Paulo Pierry

Doutorando do programa de pós-graduação em Bioquímica do Instituto de Química (IQ) da Universidade de São Paulo. O título da minha tese é “Análise do transcritoma de cepas de *Xylella fastidiosa* isoladas de diferentes hospedeiros e busca por RNAs regulatórios”, orientado pela Profa. Dra. Aline Maria da Silva. Tenho interesse em aprender a linguagem R pois irei trabalhar com análises de RNA-Seq e colaborar em trabalhos de metabolômica e secretômica, os quais utilizam desta linguagem para análises estatísticas.

exec

Trabalho final

Contextualização

Em meu projeto de doutorado irei trabalhar com dados de expressão gênica obtidos a partir de técnicas de sequenciamento de RNA em larga escala, também conhecido como RNA-seq. Até o momento não tenho dados, o que provavelmente ocorrerá por volta de maio ou junho, período em que o equipamento de sequenciamento irá chegar ao departamento de bioquímica do IQ/USP.

A técnica consiste basicamente da extração das moléculas de RNAs totais de um organismo, construir uma “biblioteca” de transcritos (fragmentação dos transcritos e preparação da amostra para inserir no sequenciador) e realizar o sequenciamento com os reagentes necessários. O output do equipamento são milhões de reads (leituras, sequências de nucleotídeos) que correspondem aos fragmentos de transcritos sequenciados. Como no nosso caso, já temos todos os genomas dessas cepas sequenciados, para a montagem do transcritoma, bastaria utilizar softwares que alinhem essas sequências dos transcritos ao genoma de referência, montando o “quebra-cabeça” de fragmentos de nucleotídeos e observando quais genes foram expressos. Além de ser uma técnica qualitativa, ela também é uma técnica quantitativa. Se um determinado gene tiver ancorado em sua sequência um maior número de reads, significa que ele foi transcrito em maior quantidade e, portanto, se encontraria mais abundante na célula. Dessa forma, é possível quantificar a expressão gênica e avaliar quais genes estão sendo estimulados ou reprimidos em determinada condição de cultivo em comparação com outra. No meu projeto, a princípio, pretendemos comparar a expressão gênica da bactéria crescida em meio de cultura rico em nutrientes com a expressão gênica da bactéria crescida em meio mínimo, com baixa concentração de nutrientes, ou seja, um cenário mais próximo do que a bactéria encontra na natureza (condições nutricionais do fluido xilemático, local onde a bactéria se instala na planta).

Diversos softwares estão disponíveis para a análise de dados de RNA-seq, calculando a expressão gênica pelo número de reads para cada gene e analisando, por diversos métodos estatísticos, a expressão diferencial em diferentes condições. Um deles é o software TopHat/Cufflinks (Trapnell et al, 2012). Estes softwares utilizam a linguagem R para análises estatísticas e visualização gráfica dos resultados (pacote CummeRbund), o que me motivou a cursar esta disciplina. Procurando no CRAN pacotes voltados para análise de RNA-seq, encontrei outros interessantes, como o pacote AMAP.seq.

Além disso, em outro projeto em nosso laboratório, irá ser realizada a análise do metaboloma destas mesmas cepas desse fitopatógeno e nas mesmas condições da análise de expressão gênica. A análise de metabolômica diz respeito à detecção por espectrometria de massas de todos os compostos metabólicos produzidos pela célula em determinada condição. Desta forma, pretendemos integrar essas diferentes “ômicas” em prol de analisarmos quais genes expressos também estão sendo traduzidos em proteínas, as quais efetuariam sua função na célula. Estes dados de metabolômica, após utilização de um software específico para a deconvolução dos espectros de leitura do equipamento, fornece dados que também são avaliados utilizando a linguagem R.

Desculpem-me por este texto imenso, mas como sou de uma área completamente diferente, quis contextualizá-los de uma forma bem clara. Abaixo duas ideias para a função, embora não saiba se serão factíveis.

Plano A

Integrar dados de transcritoma e metobolômica

Como dito acima, pretendemos realizar a análise do transcritoma e metoboloma do nosso organismo de interesse. Dessa forma, pretendemos integrar os dois resultados. Minha ideia seria de partirmos de dataframes de genes e seus respectivos valores de expressão (calculados de acordo com a quantidade de reads para cada um deles) e de dataframes de metabólitos, também com seus respectivos valores de detecção no espectrômetro de massas. Isso seria feito para diferentes condições de cultivo com oito cepas desta bactéria. Pretendo analisar para a cepa “A” a relação entre os valores de expressão gênica e os metabólitos produzidos por suas células. Em seguida, comparar essa relação entre dois tipos de condições de cultivo diferentes para a mesma cepa. Por último, comparar a relação encontrada entre a expressão gênica e os metabólitos encontrada em cada um dos dois tratamentos feitos com cada uma das 8 cepas. Com isso, poderíamos obter um output de quais genes e metabólitos foram mais ou menos representados nas condições estudadas e em cada cepa.

Entretanto, infelizmente, não tenho um bom background em estatística, e com certeza precisarei da ajuda de vocês. Pesquisando sobre possíveis softwares na rede que poderiam fazer isso, encontrei o software IMPaLA (http://impala.molgen.mpg.de/) (Kamburov et al, 2011) que permite a entrada de dados de genes e/ou proteínas para comparação com dados de metabólitos e analisar as vias metabólicas mais representadas, a partir de bancos de dados de metabolismo como o KEGG. Foi nesse site que observei que pode ser feito um teste de Wilcoxon para análises de enriquecimento, embora não saiba muito bem se seria um teste adequado para esse caso.

Resumindo:

Input: dataframes de genes e seus dados de expressão e dataframes de metabólitos e seus dados de detecção.

Output: listas de genes e metabólitos e seus respectivos valores de representação na amostra.

Plano B

Heatmap

O plano neste caso é um pouco mais simples, pois pretenderia trabalhar somente com dados de RNA-seq. A ideia seria criar uma função que tivesse como input dataframes com uma lista de genes expressos e seus respectivos dados de expressão gênica em duas ou mais condições de cultivo da bactéria. Como output, obteríamos um gráfico de Heatmap (exemplo figura abaixo retirada de software da empresa do sequenciador, Illumina) com as linhas contendo cada gene observado e nas colunas os dados de expressão nas diferentes condições de cultivo. Os retângulos em vermelho representariam genes com expressão reprimidas em comparação com a(s) outra(s) condição(ões) e retângulos em verde, genes com expressão aumentada. Diferentes tons das cores representariam uma gradação em relação aos níveis de expressão dos genes.

Resumindo:

Input: dataframes de genes e seus dados de expressão.

Output: gráfico do tipo heatmap, mostrando os níveis de expressão gênica para cada gene observado.

Plano C

Análises de dados brutos de expressão gênica

Esta proposta surge como uma modificação da proposta do plano B. Esta última propunha utilizar de dados de expressão gênica já trabalhados e normalizados para a construção de um gráfico de HeatMap. Entretanto, não é trivial o processamento dos dados obtidos a partir de plataformas de sequenciamento em larga escala, as quais geram grandes quantidades de leituras de sequências gênicas. Dessa forma, decidi modificar a função com o intuito de partir de dados brutos de expressão gênica (dados de contagem de genes), trabalhá-los e normalizá-los e, em seguida gerar gráficos de HeatMap, gráficos de dispersão entre variáveis tomadas duas a duas, além de tabelas com outros dados que serão úteis nas análises a posteriori.

Input: dataframes com dados brutos de contagem de genes para cada condição.

Output: gráficos do tipo HeatMap, gráficos dispersão entre variáveis tomadas duas a duas, tabela com dados de RPKM (uma método de normalização dos dados), tabelas com dados de comparação de variáveis tomadas duas a duas.

Referências

Trapnell et al. (2012) Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nature Protocols, 7:3, pages 562-578.

Kamburov et al. (2011) Integrated pathway-level analysis of transcriptomics and metabolomics data with IMPaLA. Bioinformatics, 27:20, pages 2917-2918.

Comentários

Oi Paulo, não tenho muita (nenhuma!) afinidade com o que você quer fazer, mas me parece que seu plano B é mais interessante e geral que seu plano A. Roubando as palavras do Chalom, “a idéia de criar uma função em R é escrever um código que possa ser aproveitado em diversas situações: ou por você mesma, mais vezes, ou por outros pesquisadores com problemas semelhantes”.

Marília

Função final

HELP

```
RNA.seq                        pacote:unknown                                 R
Documentation

~~ Processa e analisa dados brutos de contagem de genes e compara diferentes
condições experimentais~~

Descrição:
Utiliza dados brutos de contagem de genes, oriundos de experimentos de
sequenciamento de RNA em larga escala, analisando-os e comparando diferentes
condições experimentais.
Uso:
RNA.seq(data, ncon)

Argumentos:
data        data frame ou matriz com os dados brutos de contagem de genes.
ncon        número de condições experimentais a serem analisadas.

Detalhes:
O data frame com os dados deve ser importado utilizando header = TRUE, sendo
que os genes devem estar indicados nas linhas e as condições experimentais
nas colunas.

Valor:
“HeatMaps.pdf”: retorna um arquivo do tipo .pdf com dois gráficos do tipo
HeatMap, sendo o primeiro comparando as condições experimentais baseando-se
nos valores de uma matriz de distância e o segundo comparando as condições
baseando-se nos seus respectivos valores normalizados de contagem de genes.

“plotMA.pdf”: retorna um arquivo do tipo .pdf com gráficos de dispersão
utilizando valores de comparação de dados de contagem normalizados entre
duas variáveis; todas as variáveis serão comparadas entre si, duas a duas.
```

```
“baseMeans.txt”: retorna um arquivo do tipo .txt contendo uma tabela com os valores de baseMeans e baseVar para cada gene analisado.

“RPKM.txt”: retorna um arquivo do tipo .txt contendo uma tabela com os valores de contagem de genes normalizados segundo o cálculo de RPKM (Reads Per Kilobase per Million mapped reads).

“binomtest[i,j].txt”: retorna um arquivo do tipo .txt contendo uma tabela com dados obtidos de comparações entre duas das variáveis analisadas. Será gerado um arquivo para cada comparação separadamente; ”i” e “j” correspondem às variáveis a serem comparadas.
Autor:
Paulo Marques Pierry
Doutorando do Programa de Pós-graduação em Ciências, Área Bioquímica, do IQ/USP.
pmpierry@gmail.com

Agradecimentos:
Agradeço aos monitores da disciplina Marilia Gaiarsa, Danilo Muniz e Hamanda Cavalheri pela fundamental ajuda durante o desenvolvimento desta função.

Referências:
Gentleman, R.C., Carey, V.J., Bates, D.M., Bolstad, B., Dettling M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J., Hornik, K., Hothorn, T., Huber, W., Iacus, S., Irizarry, R., Leisch, F., Li, C., Maechler, M., Rossini, A.J., Sawitzki, G., Smith, C., Smyth, G., Tierney, L., Yang, J.Y.H., Zhang, J. 2004. Bioconductor: open software development for computational biology and bioinformatics. Genome Biology, 5:R80.

Anders, S., Huber, W. 2010. Differential expression analysis for sequence count data. Genome Biology, 11:R106.

Exemplo:
# Baixar os arquivos "teste_vibrio_mean.csv" e "coord.CDS.csv" e salvá-los no diretório de trabalho que será usado no R.
# “teste_vibrio_mean.csv”: contagem de genes expressos em diferentes condições de crescimento da bactéria patogênica //Vibrio cholerae// tanto in vivo como in vitro.
# “gene.coord.csv”: coordenadas gênicas de todos os genes analisados da bactéria //Vibrio cholerae//.

data <- read.table("teste_vibrio_mean.csv",  header=T,  sep=";")
gene.coord <- read.table("coord.CDS.csv",  header=T,  sep=";")

#Analisar os dados de expressão gênica de cinco condições experimentais de crescimento da bactéria //Vibrio cholerae//.
RNA.seq(data, 5)

Referência dos dados do exemplo:
Mandlik, A., Livny, J., Robins, W.P., Ritchie, J.M., Mekalanos, J.J., Waldor, M.K. 2011. RNA-Seq-Based Monitoring of Infection-Linked Changes in
```

```
//Vibrio cholerae// Gene Expression. Cell Host & Microbe, 10:165-174.
```

CÓDIGO DA FUNÇÃO

```
RNA.seq <- function(data, ncon)
{
  # Instalar e abrir os pacotes necessários
  source("http://www.bioconductor.org/biocLite.R")
  biocLite("DESeq")
  library(DESeq)
  library(gplots)
  # Verificar a existência de pelo menos duas condições distintas para serem
comparadas
  if (ncon<2)
  {
  stop("Inserir dados de pelo menos 2 condições distintas!")
  }

  # Buscar NAs no seu conjunto de dados e excluir as observações nos quais
foram encontrados
  data <- na.omit(data)
  # Adicionar nomes das observações (linhas)
  rownames(data) <- paste("gene", 1:dim(data)[1], sep="")
  # Criar objeto com as condições (variáveis) analisadas
  conditions <- unlist(dimnames(data)[2])
  # Criar objeto com seus dados de contagem e suas condições analisadas
  count.data <- newCountDataSet(data, conditions)
  # Criar objeto com seus dados normalizados
  norm.read.count <- estimateSizeFactors(count.data)
  # Criar objeto para extrair valores de normalização dos dados
  size.factors <- sizeFactors(norm.read.count)
  # Criar objeto com valores de variância dos dados normalizados
  var.norm.read.count <- estimateDispersions(norm.read.count,
method="blind", sharingMode="fit-only")
  # Criar objetos com os dados de variância normalizados
  vst.data <- getVarianceStabilizedData(var.norm.read.count)
  # Criar objeto com valores de distância entre as condições (variáveis)
  dists <- dist(t(vst.data))
  # Criar um arquivo .pdf com dois gráficos de heatmap
  pdf("HeatMaps.pdf")
  # HeatMap dos valores de distância entre variáveis
  heatmap.2(as.matrix(dists), cexCol=1.0, cexRow=0.05, main="HeatMap
distâncias entre condições", xlab="Condições", ylab="Distâncias entre
variáveis")
  # HeatMap das variâncias normalizadas entre genes
  heatmap.2(as.matrix(vst.data), cexCol=1.0, cexRow=0.05, main="HeatMap
variâncias entre genes", xlab="Condições", ylab="Valores normalizados das
variâncias dos genes")
  dev.off()
```

```
  # Adicionar coluna no objeto gene.coord com os valores dos comprimentos
dos genes
  gene.coord$length <- abs(gene.coord$End.coordinate-
gene.coord$Start.coordinate)
  # Criar objeto com os cálculos dos valores de RPKM, outra forma de
normalizar seus dados, e salvar em arquivo .txt
  RPKM <- (data/gene.coord$length/1000)/(apply(data,2,sum)/1000000)
  write.table(RPKM, "RPKM.txt")
  # Criar objeto com o mesmo conjunto de dados mas, dessa vez,
transformando-os em matriz
  data.plot <- as.matrix(data)
  # Criar objeto com valores de normalização dos dados para uma matriz
  size.factors.plot <- estimateSizeFactorsForMatrix(data.plot)
  # Criar objeto com valores de baseMeans e variâncias e salvar em arquivo
.txt
  base.means <- getBaseMeansAndVariances(data.plot, size.factors.plot)
  write.table(base.means, "baseMeans.txt")
  # Criar um arquivo .pdf com todos os gráficos de dispersão
  pdf("plotMA.pdf", paper="a4")
  # Comparar todas as variáveis, tomadas duas a duas, usando a função
nbinomTest; plotar valores de cada comparação em gráficos de dispersão;
salvar arquivos .txt com a tabela de resultados para cada comparação.
  for(i in 1:(length(conditions)-1))
  {
    for(j in (i+1):length(conditions))
    {
      fold <- na.omit(nbinomTest(var.norm.read.count, conditions[i],
conditions[j]))
      plotMA(fold, fold$baseMean, fold$log2FoldChange, ylim=c(-10,10),
main=paste("plotMA[",i,",",j,"]",sep="",""))
      write.table(fold,file=paste("binomtest[",i,",",j,"]",sep="",".txt"))
    }
  }
    dev.off()
}
```

Arquivos

rna.seq.r

help_rna.seq.txt

Arquivos usados no exemplo

Dados brutos de contagem de genes:

teste_vibrio_mean.csv

Coordenadas gênicas de todos os genes:

coord.cds.csv