y#######################################
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

The Etherpad of Day 1 is archived here:
https://aschuerch.github.io/Microbial-Genomics-2018/files/2018-04-03-Utrecht-Day1.pdf

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

https://software-carpentry.org/
http://www.datacarpentry.org/
Carpentries-NL mailing list:
https://groups.google.com/forum/#!forum/carpentries-nl

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:

Ruben
Ramon
- Reinder
- Tony
- Sarah
- Fabian
- Rozemarijn
-Jorik
Sam
Cindy
Lotte
Ethel
Div

Setup

Please install these tools for the lessons tomorrow:
Figtree http://tree.bio.ed.ac.uk/software/figtree/
Bandage http://rrwick.github.io/Bandage/
##################

##

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
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.comA
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

##

Explanation:

Our goal; Compare our citrate metabolizing E.Coli strain against the E.Coli REL606 reference strain and determine what genetic adaptation allows our strain to do this.

Our method;

1. index our reference strain (bwa index)
2. align read against e.coli reference (bwa aln --> outputs: sai-file)
3. sai is an older format, we like to work on SAM/BAM files. Conversion happens with (bwa samse --> outputs: sam-file)
4. Sam is human readable, computers are more efficient with binary. So convert sam to bam (binary equivalent of sam) using (samtools view -S -b --> outputs: bam-file)
5. Sort the file to make subsequent processes more efficient, using (samtools sort --> outputs: sorted.bam-file)
6. Identify for each nt-position in the reference, what data aligns there, e.g. at position 21 we have 32 A's and 2 T's, do this for this the entire genome. (samtools mpileup --> bcf-file)
7. Now we have to factor in the quality score of these basecalls (from mpileup), this is done using (bcftools view -bvcg --> outputs: variants.bcf)
8. Filter the SNPs identified in step 7 based on quality criteria. (bcftools view | vcftutils.pl varFilter --> outputs: final_variants.vcf)
9. Visualize the data using IGV, for this you require the reference genome .fasta, the alignment .bam-file (and it's index --> .bam.bai), and the final .vcf-file.

Script:

```
ls

Coffie break until 11.00


################################
######## SCRIPT          ########
################################
cd ~/dc_workshop/results

mkdir -p sai sam bam bcf vcf

genome=~/dc_workshop/data/ref_genome/ecoli_rel606.fasta

bwa index $genome

for fq in ~/dc_workshop/data/trimmed_fastq_small/*.fastq
do
    base=$(basename $fq .fastq_trim.fastq)
    echo "Working with file $fq"
    echo "Basename is $base"

    fq=~/dc_workshop/data/trimmed_fastq_small/${base}.fastq_trim.fastq
    sai=~/dc_workshop/results/sai/${base}_aligned.sai
```

```
    sam=~/dc_workshop/results/sam/${base}_aligned.sam
    bam=~/dc_workshop/results/bam/${base}_aligned.bam
    sorted_bam=~/dc_workshop/results/bam/${base}_aligned_sorted.bam
    raw_bcf=~/dc_workshop/results/bcf/${base}_raw.bcf
    variants=~/dc_workshop/results/bcf/${base}_variants.bcf
    final_variants=~/dc_workshop/results/vcf/${base}_final_variants.vcf

    bwa aln $genome $fq > $sai
    bwa samse $genome $sai $fq > $sam
    samtools view -S -b $sam > $bam
    samtools sort -f $bam $sorted_bam
    samtools index $sorted_bam
    samtools mpileup -g -f $genome $sorted_bam > $raw_bcf
    bcftools view -bvcg $raw_bcf > $variants
    bcftools view $variants | /usr/share/samtools/vcfutils.pl varFilter - > $final_variants
done


#########################
##R for microbial genomics###
#########################

download.file("https://raw.githubusercontent.com/datacarpentry/R-genomics/gh-pages/data/
Ecoli_metadata.csv",
  "data/Ecoli_metadata.csv")

metadata <- read.csv("data/Ecoli_metadata.csv")

1) #What is the class of the object metadata?
dataframe
dataframe
class = data.frame
data.frame
data.frame
data.frame
data frame
class = "data.frame"
dataframe

2) How may rows and how many columns are in this object?
7, 30
30 rows, 7 columns
rows = 30, colums = 7
7 columns, 30 rows
7 columns, 30 rows
30x7
row 30, col 7
7,30
```

row = 30, column =7
row=30, column=7

3) How many citrate+ mytants have been recorded in this population?
cit plus mutants: 9

(C1,C2)	C1	C2	C3	Cit+	UC	unknown	NA's
3	5	6	2	9	2	2	1

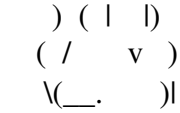
9
9 cit plus, 9 cit minus, rest is unknown
9 plus, 9 minus, 12 unknown
9 citrate+ mutants (total; 9 minus, 9 plus, 12 unkown
cit+ 9
c2 :6
C1:5
C1,C2: 3
C3:2
other 4
NA's 1
cit+: 9
summary(metadata) ---> 9 cit+ mutants

If you have answered all the questions it is time for the:

Lunch break until 13.15
really???? thats great sorry, typo :-(this really broke my heart
time to eat <>< ~~~~~<^><
yeah food

```
ggplot(metadata) +
  geom_boxplot(aes(x = cit, y = genome_size, fill = cit)) +
  ggtitle('Boxplot of genome size by citrate mutant type') +
  xlab('citrate mutant') +
  ylab('genome size') +
  theme(panel.grid.major = element_line(size = .5, color = "yellow"),
        axis.text.x = element_text(angle=45, hjust=1),
        axis.title = element_text(size = rel(1.5)),
        axis.text = element_text(size = rel(1.25)))
```

A Quick Guide to Organizing Computational Biology Projects
http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1000424
Genomic analysis of a key innovation in an experimental Escherichia coli population by Blount ZD, Barrick JE, Davidson CJ, and Lenski RE.
https://www.nature.com/articles/nature11514


Thank you!