Advanced Bioinformatics Workshop

Date: Monday, August 19th – Friday, August 23rd, 2019
Venue: Adriatico Guest House - Denardo Lecture Hall
International Centre for Theoretical Physics
Trieste, Italy
Course URL: http://indico.ictp.it/event/8847/
Material: https://codata-rda-advanced-bioinformatics-2019.readthedocs.io

General theme

Building Machine Learning workflows using NGS Data

This advanced bioinformatics course will provide an overview of the current status of different NGS workflows (variant calling, RNA-Seq, ChIP-Seq, Metagenomics etc), and combine them with the appropriate Machine Learning and Data Mining approaches. The course will heavily rely on hand-on exSercises and tutorials, and attempt to provide a strong foundation on the underlying theory.

Instructors

- **Fotis Psomopoulos**, Institute of Applied Biosciences (INAB|CERTH) / ELIXIR GR
  Contact details: email, website, twitter
- **Amel Ghouila**, H3BioNet
  Contact details: email, website, twitter
- **Phelelani Mpangase**, Sydney Brenner Institute for Molecular Bioscience, University of the Witwatersrand, Johannesburg
  Contact details: email, website

Course Schedule

	Topic
	Day 1
11:30 – 12:30	Experiments: Design and Analysis
14:00 – 15:00	Components of an Experiment. What is a good experiment design?
15:00 – 16:00	Data Distributions and Multiple Hypotheses Adjustment Methods
16:15 – 18:00	Introduction to basic NGS pipelines
	Day 2
09:00 – 10:00	Introduction to basic NGS pipelines
10:00 – 11:00	Short read quality and trimming (part 1)
11:30 – 12:30	Short read quality and trimming (part 2)
14:00 – 15:00	Mapping

15:00 – 16:00	Variant calling (part 1)
16:15 – 18:00	Variant calling (part 2)
	Day 3
09:00 – 10:00	Introduction to DM and ML, Machine Learning basic concepts
10:00 – 11:00	Taxonomy of ML and examples of algorithms
11:30 – 12:30	Applications of ML in Bioinformatics (genomics and metagenomics)
14:00 – 15:00	Practicing using the built-in R data set iris
15:00 – 16:00	RNASeq analysis using clustering in R
16:15 – 18:00	RNASeq analysis in R to be continued
	Day 4
09:00 – 10:00	Introduction to Nextflow
10:00 – 11:00	Parameters, Channels and Processes
11:30 – 12:30	Docker, Executors and Channel Operations
14:00 – 15:00	Practical exercises in Nextflow
15:00 – 16:00	Practical exercises in Nextflow
16:15 – 18:00	Practical exercises in Nextflow
	Day 5
09:00 – 11:00	Q & A and bonus examples
11:30 – 12:30	Closing, Final Remarks, Post-workshop survey

Topics

1. Experiments: Design and Analysis.
    a. Components of an Experiment.
    b. What is a good experiment design?
    c. Batch effects and Confounding factors
    d. Statistical inference
    e. Data Distributions and Multiple Hypotheses Adjustment Methods
    f. Correlation and Linear Regression in Life Sciences
2. Introduction to NGS Data Analysis
    a. Quality Assessment, Trimming and Filtering
    b. Mapping
    c. Case studies (e.g. Variant Calling, Metagenomics, etc)
3. Introduction to Machine Learning
    a. Basic concepts
    b. Taxonomy of ML and examples of algorithms
    c. Applications of ML in Bioinformatics
4. Introduction to Nextflow
    a. Use of workflow systems for automation / reproducibility
    b. Basic syntax of Nextflow
    c. Transform and execute a workflow in Nextflow
5. Machine learning in NGS
    a. RNA-Seq analysis using clustering in R
    b. Metagenomics and Machine Learning

Learning Objectives

Experimental Design

- Importance of a good experimental design
- Identify the key components of an experimental design
- Understand and identify confounding factors
- Understand p-value, confidence interval and power, including the underlying assumptions
- Understanding of the difference between pearson, spearman and kendall correlation
- Understanding of the underlying assumptions of pearson correlation

Introduction to NGS data analysis

- Use fastqc and multiqc
- Visualize read quality
- Quality filter and trim reads
- Distinguishing good/bad quality reads
- Run one end-to-end NGS data analysis pipeline

Introduction to Machine Learning

- Learn Machine Learning basic concepts and jargon
- Understand the Taxonomy of Machine Learning algorithms and differences between basic algorithms categories
- Get familiar with the basic Machine Learning algorithms in supervised and unsupervised learning categories
- Understand different parameters to take into consideration to choose the right Machine Learning technique for a given problem
- Understand how to evaluate Machine Learning results in supervised and unsupervised classification

- **Machine learning in NGS**
  Learn about some applications of Machine Learning in Bioinformatics
- Explore and apply some basic R packages to perform supervised and unsupervised classification
- Overview of the different Machine Learning techniques / tools useful in example NGS pipelines (such as RNA-Seq, metagenomics, etc)

Introduction to Nextflow

- Find and use Nextflow tool definitions online
- Understand how to write Nextflow definitions for command line tools
- Use Docker with Nextflow to provide software dependencies and ensure reproducibility
- Join Nextflow tools into a workflow
- Run Nextflow workflows on local and HPC systems

Expected Background

Overall course aimed for novices (no prior knowledge / expectations in NGS data analysis or Machine learning). However, participants should be somewhat familiar with:

- R (base commands)
- Unix shell (running basic commands)

Participants should ideally bring their own laptops - if that is not a possibility, Desktop PCs will be available during the course. Instructions on software / libraries installation will be provided to all participants prior to the workshop.

Course Material

All material, slides and exercises are available through the following URL (CC-BY-SA license)

https://codata-rda-advanced-bioinformatics-2019.readthedocs.io/en/latest/