THE CARPENTRIES

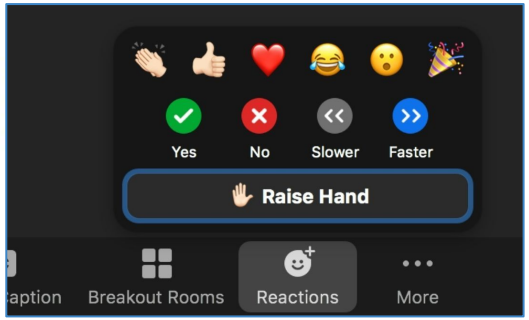
مدخل في لغة الآر (R) والعلوم المفتوحة لعلماء المعلوماتية الحيوية

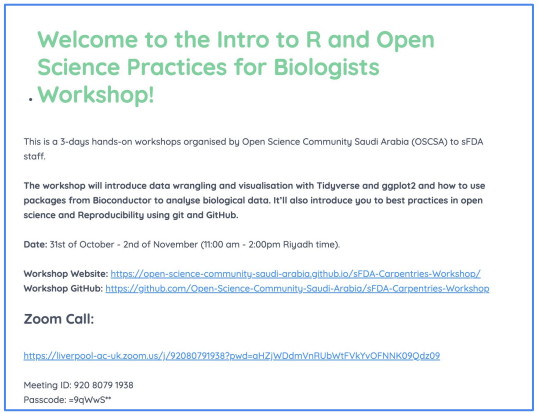
Batool Almarzouq
Monah Abou Alezz
Annajiat Alim Rase
Abdulrahman Dallak
Mona Alsharif

OSCSA
Open Science Community Saudi Arabia
مجتمع العلوم المفتوحة في المملكة العربية السعودية

Day 2
اليوم الثاني

قواعد الدورة




Welcome to the Intro to R and Open Science Practices for Biologists Workshop!

This is a 3-days hands-on workshops organised by Open Science Community Saudi Arabia (OSCSA) to sFDA staff.

The workshop will introduce data wrangling and visualisation with Tidyverse and ggplot2 and how to use packages from Bioconductor to analyse biological data. It'll also introduce you to best practices in open science and Reproducibility using git and GitHub.

Date: 31st of October - 2nd of November (11:00 am - 2:00pm Riyadh time).

Workshop Website: https://open-science-community-saudi-arabia.github.io/sFDA-Carpentries-Workshop/
Workshop GitHub: https://github.com/Open-Science-Community-Saudi-Arabia/sFDA-Carpentries-Workshop

Zoom Call:

https://liverpool-ac-uk.zoom.us/j/92080791938?pwd=aHZjWDdmVnRUbWtFVkYvOFNNK09Qdz09

Meeting ID: 920 8079 1938
Passcode: =9qWwS**

- كتابة الاسم في Etherpad
- إطفاء الجوال
- القواعد السلوكية
- لاتتردد في طرح **أي سؤال** خلال الجلسة
- الإجابة على التمارين خلال الدورة

	Online (online) ** **Instructors:** Batool Almarzouq, Monah Abou Alezz, Annajiat Alim Rasel	Oct 31 - Nov 2, 2022
	Griffith University **Instructors:** Emilia Decker, Amanda Miotto, Belinda Weaver	Nov 1 - Nov 3, 2022
	University of Idaho **Instructors:** Bernard Ricca, Vicki M. Zhang, Max Czapanskiy	Nov 1 - Nov 3, 2022
	Aarhus University ** **Instructors:** Adela Sobotkova **Helpers:** Max Odsberg	Nov 1 - Dec 9, 2022
	University of Texas at Austin **Instructors:** Emily Beagle, Meryl Brodsky, Dianna Morganti, Michael Shensky, Lydia Tressel **Helpers:** Jessica Simpson, Allyssa Guzman, Jeremy Thompson, Ian Goodale	Nov 2 - Nov 4, 2022
	University of Oslo ** **Instructors:** Olga Silantyeva, Désirée Treichler **Helpers:** Frederico Bianchini	Nov 4 - Nov 4, 2022
	Stockholm Trio university libraries ** **Instructors:** Thomas Lind, Stefan Wiens, Glenn Haya, Lina Andrén **Helpers:** Joakim Philipson, Erik Hedman	Nov 7 - Nov 8, 2022
	Friedrich Schiller University Jena **Instructors:** Cora Assmann, Christian Knüpfer, Philipp Matthias Schäfer	Nov 8 - Nov 22, 2022
	University College London ** **Instructors:** Heather Kelly, Sarah Jaffa **Helpers:** Brian Alston	Nov 8 - Nov 10, 2022
	Lancaster University ** **Instructors:** Nilani Ganeshwaran, Phil Reed, Carlene Barton	Nov 9 - Nov 9, 2022

Interaction is important in virtual settings

Today's Session Objective

- Be able to create the most common **R objects** including vectors
- Understand that **vectors** have modes, which correspond to the type of data they contain
- Be able to use **arithmetic operators** on R objects
- Be able to **retrieve (subset), name, or replace, values** from a vector
- Be able to use **logical operators** in a subsetting operation
- Understand that lists can hold data of more than one mode and can be **indexed**
- Explain the basic principle of **tidy datasets**
- Be able to **load a tabular dataset** using base R functions
- Be able to **determine the structure of a data** frame including its dimensions and the datatypes of variables
- Be able to subset/retrieve values from a data frame
- Understand how R may coerce data into different modes
- Be able to change the mode of an object
- Understand that **R uses factors** to store and manipulate categorical data

nature

Explore content ⌄ About the journal ⌄ Publish with us ⌄

nature > articles > article

Published: 01 August 2016

Tempo and mode of genome evolution in a 50,000-generation experiment

Olivier Tenaillon, Jeffrey E. Barrick, Noah Ribeck, Daniel E. Deatherage, Jeffrey L. Blanchard, Aurko Dasgupta, Gabriel C. Wu, Sébastien Wielgoss, Stéphane Cruveiller, Claudine Médigue, Dominique Schneider & Richard E. Lenski ✉

Nature **536**, 165–170 (2016) | Cite this article

12 populations of E-coli were propagated for
more than **50,000 generations** in a **glucose-limited minimal medium + citrate**

12 populations of E-coli were propagated for
more than **50,000 generations** in a **glucose-limited minimal medium + citrate**

citrate-using mutants (Cit+) appeared in a population of E.coli (designated **Ara-3**)
at around **31,000 generations**

12 populations of E-coli were propagated for
more than **50,000 generations** in a **glucose-limited minimal medium + citrate**

citrate-using mutants (Cit+) appeared in a population of E.coli (designated **Ara-3**)
at around **31,000 generations**

12 populations of E-coli were propagated for
more than **50,000 generations** in a **glucose-limited minimal medium + citrate**

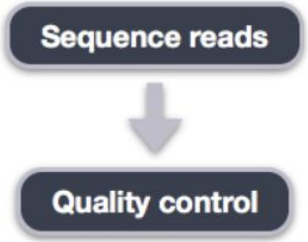
citrate-using mutants (Cit+) appeared in a population of E.coli (designated **Ara-3**)
at around **31,000 generations**

Ara-4 population appeared (**hypermutable**)

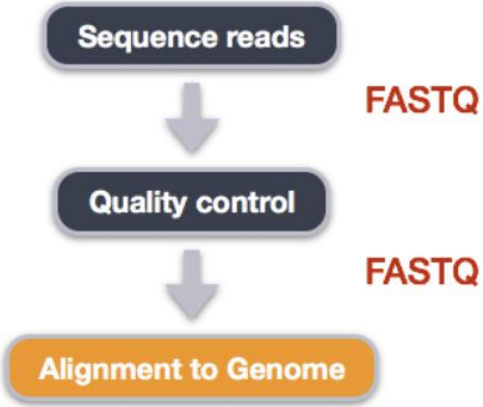
Questions

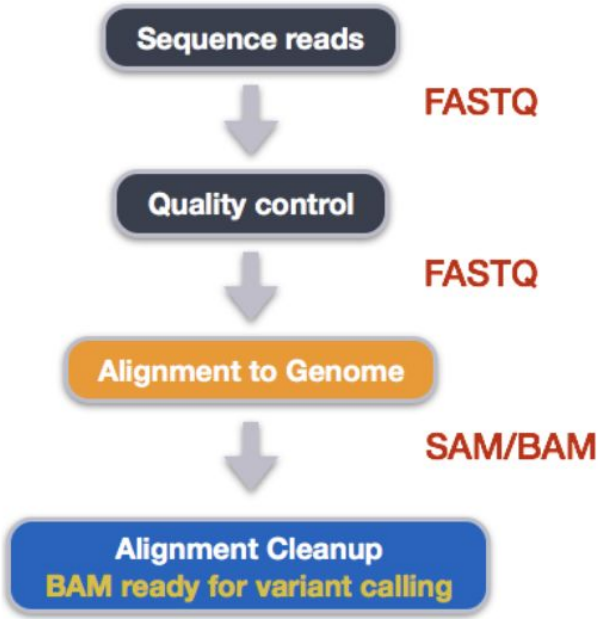
- How many base pair changes are there between the Cit+ and Cit- strains?
- What are the base pair changes between strains?

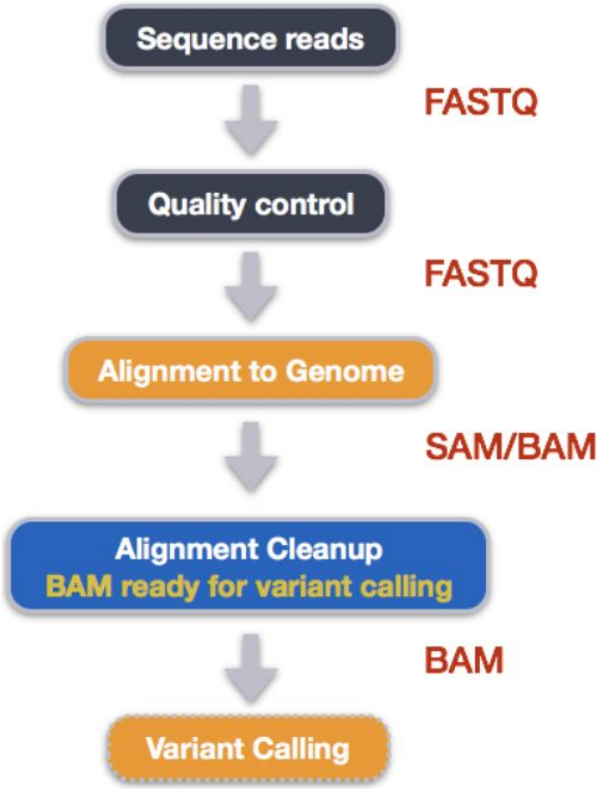
Sequence reads



FASTQ


Sequence reads
FASTQ
Quality control
FASTQ
Alignment to Genome


Sequence reads
FASTQ
Quality control
FASTQ
Alignment to Genome
SAM/BAM
Alignment Cleanup
BAM ready for variant calling


Sequence reads
FASTQ
Quality control
FASTQ
Alignment to Genome
SAM/BAM
Alignment Cleanup
BAM ready for variant calling
BAM
Variant Calling
VCF

```
CHROM POS ID REF ALT QUAL FILTER   INFO   FORMAT     Sample1
```

CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	Sample1
20	17330	.	T	A	67	PASS	DP=27;AF=0.444	GT:AD:DP	0/1:15,12:27

Code

```
##fileformat=VCFv4.2
##FILTER=<ID=PASS,Description="All filters passed">
##bcftoolsVersion=1.8+htslib-1.8
##bcftoolsCommand=mpileup -O b -o results/bcf/SRR2584866_raw.bcf -f data/ref_genome/ecoli_rel606.fasta results/bam/SRR2584866.aligned.sorted.
bam
##reference=file://data/ref_genome/ecoli_rel606.fasta
##contig=<ID=CP000819.1,length=4629812>
##ALT=<ID=*,Description="Represents allele(s) other than observed.">
##INFO=<ID=INDEL,Number=0,Type=Flag,Description="Indicates that the variant is an INDEL.">
##INFO=<ID=IDV,Number=1,Type=Integer,Description="Maximum number of reads supporting an indel">
##INFO=<ID=IMF,Number=1,Type=Float,Description="Maximum fraction of reads supporting an indel">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Raw read depth">
##INFO=<ID=VDB,Number=1,Type=Float,Description="Variant Distance Bias for filtering splice-site artefacts in RNA-seq data (bigger is bette
r)",Version=
##INFO=<ID=RPB,Number=1,Type=Float,Description="Mann-Whitney U test of Read Position Bias (bigger is better)">
##INFO=<ID=MQB,Number=1,Type=Float,Description="Mann-Whitney U test of Mapping Quality Bias (bigger is better)">
##INFO=<ID=BQB,Number=1,Type=Float,Description="Mann-Whitney U test of Base Quality Bias (bigger is better)">
##INFO=<ID=MQSB,Number=1,Type=Float,Description="Mann-Whitney U test of Mapping Quality vs Strand Bias (bigger is better)">
##INFO=<ID=SGB,Number=1,Type=Float,Description="Segregation based metric.">
##INFO=<ID=MQ0F,Number=1,Type=Float,Description="Fraction of MQ0 reads (smaller is better)">
##FORMAT=<ID=PL,Number=G,Type=Integer,Description="List of Phred-scaled genotype likelihoods">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##INFO=<ID=ICB,Number=1,Type=Float,Description="Inbreeding Coefficient Binomial test (bigger is better)">
##INFO=<ID=HOB,Number=1,Type=Float,Description="Bias in the number of HOMs number (smaller is better)">
##INFO=<ID=AC,Number=A,Type=Integer,Description="Allele count in genotypes for each ALT allele, in the same order as listed">
##INFO=<ID=AN,Number=1,Type=Integer,Description="Total number of alleles in called genotypes">
##INFO=<ID=DP4,Number=4,Type=Integer,Description="Number of high-quality ref-forward , ref-reverse, alt-forward and alt-reverse bases">
##INFO=<ID=MQ,Number=1,Type=Integer,Description="Average mapping quality">
##bcftools_callVersion=1.8+htslib-1.8
##bcftools_callCommand=call --ploidy 1 -m -v -o results/bcf/SRR2584866_variants.vcf results/bcf/SRR2584866_raw.bcf; Date=Tue Oct  9 18:48:10
2018
```