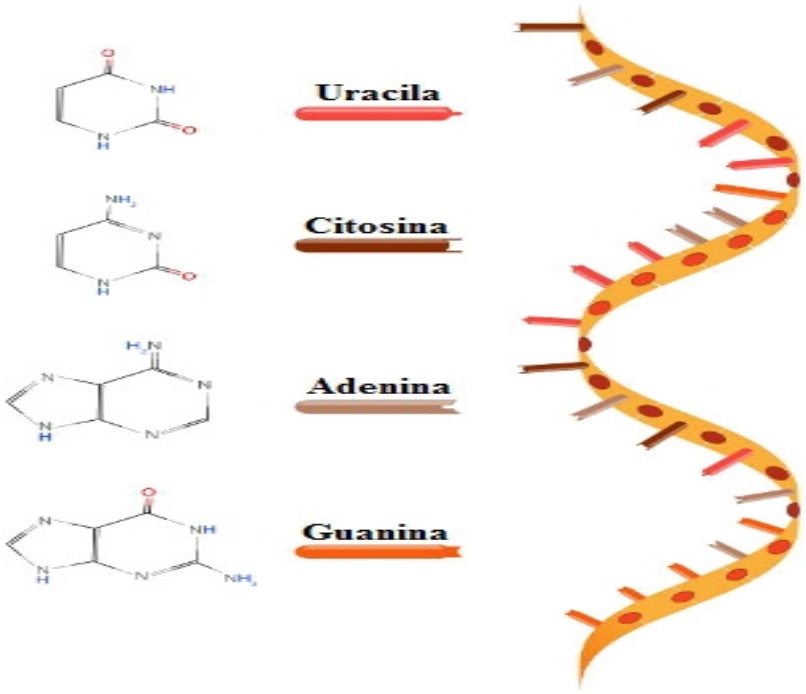
CURSO DE CAPACITAÇÂO EM BIOINFORMÁTICA

Introdução à Análise de Dados de RNA-seq

Kenny Pinheiro



Recursos utilizados:

Tophat:

http://ccb.jhu.edu/software/tophat/index.shtml

Cufflinks:

http://cole-trapnell-lab.github.io/cufflinks/

Samtools:

http://samtools.sourceforge.net/

Navegador genoma IGV:

https://software.broadinstitute.org/software/igv/download

DESeq2:

http://bioconductor.org/packages/release/bioc/html/DESeq2.html

E-mail: kennybiotec@gmail.com

Introdução

O sequenciamento da molécula de RNA (RNA-Seq), utiliza a capacidade dos sequenciadores de nova geração (NGS) para prover informações a respeito to transcriptoma de uma célula em um determinado momento. Um dos objetivos desta técnica é quantificar a expressão do conjunto de genes de uma determinada célula. Como a expressão gênica varia de acordo com as condições na qual o organismo se encontra, é possível utilizar as análises de RNA-Seq para avaliar o efeito de diferentes tratamentos na expressão de diferentes genes e desta forma, identificar genes diferencialmente expressos em diferentes condições experimentais para estudar o perfil transcricional.

A análise de expressão gênica diferencial em experimentos de RNA-Seq é relevante em experimentos controlados, nos quais se deseja verificar o efeito de um tratamento sobre a expressão gênica de tecidos e linhagens celulares de um organismo. Esses experimentos são delineados utilizando-se ao menos duas amostras (uma amostra controle e outra tratamento), podendo haver mais de uma réplica para o controle, tal como mais de uma réplica para o tratamento. Vale ressaltar que a utilização de réplicas aumenta a acurácia e garante a reprodutibilidade do experimento, uma vez que permite considerar a variação dentro dos tratamentos, além da variação entre tratamentos.

As etapas da análise podem ser sumarizadas da seguinte maneira:

Mapeamento: É necessário quantificar a expressão dos genes estudados e para isto realiza-se o mapeamento das leituras contra um genoma de referência e assim identifica-se o posicionamento das leituras em distintas regiões do genoma ou genes. Contudo, esse posicionamento muitas vezes não é exato, sendo necessário considerar não correspondências entre as leituras e a referência, ou seja, bases diferentes entre a leitura e o genoma de referência (*mismatches*). Para isto, o software que realiza o mapeamento deve ser capaz de considerar erros e variações estruturais bem como o splicing no caso de células eucarióticas.

Normalização: Essa etapa permite que comparações de nível de expressão dentro de amostras ou entre amostras sejam feitas de maneira mais acurada, devido a retirada de tendenciosidades inerentes ao tamanho dos dados. Para comparações dentro de amostras a normalização pode ser feita dividindo-se o número de leituras mapeadas pelo tamanho do gene, uma vez que genes maiores tendem a ter mais leituras mapeadas em relação a genes menores mesmo em um nível semelhante de expressão. Para comparações entre amostras, a normalização é feita pelo número de leituras presentes na biblioteca, uma vez que bibliotecas maiores tendem a ter mais leituras mapeadas no genoma.

Expressão diferencial: Um dos objetivos principais desta etapa é identificar diferenças no nível de expressão entre as condições experimentais analisadas. Portanto, aplicam-se testes estatísticos sobre os dados de contagem para cada biblioteca. Como dados de contagem são variáveis discretas (número de leituras mapeadas para cada gene), algumas distribuições de probabilidade discretas são mais adequadas para se extrair informações a partir dos dados de RNA-Seq, como a distribuição de Poisson e a distribuição Binomial Negativa. Entretanto, a variabilidade biológica não é bem captada através da distribuição de Poisson, pois esta possui um único parâmetro determinado pela sua média e isto acaba tornando a variância igual a sua média. Assim a distribuição de Poisson pode estar sujeita a uma alta taxa de falsos positivos em conjunto de dados com repetições biológicas. Essa variabilidade pode ser melhor explicada ao se utilizar a distribuição Binomial Negativa que usa um parâmetro de dispersão adicional no modelo.

De acordo com a sequência dessas etapas, primeiramente as leituras originadas do sequenciamento são mapeadas contra um genoma ou um transcriptoma de referência. Assim, uma

tabela de contagem de leituras mapeadas para cada gene pode ser gerada. Logo após, os dados são normalizados e testados estatisticamente para que a expressão diferencial seja inferida. Uma das formas de se normalizar tais dados de contagem é realizada através do cálculo do RPKM e FPKM que são idênticos na equação. A diferença é que utiliza-se RPKM para leituras single-end e FPKM para leituras paired-end. A fórmula abaixo mostra como é feito o cálculo do RPKM e FPKM:

$$FPKM = RPKM = 10^9 \frac{C}{N * L}$$

Onde: C = Número de leituras mapeadas em determinado gene.
N = Número total de leituras mapeadas.
L = Comprimento do gene (Número de bases).
10^9 = Fator de escala.

Após o cálculo dos valores de expressão em 2 condições distintas, torna-se possíve avaliar a expressão diferencial através do cálculo do fold-change (FC) que é uma medida que descreve o quanto uma variável muda de um valor inicial para um valor final. Em outras palavras, o fold-change descreve a proporção entre 2 valores e pode ser obtido pela fórmulas abaixo:

$$Fold\ change = \frac{Condição\ A}{Condição\ B} \quad \text{ou} \quad Fold\ change = \log_2\left(\frac{Condição\ A}{Condição\ B}\right)$$

Se os valor de FC calculado para um determinado gene for igual a 2, isso indica que aquele gene na condição A está 2 vezes mais expresso que na condição B.

Entretanto, os valores são frequentemente transformados para uma escala logarítmica e usados para análise e visualização dos resultados. O logaritmo na base 2 foi escolhido por ser de fácil interpretação, pois uma duplicação na escala original equivale a uma mudança de 1 na escala logarítmica. Geralmente defini-se como diferencialmente expressos todos os genes onde, na escala logarítmica, o FC é maior ou igual 1 ou menor ou igual a -1. Sendo que os valores maiores ou iguais a 1 são considerados hiperexpressos e os genes com valores de FC menores ou iguais a -1 são considerados hipoexpressos.

Desse modo, tendo-se em mãos a lista contendo os genes diferencialmente expressos, interpretações biológicas podem ser realizadas sobre os dados dos experimentos de RNA-Seq.

PRÁTICA.

O objetivo desta sessão prática é realizar algumas tarefas básicas na análise dos dados do RNA-seq. Começaremos a partir dos dados do RNA-seq alinhados com o zebrafish genoma usando Tophat. Vamos fazer a reconstrução do transcriptoma usando Cufflinks e comparar a expressão gênica entre duas condições diferentes, a fim de identificar genes diferencialmente expressos usando Cuffdiff.

Prepare o ambiente

UUtilizaremos um conjunto de dados derivado da sequência de mRNA de embriões de *Danio rerio* (Zebrafish) em duas fases de desenvolvimento diferentes. A sequenciação foi realizada na plataforma Illumina e gerou leituras de 76 pb - Paired-end. Devido às limitações de tempo da prática, usaremos apenas um subconjunto das leituras.

Os arquivos de dados estão contidos no subdiretório chamado dados e são os seguintes:

- 2cells_1.fastq e 2cells_2.fastq: estes arquivos são baseados em dados RNA-seq de um embrião zebrafish de 2 células.

- 6h_1.fastq e 6h_2.fastq: estes arquivos são baseados em dados RNA-seq do embrião de zebrafish **6h após a fertilização.**

Alinhamento

Existem numerosas ferramentas que realizam alinhamento e a escolha do alinhador deve ser feito cuidadosamente de acordo com as metas/necessidades da análise. Aqui nós vamos usar o Tophat, um alinhador ultra-rápido amplamente utilizado que realiza alinhamentos levando em consideração os exons e introns. Tophat é baseado no Bowtie (um alinhador que não leva em consideração o splicing) para realizar alinhamentos e utiliza um genoma de referência indexado. Para isso, faremos o índice apenas para um cromossomo do zebrafish.

Para isso, precisamos da sequência de cromossomo em formato fasta. Este é armazenado em um arquivo chamado Danio_rerio.Zv9.66.dna.fa

Criar o genoma indexada usando o seguinte comando:

```
bowtie2-build Danio_rerio.Zv9.66.dna.fa reference
```

Esta comando dá como saída 6 arquivos que compõem o índice. Estes arquivos que têm o prefixo reference.

Agora que o genoma está indexado podemos passar para o alinhamento real. *Tophat* tem um conjunto de parâmetros, para auxiliar no alinhamento. Para visualizá-los digite o comando abaixo:

```
tophat2 --help
```

O formato geral do comando tophat é:

```
tophat2 [ options] * <index_base> <reads_1> <reads_2>
```

Onde os dois últimos argumentos são as leituras fastq paired-end e o argumento anterior é o nome banco de dados do genoma indexado. Alguns parâmetros que vamos usar para executar *o Tophat* estão listados abaixo:

-g número máximo de multihits permitido. Leituras curtas são susceptíveis de mapear em mais de 1 local no genoma. Em RNA-seq que permitem a um número restrito de multihits, e neste caso nós pedimos para o Tophat permitir, no máximo, leituras que alinham em 2 loci diferentes.

-p número de threads permitido

- -library-type antes de realizar qualquer tipo de análise de RNA-seq você precisa saber algumas coisas sobre a preparação da biblioteca.
Foi feita usando um protocolo strand specific ou não?
Em nossos dados o protocolo NÂO foi strand specific.

-j melhorar o alinhamento fornecendo um arquivo anotado com as junções do splicing
As informações serão salvas em um arquivo chamado `transcript.junctions`.

-o especifica em que diretório deve ser salvo os resultados

Questões

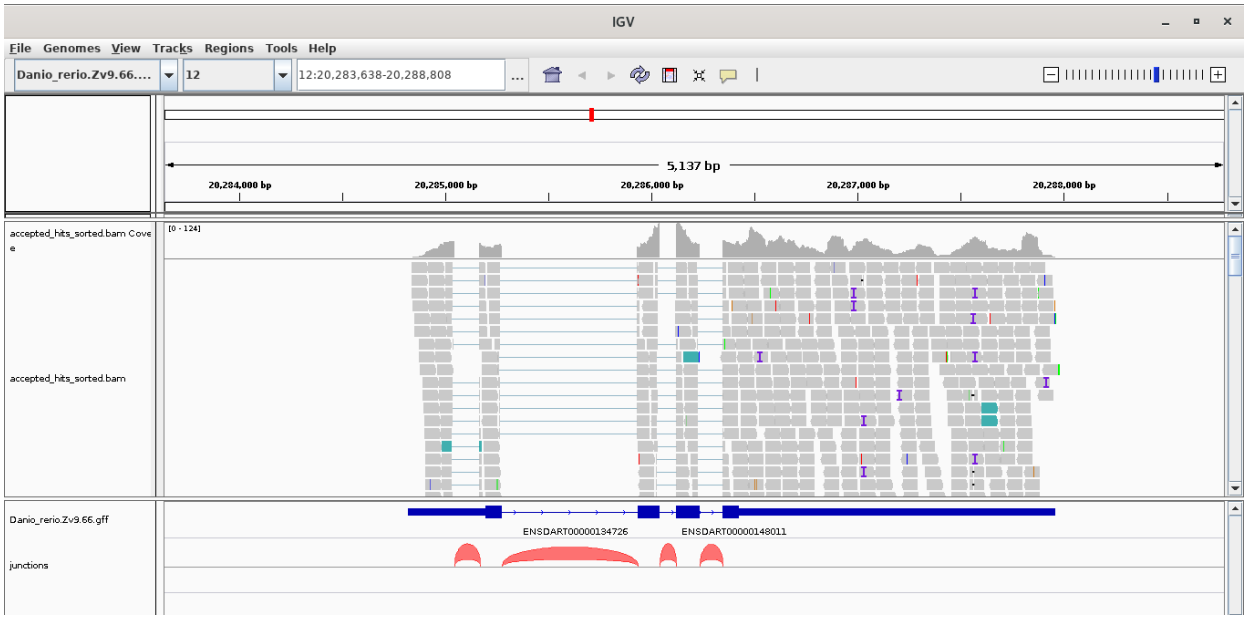
1. Tendo em conta que nós usamos o seguinte comando para alinhar o conjunto de dados 2cells:

```
tophat --solexa-quals -g 2 - p 4 --library-type fr-unstranded -j transcript.junctions -o tophat_saida/ reference reads_1.fastq reads_2.fastq
```

Qual é o comando para alinhar o conjunto de dados '6h'? Executar este comando no terminal.

Nota: Você vai ter que alterar os arquivos fastq de entrada e a pasta de saída. Se você não mudar a pasta de saída, estes resultados irão substituir aqueles feitos anteriormente para o conjunto de dados 2cells.

O *Tophat* reporta os alinhamentos em um arquivo BAM chamada accepted_hits.bam. Entre outros, ele também cria um junctions.bed arquivos que armazena as coordenadas das junções do splicing presente em seu conjunto de dados, uma vez que estes foram extraídos dos alinhamentos emendados. Agora vamos carregar o arquivo BAM e o arquivo de junções no genome browser *IGV* para visualizar os alinhamentos reportados pelo *Tophat*.



Questões

1. **Você consegue identificar as junções de emenda do arquivo BAM?** _____

_____

2. **As junções do splicing anotadas para CBY1 consiste com a anotação?** _____

_____

3. **Os genes anotados são todos expressos?** _____

_____

_____

Nós já sabemos que, a fim de carregar um arquivo BAM para IGV precisamos ter esse arquivo ordenado e indexado. Aqui está um lembrete dos comandos para executar estes passos através do samtools.

Ordenar o arquivo BAM usando:

```
samtools sort [ arquivo bam] -o [ Nome do arquivo de saída]
```

Índice do arquivo ordenado.

```
samtools index [ arquivo bam ordenada]
```

Expressão dos transcritos.

Há uma série de ferramentas que realizam reconstrução do transcriptoma e para esta finalidade estamos usando o *Cufflinks.* Ele também quantifica a expressão da isoforma em FPKM.

Para visualizar todos os parâmetros utiizados pelo Cufflinks:

```
cufflinks --help
```

O formato geral do comando cufflinks é:

```
cufflinks [ opções] *  <aligned_reads. (sam / bam) >
```

Onde a entrada é o arquivo de alinhamento (SAM ou formato BAM). Alguns dos parâmetros que vamos usar para executar *Cufflinks* estão listados abaixo:

-o diretório de saída

-G arquivo de anotação GTF ou GFF

-b detecção de viés e correção para melhorar a precisão da quantificação dos transcritos. Para isto, requer o arquivo de referência contra o qual alinhamos nossos dados.

-u um procedimento de estimativa inicial para medir com maior precisão o mapeamento de leituras em múltiplos locais do genoma.

--library-type fr-unstranded

-p número de threads.

```
cufflinks -o saida/ - G Danio_rerio.Zv9.66.gtf - p 8 -b Danio_rerio.Zv9.66.dna.fa -u --library-type fr-unstranded accepted_hits.bam
```

Questões

1. **Dado o comando anterior para 2cells, como você executaria esse comando para o outro conjunto de dados 6h? Não se esqueça de alterar a pasta de saída. Caso contrário, o segundo comando irá substituir os resultados da execução anterior.**

_____

_____

Dê uma olhada nas pastas de saída que foram criados. Os resultados são armazenados em 4 arquivos diferentes nomeados:

- **genes.fpkm_tracking**
- **isoforms.fpkm_tracking**
- **skipped.gtf**
- **transcripts.gtf**

Aqui está uma breve descrição desses arquivos:

- **genes.fpkm_tracking: contém os valores de nível de expressão de genes estimados.**
- **isoforms.fpkm_tracking: contém os valores de nível de expressão das isoformas.**

- **transcripts.gtf: Este arquivo contém GTF contém as as coordenadas e expressão das isoformas que foram detectadas (Assembled).**

Agora, vamos mudar o -G da opção do comando anterior. Em seu lugar, vamos utilizar o -g opção que diz para o Cufflinks usar a anotação como um guia e assim, permitindo a detecção de novos transcritos.

Questões

1.Na área caixa de pesquisa do IGV, digite ENSDART00000082297 para que o navegador possa ampliar o gene de interesse. Compare entre os transcritos já anotados e os montados pelo cufflinks usando o parâmetro -G e -g. Você observa alguma diferença?

Expressão diferencial

Uma das ferramentas que realiza a análise de expressão diferencial é *Cuffdiff.*

Usamos essa ferramenta para comparar entre duas condições. No nosso caso, queremos identificar genes que são diferencialmente expressos entre dois estágios de desenvolvimento; um embrião de 2 células e 6h pós-fertilização. O formato geral do comando é:

```
cuffdiff [ options] *  <transcripts.gtf>
      <Sample1_replicate1.sam [, ..., sample1_replicateM] >
      <Sample2_replicate1.sam [, ..., sample2_replicateM.sam] >
```

Onde a entrada inclui um arquivo transcripts.gtf, que é um arquivo de anotação do genoma de interesse, além das leituras alinhadas (SAM ou formato BAM) para as distintas condições.

Algumas das opções do Cuffdiff que vamos usar para executar o programa são:

-o diretório de saída
-L Labels para as diferentes condições
-T Sinaliza que as leituras são de um experimento de séries temporais
-b, -u, --library-type : Igual ao Cufflinks.

```
cuffdiff -o cuffdiff/ -L 2cells,6h -T -b Danio_rerio.Zv9.66.dna.fa -u --library-type fr-unstranded Danio_rerio.Zv9.66.gtf ZV9_2cells/accepted_hits.bam ZV9_6h/accepted_hits.bam
```

No comando acima assumimos que a pasta onde você armazenou os resultados do *Tophat* para o conjunto de dados 6h foi chamado ZV9_6h. Se este não é o caso, por favor mudar o comando anterior com o diretório correto, caso contrário você terá um erro.

Estamos interessados na expressão diferencial ao nível do gene. Os resultados são relatados pelo Cuffdiff no arquivo gene_exp.diff.

Olhe para as primeiras linhas do arquivo usando o seguinte comando no terminal:

```
head -n 20 gene_exp.diff
```

Nós gostaríamos de ver quais são os genes mais significativos diferencialmente expressos. Portanto, vamos ordenar o arquivo acima de acordo com o q-value (p-value corrigido para testes de múltiplos). O resultado será armazenado em um arquivo diferente, chamado: gene_exp_qval.sorted.diff.

```
sort - t$'\ t' -g -k 13 gene_exp.diff > gene_exp_qval.sorted.diff
```

Olhe novamente para as primeiras linhas do arquivo ordenado, digitando:

```
head -n 20 gene_exp_qval.sorted.diff
```

Copie o identificador Ensembl de um desses genes. Agora volte para o *IGV* navegador e cole-o na caixa de pesquisa.

Olhe para os dados alinhados brutos para os dois conjuntos de dados.

Questões

Você vê alguma diferença na cobertura do gene entre as duas condições que justifiquem este gene ter sido identificado como diferencialmente expresso? _____

Avaliação gráfica dos resultados

Primeiro, iremos abrir o software Rsudio. Em seguida iremos instalar o pacote cummeRbund utilizando os comandos abaixo:

```
if (!requireNamespace("BiocManager", quietly = TRUE))
   install.packages("BiocManager")

BiocManager::install("cummeRbund")
```

Agora, iremos setar a pasta onde os resultados estão salvos. Em seguida iremos digitar os seguintes comandos:

```
> library(cummeRbund)
> cuff<-readCufflinks()
> cuff
```

Boxplots da expressão de ambas as condições podem ser visualizados da seguinte forma:

```
>b<-csBoxplot(genes(cuff))
>b
>brep<-csBoxplot(genes(cuff),replicates=T)
>brep
```

Dendogramas:

```
> dend<-csDendro(genes(cuff))
> dend.rep←csDendro(genes(cuff),replicates=T)
```

Outras formas de avaliar e expressão diferencial.

Iremos agora, utilizar o software featureCounts para realizar a contagem das leituras no arquivo bam de alinhamento.

(https://sourceforge.net/projects/subread/files/subread-2.0.0/)

```
featureCounts -a Danio_rerio.Zv9.66.gtf -o featurecounts_results
tophat_saida_2cells/accepted_hits_sorted.bam tophat_saida_6h/accepted_hits_sorted.bam
```

Iremos agora, utilizar o pacote DESeq2 para realizar a expressão diferencial nos resultados obtidos pelo featureCounts. Siga os comandos abaixo:

```
data <- read.table("featureCounts_results", header=T, sep="\t", dec=".", row.names=1)

countdata <- read.table("featureCounts_results", header=T, sep="\t", dec=".", row.names=1)

countdata <- countdata[ ,6:ncol(countdata)]

condition <- factor(c(rep("normal", 2), rep("tumor", 2)))

coldata <- data.frame(row.names=colnames(countdata), condition)

dds <- DESeqDataSetFromMatrix(countData = countdata, colData = coldata, design = ~ condition)

#Expressão Diferencial:
dds <- DESeq(dds)
res <- results(dds)
res

summary(res)
```

PLOT MA:

```
plotMA(res, ylim=c(-2,2))
```

Identificando os genes no gráfico:

```
idx <- identify(res$baseMean, res$log2FoldChange)
rownames(res)[idx]
```

Volcano Plot:

```
par(mar=c(5,5,5,5), cex=1.0, cex.main=1.4, cex.axis=1.4, cex.lab=1.4)
topT <- as.data.frame(res)

#Adjusted P values (FDR Q values)
with(topT, plot(log2FoldChange, -log10(padj), pch=20, main="Volcano plot",
cex=1.0, xlab=bquote(~Log[2]~fold~change), ylab=bquote(~-log[10]~Q~value)))

with(subset(topT, padj<0.05 & abs(log2FoldChange)>2), points(log2FoldChange, -
log10(padj), pch=20, col="red", cex=0.5))

#with(subset(topT, padj<0.05 & abs(log2FoldChange)>2), text(log2FoldChange, -
log10(padj), labels=subset(rownames(topT), topT$padj<0.05 &
abs(topT$log2FoldChange)>2), cex=0.8, pos=3))

#Add lines for absolute FC>2 and P-value cut-off at FDR Q<0.05
abline(v=0, col="black", lty=3, lwd=1.0)
abline(v=-2, col="black", lty=4, lwd=2.0)
abline(v=2, col="black", lty=4, lwd=2.0)
abline(h=-log10(max(topT$pvalue[topT$padj<0.05], na.rm=TRUE)), col="black", lty=4,
lwd=2.0)
```

PARABÉNS! Você fez isso até o fim da prática. Nós esperamos que você tenha apreciado!

Não hesite em perguntar qualquer dúvida e não hesite em contactar-nos a qualquer momento (endereços de email na primeira página).