7。短读质量与修整¶

学习目标：

通过Conda安装软件(FastQC，Multiqc)

下载数据

可视化阅读质量

质量过滤器和修剪读数

启动Jetstream m1.中等或更大按照Jetstream启动指令.

您现在应该登录到您的Jetstream计算机！你应该看看这样的东西

titus@js-17-71:~$

7.1.开始

更改到主目录：

cd ~/

并安装FastQC、MultiQC和微调：

conda install -y fastqc multiqc trimmomatic

7.2.数据源

创建一个“数据”目录：

cd ~/

mkdir -p data

cd data

并从Schutch等人，2016年酵母RNAseq研究：

curl -L https://osf.io/5daup/download -o ERR458493.fastq.gz

curl -L https://osf.io/8rvh5/download -o ERR458494.fastq.gz

curl -L https://osf.io/2wvn3/download -o ERR458495.fastq.gz

curl -L https://osf.io/xju4a/download -o ERR458500.fastq.gz

curl -L https://osf.io/nmqe6/download -o ERR458501.fastq.gz

curl -L https://osf.io/qfsze/download -o ERR458502.fastq.gz

让我们确保下载我们所有的数据使用md5和。md5sum散列是允许您与原始文件进行比较的指纹。一些排序工具将为传递给您的文件提供包含md5sum散列的文件。通过将md5sum与排序工具生成的原始md5sum进行比较，可以确保您已经下载了完整的文件。如果在下载过程中出现中断，文件的所有字节都不会被正确传输，您可能会丢失一些序列。

md5sum \*.fastq.gz

你应该看看这个：

2b8c708cce1fd88e7ddecd51e5ae2154 ERR458493.fastq.gz

36072a519edad4fdc0aeaa67e9afc73b ERR458494.fastq.gz

7a06e938a99d527f95bafee77c498549 ERR458495.fastq.gz

107aad97e33ef1370cb03e2b4bab9a52 ERR458500.fastq.gz

fe39ff194822b023c488884dbf99a236 ERR458501.fastq.gz

db614de9ed03a035d3d82e5fe2c9c5dc ERR458502.fastq.gz

若要检查md5sum散列是否与包含md5sum散列的文件匹配，请执行以下操作：

md5sum \*.fastq.gz > md5sum.txt

md5sum -c err\_md5sum.txt

(首先，我们创建一个包含md5sum文件的文件，然后检查它。实际上，你不会让这个文件检查。该设施将为你创造它。)

你应该看看这个：

ERR458493.fastq.gz: OK

ERR458494.fastq.gz: OK

ERR458495.fastq.gz: OK

ERR458500.fastq.gz: OK

ERR458501.fastq.gz: OK

ERR458502.fastq.gz: OK

现在，如果您键入：

ls -l

你应该看到这样的东西：

-rw-rw-r-- 1 titus titus 59532325 Jun 29 09:22 ERR458493.fastq.gz

-rw-rw-r-- 1 titus titus 58566854 Jun 29 09:22 ERR458494.fastq.gz

-rw-rw-r-- 1 titus titus 58114810 Jun 29 09:22 ERR458495.fastq.gz

-rw-rw-r-- 1 titus titus 102201086 Jun 29 09:22 ERR458500.fastq.gz

-rw-rw-r-- 1 titus titus 101222099 Jun 29 09:22 ERR458501.fastq.gz

-rw-rw-r-- 1 titus titus 100585843 Jun 29 09:22 ERR458502.fastq.gz

这是酵母研究的六个数据文件。文件。

这些文件的一个问题是它们是可写的-默认情况下，UNIX使文件所有者可以写东西。这给在原始数据中创建打印或错误带来了问题。在我们进一步讨论之前，让我们先解决这个问题：

chmod a-w \*

看看他们现在的权限-

ls -l

您应该看到原始权限字符串中的“w”(-rw-rw-r--)已从每个文件中删除，现在应该如下所示-r--r--r--.

我们将讨论这些文件下面的内容。

7.2.1.1。将数据链接到我们的工作地点

首先，创建一个新的工作目录：

mkdir -p ~/quality

cd ~/quality

现在，我们将在我们当前的工作目录中创建一个与我们的质量被裁剪的数据的“链接”：

ln -fs ~/data/\* .

然后，您将看到，当您执行以下操作时，它们将被链接到当前目录中。ls。这些链接使我们不必指定到其在计算机上的位置的完整路径(地址)，而无需实际移动或复制文件。但是请注意，在这里更改这些文件仍然会更改原始文件！

这些是FASTQ文件-让我们看看它们：

less ERR458493.fastq.gz

(使用空格键向下滚动，键入‘q’退出less)

问题：

文件名是从哪里来的？

链接：

FASTQ格式

7.2.2.2。FastQC

我们要用FastQC总结数据。我们已经安装了上面的‘FastQC’，使用Conda命令。

现在，对两个文件运行FastQC：

fastqc ERR458493.fastq.gz

fastqc ERR458500.fastq.gz

现在输入‘ls’：

ls -d \*fastqc.zip\*

要列出这些文件，您应该看到：

ERR458493\_fastqc.zip

ERR458500\_fastqc.zip

在每个FastQC目录中，您将从FastQC项目中找到报告。如果愿意，可以使用RStudioServer控制台下载这些文件。(@CTB)

或者你可以看看这些副本：

ERR 458493\_FAST qc.html

ERR 458500\_FAST qc.html

问题：

在FastQC报告中你应该注意什么？

链接：

FastQC

FastQC教程视频

FastQC有几个注意事项，主要是它只计算数据子集的某些统计数据(如重复序列)(例如，重复序列只对每个文件中的前100，000个序列进行分析)。

7.2.3.3。三毛

现在我们要做一些修剪！我们会用三毛，我们已经通过Conda安装了它(和FastQC一样)。

我们需要的第一件事是剪掉适配器：

cp /opt/miniconda/pkgs/trimmomatic-\*/share/trimmomatic-\*/adapters/TruSeq2-PE.fa .

(您可以使用cat TruSeq2-PE.fa)

现在，对这两个人运行Trimmoatic：

trimmomatic SE ERR458493.fastq.gz \

ERR458493.qc.fq.gz \

ILLUMINACLIP:TruSeq2-PE.fa:2:40:15 \

LEADING:2 TRAILING:2 \

SLIDINGWINDOW:4:2 \

MINLEN:25

trimmomatic SE ERR458500.fastq.gz \

ERR458500.qc.fq.gz \

ILLUMINACLIP:TruSeq2-PE.fa:2:40:15 \

LEADING:2 TRAILING:2 \

SLIDINGWINDOW:4:2 \

MINLEN:25

您应该看到如下所示的输出：

...

Input Reads: 1093957 Surviving: 1092715 (99.89%) Dropped: 1242 (0.11%)

TrimmomaticSE: Completed successfully

我们还可以更有效地对所有6个示例运行相同的过程。for循环，如下所示：

for filename in \*.fastq.gz

do

# first, make the base by removing fastq.gz

base=$(basename $filename .fastq.gz)

echo $base

trimmomatic SE ${base}.fastq.gz \

${base}.qc.fq.gz \

ILLUMINACLIP:TruSeq2-PE.fa:2:40:15 \

LEADING:2 TRAILING:2 \

SLIDINGWINDOW:4:2 \

MINLEN:25

done

此脚本将对以fastq.gz为它运行Trimmoatic。

问题：

你怎么知道这些参数是什么意思？

如何确定要使用哪些参数？

你用什么适配器？

我们在这里使用的是什么版本的Trimmoatic？(FastQC呢？)

你认为RNAseq和基因组数据集的参数不同吗？

这些烦人的长而复杂的文件名是怎么回事？

有关最优修剪策略的讨论，请参见MacManes，2014年-这是关于RNAseq的，但类似的论点应该适用于Metagenome程序集。

链接：

三毛

7.2.4.4。再来一次FastQC

再次对修剪过的文件运行FastQC：

fastqc ERR458493.qc.fq.gz

fastqc ERR458500.qc.fq.gz

现在查看这些文件的副本：

ERR 458493.qc\_Quickqc.html

ERR 458500.qc\_Quickqc.html

让我们看看输出文件：

less ERR458493.qc.fq.gz

(同样，使用空格滚动，‘q’退出较少)。

7.2.5.5。多QC

多重QC将多个样本的结果聚合到一个单一的报告中，以便于比较。

对未修剪和已裁剪的文件运行Mulitqc

multiqc .

现在您应该看到如下所示的输出：

[INFO ] multiqc : This is MultiQC v1.0

[INFO ] multiqc : Template : default

[INFO ] multiqc : Searching '.'

Searching 15 files.. [####################################] 100%

[INFO ] fastqc : Found 4 reports

[INFO ] multiqc : Compressing plot data

[INFO ] multiqc : Report : multiqc\_report.html

[INFO ] multiqc : Data : multiqc\_data

[INFO ] multiqc : MultiQC complete

可以查看输出html文件。multiqc\_report.html通过转到RStudio，选择文件，并说“Web浏览器中的视图”。