Welcome to Software Carpentry Etherpad!

This pad is synchronized as you type, so that everyone viewing this page sees the same text. This allows you to collaborate seamlessly on documents.

Use of this service is restricted to members of the Software Carpentry and Data Carpentry community; this is not for general purpose use (for that, try etherpad.wikimedia.org).

Users are expected to follow our code of conduct: http://software-carpentry.org/conduct.html

##
Microbial Genomics course 2018
##
Website **Microbial Genomics course**: https://aschuerch.github.io/Microbial-Genomics-2018/
This **collaborative document**: http://pad.software-carpentry.org/2018-04-03-Utrecht
Week 1: Anita Schürch, Aldert Zomer and Bas Dutilh
Week 2: Jerome Collemare, Ronnie de Jonge, and Robin Ohm
##

Day1 and 2: Data Carpentry Genomics

Pre-workshop Survey: https://www.surveymonkey.com/r/dcpreworkshopassessment?workshop_id=2018-04-03-Utrecht
Post-workshop Survey: https://www.surveymonkey.com/r/dcpostworkshopassessment?workshop_id=2018-04-03-Utrecht

Data Carpentry Instructors:
Marieke Dirksen,
Dennis Schmitz, Dennis.Schmitz (at) rivm.nl
Sam Nooij, sam.nooij@rivm.nl
Anita Schürch, a.c.schurch@umcutrecht.nl

Participants:
DONT ADD YOUR ANSWERS HERE, thanks :) add them at the bottom of the etherpad, at their respective assignments
Sam
Reinder
Div
Ethel
Cindy
Tom
Rozemarijn
Ruben
Ramon

Timo
Tony
Fabian
Sarah
Jorik

##
Pre-workshop survey results:
https://docs.google.com/spreadsheets/d/
e/2PACX-1vSFFj8Ms6gt28aVOs3db2mjHrmzDNG1_u1pXxGXjIQsUzPGIWja4aRpO8iKg
5Y1JYCI5fjTCiyhxt0L/pubchart?oid=1583735864&format=image

Instances:

- ec2-54-196-53-101.compute-1.amazonaws.com** - Div

ec2-107-21-37-31.compute-1.amazonaws.com** - Reinder
ec2-54-173-248-200.compute-1.amazonaws.com
ec2-54-224-174-220.compute-1.amazonaws.com ** Tony /home/dcuser/
dc_sample_data/.hidden
ec2-54-162-255-112.compute-1.amazonaws.com ** Rozemarijn
ec2-18-233-100-142.compute-1.amazonaws.com
ec2-54-159-107-14.compute-1.amazonaws.com **Sarah
ec2-34-201-53-122.compute-1.amazonaws.com
ec2-54-210-240-156.compute-1.amazonaws.com ** jorik
ec2-34-238-135-213.compute-1.amazonaws.com
ec2-54-159-185-105.compute-1.amazonaws.com * * - Cindyjnuh- .hidden
ec2-54-204-207-212.compute-1.amazonaws.com* - Ruben
ec2-35-174-139-101.compute-1.amazonaws.com
ec2-54-165-21-21.compute-1.amazonaws.com- Tom
ec2-54-208-199-129.compute-1.amazonaws.com
ec2-54-234-33-250.compute-1.amazonaws.com
ec2-52-90-199-225.compute-1.amazonaws.com** Timo
ec2-52-90-37-85.compute-1.amazonaws.com
ec2-107-22-153-37.compute-1.amazonaws.com - Ethel
ec2-52-54-246-20.compute-1.amazonaws.com
ec2-35-174-137-218.compute-1.amazonaws.comLotte
ec2-35-174-153-155.compute-1.amazonaws.com
ec2-54-89-81-152.compute-1.amazonaws.com
ec2-184-72-110-183.compute-1.amazonaws.com* - Ramon
ec2-54-161-90-168.compute-1.amazonaws.com
ec2-52-90-200-217.compute-1.amazonaws.com
ec2-54-174-173-70.compute-1.amazonaws.com
ec2-52-87-152-114.compute-1.amazonaws.com
ec2-34-238-242-58.compute-1.amazonaws.com - Sam
ec2-18-233-170-56.compute-1.amazonaws.com** Fabian
ec2-34-238-154-210.compute-1.amazonaws.com - Marieke

ec2-54-89-205-213.compute-1.amazonaws.com - Sam Nooij
ec2-34-226-154-152.compute-1.amazonaws.com - Dennis
ec2-54-161-227-79.compute-1.amazonaws.com - Anita

user: dcuser
password: data4Carp

Shell commands that we have used:

clear

- clears the terminal screen

export PS1='$ '

- changes the prompt (beginning of the line) to just a dollar sign

ls

- list what is in the current directory (files, directories, links, anything)

cd

- change directory (to the directory you type after it)

pwd

- print the current working directory

ls -F

- list, and show which files are regular files, and which are directories

ls -l

- list in 'long' format

[TAB]

- the TAB key on your keyboard may be used to automatically complete file names or commands, or when pressed twice will show a list of available options

~

- the 'tilde' sign is short for 'home'

*

- asterisk or star is a wildcard that means 'anything'

echo

- 'echoes' back what you typed, can be used to show how '*' (asterisk) is interpreted

cat

- short for 'concatenate'; can be used to read the contents of files (usually text files)

head

- shows the top 10 lines of a file (by default)

tail

- shows the bottom 10 lines of a file

head -n 1

- shows only the first line of a file

tail -n 1

- shows only the last line of a file

mkdir

- make a new directory

cp

- copy a file

mv

- 'move' a file (can also be used to rename)

rm

- remove a file (delete it, NOT move to trashcan)

chmod

- change the persmissions of a file

history

- get a complete history of your commands (or at least very long)

[CTRL] + c

- stop a command (one that is running or the one you are typing)

grep

- search in a file for a word or 'string' of letters

grep -B1 -A2

- find also 1 line Before the match and 2 lines After

>

- redirect the output of your command to a file

wc

- 'word count' counts the number of lines, words and bytes of a file

wc -l

- count only lines

|

- 'pipe' can be used to link different commands together: the output of the first will be used by the second

>>

- redirect to a file and append to it instead of overwriting

cut

- cuts columns out a table (or fields from any text)

nano

- simple command-line text editor. Use CTRL + X to close

fastqc

- quality control program for (next-generation) sequencing reads

less

- open and read a (text) file on the command line

Shell command history:

```
cd dc_sample_data
ls
cd untrimmed_fastq
ls
cd
pwd
cd_dc_sample_data/
cd untrimmed_fastq/
ls
ls SRR[TAB]
ls SRR09[TAB][TAB]
ls SRR097977.fastq
pw[TAB][TAB]
pwd
cd ..
pwd
cd ..
pwd
cd dc_sample_data/untrimmed_fastq/
pwd
cd ../..
pwd
cd /home/dcuser/dc_sample_data/
cd ../../
pwd
cd
pwd
cd dc_sample_data/
man ls
ls -a
```

```
ls .hidden
cd
pwd
cd dc_sample_data/
ls ~
cd ~
cd ~/dc_sample_data/untrimmed_fastq/
ls *.fastq
ls *977.fastq
echo *.fastq
ls /usr/bin/*.sh
ls
cat SRR098026.fastq
head SRR098026.fastq
tail SRR098026.fastq
cd
head -n 1 dc_sample_data/untrimmed_fastq/SRR098026.fastq
tail -n 1 dc_sample_data/untrimmed_fastq/SRR098026.fastq
cd dc_sample_data/untrimmed_fastq
mkdir backup
cp SRR097977.fastq SRR097977-copy.fastq
cp SRR098026.fastq SRR098026-copy.fastq
ls -F
mv SRR097977-copy.fastq backup/
mv SRR098026-copy.fastq SRR098026-backup.fastq
ls backup/
ls
mv SRR098026-backup.fastq backup/
cd backup
ls
rm SRR097977-copy.fastq
ls -lat
ls -l
chmod -w SRR098026-backup.fastq
cd ..
ls
rm -r backup/
 (yes)
ls
history
clear
mkdir backup
cp SRR097977.fastq backup/SRR097977-backup.fastq
cp SRR097977.fastq backup/SRR097977-backup.fastq
chmod -w backup/*.fastq
ls -l
pwd
grep NNNNNNNNNN SRR098026.fastq
```

```
grep -B1 -A2 NNNNNNNNNN SRR098026.fastq
grep -B1 -A2 GNATNACCACTTCC SRR098026.fastq
grep -B1 AAGTT *.fastq
grep -B1 -A2 NNNNNNNNNN SRR098026.fastq > bad_reads.txt
head bad_reads.txt
wc bad_reads.txt
wc -l bad_reads.txt
grep NNN SRR098026.fastq | wc -l
grep -c NNN SRR098026.fastq
grep -B1 -A2 NNNNNNNNNN SRR097977.fastq > bad_reads.txt
wc -l bad_reads.txt
grep -B1 -A2 NNNNNNNNNN SRR098026.fastq > bad_reads.txt
grep -B1 -A2 NNNNNNNNNN SRR097977.fastq >> bad_reads.txt
wc -l bad_reads.txt
grep -B1 -A2 NNNNNNNNNN *.fastq >> bad_reads.txt
wc -l
grep -B1 -A2 NNNNNNNNNN *.fastq > bad_reads.fastq  # throws an error! at least when run twice
cd
cd dc_sample_data/sra_metadata/
ls
head -n 1 SraRunTable.txt
head -1 SraRunTable.txt | cut -f1-4
head -1 SraRunTable.txt | cut -f3
cut -f3 SraRunTable.txt
cut -f3 SraRunTable.txt | head -n 10
cut -f3 SraRunTable.txt | grep PAIRED | wc -l
cut -f3 SraRunTable.txt | grep SINGLE | wc -l
cut -f3 SraRunTable.txt | grep -v LibraryLayout_s | sort | uniq -c
cd
cd dc_sample_data/untrimmed_fastq/
nano README.txt
head README.txt
grep -B1 -A2
grep -B1 -A2 NNNNNNNNNN *.fastq > scripted_bad_reads.txt
nano bad-reads-script.sh
bash bad-reads-script.sh
nano bad-reads.script.sh
# write: echo "Script finished!"
bash bad-reads-script.sh
ls -l bad-reads-script.sh
chmod +x bad-reads-script.sh #add execute permissions
./bad-reads-script.sh  # run as a computer program
cd
mkdir dc_workshop
mkdir dc_workshop/docs
mkdir dc_workshop/data
mkdir dc_workshop/results
```

```
ls -R
ls -R dc_workshop/
history | tail -n 7 >> dc_workshop/docs/dc_workshop_log_20180403.txt
nano dc_workshop/docs/dc_workshop_log_20180403.txt
mv dc_workshop/docs/dc_workshop_log_20180403.txt .
rm -r dc_workshop
bash dc_workshop_log_20180403.txt
ls -R dc_workshop
```

PART 2: sequence analysis (read quality, trimming, variant calling)

```
ls dc_sample_data/untrimmed_fastq
cd
pwd
mv ~/.dc_sampledata_lite/untrimmed_fastq/ ~/dc_workshop/data/
ls dc_workshop/data/
ls dc_workshop/data/untrimmed_fastq # should display 6 fastq files!
dc_workshop/docs/dc_workshop_log_20180403.txt
~/FastQC/fastqc *.fastq
ls
mkdir ~/dc_workshop/results/fastqc_untrimmed_reads
mv *.zip ~/dc_workshop/results/fastqc_untrimmed_reads
mv *.html ~/dc_workshop/results/fastqc_untrimmed_reads/
# copy the files to your local machine using Filezilla and view them in a webbrowser
cd ~/dc_workshop/results/fastqc_untrimmed_reads
unzip *.zip # unzip all zip files..., or does it?
for filename in *.zip  # Press Shift + Enter when not on a Linux computer
do
unzip $filename
done
ls -F # see the unzipped folders
ls SRR097977_fastqc/
less SRR097977_fastqc/summary.txt # press Q to quit 'less'
cat */summary.txt > ~/dc_workshop/docs/fastqc_summaries.txt
less ~/dc_workshop/docs/fastqc_summaries.txt
grep FAIL ~/dc_workshop/docs/fastqc_summaries.txt | cut -f3 | sort | uniq -c
cd ~/dc_workshop/data/untrimmed_fastq
ls
java -jar ~/Trimmomatic-0.32/trimmomatic-0.32.jar SE SRR098283.fastq
SRR098283.fastq_trim.fastq SLIDINGWINDOW:4:20 MINLEN:20
for infile in *.fastq # Shift + Enter when not using Linux
do
outfile="${infile}"_trim.fastq
java -jar ~/Trimmomatic-0.32/trimmomatic-0.32.jar SE "${infile}" "${outfile}"
SLIDINGWINDOW:4:20 MINLEN:20
done
ls
```

```
cd ~/dc_workshop/data/untrimmed_fastq/
mkdir ../trimmed_fastq
rm SRR098283.fastq_trim.fastq_trim.fastq
mv *_trim* ../trimmed_fastq/
cd ../trimmed_fastq
ls
cd ~/dc_workshop
ln -s ~/.dc_sampledata_lite/trimmed_fastq_small/ data/
ln -s ~/.dc_sampledata_lite/ref_genome/ data/
mkdir -p results/sai results/sam results/bam results/bcf results/vcf
bwa index data/ref_genome/ecoli_rel606.fasta
bwa aln data/ref_genome/ecoli_rel606.fasta data/trimmed_fastq_small/
SRR097977.fastq_trim.fastq > results/sai/SRR097977.aligned.sai
bwa samse data/ref_genome/ecoli_rel606.fasta \ # Press Shift + Enter on non-Linux
machines
results/sai/SRR097977.aligned.sai \
data/trimmed_fastq_small/SRR097977.fastq_trim.fastq > \
results/sam/SRR097977.aligned.sam
samtools view -S -b  results/sam/SRR097977.aligned.sam > results/bam/
SRR097977.aligned.bam
samtools sort results/bam/SRR097977.aligned.bam results/bam/SRR097977.aligned.sorted
samtools mpileup -g -f data/ref_genome/ecoli_rel606.fasta results/bam/
SRR097977.aligned.sorted.bam > results/bcf/SRR097977_raw.bcf
bcftools view -bvcg results/bcf/SRR097977_raw.bcf > results/bcf/
SRR097977_variants.bcf
bcftools view results/bcf/SRR097977_variants.bcf | /usr/share/samtools/vcfutils.pl
varFilter - > results/vcf/SRR097977_final_variants.vcf
less results/vcf/SRR097977_final_variants.vcf
samtools index results/bam/SRR097977.aligned.sorted.bam
samtools tview results/bam/SRR097977.aligned.sorted.bam
samtools tview results/bam/SRR097977.aligned.sorted.bam data/ref_genome/
ecoli_rel606.fasta
```

Find hidden directory:

Sam .hidden
Reinder .hidden/youfoundit.txt
Div /home/dcuser/dc_sample_data/.hidden/youfoundit.txt
Ethel .hidden/youfoundit.txt
Cindy
Tom.hidden/youfoundit.txt
Rozemarijn - youfoundit.txt
Ruben .hidden/youfoundit.txt Here I am
Ramon .hidden/youfoundit.txt
Timo dc_sample_data/.hidden/youfoundit.txt
Tony
Fabian youfoundit.txt

Sarah .hidden/youfoundit.txt

Using the filesystem diagram below, if pwd displays /Users/thing, what will ls ../backup display?

1../backup: No such file or directory
2. 2012-12-01 2013-01-08 2013-01-27
3. 2012-12-01/ 2013-01-08/ 2013-01-27/
4. original pnas_final pnas_sub

1
2
3
4 |||| |||| \

- Break until 10.25 . \m/

ls *a* ^ *c*
ls *{a,c}*
$ls /usr/bin/*[a,c]*

chmod: how can I protect a file from being removed?

Sam
Reinder chmod -w SRR*
Div chmod go-w filename
Ethel chmod a-w SRR*
Cindy
Tom
Rozemarijn chmod -u backup/
Ruben
Ramon
Timo
Tony chmod 0444 *.fastq
Fabian
Sarah chmod -w SRR09*
Jorik

How many sequences in SRR098026.fastq contain at least 3 consecutive Ns?
sarah 249
cindy 249
Timo 249 $grep NNN SRR098026.fastq >NNN8026.txt
$wc -l NNN8026.txt
Reinder 249 sequences grep NNN SSR098026.fastq | wc -l
Ethel 249 grep -c NNN SRR098026.fastq
Div 249

Fabian grep NNN SRR098026.fastq | wc -l
Tony grep NNN SRR098026.fastq | wc -l gives 249 seqs
Ruben grep NNN SRR098026.fastq | wc -l 249 sequences
Tom grep NNN SRR098026.fastq | wc-l
Rozemarijn grep NNN SRR098026.fastq, 249 sequences

Reinder grep -c NNNNN*.NNNNN SSR098026.fastq ; 134 lines
Ethel 186 $grep NNNNN*NNNNN SRR098026.fastq | wc -l
[Tony] 186 lines
Timo 186 lines
$grep NNNNN*NNNNN SRR098026.fastq > 8026N.txt
$wc -l 8026N.txt
186 8026N.txt
Sarah 186

Back after lunch at 13.00

How to organize computational biology projects:
http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1000424

From this point on you need to have the folders and files if you would like to create the pipeline and get the analysis results.

mkdir dc_workshop
mkdir dc_workshop/docs
mkdir dc_workshop/data
mkdir dc_workshop/results

CNCTNTATGCGTACGGCAGTGANNNNNNNGGAGAT
A!@B!BBB@ABAB########!!!!!!!######
First half seems moderately good, second half is hardly read

tail -n 4 SRR098026.fastq
@SRR098026.249 HWUSI-EAS1599_1:2:1:2:1057 length=35
CNCTNTATGCGTACGGCAGTGANNNNNNNGGAGAT
+SRR098026.249 HWUSI-EAS1599_1:2:1:2:1057 length=35
A!@B!BBB@ABAB########!!!!!!!######
first half is alright, second half is bad
@SRR098026.249 HWUSI-EAS1599_1:2:1:2:1057 length=35
CNCTNTATGCGTACGGCAGTGANNNNNNNGGAGAT
+SRR098026.249 HWUSI-EAS1599_1:2:1:2:1057 length=35
A!@B!BBB@ABAB########!!!!!!!######

it's a bad read, # and ! are very low PHRED scores. Might trim it.
+SRR098026.249 HWUSI-EAS1599_1:2:1:2:1057 length=35
A!@B!BBB@ABAB########!!!!!!!######
Not a good read, many explanation marks

bad
@SRR098026.11 HWUSI-EAS1599_1:2:1:0:542 length=35
NNNNNNNNNNNNNNNNGNNNNNNNNNNNNNNNNN
+SRR098026.11 HWUSI-EAS1599_1:2:1:0:542 length=35
!!!!!!!!!!!!!!!#!!!!!!!!!!!!!!!!!!
@SRR098026.12 HWUSI-EAS1599_1:2:1:0:573 length=35
NNNNNNNNNNNNNNNNNTNNNNNNNNNNNNNNNNN
+SRR098026.12 HWUSI-EAS1599_1:2:1:0:573 length=35
!!!!!!!!!!!!!!!#!!!!!!!!!!!!!!!!!!

First half is pretty good, the rest is bad.
@SRR098026.249 HWUSI-EAS1599_1:2:1:2:1057 length=35
CNCTNTATGCGTACGGCAGTGANNNNNNNGGAGAT
+SRR098026.249 HWUSI-EAS1599_1:2:1:2:1057 length=35
A!@B!BBB@ABAB########!!!!!!!######

bad
@SRR098026.249 HWUSI-EAS1599_1:2:1:2:1057 length=35
CNCTNTATGCGTACGGCAGTGANNNNNNNGGAGAT
+SRR098026.249 HWUSI-EAS1599_1:2:1:2:1057 length=35
A!@B!BBB@ABAB########!!!!!!!######

@SRR098026.249 HWUSI-EAS1599_1:2:1:2:1057 length=35
CNCTNTATGCGTACGGCAGTGANNNNNNNGGAGAT
+SRR098026.249 HWUSI-EAS1599_1:2:1:2:1057 length=35
A!@B!BBB@ABAB########!!!!!!!######
first half is Ok, second half is bad.

bad
$ tail -4 SRR098026.fastq
@SRR098026.249 HWUSI-EAS1599_1:2:1:2:1057 length=35
CNCTNTATGCGTACGGCAGTGANNNNNNNGGAGAT
+SRR098026.249 HWUSI-EAS1599_1:2:1:2:1057 length=35
A!@B!BBB@ABAB########!!!!!!!######

bad
$ tail -4 SRR098026.fastq
@SRR098026.249 HWUSI-EAS1599_1:2:1:2:1057 length=35
CNCTNTATGCGTACGGCAGTGANNNNNNNGGAGAT
+SRR098026.249 HWUSI-EAS1599_1:2:1:2:1057 length=35
A!@B!BBB@ABAB########!!!!!!!######

bad Roos
@SRR098026.249 HWUSI-EAS1599_1:2:1:2:1057 length=35
CNCTNTATGCGTACGGCAGTGANNNNNNNGGAGAT
+SRR098026.249 HWUSI-EAS1599_1:2:1:2:1057 length=35
A!@B!BBB@ABAB########!!!!!!!######

```
$ ls -hs
total 17G
840M SRR097977.fastq  875M SRR098027.fastq  4.0G SRR098281.fastq
3.4G SRR098026.fastq  3.4G SRR098028.fastq  3.9G SRR098283.fastq

840M SRR097977.fastq  3.4G SRR098026.fastq  875M SRR098027.fastq  3.4G
SRR098028.fastq       4.0G SRR098281.fastq  3.9G SRR098283.fastq

$ls dc_workshop/data//untrimmed_fastq/
SRR097977.fastq  SRR098026.fastq  SRR098027.fastq  SRR098028.fastq
SRR098281.fastq  SRR098283.fastq
$ls -l --block-size=M
total 1M
drwxrwxr-x 2 dcuser dcuser 1M Jun  6  2014 Configuration
-rwxrwxr-x 1 dcuser dcuser 1M Jun  4  2014 fastqc
-rw-rw-r-- 1 dcuser dcuser 1M Mar 21  2012 fastqc_icon.ico
drwxrwxr-x 5 dcuser dcuser 1M Jun  6  2014 Help
-rw-rw-r-- 1 dcuser dcuser 1M May  7  2014 INSTALL.txt
-rw-rw-r-- 1 dcuser dcuser 1M Feb 24  2014 jbzip2-0.9.jar
-rw-rw-r-- 1 dcuser dcuser 1M Mar 21  2012 LICENSE.txt
drwxrwxr-x 3 dcuser dcuser 1M Jun  6  2014 net
drwxrwxr-x 3 dcuser dcuser 1M Jun  6  2014 org
-rw-rw-r-- 1 dcuser dcuser 1M Mar 21  2012 README.txt
-rw-rw-r-- 1 dcuser dcuser 1M Jun  6  2014 RELEASE_NOTES.txt
-rw-rw-r-- 1 dcuser dcuser 1M May  6  2014 run_fastqc.bat
-rw-rw-r-- 1 dcuser dcuser 1M Feb 14  2014 sam-1.103.jar
drwxrwxr-x 3 dcuser dcuser 1M Jun  6  2014 Templates
drwxrwxr-x 3 dcuser dcuser 1M Jun  6  2014 uk


-rw-r--r-- 1 dcuser dcuser 840M Jul 30  2015 SRR097977.fastq
-rw-r--r-- 1 dcuser dcuser 3.4G Jul 30  2015 SRR098026.fastq
-rw-r--r-- 1 dcuser dcuser 875M Jul 30  2015 SRR098027.fastq
-rw-r--r-- 1 dcuser dcuser 3.4G Jul 30  2015 SRR098028.fastq
-rw-r--r-- 1 dcuser dcuser 4.0G Jul 30  2015 SRR098281.fastq
-rw-r--r-- 1 dcuser dcuser 3.9G Jul 30  2015 SRR098283.fastq

ls -hs
840M SRR097977.fastq  875M SRR098027.fastq  4.0G SRR098281.fastq
3.4G SRR098026.fastq  3.4G SRR098028.fastq  3.9G SRR098283.fastq


$ du -h *
840M    SRR097977.fastq
3.4G    SRR098026.fastq
875M    SRR098027.fastq
```

3.4G SRR098028.fastq
4.0G SRR098281.fastq
3.9G SRR098283.fastq

$ ls -hl dc_workshop/data/untrimmed_fastq/
total 17G
-rw-r--r-- 1 dcuser dcuser 840M Jul 30 2015 SRR097977.fastq
-rw-r--r-- 1 dcuser dcuser 3.4G Jul 30 2015 SRR098026.fastq
-rw-r--r-- 1 dcuser dcuser 875M Jul 30 2015 SRR098027.fastq
-rw-r--r-- 1 dcuser dcuser 3.4G Jul 30 2015 SRR098028.fastq
-rw-r--r-- 1 dcuser dcuser 4.0G Jul 30 2015 SRR098281.fastq
-rw-r--r-- 1 dcuser dcuser 3.9G Jul 30 2015 SRR098283.fastq

total 16874668
-rw-r--r-- 1 dcuser dcuser 879991940 Jul 30 2015 SRR097977.fastq
-rw-r--r-- 1 dcuser dcuser 3585526358 Jul 30 2015 SRR098026.fastq
-rw-r--r-- 1 dcuser dcuser 917278492 Jul 30 2015 SRR098027.fastq
-rw-r--r-- 1 dcuser dcuser 3587743292 Jul 30 2015 SRR098028.fastq
-rw-r--r-- 1 dcuser dcuser 4193378186 Jul 30 2015 SRR098281.fastq
-rw-r--r-- 1 dcuser dcuser 4115713378 Jul 30 2015 SRR098283.fastq

Roos:
total 17G
-rw-r--r-- 1 dcuser dcuser 840M Jul 30 2015 SRR097977.fastq
-rw-r--r-- 1 dcuser dcuser 3.4G Jul 30 2015 SRR098026.fastq
-rw-r--r-- 1 dcuser dcuser 875M Jul 30 2015 SRR098027.fastq
-rw-r--r-- 1 dcuser dcuser 3.4G Jul 30 2015 SRR098028.fastq
-rw-r--r-- 1 dcuser dcuser 4.0G Jul 30 2015 SRR098281.fastq
-rw-r--r-- 1 dcuser dcuser 3.9G Jul 30 2015 SRR098283.fastq

Run Quality Check with FastQC:

cd /home/dcuser/dc_workshop/data/untrimmed_fastq
~/FastQC/fastqc *fastq

Filezilla:
dcuser
data4Carp
port:22

Break, back at 14.30

mkdir ~/dc_workshop/results/fastqc_untrimmed_reads
mv *.zip ~/dc_workshop/results/fastqc_untrimmed_reads/
mv *.html ~/dc_workshop/results/fastqc_untrimmed_reads/

SRR098028 Best
SSR098027 Worst

Files FastQC
SRR097977 good
SRR098027 nice
SRR098028 nice
SRR098026 beginning okay then error bars became worse
SRR098281 not so nice
SRR098283 not so nice.
However, it depends what sample it was and the age of the DNA sample.

SRR097977 is alright
SRR098027 is also nice even though it says it has a bad base quality
SRR098026 very large error bars
SRR098028 very large error bars
SRR098283 very large error bars
SRR098281 very large error bars

SRR097977 is best

SRR098027 and SRR097977 are usable (first part of the sequence is sufficient)

unzip zip files:
for i in *.zip

- do
- unzip $i
- done

Exercise:
Which samples failed at least one of FastQC's quality tests? What test(s) did those samples fail?

Bonus points: Remove the redudancy, and report which files failed (e.g. sample x and y instead of sample x module 1 and module 4 and sample y module 4 and 7)
Bonus points: Create a summary for all samples describing which modules failed (e.g. Kmer content module failed 5 times, etc)

Main:
```
grep FAIL fastqcsummaries.txt | cut -f2-3
```
Bonus 1:
```
grep FAIL fastqcsummaries.txt | cut -f3 | uniq
  SRR097977.fastq
  SRR098026.fastq
  SRR098027.fastq
  SRR098028.fastq
```

SRR098281.fastq
SRR098283.fastq

Bonus 2:

```
grep FAIL fastqcsummaries.txt | cut -f2 | sort | uniq -c
      5 Kmer Content
      4 Overrepresented sequences
      1 Per base sequence quality
      4 Per tile sequence quality

grep FAIL ~/dc_workshop/docs/fastqc_summaries.txt | cut -f3 | uniq
SRR097977.fastq
SRR098026.fastq
SRR098027.fastq
SRR098028.fastq
SRR098281.fastq
SRR098283.fastq


$ grep FAIL fastqc_summaries.txt | cut -f3 | sort |  uniq
SRR097977.fastq
SRR098026.fastq
SRR098027.fastq
SRR098028.fastq
SRR098281.fastq
SRR098283.fastq




$ grep FAIL ~/dc_workshop/docs/fastqc_summaries.txt | cut -f3 | sort | uniq -c


dcuser@ip-172-31-63-201:~/dc_workshop/docs$ grep PASS fastqc_summaries.txt | cut -f1,3
| uniq -c | sort -rn
      9 PASS        SRR098027.fastq
      9 PASS        SRR097977.fastq
      8 PASS        SRR098283.fastq
      8 PASS        SRR098281.fastq
      7 PASS        SRR098028.fastq
      7 PASS        SRR098026.fastq
dcuser@ip-172-31-63-201:~/dc_workshop/docs$ grep FAIL fastqc_summaries.txt | cut -f1,3
| uniq -c | sort -rn
      3 FAIL        SRR098028.fastq
      3 FAIL        SRR098027.fastq
      3 FAIL        SRR098026.fastq
      2 FAIL        SRR098283.fastq
      2 FAIL        SRR098281.fastq
      1 FAIL        SRR097977.fastq
```

```
dcuser@ip-172-31-63-201:~/dc_workshop/docs$ grep WARN fastqc_summaries.txt | cut -
f1,3 | uniq -c | sort -rn
      2 WARN	SRR098283.fastq
      2 WARN	SRR098281.fastq
      2 WARN	SRR098028.fastq
      2 WARN	SRR098026.fastq
      2 WARN	SRR097977.fastq

grep FAIL fastqc_summaries.txt | uniq -f3 | sort

$grep FAIL ~/dc_workshop/docs/fastqc_summaries.txt |cut -f3 |uniq -c
      1 SRR097977.fastq
      3 SRR098026.fastq
      3 SRR098027.fastq
      3 SRR098028.fastq
      2 SRR098281.fastq
      2 SRR098283.fastq
$grep PASS ~/dc_workshop/docs/fastqc_summaries.txt |cut -f3 |uniq -c
      9 SRR097977.fastq
      7 SRR098026.fastq
      9 SRR098027.fastq
      7 SRR098028.fastq
      8 SRR098281.fastq
      8 SRR098283.fastq
$grep WARN ~/dc_workshop/docs/fastqc_summaries.txt |cut -f3 |uniq -c
      2 SRR097977.fastq
      2 SRR098026.fastq
      2 SRR098028.fastq
      2 SRR098281.fastq
      2 SRR098283.fastq

$ grep FAIL fastqc_summaries.txt | cut -f3 | sort | uniq -c
      1 SRR097977.fastq
      3 SRR098026.fastq
      3 SRR098027.fastq
      3 SRR098028.fastq
      2 SRR098281.fastq
      2 SRR098283.fastq
$
```

Trimming:
java -jar ~/Trimmomatic-0.32/trimmomatic-0.32.jar SE SRR098283.fastq
SRR098283.fastq_trim.fastq SLIDINGWINDOW:4:20 MINLEN:20

Kept: 78.98%
Dropped: 21.02%

*drumrolls: 21564058 Surviving: 17 mln (78.98%) Dropped: 4.5 mln (21.02%)

Discarded: 21.02%
Kept reads: 78.98%

Input Reads: 21564058 Surviving: 17030985 (78.98%) Dropped: 4533073 (21.02%)

Input Reads: 21564058 Surviving: 17030985 (78.98%) Dropped: 4533073 (21.02%)

Dropped: 21.02, survived:
78.98

File: SRR098283.fastq
Dropped 21.02%
Survived 78.98%

Input Reads: 21564058 Surviving: 17030985 (78.98%) Dropped: 4533073 (21.02%)

Input Reads: 21564058 Surviving: 17030985 (78.98%) Dropped: 4533073 (21.02%)

```
for infile in *.fastq
do
outfile="${infile}"_trim.fastq
java -jar ~/Trimmomatic-0.32/trimmomatic-0.32.jar SE "${infile}" "${outfile}"
SLIDINGWINDOW:4:20 MINLEN:20
done
```

Look at the first read in SRR098026.fastq file. After filtering out the bad reads, what is the first remaining read for this sample? What does the quality look like?

Exercise
Earlier we looked at the first read in our SRR098026.fastq file and saw that it was very poor quality.

```
$ head -n4 SRR098026.fastq
```

```
@SRR098026.1 HWUSI-EAS1599_1:2:1:0:968 length=35
NNNNNNNNNNNNNNNNCNNNNNNNNNNNNNNNNN
+SRR098026.1 HWUSI-EAS1599_1:2:1:0:968 length=35
!!!!!!!!!!!!!!!#!!!!!!!!!!!!!!!!!!
```

After filtering out bad reads, what is the first remaining read for this sample? What does its quality look like?

first read before trimming:
@SRR098026.1 HWUSI-EAS1599_1:2:1:0:968 length=35
NNNNNNNNNNNNNNNNCNNNNNNNNNNNNNNNNN
+SRR098026.1 HWUSI-EAS1599_1:2:1:0:968 length=35
!!!!!!!!!!!!!!!!#!!!!!!!!!!!!!!!!!!

first read after trimming:
@SRR098026.342 HWUSI-EAS1599_1:2:1:3:655 length=35
GGATNGGCCTTGTATTTATGATTCTCNGAGTCTGT
+SRR098026.342 HWUSI-EAS1599_1:2:1:3:655 length=35
BB@B!B@AACBBABCCCCBBBBBB@@!B?B<ABB@

$ head SRR098026.fastq_trim.fastq
@SRR098026.342 HWUSI-EAS1599_1:2:1:3:655 length=35
GGATNGGCCTTGTATTTATGATTCTCNGAGTCTGT
+SRR098026.342 HWUSI-EAS1599_1:2:1:3:655 length=35
BB@B!B@AACBBABCCCCBBBBBB@@!B?B<ABB@
@SRR098026.343 HWUSI-EAS1599_1:2:1:3:1865 length=35
CCCGNATCTGGCGTTTGTTGATGG
+SRR098026.343 HWUSI-EAS1599_1:2:1:3:1865 length=35
B?A@!BBAB?AAA@BBC@@;>A>;
@SRR098026.344 HWUSI-EAS1599_1:2:1:3:1322 length=35
ATCANGGCACTGATGTCTTCAGTACG
$ head SRR098028.fastq
@SRR098028.1 HWUSI-EAS1599_1:1:1:0:45 length=35
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
+SRR098028.1 HWUSI-EAS1599_1:1:1:0:45 length=35
!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!
@SRR098028.2 HWUSI-EAS1599_1:1:1:0:435 length=35
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
+SRR098028.2 HWUSI-EAS1599_1:1:1:0:435 length=35
!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!
@SRR098028.3 HWUSI-EAS1599_1:1:1:0:807 length=35
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
$

 head -n 4 SRR098026.fastq*
==> SRR098026.fastq <==
@SRR098026.1 HWUSI-EAS1599_1:2:1:0:968 length=35
NNNNNNNNNNNNNNNNCNNNNNNNNNNNNNNNNN
+SRR098026.1 HWUSI-EAS1599_1:2:1:0:968 length=35
!!!!!!!!!!!!!!!!#!!!!!!!!!!!!!!!!!!
==> SRR098026.fastq_trim.fastq <==
@SRR098026.342 HWUSI-EAS1599_1:2:1:3:655 length=35
GGATNGGCCTTGTATTTATGATTCTCNGAGTCTGT
+SRR098026.342 HWUSI-EAS1599_1:2:1:3:655 length=35
BB@B!B@AACBBABCCCCBBBBBB@@!B?B<ABB@

```
$ head -n4 SRR098026.fastq
@SRR098026.1 HWUSI-EAS1599_1:2:1:0:968 length=35
NNNNNNNNNNNNNNNNCNNNNNNNNNNNNNNNNN
+SRR098026.1 HWUSI-EAS1599_1:2:1:0:968 length=35
!!!!!!!!!!!!!!!!#!!!!!!!!!!!!!!!!!!
$ head -n4 SRR098026.fastq_trim.fastq
@SRR098026.342 HWUSI-EAS1599_1:2:1:3:655 length=35
GGATNGGCCTTGTATTTATGATTCTCNGAGTCTGT
+SRR098026.342 HWUSI-EAS1599_1:2:1:3:655 length=35
BB@B!B@AACBBABCCCCBBBBBB@@!B?B<ABB@

dcuser@ip-172-31-63-201:~/dc_workshop/data/untrimmed_fastq$ head -4
SRR098026.fastq_trim.fastq
@SRR098026.342 HWUSI-EAS1599_1:2:1:3:655 length=35
GGATNGGCCTTGTATTTATGATTCTCNGAGTCTGT
+SRR098026.342 HWUSI-EAS1599_1:2:1:3:655 length=35
BB@B!B@AACBBABCCCCBBBBBB@@!B?B<ABB@
dcuser@ip-172-31-63-201:~/dc_workshop/data/untrimmed_fastq$ head -4 SRR098026.fastq
@SRR098026.1 HWUSI-EAS1599_1:2:1:0:968 length=35
NNNNNNNNNNNNNNNNCNNNNNNNNNNNNNNNNN
+SRR098026.1 HWUSI-EAS1599_1:2:1:0:968 length=35
!!!!!!!!!!!!!!!!#!!!!!!!!!!!!!!!!!!

$ head -n4 SRR098026.fastq
@SRR098026.1 HWUSI-EAS1599_1:2:1:0:968 length=35
NNNNNNNNNNNNNNNNCNNNNNNNNNNNNNNNNN
+SRR098026.1 HWUSI-EAS1599_1:2:1:0:968 length=3
!!!!!!!!!!!!!!!!#!!!!!!!!!!!!!!!!!!
$ head -n4 SRR098026.fastq_trim.fastq
@SRR098026.342 HWUSI-EAS1599_1:2:1:3:655 length=35
GGATNGGCCTTGTATTTATGATTCTCNGAGTCTGT
+SRR098026.342 HWUSI-EAS1599_1:2:1:3:655 length=35
BB@B!B@AACBBABCCCCBBBBBB@@!B?B<ABB@

$ head -4 SRR098026.fastq
@SRR098026.1 HWUSI-EAS1599_1:2:1:0:968 length=35
NNNNNNNNNNNNNNNNCNNNNNNNNNNNNNNNNN
+SRR098026.1 HWUSI-EAS1599_1:2:1:0:968 length=35
!!!!!!!!!!!!!!!!#!!!!!!!!!!!!!!!!!!
$ head -4 SRR098026.fastq_trim.fastq
@SRR098026.342 HWUSI-EAS1599_1:2:1:3:655 length=35
GGATNGGCCTTGTATTTATGATTCTCNGAGTCTGT
+SRR098026.342 HWUSI-EAS1599_1:2:1:3:655 length=35
BB@B!B@AACBBABCCCCBBBBBB@@!B?B<ABB@
```

```
#move trimmed data
cd ~/dc_workshop/data/untrimmed_fastq
mkdir ../trimmed_fastq
rm SRR098283.fastq_trim.fastq_trim.fastq
mv *_trim* ../trimmed_fastq
cd ../trimmed_fastq
ls
cd ~/dc_workshop/
#make symbolic links to trimmed data
ln -s ~/.dc_sampledata_lite/trimmed_fastq_small/ data/
#to reference genome
ln -s ~/.dc_sampledata_lite/ref_genome/ data/
#make new directories for output files
mkdir -p results/sai results/sam results/bam results/bcf results/vcf
#make an index with bwa
bwa index data/ref_genome/ecoli_rel606.fasta

#make an alignment
bwa aln data/ref_genome/ecoli_rel606.fasta data/trimmed_fastq_small
/SRR097977.fastq_trim.fastq > results/sai/SRR097977.aligned.sai

#convert to sam format
bwa samse data/ref_genome/ecoli_rel606.fasta \
  results/sai/SRR097977.aligned.sai \
  data/trimmed_fastq_small/SRR097977.fastq_trim.fastq > \
  results/sam/SRR097977.aligned.sam

#convert to bam format for downstream analysis
samtools view -S -b results/sam/SRR097977.aligned.sam > results/bam/
SRR097977.aligned.bam

#sort by coordinates
samtools sort results/bam/SRR097977.aligned.bam results/bam/SRR097977.aligned.sorted

#snp calling
samtools mpileup -g -f data/ref_genome/ecoli_rel606.fasta \
 results/bam/SRR097977.aligned.sorted.bam > results/bcf/SRR097977_raw.bcf

 bcftools view -bvcg results/bcf/SRR097977_raw.bcf > results/bcf/SRR097977_variants.bcf

bcftools view results/bcf/SRR097977_variants.bcf | /usr/share/samtools/vcfutils.pl varFilter
> results/vcf/SRR097977_final_variants.vcf

##########
# Exercise #
##########
```

Use the grep, cut, and less commands you've learned to extract the POS and QUAL columns from your output file (without the header lines). What is the position of the first variant to be

called with a QUAL value of less than 4?

$ tail -n50 results/vcf/SRR097977_final_variants.vcf | cut -f2,6
First position with Phred below 4: 1294137

tail -n +37 SRR097977_final_variants.vcf | cut -f2,6
Position 1294137 has a quality of 3.54

1294137
$ grep -v '#' results/vcf/SRR097977_final_variants.vcf | cut -f2,6

dcuser@ip-172-31-63-201:~/dc_workshop$ grep -v "#" results/vcf/
SRR097977_final_variants.vcf | cut -f2,6 | awk '$2<=4'| head
1294137 3.54
1505898 3.01
3075282 3.01
4439380 3.02

1294137 with a Qual 3.54
cut results/vcf/SRR097977_final_variants.vcf -f 6,2 | grep -v "##" | sort -k2 -n

#determine columns of POS and QUAL and write to new file
cut -f 2,6 SRR097977_final_variants.vcf > pos_qual.vcf
#Delete header lines by deleting lines containg "#"
sed -i.bak '/#/d' ./pos_qual.vcf

#visualization
samtools index results/bam/SRR097977.aligned.sorted.bam
samtools tview results/bam/SRR097977.aligned.sorted.bam data/ref_genome/
ecoli_rel606.fasta

files for IGV, to be copied by FileZilla:

~/dc_workshop/results/bam/SRR097977.aligned.sorted.bam
~/dc_workshop/results/bam/SRR097977.aligned.sorted.bam.bai
~/dc_workshop/data/ref_genome/ecoli_rel606.fasta
~/dc_workshop/results/vcf/SRR097977_final_variants.vcf

IGV:
1. Open IGV.
2. Load our reference genome file (ecoli_rel606.fasta) into IGV using the **“Load Genomes from File…“** option under the **“Genomes”** pull-down menu.
3. Load our BAM file (SRR097977.aligned.sorted.bam) using the **“Load from File…“** option under the **“File”** pull-down menu.

4. Do the same with our VCF file (SRR097977_final_variants.vcf).

NC_012967.1:768,349-768,405

##########
##DAY 2####
##########