

Aula prática: R-Peridot.

Encontrando expressão diferencial.

27/06/2017

Professor: Jorge Estefano Santana de Souza, jorge@imd.ufrn.br;

Monitores: Danilo Lopes Martins, danilolmartins@gmail.com;
Luan Pereira, luanpereira00@outlook.com.

Objetivos:

Utilizar ferramenta para a análise de expressão gênica a partir do resultado do miRDeep2.

Ferramentas:

- 1- Linux.
- 2- WebServer.
- 3- miRDeep2
- 4- R-Peridot

Comandos Básicos:

Durante a execução dos tutoriais necessitaremos saber alguns comandos básicos do Linux. Podem procurar mais informação no site:

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

Login Servidor:

Inicialmente vamos fazer o login no servidor, abra um terminal no linux e digite:

```
ssh -p 4422 bif@10.7.5.38
```

Irá pedir uma senha, digite:

```
bif0003
```

Regras para login no servidor:

Interno à UFRN:

```
ssh -p 4422 bif@10.7.5.38
```

Senha: bif0003

Externo à UFRN:

```
ssh -p 4422 bif@177.20.147.141
```

Senha: bif0003

Dados brutos (raw data):

Durante a execução dos tutoriais necessitaremos de alguns dados iniciais, em via de regra estarão disponíveis no diretório:

```
/home/treinamento/NGS/RNAseq/
```

Servidor WEB:

Como os trabalhos realizados no servidor são de difícil visualização, iremos necessitar de uma área web para facilitar nossa tarefa, todos os arquivos copiados para o diretório:

```
/home/bif/public_html/
```

Estarão disponíveis via navegador web em:

```
http://177.20.147.141/~bif/
```

Iniciando o Workflow:

1) Vamos começar pelo básico, certifique-se de que a pasta atual é:

```
/home/bif
```

Para isso digite o comando:

```
pwd
```

2) Crie um diretório contendo o seu nome, digite o comando:

```
mkdir SeuNome
```

3) Entre no diretório criado:

```
cd SeuNome
```

4) Crie um diretório chamado rna2:

```
mkdir rna2
```

5) Entre no diretório criado:

```
cd rna2
```

6) certifique-se de que a pasta atual é a correta:

```
pwd
```

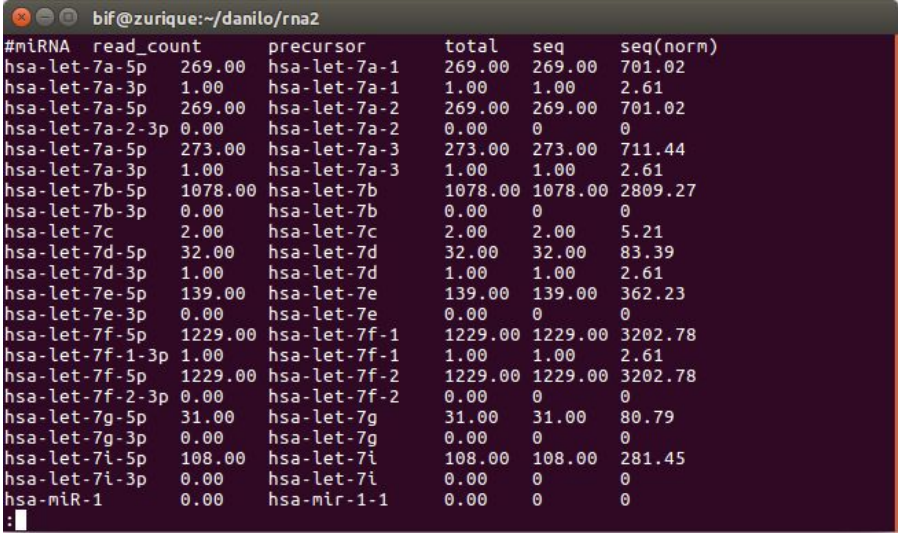
O diretório atual deve ser: **/home/bif/SeuNome/rna2**

7) Crie links simbólicos para os arquivos:

```
ln -s
/home/treinamento/NGS/biome/rna2/miRNAs_expressed_all_samples_07_07_
2017_t_23_25_06.csv .
ln -s
/home/treinamento/NGS/biome/rna2/miRNAs_expressed_all_samples_07_07_
2017_t_23_25_32.csv .
ln -s
/home/treinamento/NGS/biome/rna2/miRNAs_expressed_all_samples_07_07_
2017_t_23_25_49.csv .
ln -s
/home/treinamento/NGS/biome/rna2/miRNAs_expressed_all_samples_07_07_
2017_t_23_26_04.csv .
```

8) Agora vamos ver como esse arquivo csv resultado do miRDeep é.
comando:

```
less -S miRNAs_expressed_all_samples_07_07_2017_t_23_25_06.csv
```



#miRNA	read_count	precursor	total	seq	seq(norm)
hsa-let-7a-5p	269.00	hsa-let-7a-1	269.00	269.00	701.02
hsa-let-7a-3p	1.00	hsa-let-7a-1	1.00	1.00	2.61
hsa-let-7a-5p	269.00	hsa-let-7a-2	269.00	269.00	701.02
hsa-let-7a-2-3p	0.00	hsa-let-7a-2	0.00	0	0
hsa-let-7a-5p	273.00	hsa-let-7a-3	273.00	273.00	711.44
hsa-let-7a-3p	1.00	hsa-let-7a-3	1.00	1.00	2.61
hsa-let-7b-5p	1078.00	hsa-let-7b	1078.00	1078.00	2809.27
hsa-let-7b-3p	0.00	hsa-let-7b	0.00	0	0
hsa-let-7c	2.00	hsa-let-7c	2.00	2.00	5.21
hsa-let-7d-5p	32.00	hsa-let-7d	32.00	32.00	83.39
hsa-let-7d-3p	1.00	hsa-let-7d	1.00	1.00	2.61
hsa-let-7e-5p	139.00	hsa-let-7e	139.00	139.00	362.23
hsa-let-7e-3p	0.00	hsa-let-7e	0.00	0	0
hsa-let-7f-5p	1229.00	hsa-let-7f-1	1229.00	1229.00	3202.78
hsa-let-7f-1-3p	1.00	hsa-let-7f-1	1.00	1.00	2.61
hsa-let-7f-5p	1229.00	hsa-let-7f-2	1229.00	1229.00	3202.78
hsa-let-7f-2-3p	0.00	hsa-let-7f-2	0.00	0	0
hsa-let-7g-5p	31.00	hsa-let-7g	31.00	31.00	80.79
hsa-let-7g-3p	0.00	hsa-let-7g	0.00	0	0
hsa-let-7i-5p	108.00	hsa-let-7i	108.00	108.00	281.45
hsa-let-7i-3p	0.00	hsa-let-7i	0.00	0	0
hsa-miR-1	0.00	hsa-miR-1-1	0.00	0	0

9) Faça o mesmo para os outros arquivos.

```
less -S miRNAs_expressed_all_samples_07_07_2017_t_23_25_32.csv
```

```
less -S miRNAs_expressed_all_samples_07_07_2017_t_23_25_49.csv
```

```
less -S miRNAs_expressed_all_samples_07_07_2017_t_23_26_04.csv
```

10) Para a utilização do R-Peridot precisamos apenas das colunas “#miRNA” e “read_count” fazendo:

```
cut -f 1,2 miRNAs_expressed_all_samples_07_07_2017_t_23_25_06.csv >  
SRR326279.csv
```

11) Faça o mesmo para os outros arquivos gerados:

```
cut -f 1,2 miRNAs_expressed_all_samples_07_07_2017_t_23_25_32.csv >  
SRR326280.csv
```

```
cut -f 1,2 miRNAs_expressed_all_samples_07_07_2017_t_23_25_49.csv >  
SRR326281.csv
```

```
cut -f 1,2 miRNAs_expressed_all_samples_07_07_2017_t_23_26_04.csv >  
SRR326282.csv
```

12) Precisamos também remover nomes repetidos da coluna de nomes dos miRNA, para isso fazemos:

```
(head -n 1 SRR326279.csv && tail -n +3 SRR326279.csv | sort -u -d  
-k1,1) > SRR326279.unique.csv
```

13) Faça o mesmo para os outros arquivos:

```
(head -n 1 SRR326280.csv && tail -n +3 SRR326280.csv | sort -u -d  
-k1,1) > SRR326280.unique.csv
```

```
(head -n 1 SRR326281.csv && tail -n +3 SRR326281.csv | sort -u -d  
-k1,1) > SRR326281.unique.csv
```

```
(head -n 1 SRR326282.csv && tail -n +3 SRR326282.csv | sort -u -d  
-k1,1) > SRR326282.unique.csv
```

14) Agora vamos fazer a união de todos os dados em um único arquivo para a análise:

```
join -j 1 -t$'\t' SRR326279.unique.csv SRR326280.unique.csv |  
join -j 1 -t$'\t' - SRR326281.unique.csv |  
join -j 1 -t$'\t' - SRR326282.unique.csv >  
mergedAll.tsv
```

*obs: Digitar tudo na mesma linha

15) E renomear a coluna de nomes(primeira linha do arquivo) para facilitar o entendimento posterior:

```
sed -i '1s/.*/gene_id\tSRR326279\tSRR326280\tSRR326281\tSRR326282/'  
mergedAll.tsv
```

16) Vamos ver o resultado da união dos dados:

```
more mergeAll.tsv
```

17) Agora vamos copiar o arquivo para nossa área web:

Primeiro crie um diretório:

```
mkdir /home/bif/public_html/SeuNome
```

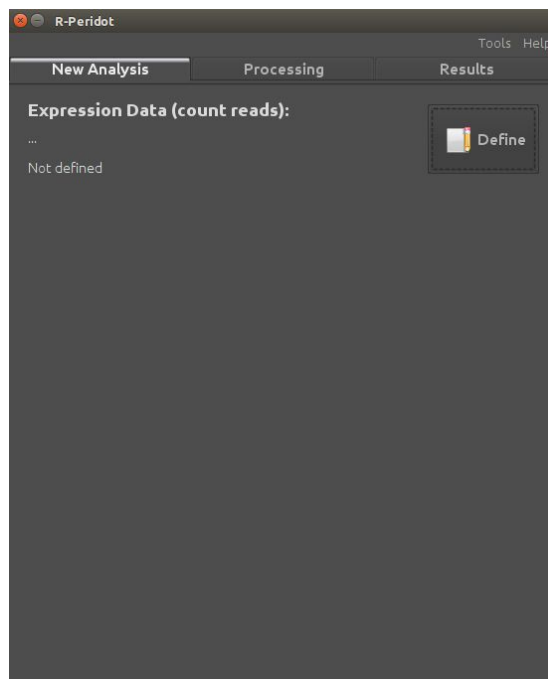
Depois copie:

```
cp mergeAll.tsv /home/bif/public_html/SeuNome
```

Agora entre no navegador web e visualize no endereço:

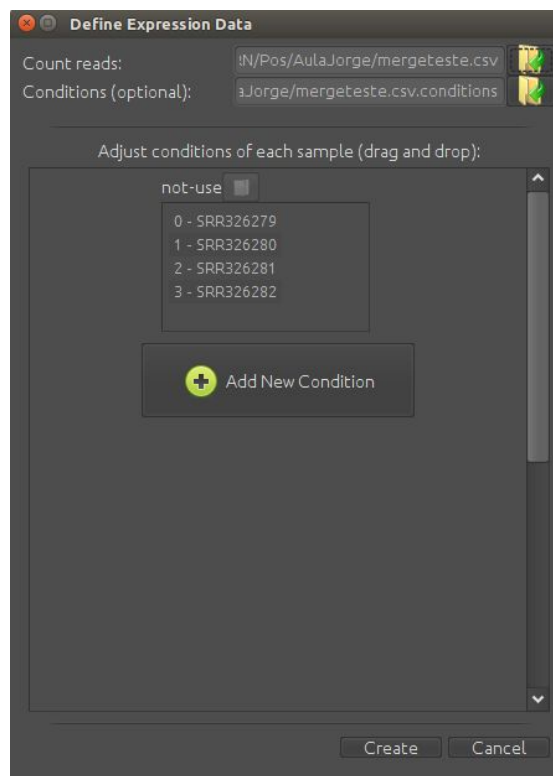
```
http://177.20.147.141/~bif/SeuNome/
```

18) Agora baixamos o arquivo mergeAll.tsv e abrimos o R-peridot:



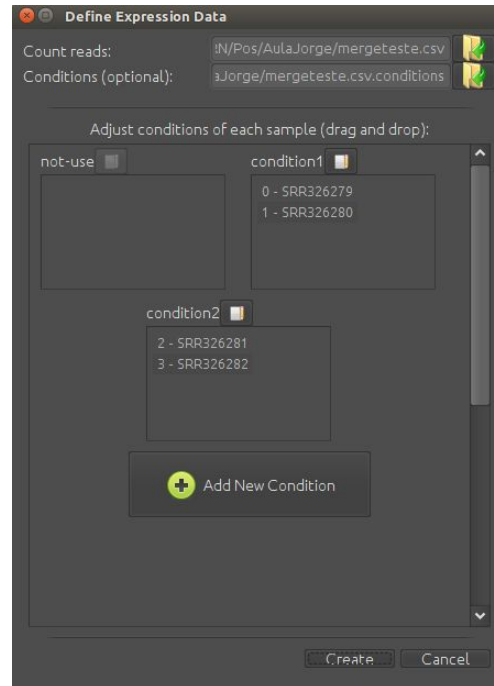
Tela inicial do R-Peridot

19) Clique no botão Define e navegue até onde se encontra o arquivo mergeAll.tsv. Como nosso arquivo possui uma coluna de nomes que definimos na primeira linha e os nomes dos miRNA na primeira coluna só precisamos apertar no botão ok:



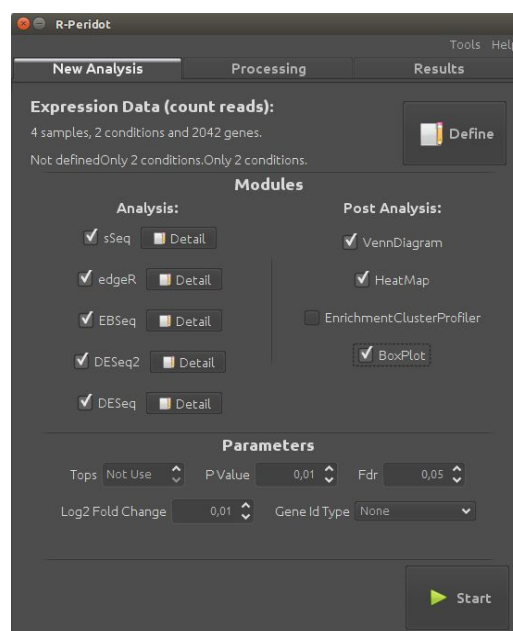
Tela de atribuição de condições

20) Adicione duas novas condições, apertando no botão Add New Condition para separarmos os grupos de amostras a serem analisadas. Após criar as condições arraste as amostras SRR326279 e SRR326280 para a caixa condition 1 e as amostras SRR326281 e SRR326282 para a caixa condition 2 e clique no botão create:



Condições atribuídas

21) Selecione todos os pacotes de análise: sSeq, edgeR, EBSeq, DESeq2 e DESeq. Selecione também os pacotes de pós análise VennDiagram, HeatMap e BoxPlot, deixaremos o EnrichmentClusterProfiler para o caso de precisarmos analisar a ontologia e aperte o botão Start:

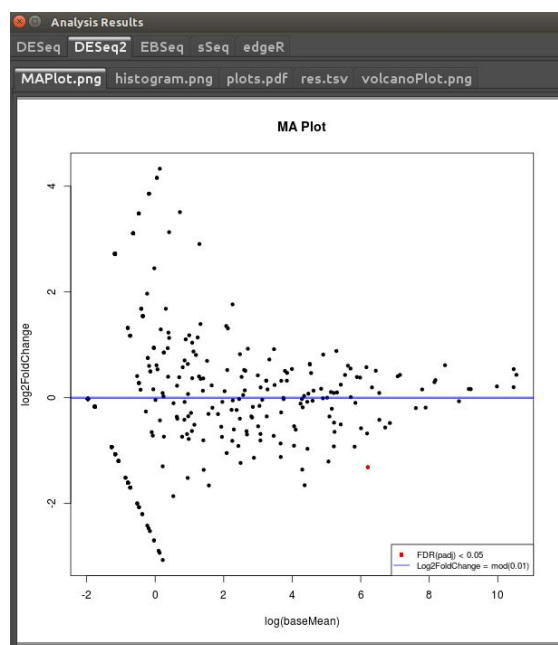


Todos os pacotes de análises selecionados.

22) Com o término do processamento da análise na aba de Resultados(Results) podemos visualizar todos os gráficos e tabelas de cada pacote e os gráficos dos resultados das pós análises:



Aba de Resultados do R-Peridot



Exemplo de Gráfico de resultado do R-Peridot(MA Plot do DESeq2)

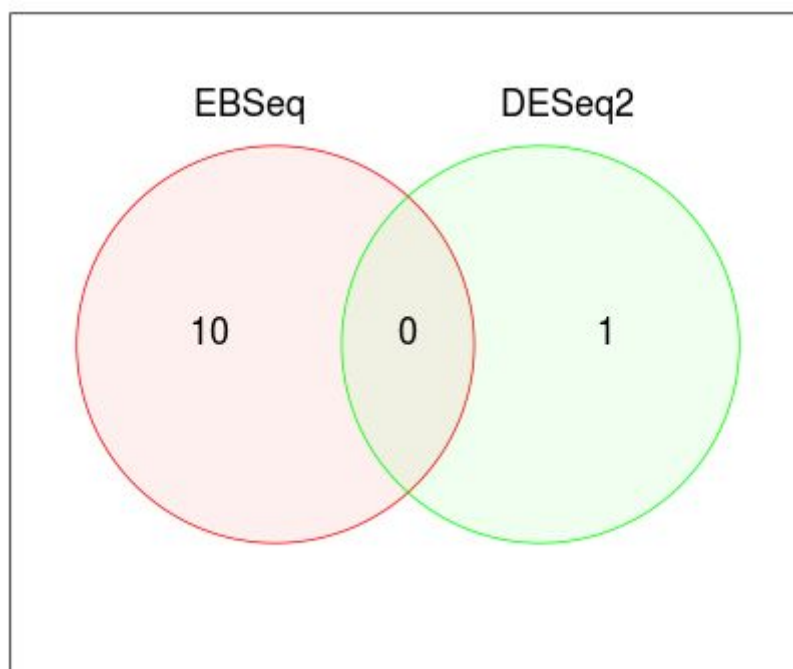


Diagrama de Venn do resultado da análise. Ausência de pacotes indica que estes não encontram miRNA diferencialmente expressos.

Referências:

1. Anders, S and Huber W (2010). "Differential expression analysis for sequence count data." *Genome Biology*, 11, pp. R106;
2. Anders S, Huber W (2016). "Differential expression of RNA-Seq data at the gene level" - the DESeq package. Available: <http://bioconductor.org/packages/release/bioc/vignettes/DESeq/inst/doc/DESeq.pdf>. Accessed em 25/07/2016;
3. Leng, N., J.A. Dawson and C. Kendzierski (2015). "EBSeq: An R package for differential expression analysis using RNA-seq data." Available: http://bioconductor.org/packages/release/bioc/vignettes/EBSeq/inst/doc/EBSeq_Vignette.pdf. Accessed 25/07/2016;
4. Chen Y, McCarthy DJ, Ritchie M, Robinson MD and Smyth GK. "edgeR: differential expression analysis of digital gene expression" data User's Guide. Available: <https://bioconductor.org/packages/release/bioc/vignettes/edgeR/inst/doc/edgeUsersGuide.pdf>. Accessed 25/07/2016;
5. Robinson MD, McCarthy DJ and Smyth GK (2010). "edgeR: a Bioconductor package for differential expression analysis of digital gene expression data." *Bioinformatics*, 26, pp. -1;

6. Suzuki, R. and Shimodaira, H. (2006) "Pvclust: an R package for assessing the uncertainty in hierarchical clustering", *Bioinformatics*, 22 (12): 1540-1542;
7. Yu D, Huber W and Vitek O (2013). "sSeq: Shrinkage estimation of dispersion in Negative Binomial models for RNA-seq experiments with small sample size" *Bioinformatics*. 15;29(10):1275-82;
8. Love MI, Huber W and Anders S (2014). "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2." *Genome Biology*, 15, pp. 550. doi: 10.1186/s13059-014-0550-8
9. Yu G, Wang L, Han Y and He Q (2012). "clusterProfiler: an R package for comparing biological themes among gene clusters." *OMICS: A Journal of Integrative Biology*, 16(5), pp. 284-287. doi: 10.1089/omi.2011.0118
10. Wickham, H; Chang, W.; "ggplot2", (2016). GitHub repository: <https://github.com/tidyverse/ggplot2>, Accessed on 25/072017;
11. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W and Smyth GK (2015). "limma powers differential expression analyses for RNA-sequencing and microarray studies." *Nucleic Acids Research*, 43(7), pp. E47;