CURSO DE CURTA DURAÇÃO - 2017

BIOINFORMÁTICA

BioME – CENTRO MULTIUSUÁRIO DE BIOINFORMÁTICA - UFRN

NEXT GENERATION SEQUENCING

Análise de Dados de Sequenciadores de Segunda Geração

Prof. Dr. JORGE ESTEFANO SANTANA DE SOUZA

E-mail: jorge@imd.ufrn.br

RNA-seq I

BioME Bioinformatics Multidisciplinary Environment | Centro Multiusuário de Bioinformática

2Bio Instituto de Bioinformática e Biotecnologia

IMD Instituto Metrópole Digital | PPG BIOINFO UFRN | UFRN Universidade Federal do Rio Grande do Norte

Objetivo:

Utilizar as ferramentas básicas de alinhamento e montagem de transcriptoma para obter
os genes diferencialmente expressos entre duas amostras

Comandos Básicos de Linux:

Para trabalhar com nossos dados, vamos precisar saber alguns comandos básicos do Linux. Podem procurar mais informação no site:

http://wiki.ubuntu-br.org/ComandosBasicos

Ferramentas:

1- Linux.
2- WebServer.

3- **fastqc**
4- **tophat2**
5- **cufflinks**
6- **cuffmerge**
7- **cuffdiff**
8- **trimmomatic**

Inicial:

Login maquina local:
Login:
Senha:

Login no server:

ssh –p 4422 bif@10.7.5.38
Senha: bif0003

Inicial:

Pasta com dados iniciais:

/home/treinamento/NGS/RNAseq/

Pasta servidor WEB:

/home/bif/public_html/

Preparando os dados iniciais:

1) Vamos comecar pelo básico, certifique-se de que a pasta atual é :
/home/bif/aluno

Para isso digite o comando
pwd

2) Crie um diretório com o seu nome no seguinte formato:
mkdir aluno

3) Entre no diretório criado
cd aluno

4) Crie um link simbólico dos arquivos
ln -s /home/treinamento/NGS/RNAseq/adrenal_1.fastq .
ln -s /home/treinamento/NGS/RNAseq/adrenal_2.fastq .
ln -s /home/treinamento/NGS/RNAseq/brain_1.fastq .
ln -s /home/treinamento/NGS/RNAseq/brain_2.fastq .

Olhando:

5) Agora de uma olhada nos arquivos fastq. Utlize:

less -S adrenal_1.fastq
(para sair digite a letra Q)

Faça o mesmo para os outros arquivos.

Less -S adrenal_2.fastq
(para sair digite a letra Q)

Nota: mais informação do formato FASTQ em:

https://en.wikipedia.org/wiki/FASTQ_format

Olhando:

6) Seria interessante conseguir saber o numero de sequências totais nos arquivos. Temos algo para isso. O comando **wc** nome_do_arquivo (word count)

wc adrenal_2.fastq

O problema aqui é que o comando conta o número de linhas totais. Mas podemos utilizar uma união com o comando **grep** (para procurar apenas o cabeçalho das reads. E só depois fazer a contagem.) Vamos lá!

grep '@ERR' adrenal_1.fastq | wc

NEXT

Filtragem:

7) Vamos analisar as sequências de entrada.

Use o comando **fastqc** no terminal para checar a qualidade do sequenciamento.

8) No próximo passo vamos filtrar os arquivos fastq. Essa etapa é importante para a diminuição dos erros gerados durante o sequenciamento.

trimmomatic:

9) A ferramenta que iremos nessa etapa será o trimmomatic. O comando abaixo faz:
*Remoção de adaptadores
*Remoção de bases do início com baixa qualidade ou Ns
*Remoção de bases do fim com baixa qualidade ou Ns
*Percorre o read com uma janela de 4, removendo quando a qualidade média por base é menor do que 15
*Descarta reads com cumprimento menor do que 20 bases

ln -s /root/Trimmomatic-0.36/adapters/TruSeq3-PE-2.fa .

trimmomatic PE -threads 1
adrenal_1.fastq adrenal_2.fastq
adrenal_1_paired.fastq.gz adrenal_1_unpaired.fastq.gz
adrenal_2_paired.fastq.gz adrenal_2_unpaired.fastq.gz
ILLUMINACLIP:TruSeq3-PE-2.fa:2:30:10
LEADING:3
TRAILING:3
SLIDINGWINDOW:4:15
MINLEN:20

Repita o passo com os dados do cerebro

Descompacte:

Descompacte os arquivos que foram considerados filtrados e que mantiveram os pares após a filtragem.

gzip -d *_paired.fastq.gz

Genoma Referência:

10) Para realizar o mapeamento, primeiro temos que criar links simbólicos do genoma de referência e de nosso arquivo .gtf (*lembrese de estar no diretório "aluno");*

ln -s /home/databases/hg19/ .

ln -s /home/treinamento/NGS/RNAseq/gene19_annotation.gtf .

É possível olhar o conteúdo da pasta com 'ls -lhrt'

Mapeamento:

```
tophat -p 1 -G gene19_annotation.gtf
        -o thout_adrenal
        hg19/hg19
        adrenal_1_paired.fastq
        adrenal_2_paired.fastq
```

Faremos o mesmo com a amostra de cérebro

Montagem:

12) A montagem dos transcritos pode ser feita com a ferramenta Cufflinks

cufflinks -p 4 -o clout_adrenal thout_adrenal/accepted_hits.bam

Repita o passo com a amostra de cérebro

Expressão diferencial:

13) Quais genes estão diferencialmente expressos? para responder a pergunta, primeiro vamos criar um arquivo de anotação de referência com o cuffmerge:

nano assemblies.txt

nano é um editor de texto, e esse comando abre o editor criando o arquivo assembies.txt. Dentro desse arquivo vamos escrever:

./clout_adrenal/transcripts.gtf
./clout_brain/transcripts.gtf

(para sair pressionar Ctrl + x) Salvar pressionando Y

Expressão diferencial:

```
cuffmerge -g gene19_annotation.gtf -s hg19/hg19.fa -p 1 assemblies.txt


cuffdiff -o diff_out
        -b hg19/hg19.fa
        -p 1
        -L A,B
         -u  merged_asm/merged.gtf
         ./thout_adrenal/accepted_hits.bam
         ./thout_brain/accepted_hits.bam
```

Explorando os resultados:

14) Dentro da pasta diff_out encontramos vários resultados interessantes. Com **ls** podemos checar alguns desses arquivos gerados.

ls diff_out

Vamos manter o foco no arquivo gene_exp.diff .

more diff_out/gene_exp.diff

Agora vamos selecionar apenas aqueles genes dados como diferencialmente expressos pelos testes estatísticos do cuffdiff.

grep 'yes$' diff_out/gene_exp.diff

CURSO DE CURTA DURAÇÃO - 2017

BIOINFORMÁTICA

BioME – CENTRO MULTIUSUÁRIO DE BIOINFORMÁTICA - UFRN

Obrigado.

E-mail: jorge@imd.ufrn.br

Bioinformatics Multidisciplinary Environment | Centro Multiusuário de Bioinformática

2Bio Instituto de Bioinformática e Biotecnologia