**Data analysis and interpretation in molecular medicine:**

**from databases to artificial intelligence**

***Module 1:***

***GENOMICS***

**Organizers**

- Prof. Enza Maria Valente (coordinator, PhD course in Translational Medicine)

- Prof. Riccardo Bellazzi (coordinator, PhD course in Bioengineering and Bioinformatics)

**Objectives of the course**

On day 1, participants will learn how to evaluate NGS (e.g. whole exome) data by bio-informatic inspection of NGS files and how to interpret variants for clinical purposes. Some popular and freely-available tools will be used; examples of variants identified in real clinical cases will be provided.

On day 2, participants will learn the basics of machine learning and data analytics. Teaching will be performed following a learning-by-doing approach thanks to the use of Orange, a visual programming environment. A focus will be performed on representing genomics information as features of machine learning algorithms.

On day 3 (morning only), participants will finally have the chance to put in practice what they have learnt in the two previous days, by confronting themselves with practical exercises.

***Instructors:***

**Day 1**

Enza Maria Valente – Dept. of Molecular Medicine, University of Pavia

Edoardo Errichiello – Dept. of Molecular Medicine, University of Pavia

Sara Nuovo – IRCCS Santa Lucia Foundation, Rome

Tommaso Pippucci – Policlinico S. Orsola-Malpighi, Bologna

**Day 2**

Riccardo Bellazzi – Dept. of Electrical, Computer and Biomedical Engineering

Lucia Sacchi – Dept. of Electrical, Computer and Biomedical Engineering

Elisabetta Sauta – Dept. of Electrical, Computer and Biomedical Engineering

**Dates and venues**

- Wednesday June 19, 2019

*9am-1pm – room D8, floor D, Engineering Faculty, University of Pavia, Via Ferrata 5, Pavia*

*2pm-6pm – room B2, floor B, Engineering Faculty, University of Pavia, Via Ferrata 5, Pavia*

- Thursday June 20, 2019

*9am-6pm - room B2, floor B, Engineering Faculty, University of Pavia, Via Ferrata 5, Pavia*

- Friday June 21, 2019

*9am-1pm – room B2, floor B, Engineering Faculty, University of Pavia, Via Ferrata 5, Pavia*

**Course Program**

**Day 1**

***Morning session (9am-1pm)***

1) Enza Maria Valente – Riccardo Bellazzi (9am-9.30am)

* Introduction to the course

2) Tommaso Pippucci (9.30am-10.30am)

* Quality assessment - How to evaluate quality of an exome sequencing dataset from alignment and variant data
* File types: BAM / VCF
* Softwares: FastQC/MultiQC/Qualimap2 and IGV

3) Tommaso Pippucci (11.00am-1pm)

* Clinical annotation (including importance of using appropriate HPO terms) - What are the most popular/important sources of functional annotation for clinical variant interpretation
* File types: VCF
* Softwares: Internet browser (Ensembl VEP/ANNOVAR/VARAFT); population databases for germline (dbSNP, 1000G, ESP6500, ExAC, gnomAD, TOPMed, SHGP) and somatic SNVs (COSMIC, ICGC), and CNVs (Decipher, DGV); prediction/conservation tools (CADD, GERP, M-CAP, etc.); splice site prediction (HSF, MaxEntScan, etc.)

***Afternoon session (2pm-6pm)***

4) Edoardo Errichiello (2pm-3.30pm)

* Guidelines for the interpretation of Sequence Variants (ACMG general and disease-specific criteria) - practical examples.
* Softwares to assess pathogenicity by ACMG criteria (VarSome, ClinGen, eVAI, ClinVar, HGMD, LOVD/Locus Specific Mutation Databases, etc).
* Nomenclature Guidelines (HGVS) - Nomenclature tools (Mutalyzer, etc.)

5) Sara Nuovo (4pm-6pm)

* Exome interpretation and gene identification strategies for exome sequencing (linkage based, homozygosity based, double-hit based, overlap based, *de novo* based, etc.) – How to explore exome data, prioritize candidate genes and pick up clinically relevant variants
* File types - Annotated text files (Excel)
* Softwares - Microsoft Excel and IGV, cross-species phenotyping for novel gene discovery (e.g. MGI), interactome/pathways databases (e.g. STRING, HuRI), prioritization databases (Exomiser/Phenolyzer, Phenomizer, Endeavour, GeneDistiller)

**Day 2**

***Morning session (9am-1pm)***

1) Riccardo Bellazzi (9am-10.30am)

* Visual programming, data visualisation, introduction to explorative analysis
* Data projection (PCA, MDS, tSNE)
* Clustering, cluster scoring and evaluation

2) Lucia Sacchi (11am-1pm)

* Classification, overfitting, model scoring and evaluation
* Regression, regularization, model scoring

***Afternoon session (2pm-6pm)***

3) Riccardo Bellazzi (2pm-3.30pm)

* Variant representation and coding genomics features
* Novel machine learning methods
* Text and Image Analytics

4) Elisabetta Sauta (4pm-6pm)

* Hands on with Orange analytics

**Day 3**

Hands-on exercises on course topics, with all instructors.