Global Pneumococcal Sequencing Project

CDC Centers for Disease Control and Prevention

wellcome sanger institute

Hello participants,

We are looking forward to welcoming you to the **Pneumococcal Genomics A to Z workshop**. This workshop is a combination of theory and practical exercises. Please see below for the agenda.

| Time | Content |
|---|---|
| 13:00 - 13:20 | Landscape of whole-genome sequencing technology |
| 13:20 - 13:40 | From DNA to FASTQ |
| 13:40 - 14:00 | GPS Pipeline for data analysis **(Task 1)** |
| 14:00 - 14:15 | Tea break (while computer runs the pipeline) |
| 14:15 - 14:40 | Create Microreact instance to interactively view phylogeny and epidemiological data **(Task 2)** |
| 14:40 - 15:00 | Statistical test to evaluate vaccine impact **(Task 3)** |
| 15:00 - 15:20 | Plot figure to evaluate vaccine impact **(Task 4)** |
| 15:20 - 15:30 | Anatomy of a manuscript |
| 15:30 - 15:45 | Online bioinformatics training resources |
| 15:45 - 16:00 | Finishing up |

In the practical, we will complete four tasks based on a real-world study reported by Argentinian scientists. Please try to read the paper Population genetic structure, serotype distribution and antibiotic resistance of *Streptococcus pneumoniae* causing invasive disease in children in Argentina (DOI: 10.1099/mgen.0.000636) in advance of the workshop. We will aim to reproduce the figures and tables, with statistical supports.

**In advance of the workshop, please:**

1. Read the Argentina paper
2. Download the materials: https://github.com/sanger-bentley-group/AtoZ_PneumoGenomics/archive/refs/heads/main.zip
3. Setup GPS Pipeline (Page 2 - Section 1)
4. Install R and RStudio, and several R packages (Page 3 - 4)
5. Be familiar with the practical workflow (Page 5)

Global Pneumococcal Sequencing Project

CDC Centers for Disease Control and Prevention

wellcome sanger institute

# GPS PIPELINE

# Quickstart Guide

## Requirements

- Compatible with most operating systems: Linux, Windows (running Linux with WSL2), macOS
- Java 11+ or OpenJDK 11+
- Docker or Singularity/Apptainer
- Have at least 16GB of RAM and 50GB of free storage

## Setup (Internet connection required)

1. Download or Git Clone the pipeline core files from its GitHub Repository
   a. Download from: https://github.com/sanger-bentley-group/gps-pipeline/releases
   b. To clone, run: `git clone https://github.com/sanger-bentley-group/gps-pipeline.git`
2. Initialise the pipeline after changing directory (`cd`) into the pipeline directory:
   a. Using Docker: `./run_pipeline --init`
   b. Using Singularity: `./run_pipeline --init -profile singularity`
3. This can take a while, as it will download 13GB of container images and 8GB of databases

## Run (No internet connection required after initialisation)

1. Run the pipeline with the directory containing your FASTQ files as the input using `--reads`
   a. Using Docker: `./run_pipeline --reads /path/to/reads-dir`
   b. Using Singularity: `./run_pipeline --reads /path/to/reads-dir -profile singularity`
2. Grab a cup of tea and wait

**Tip 1:**
If you have not specified output path with `--output`, the default is the `output` directory in the pipeline directory.

**Tip 2:**
Each input sample will generate ~2GB intermediate files on average. You might need to process your samples in batches if the storage space is limited on your system. The `clean_pipeline` helper script of the pipeline may be useful after each successful run.

## Documentation

- GitHub Repository: https://github.com/sanger-bentley-group/gps-pipeline

## Notice

- The current release of the pipeline only works with Illumina paired-end short reads
- Use a specific version of the pipeline to ensure consistent output for the same study

Global Pneumococcal Sequencing Project

CDC Centers for Disease Control and Prevention

wellcome sanger institute

R & RStudio®

# Installation Guide

## Requirements

- Have admin access (administrative privileges) for your machine to be able to install programs.
- Have at least 2GB of RAM

## Install R

### For a macOS computer

1. Click on the Apple logo in the top left of the screen to open the Apple menu. Click "About This Mac"
   a. If you have "Chip" listed here, you have a Apple Silicon Mac.
   b. If you have "Processor" listed here, you have an Intel Mac.
2. Go to https://cran.rstudio.com/bin/macosx/
3. Of the top two links on the left of the page, click the one that applies to your Mac (Intel or Apple Silicon) to download the installer. Once this has downloaded, double click the downloaded installer to run it.
4. The installer will walk you through installing R. The default settings will work perfectly!

### For a Windows computer

1. Go to https://cran.rstudio.com/bin/windows/base/
2. Click "Download R-x for Windows" to download the installation file, R-x-win.exe, where X is the latest version of R. Once this has downloaded, double click on R-x-win.exe to run it.
3. Click through the installation wizard to install R.

### For a Linux (Ubuntu*) computer

1. Go to https://cran.rstudio.com/bin/linux/ubuntu/
2. Run the following commands to install R:

```
sudo apt update -qq
```

```
sudo apt install --no-install-recommends software-properties-common dirmngr
```

```
wget -qO- https://cloud.r-project.org/bin/linux/ubuntu/marutter_pubkey.asc |
sudo tee -a /etc/apt/trusted.gpg.d/cran_ubuntu_key.asc
```

```
sudo add-apt-repository "deb https://cloud.r-project.org/bin/linux/ubuntu $(lsb_
release -cs)-cran40/"cran_ubuntu_key.asc
```

```
sudo apt install --no-install-recommends r-base
```

## Install RStudio

1. Go to https://posit.co/download/rstudio-desktop/
2. Scroll down to "All Installers and Tarballs". Below this, there is a list of different Operating Systems (OSs). Find your OS and click the link in the "Download" column.
   a. For **macOS**, double click the downloaded file to open the installer. Drag and drop the "RStudio" logo into the "Applications" folder in the window that pops up.
   b. For **Windows**, double click the downloaded file to run the RStudio installer. Follow the steps and use the default settings.
   c. For **Linux (Ubuntu)**, double click on the downloaded file to install RStudio.

* If you're not using Ubuntu, guides to install R for other Linux distributions can be found on https://cran.rstudio.com/bin/linux/

# R Packages Installation

Several R packages are required for this workshop. Run the following commands in RStudio to install the packages:

```
install.packages("ggplot2")
install.packages("ggpattern")
install.packages("plyr")
install.packages("dplyr")
install.packages("tidyr")
install.packages("gridExtra")
install.packages("scales")
install.packages("cowplot")
install.packages("pwr")
install.packages("readr")
```

After Installing the packages, you can load the packages by running the following commands:

```
library("ggplot2")
library("ggpattern")
library("plyr")
library("dplyr")
library("tidyr")
library("gridExtra")
library("scales")
library("cowplot")
library("pwr")
library("readr")
```

# Practical Workflow

We will aim to reproduce the figures and tables, with statistical supports in the Argentina paper.

**Task 1**

- Analyse three pneumococcal genomes using GPS Pipeline

**Task 2**

a. Combine output from the GPS Pipeline and epidemiological data (.csv)

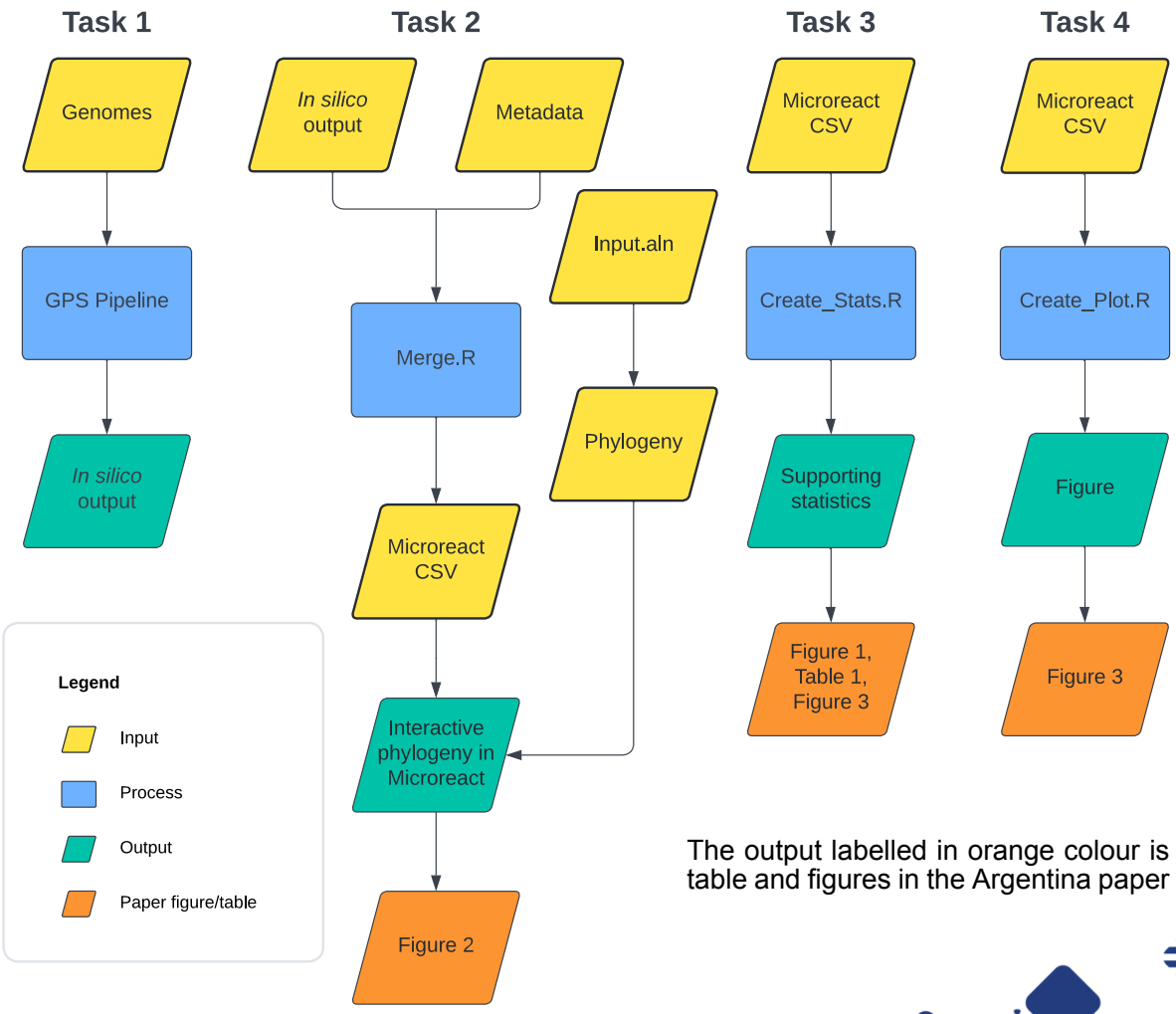
b. Construct a phylogeny using provided alignment (.nwk). Upload .csv and .nwk to Microreact to create an interactive visualisation

**Task 3**

- Evaluate the vaccine impact by detecting changes in serotypes, pneumococcal lineages (GPSC) and antimicrobial resistance using RStudio

**Task 4**

- Plot the vaccine impact in a paper-quality figure using RStudio




The output labelled in orange colour is table and figures in the Argentina paper

# Online Bioinformatics Training

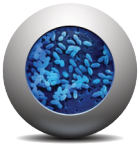
If you are interested in learning more, we offer an online bioinformatics training website at https://training.bactgen.sanger.ac.uk/

We have recently added **Advanced Bioinformatics Course** to the training website, it is designed to analyse *Streptococcus pneumoniae*.

In this course, you will learn about:

i. Setting up your computer for bioinformatics analysis

ii. Genomic data analysis with a focus with on NGS data quality control, assembly and genome annotation, reference mapping and variant calling, *in silico* isolate characterisation (i.e., AMR profiling, serotyping, MLST and lineage classification using PopPUNK), and downstream analyses such as creating a phylogenetic trees and identifying regions of recombination using Gubbins. In each of these sessions, you will learn why each of these aspects are important and how to use individual bioinformatics tools for practical analysis.



Global Pneumococcal Sequencing Project

CDC
Centers for Disease Control and Prevention

wellcome sanger institute

# Contributors

(in alphabetical order by surname)

**Sophie Belman**

**Stephen Bentley**

**Raymond Cheng**

**Sopio Chochua**

**Ana Ferreira**

**Harry Hung**

**Alannah King**

**Yuan Li**

**Stephanie Lo**

**Oliver Lorenz**

**Lesley McGee**

**Jolynne Mokaya**