escolher entre o método Benjamini-Hochberg, Q-value ou Bayes empírico (Morota et al., 2016). A prática usual é estabelecer um limiar para p-valor, ou q-valor, e identificar todos os termos que estão acima deste limiar. Neste caso, utilizamos um p-valor brando entre 0,05 e 0,10 para os dados ilustrativos.

Neste exemplo (Xia et al., 2016), foi possível associar na tabela 1 especialmente o termo nomeado “gordura subcutânea” (Subcutaneous fat-D050151), que representa o gene LBP- Proteína de ligação a lipopolissacarídeos, às características de qualidade de carne testadas. Contudo, o termo “Lente Cristalina” (Lens, Crystalline- D007908), que parece fora do contexto de formação de fibras musculares ou deposição de gordura, apresenta o gene AKR1C4 (Aldo-keto reductases 1C4), que está relacionado à obesidade (Ilozumba, 2018). Desta forma, a revisão de literatura é crucial para definir quais termos apontados pelo MeSH, e também por outras ferramentas para análise funcional in silico, podem elucidar mecanismos genéticos ou sugestões de algum possível gene candidato relacionado à característica estudada.

```
#Para salvar a tabela de resultados

write.csv(x, file="tabela_.csv", append = T, quote = T,sep=";",dec=".",

          col.names = T, row.names=F, fileEncoding = "", na="")


#Visualização gráfica

cnetplot(x, showCategory=20)

emapplot(y, showCategory=20, color="pvalue")
```

O gráfico gerado pela função “cnetplot” apresenta a conexão e sobreposição dos genes pelos termos indicados dentro da categoria A (Figura 1).

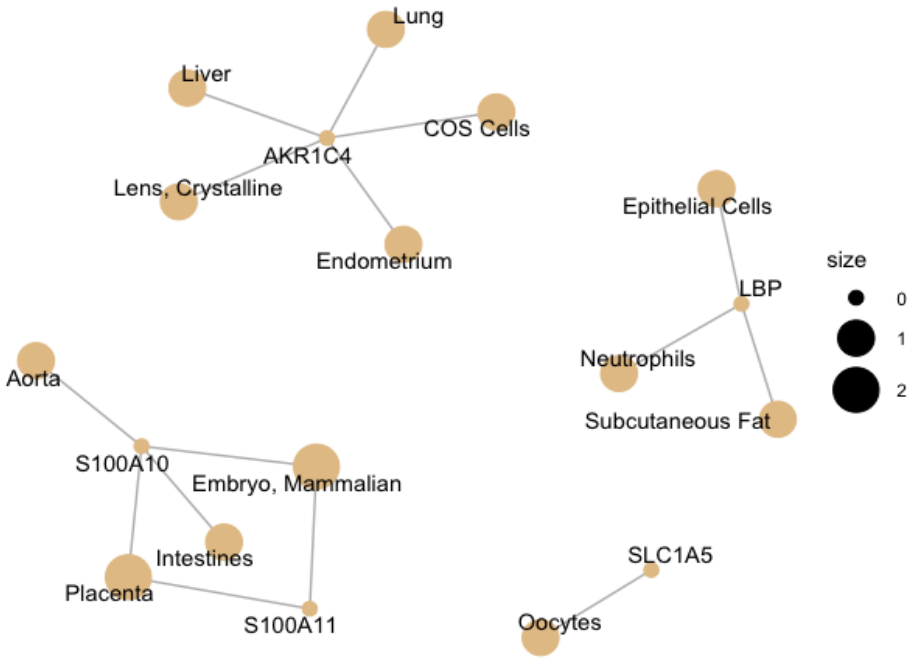


Figura 1. Gráfico representando os genes e termos interligados dentro da categoria A.

Para representar os resultados da categoria D, a função “emapplot” gerou um gráfico que representa todos os termos relacionados a esta categoria investigada pela ORA (significativos ou não) (Figura 2). Os conjuntos de genes que se sobrepõem mutualmente (indicados em mais de um termo) tendem a se agrupar, facilitando a identificação do módulo funcional (termo). Visualmente, o gradiente de cores identifica aqueles termos de maior ou menor grau de significância.

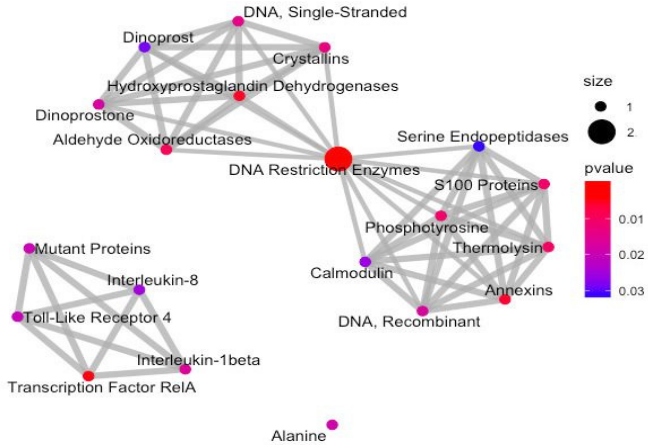


Figura 2. Gráfico representante dos termos MeSH elencados na categoria D, ilustrado por número de genes (size) e p-valor (gradiente de cores).

Outra fonte de dados utilizada para demonstrar esta ferramenta foi os 170 SNPs identificados como mais informativos para a característica idade à puberdade de fêmeas da raça Brahman, em um estudo de associação genômica ampla (Hawken et al., 2012). A partir do mapeamento dos genes (UMD3.1) ao redor destes marcadores utilizando este pipeline, foram extraídas informações importantes quanto às vias metabólicas e potenciais genes candidatos atuantes. Dentre os 250 genes mapeados (como apresentados no ítem 3.1), para a análise da categoria C (doença), foram encontrados os seguintes resultados:

Tabela 2. Descrição de todos os termos MeSH (categoria C) e parâmetros gerados pelo teste hipergeométrico associados a idade à puberdade de fêmeas da raça Brahman (Hawken et al., 2012).

ID	Descrição	GeneRatio	BgRatio	p-valor	p.adjust	q-valor	geneID
D010051	Ovarian Neoplasms	1/21	11/4380	0,05155	0,22627	0,22627	GPX3
D001424	Bacterial Infections	1/21	13/4380	0,06064	0,22627	0,22627	CD14
D006566	Herpesviridae Infections	1/21	17/4380	0,07859	0,22627	0,22627	CAP1
D012174	Retinitis Pigmentosa	1/21	17/4380	0,07859	0,22627	0,22627	RP1
D006984	Hypertrophy	1/21	18/4380	0,08302	0,22627	0,22627	IGF1R
D010048	Ovarian Cysts	1/21	19/4380	0,08744	0,22627	0,22627	IGF1R
D004716	Endometritis	1/21	20/4380	0,09183	0,22627	0,22627	CD14
D013203	Staphylococcal Infections	1/21	22/4380	0,10056	0,22627	0,22627	CSN3
D016643	Encephalopathy, Bovine Spongiform	1/21	25/4380	0,11350	0,22701	0,22701	CHST8
D004927	Escherichia coli Infections	1/21	39/4380	0,17159	0,25853	0,25853	CASP3
D009389	Neovascularization, Pathologic	1/21	40/4380	0,17560	0,25853	0,25853	SERPINF1
D020964	Embryo Loss	1/21	41/4380	0,17959	0,25853	0,25853	IGF1R
D014380	Tuberculosis, Bovine	1/21	46/4380	0,19926	0,25853	0,25853	CD14
D018457	Placenta, Retained	1/21	48/4380	0,20700	0,25853	0,25853	CD14
D004195	Disease Models, Animal	1/21	53/4380	0,22605	0,25853	0,25853	IGF1R
D007249	Inflammation	1/21	54/4380	0,22980	0,25853	0,25853	GPX3
D060467	Disease Resistance	1/21	111/4380	0,41742	0,44198	0,44198	CD14
D014777	Virus Diseases	1/21	260/4380	0,72421	0,72421	0,72421	MIR2461

Termos significativos (p-valor<0,10).

Percebe-se que apesar de não ter sido identificado p-valor menor que 0,05, a maioria dos termos apresentados estão relacionados à característica. Vale ressaltar que as funções deste pacote apresentam todas as associações identificadas entre a lista de genes selecionados e universal com a base de dados MeSH, independentemente do nível de significância encontrado. Entretanto, podem ocorrer casos em que nenhum gene selecionado é encontrado dentro de determinada categoria e termo, respectivamente, o que gera ausência de resultado.

Utilizando os mesmos dados para a categoria G (fenômeno e processo), vários outros termos significativos (p<0.05) foram indicados e os resultados foram representados utilizando o gráfico em barras (figura 3), que pode ser gerado pela função “barplot”:

```
barplot(x, showCategory=21, color="pvalue")
```

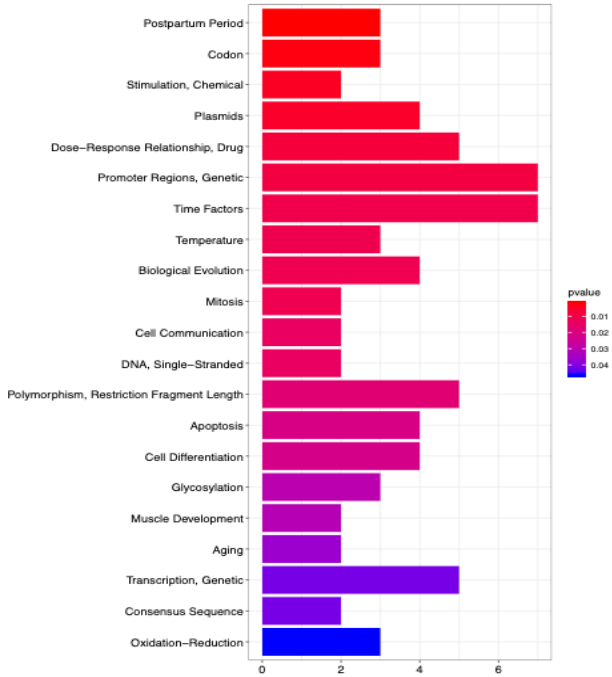


Figura 3. Gráfico representando os termos MeSH indicados pelo teste hipergeométrico em ordem crescente do p-valor associado e visualmente representado pelo gradiente de cor. O eixo x apresenta o número de genes relacionado a cada termo na categoria G.

3.3 GSEA MeSH – Dados transcriptômicos

Análises de enriquecimento funcional envolvem tanto “Over Representation Analysis” (ORA) como “Gene Set Enrichment Analysis” (GSEA). Ambos métodos identificam genes mais significativos/informativos/interessantes, sendo o primeiro, como visto no item 3.2, determinante da função biológica ou processo mais representado num experimento de um grupo de genes específicos em relação a todos os genes anotados ou considerados (lista universal) de uma determinada espécie. O segundo método a ser abordado aqui, relevante para estudos de expressão gênica, permite detectar situações em que todos os genes em um conjunto pré-definido mudam (em nível de expressão), ainda que minimamente, mas coordenadamente. Desta forma, tem-se um ranqueamento de genes com base em um contraste definido e nos valores de expressão diferencial (normalmente apresentado em “fold change”).

Para exemplificar uma GSEA, foram utilizados os dados de um experimento de sequenciamento de RNA, no qual bovinos mais resistentes e susceptíveis ao carrapato foram comparados (Moré et al., 2019). Utilizou-se a lista de genes diferencialmente expressos que representaram o contraste entre bovinos classificados como resistentes, antes e após a infestação artificial por carrapato, para exemplificar as funções. Normalmente, estas análises são feitas em todos os contrastes do experimento e os comparam para entender a dinâmica dos genes nas diferentes situações, por exemplo através de “heatmaps” (Gu et al., 2016).

O pacote meshes citado na seção anterior também permite a análise de “gene set enrichment” (GSEA), e conta com opções de visualização para ajudar na interpretação dos resultados transcriptômicos.

Obs.: As funcionalidades do pacote meshes podem ser aprimoradas por outros pacotes R, como ChIPseeker (Yu et al., 2015) e clusterProfiler (Yu et al., 2012), dependendo do experimento.

Para implementar a função GSEA, é preciso montar o arquivo de entrada “geneList”, o qual contém três recursos:

- vetor numérico: “fold change” (diferencial de expressão) ou outro tipo de variável numérica;

- vetor nomeado: cada variável numérica é nomeada pelo ID do gene correspondente

- vetor ordenado: a variável numérica deve estar em ordem decrescente

```
require(“meshes”)
# Leitura do arquivo
d<-read.csv(file=”RNAseq.csv”, sep=”,”, header=TRUE)# Moré et al. (2019)

#### Obter Entrez Gene ID a partir dos símbolos dos genes
library(clusterProfiler)
id<-clusterProfiler::bitr(d1$Ensembl.Gene.ID, “ENSEMBL”, “ENTREZID”, “org.Bt.eg.db”)
all<- merge(d,id, by.x=” Ensembl.Gene.ID “, by.y=”SYMBOL”, all=F)

# Seleciona as colunas necessárias para montar o arquivo geneList (ID/FC)
db<-all[,c(2,4)]
geneList = db[,1]
geneList<-as.numeric(geneList)
names(geneList) = as.character (db[,2])

# FC em ordem descrescente
geneList = sort(geneList, decreasing=T)
```

Vale lembrar que em alguns casos em que a lista de genes significativos é muito extensa, esta poderá ser composta de outra forma, que não por todos os genes identificados como significativamente (p-valor, q-valor, FDR etc.) e diferencialmente expressos no contraste, como apresentado aqui, mas por uma sub-lista contemplando somente aqueles significativos e com “fold chan-

ge" superior a um limiar estipulado.como significativamente (p-valor, q-valor, FDR etc.) e diferencialmente expressos no contraste, como apresentado aqui, mas por uma sub-lista contemplando somente aqueles significativos e com "fold change" superior a um limiar estipulado. Isto depende do número de genes significativos obtidos em determinado contraste, da complexidade da característica estudada e do cientista; alguns preferem considerar somente o q-valor para submeter às análises funcionais.

Nestas análises, será usado novamente o pacote MeSH.Bta.eg.db, que permite o mapeamento dos genes IDs de Bos taurus ao MeSH.

```
library(MeSH.Bta.eg.db)
library(org.Bt.eg.db)
library(MeSH.db)
```

Segue então, a função "gseMeSH" (GSEA) aplicada neste conjunto de dados (Moré et al., 2019) para duas categorias (C, D).

```
require(meshes)
# GSEA para categoria C
y <- gseMeSH(geneList, MeSHDb = "MeSH.Bta.eg.db", database="gene2pubmed", category =
     "C", nPerm=1000)

# Substitui geneID para símbolo do gene
y <- DOSE::setReadable(y, "org.Bt.eg.db", "ENTREZID")

# GSEA para categoria D
Y2 <- gseMeSH(geneList, MeSHDb = "MeSH.Bta.eg.db", database="gene2pubmed", category =
     "C", nPerm=1000)
# Substitui geneID para símbolo do gene
Y2 <- DOSE::setReadable(y2, "org.Bt.eg.db", "ENTREZID")
```

Tabela 3. Descrição de termos MeSH (categoria C), genes associados e parâmetros gerados pela função de enriquecimento.

ID	Descrição	set-Size	ES	NES	p-valor	p.adjust	q-valor	rank	leading_edge	core_enrichment
D007249	Inflammation	12	0,79253	2,25113	0,00161	0,00970	0,00340	109	tags=58%, list=5%, signal=56%	MMP1/CXCL8/IL6/LPO/HP/LBP/CCR7
D002418	Cattle Diseases	44	0,45898	1,84074	0,00410	0,01232	0,00432	265	tags=27%, list=11%, signal=25%	MMP1/CXCL8/IL6/PGLYRP1/HP/ARG2/GAL/MMP9/SOCS3/PAPPA/S100A9/TNNI3
D008414	Mastitis, Bovine	23	0,52203	1,76888	0,00771	0,01543	0,00541	641	tags=61%, list=27%, signal=45%	CXCL8/PGLYRP1/HP/LBP/CCR1/SLC2A3/CD4/LTF/NCF1/HIF1A/PRKDC/GAPDH/CD14/NOD1
D020022	Genetic Predisposition to Disease	16	0,52259	1,60143	0,03205	0,04807	0,01686	515	tags=44%, list=22%, signal=34%	CXCL8/OLR1/LBP/SOCS3/LTF/ITGB6/FGD3

Termos significativos (p-valor<0,10).

O escore de enriquecimento (enrichment score- ES) representa o grau que um grupo de genes é enriquecido no topo ou na base da lista total de genes ranqueada (todos os genes diferencialmente expressos no contraste). O escore é calculado para cada termo, de cima para baixo, e o valor é aumentado à medida em que se encontra um gene enriquecido, ou diminuído quando determinado gene, na sequência, não está associado ao termo. O teste hipergeométrico da lista ranqueada (“ranked list”) calcula p-valores utilizando genes a partir do topo, testando diferentes combinações até encontrar o menor p-valor (FDR, posteriormente) para cada termo. Por sua vez, o escore de enriquecimento normalizado (NES) indica a distribuição das categorias de ontologia gênica; se, em média, a regulação do termo é positiva ou negativa (regulação up/down).

O parâmetro denominado “leading edge” apresenta o subconjunto de genes mais enriquecidos dentro do termo: tags - indica a porcentagem de genes selecionados que contribuíram para a pontuação de enriquecimento; list - indica a posição na lista em que a pontuação de enriquecimento é máxima; signal - indica a força do sinal de enriquecimento.

Ainda a partir dos resultados da função “gseMeSH”, são apresentados dois tipos de gráficos que podem ser gerados.

```
require(enrichplot)


# Verifica ranqueamento e “running score” para cada “mesh term”- categoria C

gseaplot2(y, geneSetID = 1, title = y$Description[1])

# Teste também: gseaplot2(y, geneSetID = 2, title = y$Description[2])

## Heatplot - representação da Categoria D

heatplot(y2, foldChange=geneList)
```

A figura 4 representa o resultado de “ranking” e “running score” para os dois primeiros termos mais significativos da categoria (tabela 3). No “gseaplot”, o pico da curva de enriquecimento (distribuição) representada pela linha verde refere-se ao maior valor de enrichment score (ES) para o termo “inflamação” (Inflammation) e “Doenças em bovinos” (cattle diseases). Indica, também, o quão positiva (up) ou negativa (down) está a regulação da expressão dos genes em relação à lista de todos aqueles diferencialmente expressos (ranked list) dentro de cada termo.

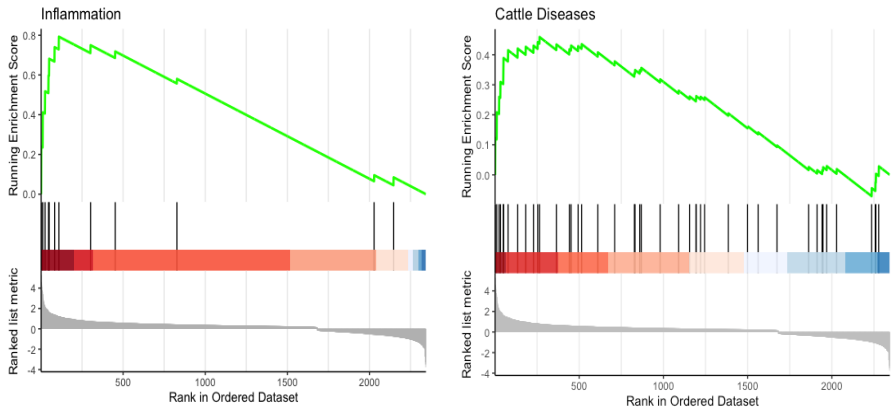


Figura 4. Exemplo de um “gseaplot”, apresentando a curva de enriquecimento sobreposta à lista ordenada (ranqueada) de todos os genes diferencialmente expressos no contraste (ranked list metric) para os termos (“Inflammation” e “cattle disease”) mais significativos na categoria doenças (C). As barras pretas indicam o posicionamento dos genes que contribuíram para o escore de enriquecimento em cada termo, com ênfase para aqueles dentro da graduação (do vermelho para o azul) representada pela cor vermelha.

Ainda, foi utilizada a ferramenta de visualização “heatplot” (Figura 5) para apresentar os resultados testados na categoria D. Este gráfico apresenta a distribuição da expressão dos genes de cada termo enriquecido (eixo y) para facilitar a interpretação. De acordo com o contraste testado, tem-se a distribuição dos fold changes (regulação up/down) dos genes representados no eixo x. Os valores de expressão associados a cada gene são distintos por um gradiente de cor (vermelho/azul).

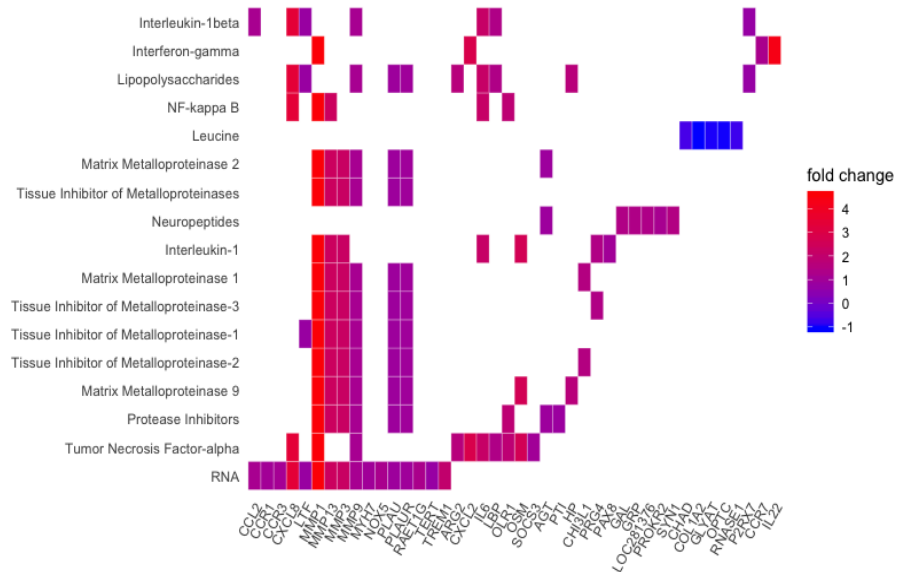


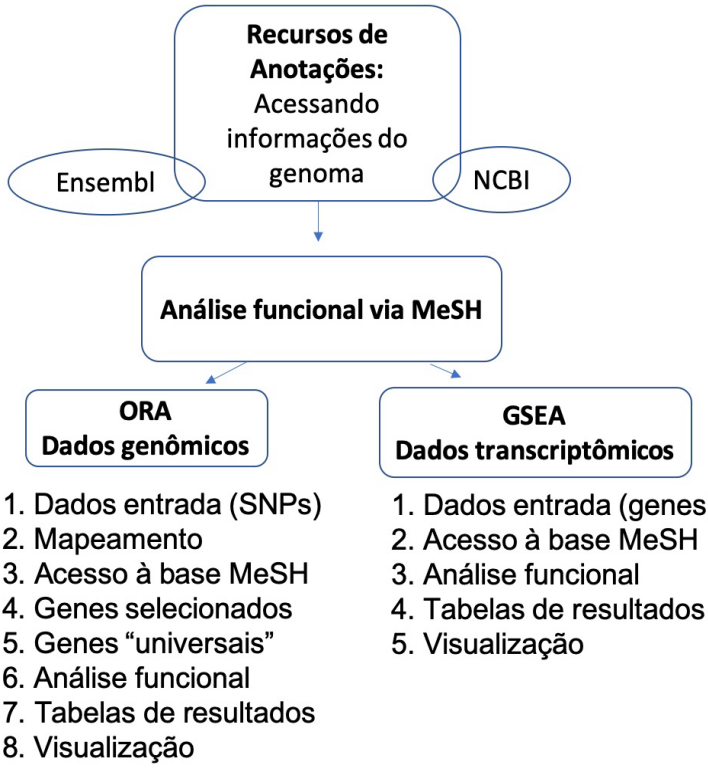
Figura 5. Heatplot dos resultados para a categoria D (drogas e químicos) representando os 21 termos significativos para a lista de genes diferencialmente expressos no contraste (após vs antes infestação).

Neste exemplo (Moré et al., 2019), para as categorias C e D, a maioria dos genes significativos (gene set) correspondentes ao contraste analisado teve maior expressão após o desafio.

Obs.: A função que permite a análise de “over-representation” (ORA) pode ser realizada com este tipo de dado (transcriptômico) também. Basta separar a lista de todos os genes presentes no contraste daqueles significativos (lista de genes universais vs lista de genes selecionados), como já apresentado na seção anterior para dados genômicos.

Anexo I

Diagrama apresentando o resumo do pipeline:



Referências

BEISSINGER, T. M.; MOROTA, G. Medical Subject Heading (MeSH) annotations illuminate maize genetics and evolution. **Plant Methods**, v. 13, n. 8, Feb. 2017. DOI: https://doi.org/10.1186/s13007-017-0159-5

CAMPOS, G. S.; SOLLERO, B. P.; REIMANN, F. A.; JUNQUEIRA, V. S.; CARDOSO, L. L.; YOKOO, M. J. I.; BOLIGON, A. A.; BRACCINI, J.; CARDOSO, F. F. Tag-SNP selection using Bayesian genomewide association study for growth traits in Hereford and Braford cattle. **Journal of Animal Breeding and Genetics**, 27 Nov. 2019. DOI: 10.1111/jbg.12458.

DUROUX D.; CLIMENTE-GONZÁLEZ H.; WIENBRANDT L.; VAN STEEN K. Network aggregation to enhance results derived from multiple analytics. In: IFIP WG 12.5 INTERNATIONAL CONFERENCE ON ARTIFICIAL INTELLIGENCE APPLICATIONS AND INNOVATIONS, 16., 2020, Neos Marmaras. **Proceedings**… Cham: Springer, 2020. Part I, p. 128-140. (IFIP advances in information and communication technology, 583). Ilias Maglogiannis; Lazaros Iliadis; Elias Pimenidis (ed.). AIAI 2020.

FERNÁNDEZ-SUÁREZ, X. M.; RIGDEN, D. J.; GALPERIN, M. Y. The 2014 Nucleic Acids Research Database Issue and an updated NAR online Molecular Biology Database Collection. **Nucleic Acids Research**, v. 42, n. D1, p. D1-D6, Jan. 2014. DOI: https://doi.org/10.1093/nar/gkt1282.

GU, Z.; EILS, R.; SCHLESNER, M. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. **Bioinformatics**, v. 32, n. 18, p. 2847-2849, Sept. 2016. DOI: 10.1093/bioinformatics/btw313.

HAWKEN, R. J.; ZHANG, Y. D.; FORTES, M. R. S.; COLLIS, E.; BARRIS, W. C.; CORBET, N. J.; WILLIAMS, P. J.; FORDYCE, G.; HOLROYD, R. G.; WALKLEY, J. R. W.; BARENDSE, W.; JOHNSTON, D. J.; PRAYAGA, K. C.; TIER, B.; REVERTER, A.; LEHNERT, S. A. Genome-wide association studies of female reproduction in tropically adapted beef. **Journal of Animal Science**, v. 90, n. 5, p. 1398-1410, May 2012. DOI: 10.2527/jas.2011-4410.

ILOZUMBA, M. N. **Impact of obesity and expression of obesity-related genes in the progression of prostate cancer in African American men**. 2018. 59 f. Dissertation (Master of Science in Public Health) - University of South Florida, Tampa, 2018.

LEVIN, C.; DYNOMANT, E.; GONZALEZ, B. A data-supported history of bioinformatics tools. **Arxiv**, n. 1807.06808, 2018. DOI: https://arxiv.org/abs/1807.06808.

MORÉ, D. D.; CARDOSO, F. F.; MUDADU, M. de A.; MALAGO JUNIOR, W.; GULIAS GOMES, C. C.; SOLLERO, B. P.; IBELLI, A. M. G.; COUTINHO, L. L.; REGITANO, L. C. de A. Network analysis uncovers putative genes affecting resistance to tick infestation in Braford cattle skin. **BMC Genomics**, v. 20, n. 998, p. 1-10, 2019. DOI: https://doi.org/10.1186/s12864-019-6360-3.

MOROTA, G.; BEISSINGER, T. M.; PEÑAGARICANO, F. MeSH-Informed Enrichment Analysis and MeSH-Guided Semantic Similarity Among Functional Terms and gene products in chicken. **G3**, v. 6, n. 8, p. 2447-2453, Aug. 2016. DOI: 10.1534/g3.116.031096.

MOROTA, G.; PENAGARICANO, F.; PETERSEN, J. L.; CIOBANU, D. C.; TSUYUZAKI, K.; NIKAIDO, I. An application of MeSH enrichment analysis in livestock. **Animal Genetics**, v. 46, n. 4, p. 381-387, Aug. 2015. DOI: 10.1111/age.12307 http://onlinelibrary.wiley.com/doi/10.1111/age.12307/abstract.

MOTA, R. R.; TEMPELMAN, R. J.; LOPES, P. S.; TEMPELMAN, R. J.; SOLLERO, B. P.; AGUILAR, I.; CARDOSO, F. F. Analyses of reaction norms reveal new chromosome regions associated with tick resistance in cattle. **Animal**, v. 12, n. 2, p. 205-214, Feb. 2018. DOI: https://doi.org/10.1017/S1751731117001562.

OLIVEIRA JÚNIOR, G. A.; SANTOS, D. J. A.; CESAR, A. S. M.; BOISON, S. A.; VENTURA, R. V.; PEREZ, B. C.; GARCIA, J. F.; FERRAZ, J. B. S.; GARRICK, D. J. Fine mapping of genomic regions associated with female fertility in Nellore beef cattle based on sequence variants from segregating sires. **Journal of Animal Science and Biotechnology**, v. 10, n. 97, Dec. 2019. DOI: https://doi.org/10.1186/s40104-019-0403-0

SOLLERO, B. P.; JUNQUEIRA, V. S.; GULIAS GOMES, C. C.; CAETANO, A. R.; CARDOSO, F. F. Tag SNP selection for prediction of tick resistance in Brazilian Braford and Hereford cattle breeds using Bayesian methods. **Genetics Selection Evolution**, v. 49, 15 June 2017. Article 49.

TANG, Z.; XU, J.; YIN, L.; YIN, D.; ZHU, M.; YU, M.; LI, X.; ZHAO, S.; LIU, X. Genome-wide association study reveals candidate genes for growth relevant traits in pigs. **Frontiers in Genetics**, v. 10, n. 302, Apr. 2019.

TIMMONS, J. A.; SZKOP, K. J.; GALLAGHER, I. J. Multiple sources of bias confound functional enrichment analysis of global -omics data. **Genome Biology**, v. 16, n. 186, 2015. DOI: https://doi.org/10.1186/s13059-015-0761-7.

TSUYUZAKI, K.; MOROTA, G.; ISHII, M.; NAKAZATO, T.; MIYAZAKI, S.; NIKAIDO, I. MeSH ORA framework: R/Bioconductor packages to support MeSH over-representation analysis. **BMC Bioinformatics**, v. 16, n. 45, Feb. 2015. DOI: 10.1186/s12859-015-0453-z, http://www.biomedcentral.com/1471-2105/16/45.

XIA, J.; QI, X.; WU, Y.; ZHU, B.; XU, L.; ZHANG, L.; GAO, X.; CHEN, Y.; LI,J.; GAO, H. Genome-wide association study identifies loci and candidate genes for meat quality traits in Simmental beef cattle. **Mammalian Genome**, v. 27, n. 5-6, p. 246-255, June 2016. DOI: 10.1007/s00335-016-9635-x.

YU, G. Using meshes for MeSH term enrichment and semantic analyses. **Bioinformatics**, v. 34, n. 21, p. 3766-3767, Nov. 2018.

YU, G.; WANG, L.-G.; HAN, Y.; HE, Q.-Y. clusterProfiler: an R package for comparing biological themes among gene clusters. **OMICS**, v. 16, n. 5, p. 284-287, May 2012. DOI: 10.1089/omi.2011.0118.

YU, G.; WANG, L.-G.; YAN, G.-R.; HE, Q.-Y. DOSE: An R/Bioconductor Package for Disease Ontology Semantic and Enrichment Analysis. **Bioinformatics**, v. 31, n. 4, p. 608-609, Feb. 2015.

CGPE 16139

MINISTÉRIO DA
AGRICULTURA, PECUÁRIA
E ABASTECIMENTO

PÁTRIA AMADA
BRASIL
GOVERNO FEDERAL