# Bioinformatic Summer School - Comparative Genomics and Metataxonomic Analysis

The overall teaching aims of this module is to create competence in the students in terms of independently perform bioinformatic analysis on the command line. This includes comparative analysis on bacterial genomes and metataxonomic analysis.

**Specifically, we will:**

For comparative genomics:

* Download all genomes ofa bacterial species.
* Annotate the genetic content of these genomes
* Profile the prophage content of these genomes
* Profile the secondary metabolism of these genomes
* Build a whole-genome phylogeny of the genomes
* Infer the linkage between data-driven analysis and fundamental biology

For metataxonomy:

* Download 16S data
* Demultiplex the data
* Clean, filter and denoise the data
* Taxonomically classify the data
* Statistically analyze the data

The successful student will

1. Have a working knowledge of the Linux command line
2. Be able to use suggested command line tools
3. Investigate the bioinformatic literature to find further tools relevant for the biological questions
4. Infer the biological relevance of the genetic content in bacteria
5. Infer the composition of microbiomes
6. Discuss the biological implications of phylogeny, genetic content, microbial composition and the relatedness of all the above

# Homework

## Linux install and BASH

Installation of a Linux environment. Most bioinformatics is done with the Linux operating system, rather than for example windows. Ubuntu is a modern version of Linux which is free and widely used. If you run windows, you can install an Ubuntu terminal directly with the Ubuntu subsystem – just follow these instructions: <https://ubuntu.com/tutorials/install-ubuntu-on-wsl2-on-windows-10#1-overview>

When you have successfully installed ubuntu, make sure you update it with these two commands:

**sudo apt update**

**sudo apt upgrade**

**\*\*IMPORTANT\*\***

**THE COMMANDS OF THE COURSE ARE SPECIFICALLY MADE FOR A WINDOWS SYSTEM RUNNING THE WINDOWS SUBSYSTEM FOR LINUX (WSL) AS INSTALLED ABOVE – I STRONGLY SUGGEST THAT YOU ONLY USE THAT**

**IF YOU HAVE AN OLD VERSION OF UBUNTU YOUR WINDOWS, GET RID OF IT AND INSTALL AS ABOVE**

**IF YOU HAVE UBUNTU THROUGH VIRTUAL BOX, GET RID OF IT AND INSTALL AS ABOVE**

**IF YOU HAVE ANYTHING OTHER THAN AS DESCRIBED ABOVE, GET RID OF IT AND REINSTALL AS ABOVE**

**WE ALWAYS SPEND WAY TOO MUCH TIME TRYING TO MAKE ALTERNATIVES WORK, SO PLEASE JUST DO AS INSTRUCTED**

**\*\*EVEN MORE IMPORTANT\*\***

**PLEASE JUST DO THE ABOVE, IT WILL SAVE YOU SO MUCH TIME**

For MAC OS X users, you are already running Linux underneath the OS X interface. You have access to the command line with a standard program called ‘Terminal’. The work we will be doing should all work here (but with some minor details).

Regardless of your operating system, make sure you have a working terminal and then do a Linux tutorial here:

[https://app.datacamp.com/learn/courses/introduction-to-bash-scripting](https://app.datacamp.com/learn/courses/introduction-to-bash-scripting?fbclid=IwAR0efqQIXcikI_RaKUHWFK87wBJKvHxkDkuwp1IUgiceqNkA1CQmV0zshmo).

The basic programming language of the Linux command line is called bash, and we will be using this extensively. If you show up with no knowledge of this, you will probably not learn a whole lot, so do the tutorial and play around as much as you can.

## Conda installation

One of the most tricky parts of bioinformatics is the installation of packages. Package A needs version X of package B, but package C need version Y of Package B, which might not be possible. Luckily, the conda package manager takes care of this for us by working out the details of these dependencies and allows us to make individual ‘environments’ for each set of packages for each analysis.

Follow the instructions here (for the **miniconda** installation)

<https://docs.conda.io/projects/conda/en/latest/user-guide/install/linux.html>

For MAC users, follow this:

<https://docs.conda.io/projects/conda/en/latest/user-guide/install/macos.html>

As the first exercise to check your conda installation, run this code (from your ‘base’ environment):

**conda install mamba -n base -c conda-forge**

This will install mamba, a faster implementation of conda. Conda is pretty good, but it has not scaled very well with time, and mamba solves that.

## Installation of R and Rstudio

We need the programming language R for the metataxonomy. R is a statistically minded programming language, which is great for statistics and for plotting – all my plots are made in R.

We will use a great integrated development environment (IDE) for R called Rstudio. Rstudio makes writing and running code exceptionally easy. **Rstudio runs directly in windows (or macOS) so don’t mix it up with Linux!**

First you install R:

<https://www.r-project.org/>

Then you install Rstudio

<https://posit.co/download/rstudio-desktop/>

# Course content:

## Day 1:

Introduction to the command line, the conda package manager and the bash scripting language. Downloading and annotating.

9:00 – 9:15 Introduction and course overview

9:15 – 10:00: Interactive lecture [Le1]: setting up the command line (prior Linux install is expected)

10:00 – 12:00: Interactivate Linux exercises – navigating, copying and loops

12:00 – 13:00: Lunch

13:00 – 14:00: Lecture [Le2]: Basic genome assembly and what we can download.

14:00 – 15:00: Genome download & overview – how & what software?

15:00 – 17:00: Genome annotation – new environments and our first loop

## Day 2:

Finding the functional gene content and building a phylogeny

9:00 – 9:30: Recap and follow up [Le3] – the hard parts of bioinformatics?

9:30 – 10:00: Lecture [Le4] – annotation of phages and BGCs

10:00 – 11:00: Prophage annotation – another loop, another package

11:00 – 12:00: BGC annotation – yet another loop and a different package

12:00 – 13:00: Lunch

13:00 – 14:00: Phylogeny, Trees and Genotypes [Le5]

14:00 – 16:00: Building a core genome phylogenetic tree and finding patterns

## Day 3:

River and presentations?

## Day 4:

Introduction to R and Rstudio, installing R packages. The concepts of metataxonomics and the DADA2 pipeline

9:00 – 9:15 Recap and follow up

9:15 – 10:00: Interactive lecture [Le5]: setting up Rstudio projects

10:00 – 11:00: Interactivate R exercises – navigating, installing and plotting

11:00 – 12:00: Lecture [Le6]: Metataxonomics as a concept.

12:00 – 13:00: Lunch

13:00 – 16:00: Downloading and bioinformatic analysis of data

## Day 2:

Statistical analysis of metataxonomic data.

9:00 – 9:30: Recap and follow up

930 – 12:00: Statistical analysis of metataxonomic data

12:00 – 13:00: Lunch

13:00 – 16:00: Univariate and multivariate analysis