Bioinformatics program for PhD students

Description of the modules
Academic year 2021-2022

R programming and Statistics

Material:
Laptop with R and RStudio installed, following guidelines in the Introduction to R and statistics course.

R programming and Statistics homepage:

Introduction to R and statistics (12h or 18h)

Teachers Blaise Li, Sébastien Mella, Anđela Davidovic

Dates Nov. 2, 3, 4, 5, 2021 (session 1, online : 12h)
Apr. 4, 5, 6, 2022 (session 2, on site : 18h)

Objectives: knowing:

- How to use RStudio
- How to import / export tabulated data
- Data types in R
- Commonly used R functions
- Descriptive statistics
- Descriptive analysis of biological datasets

Prerequisites Be organized with one's files, be able to type "~", "[" and "]" on the keyboard.

Content

- **R and RStudio**
  - Installation, description
- **Programming basics in R**
  - Variables, data types, scripting
  - Functions, arguments
  - How to get help
- **Descriptive statistics on vectors**
  - Statistical distributions
  - Location, dispersion
- **Manipulating more complex data structures**
  - Matrix, List, Dataframe
  - Indexing, Subsetting
  - Correlation coefficient
- **Descriptive analysis of a dataset**
  - Head, summary, table
  - Data distribution with usual statistical descriptors (mean, median, variance, …)
  - Graphical representation: barplot, boxplot, histogram, scatter plot

Hypothesis testing (12h)

Teachers Pascal Campagne, Thomas Obadia

Dates Nov. 8, 9, 10, 15, 2021 (session 1, online)
Apr. 7, 8, 2022 (session 2, on site)

Objectives:

- Be able to select the appropriate statistical test.
- Have a good interpretation and understanding of a confidence interval and a statistical test

Prerequisites R programming and basic statistics

Content

- **Confidence intervals**
  - What is a confidence interval ?
  - How to construct a confidence interval ?
  - Interpretation
- **Hypothesis testing**
  - General definition
  - Different types of risks
  - Compute and interpret a *p-value*
- **Multiple testing**
- **Practical session**

Linear Models (12h)

Teachers Emeline Perthame, Hugo Varet

Dates Nov. 17, 18, 19, 22, 2021 (session 1, online)
Apr. 11, 12, 2022 (session 2, on site)

Objectives:

- Quantify bivariate relationships.
- Understand when and how to apply ANOVA and linear models.
- Interpret the outputs of linear regression and ANOVA models.

Prerequisites R programming and basic statistics
Hypothesis testing

Content

- Correlation
- Simple linear model
- Multiple linear model
- One-way ANOVA
- Two-way ANOVA
- Two-way ANOVA with interaction
- Writing contrasts in ANOVA
- Practical session

Multivariate Analysis (12h)

Teachers Hanna Julienne, Violaine Saint-André

Dates Nov. 23, 24, 25, 26, 2021 (session 1, online)
Apr. 14, 15, 2022 (session 2, on site)

Objectives:

Perform Principal Component Analysis (PCA) and data clustering (hierarchical, k-means)

Prerequisites
Introduction to R and statistics

Content

- We will alternate theoretical and practical sessions to help you perform PCAs and basic clustering (k-means, hierarchical clustering) on multiple types of omics data.
- You will also learn how to display clustering results with PCA components and with heatmaps.

Bioinformatics

Materials
Teaching room will be equipped with computers with pre-installed softwares.

B1 - Unix basic commands (12h)

Teachers Julien Guglielmini, Nicolas Maillet

Dates February 17-18, 2022

Objectives: knowing how to

- Navigate the filesystem, identify files by their path, create, move, rename, delete files and directories
- Extract information from text files

Prerequisites None

Content

- We will alternate theoretical and practical sessions to introduce how a filesystem is organized in a Unix system, how commands can be entered, and the most important commands.

- Once you are able to easily navigate using the command-line in a Unix system, we will teach you how to connect to a distant computer and copy/paste files from your computer to the remote one, and reversely.

B2 - Introduction to sequence analysis (12h)

Teachers Thomas Bigot, Hélène Lopez-Maestre, Damien Mornico

Dates February 24-25, 2022

Objectives:

- Understand biological information available online
- Understand sequence comparisons using local alignment
- Find information about unknown sequences, using online tools

Content

- Online sequences and databanks
- Structure of the information and cross links
- Predict gene features with *ab initio* methods
- Querying data banks with BLAST to find homologues, interpret the results
- Finding resources online to characterize sequences

Prerequisites None

B3 - Proteomics data analysis (12h)

Teachers Quentin Giai Gianetto and Guests

Dates February 28 and March 1st, 2022

Objectives:

- Understand basic concepts of mass spectrometry-based proteomics
- Identify and quantify proteins using software and protein sequence database
- Conduct differential analyses and interpret results

Prerequisites None

Content

- Basic concepts of mass spectrometry-based proteomics
  - Identify peptide sequences from mass spectra
  - Quantification approaches in MS-based proteomics
- Statistical analysis of quantitative proteomics data
  - Quality control
  - Transformation/Normalisation
  - Differential analysis/Clustering
- Using web tools to interpret lists of proteins
  - Access details of a protein through Uniprot
  - Convert ID using Uniprot
  - Enrichment analysis with DAVID and webgestalt
  - Highlight interactions with STRING and Cytoscape

B4 - Refresher on utilities for HTS data analysis (6h)

Teachers Victoire Baillet, Blaise Li, Christophe Malabat

Dates March 3, 2022

Objectives:

- Attend minimal knowledge for HTS courses on
  - Unix
  - Cluster
  - R
  - Galaxy

Prerequisites None

Content

- **Unix**
  - Open a Unix terminal
  - Type a command
  - Read the results
- **Cluster**
  - Connect to cluster
  - Type a command
  - Get results
- **R**
  - Open Rstudio
  - Understand the interface,
  - Type a command,
  - See the result
- **Galaxy**
  - Create an account and connect to Galaxy
  - Upload data
  - Use basic tools

B5 - Basic concepts in HTS data analysis (6h)

Teachers Victoire Baillet, Claudia Chica, Christophe Malabat

Dates March 4, 2022

Objectives:

- Understand the key concepts underlying the quantification and comparison of OMIC datasets
- Perform the initial steps of most HTS data analysis

Prerequisites Minimum knowledge of unix command lines, galaxy and cluster usage

Content

- **How much?**
  - Sequence coverage estimation: which sequencing depth for my OMIC?
  - From the Lander-Waterman model to the reality of short reads
  - Abundance is a relative normalised value
- **How different?**
  - Facing technical bias:
    - Normalization: library size, background noise
    - Batch correction
  - Modelling read counts:
    - From binomial to negative binomial distribution
- **Getting started with the analysis:**
  - Learn basic command lines for HTS analysis
  - Assessing the quality of your sequencing data: FastQC
  - Mapping short reads

B6 -Expression, quantification, differential analysis (6h)

Teachers Emmanuelle Permal, Hugo Varet

Dates March 21, 2022

Objectives:

- Design a successful transcriptomics experiment
- How to get differential expressed genes from raw data

Prerequisites
A minimal knowledge about R would be prefered
Basic concepts in NGS data analysis

Content

- **Experimental design**
  - Definitions
  - Confounding effects
  - Interactions
- **Bioinformatics**
  - Quality control
  - Mapping
  - Counting
- **Statistics**
  - Exploratory data analysis
  - Normalization
  - Modeling
  - Differential analysis

B7 - Variant calling (6h)

Teachers Pascal Campagne, Adrien Pain, Amaury Vaysse

Dates March 22, 2022

Objectives:

- Learn the key steps to obtain genetic reliable variants from raw sequencing reads
- First applications : use filtered files of variants to perform basic analyses

Content

- **From raw sequencing data to variants** :
  - Technical and biological biases
  - Reference genome & read mapping
  - Visualisation
  - Mutation types and variant detection
  - From filtering to reliable variants
- **Getting started with data analysis** :
  - Handling output files
  - Examples of data analysis

Prerequisites

A minimal knowledge about R and command lines would be prefered
Basic concepts in NGS data analysis

B8 - Genotype data & association studies (6h)

Teachers Pascal Campagne, Adrien Pain, Amaury Vaysse

Dates March 23, 2022

Objectives:

- Getting exposed to important applications in the analysis of genetic data, genomewide
- Analysing genomic data with R and other popular software

Content

- **Genetic structure**
  - Basics
  - Linkage disequilibrium
  - Common analyses and clustering
- **Association and linkage studies**
  - Genome Wide Association Studies
  - Genetic mapping
  - Mapping mendelian traits
  - QTL mapping

Prerequisites

A minimal knowledge about R and command lines would be prefered
Basic concepts in NGS data analysis and statistics
This training shares important links with “Variant calling & genotyping”, participants may want to follow both sessions.

B9 - ChIP-seq data analysis (6h) ....

Teachers Remi Planel, Violaine Saint-André

Dates March 24, 2022

Objective:

Be able to analyze ChIP-seq data from sequencing reads to functional interpretation

Prerequisites
You will need to have an account on Galaxy Pasteur and explore how it works before the class

Content

- Epigenetics and ChIP-sequencing
- ChIP-seq QC
- Mapping
- Peaks Calling
- Visualization
- Functional annotations and metagenes
- Super-enhancers identification
- Motifs analysis and data integration

B10 - Metagenomics (6h)

Teachers Mathieu Almeida
And Amine Ghozlane

Dates March 25, 2022

Objectives:

- Learn how to process raw sequencing data to comparative analysis
- Interpret results and gain experience in the visualization of metagenomics data

Content

- **Introduction to basic concepts**
  - Overview of state-art approach
  - Targeted metagenomics
  - WGS metagenomics
- **Practice - Microbiome under antibiotic pressure**
  - OTU processing
  - Metagenomic assembly, Binning
  - Annotations

Prerequisites

Modules Introduction to R and statistics, Unix basic commands Basic in HTS data analysis, A minimal knowledge about R, Differential analysis
Software Unix / R

B11 - Single-cell Analysis (12h)

Teachers Claudia Chica, Bernd Jagla, Yann Loe-Mie, Sebastien Mella, Olivier Mirabeau

Dates March 28-29, 2022

Objectives:

- Get acquainted with the key concepts of scRNAseq raw data analysis, quantification, functional characterization and visualization.

Modules Introduction to RStudio ; Basic concepts in HTS data analysis ; Expression, quantification, differential analysis

Content

- **Single cell technologies:**
  - Biological questions
  - Technology overview
- **From raw data to count matrix:**
  - Sequencing QC
- **Cell QC and normalization**
- **Estimating the dimensionality of the dataset:**
  - Highly variable genes
  - PCA (linear dimensionality reduction)
  - Clustering
- **Data visualisation:**
  - UMAP/tSNE (non-linear dimensionality reduction)
- **Gene level meta-analysis:**
  - Marker genes
  - Functional annotation
- **Cell level meta-analysis:**
  - Cluster annotation
- **Practical:**
  - Summary/recall with SCHNAPPs

B12 - Functional analysis (12h)

Teachers Helene Lopez-Maestre, Natalia Pietrosemoli, Emeline Perthame

Dates March 30-31, 2022

Objectives

- Understand the key concepts of functional analysis
- Perform functional analysis on RNA-Seq data on a simple design using R tools

Prerequisites

Modules Introduction to RStudio ; Basic concepts in HTS data analysis ; Expression, quantification, differential analysis
Software R basic programming, RStudio

Content

- **Introductory concepts**
  - Gene annotations
  - Gene Ontology (GO)
  - Protein-protein interactions (PPIs)
  - Protein pathways and networks
  - Gene sets
- **Functional enrichment analysis theory**
  - General framework
  - Knowledge bases for gene sets and pathways
  - Methods
  - Examples of tools
- **Functional enrichment analysis practical**
  - Analysis from the count matrix to enriched gene sets and pathways
  - Result visualisation
  - Result (biological) interpretation

B13 - Advanced Unix commands (12h)

Teachers Julien Guglielmini, Nicolas Maillet

Dates May 19-20, 2022

Objectives:

- understand more complex but useful options and commands
- learn regular expressions and few commands using them,
- everything needed to create your firsts own scripts.

Prerequisites Unix basic commands

Content

- We will alternate theoretical and practical sessions to drive basic users to an advance Unix level, both in terms of command lines and scripting.

- We will teach you few more commands, some options on some already seen commands, regular expressions and all you will need to perform your own scripts.

Image Analysis

Reconstruction of Super-resolution images:

Material: Computer with Fiji installed

If your computer includes Nvidia gpu, please update your graphic driver (https://www.nvidia.fr/Download/index.aspx?lang=fr) and install a recent version of Cuda (cuda 10.0 preferably https://developer.nvidia.com/cuda-10.0-download-archive)

Link to data: https://www.dropbox.com/sh/iutuw6o06xrvhci/AABSxKl_U4guMKy6JSOppaR0a?dl=0

Getting started in Bioimage Analysis with Fiji (6h)

Teacher Jean-Yves Tinevez

Dates session 1: January 11, 2022
session 2: May 31, 2022

Objectives

This introduction is suited to scientists that have no or very little experience with image analysis.

It alternates between short lectures and practicals, and focuses on teaching image analysis rather than just the use of a software. During this workshop, we will rely on the Fiji/ImageJ software for hands-on sessions.

Prerequisites None

Content

- **Typical challenges in Image Analysis (lecture).**
  - Making sense of images.
  - Typical challenges.
- **Fundamentals of Image Analysis (practicals).**
  - Bit-depth and encoding.
  - Histograms.
  - Display adjustments and LUTs.
  - View of the data *vs* the data.
  - Digital image ethics.
  - From 2D images to 3D images.
  - Pixel size and calibration issues.
- **Spatial filters (lecture).**
- **Detection and segmentation of objects (practicals).**
- **Analysis of data over time (practicals).**

Using Icy for Bioimage Analysis (3h)

Teacher Stéphane Rigaud

Dates session 1: January 12, 2022 morning
session 2: June 1, 2022 morning

Objectives

Understand how Icy software works and what you can do with it.

Interface overview, basic image operation, region of interest. Advance operation for image analysis.

Prerequisites Bioimage Analysis with Fiji

Content

- What is Icy
- Graphical User Interface (GUI)
  - Histogram, colormap, look up table
  - Basic operations
  - Layers
  - 3D view
- Investigate an image
- Region of Interest (RoI)
- Analysis examples
  - Spot detection
  - Tracking
  - Active contours
  - …

Advanced Icy features: scripting and protocols (3h)

Teacher Stéphane Rigaud

Dates session 1: January 12, 2022 afternoon
session 2: June 1, 2022 afternoon

Objectives

Understanding protocols in Icy

Pipeline construction, Complex image processing, task automatisation.

Come with your own images/projects if you have

Prerequisites Bioimage Analysis with Fiji

Content

- What is an Icy protocol
- Blocks
  - Types
  - Inputs / Outputs
  - RoI
  - Processing
  - Quantification
  - Loop
- Download new blocks
- Let's build a pipeline
- Scripting
- ...

Reconstruction of super-resolution images (3h)

Teacher Benoit Lelandais

Dates session 1: January 13, 2022 morning
session 2: June 2, 2022 morning

Objectives

Understand how images are formed in microscopy

Be able to reconstruct 2D and 3D super resolution images with Fiji (Single Molecule Localization Microscopy) whose imaging system will be available in Pasteur imaging platform

Prerequisites Bioimage Analysis with Fiji

Content

- Single Molecule Localization Microscopy
  - PALM/STORM/PAINT imaging
  - (This course is not about SIM STED or MINFLUX imaging)
- 2D localization microscopy
- Image formation model in microscopy
- 3D localization microscopy using ZOLA-3D[1]
  - Phase retrieval
  - Detection by cross-correlation
  - Localization by Maximum likelihood
  - Drift correction
  - Filtering
  - …

(1) Aristov et al, nat. Comm, 2018 (https://github.com/imodpasteur/ZOLA-3D)

Using Machine Learning for BioImage analysis (3h)

Teacher Dmitri Ershov

Dates session 1: January 13, 2022 afternoon
session 2: June 2, 2022 afternoon

Objectives

Understand main concepts of Machine Learning

Learn how to use ilastik, ML-based tool that allows segmentation, classification, counting cells or other experimental data.

Prerequisites: Install ilastik

Content

- Introduction to Machine Learning in Image Analysis
  - What is ML
  - Applications in (Bio)Image Analysis
  - Details: “feature space”, “training”, “classifiers”...
  - Limitations
  - Examples of plugins in FIJI
  - Intro to Ilastik (state of the art of ML in IA)
- Hands-on tutorial on Ilastik
  - Pixel Classification
  - Pixel + Object Classification