Programming and database management for bioinformatics

The aim of this course is to provide some important programming and database management skills that are essential for bioinformatics: using Linux operating system, managing databases, programming in Python, using and writing statistical functions in R.

**Program:**

- Linux
- Programming Languages for Bioinformatics: Python
- Database Management
- R and Bioconductor

**Lecturers and trainers:**
Alfons Nonell, Mindthe Byte
Victor Urrea, Systems Biology Dept, EPS, Uvic
Statistical and data-mining methods for omics data analysis

The aim of this course is to introduce the most important statistical and data mining methods for bioinformatics and omics data analysis. The course combines lectures with hands-on sessions using R for illustration of the different methodologies.

Program:

1. Exploratory data analysis – Descriptive statistics
   a. Usual data structures
   b. Gene expression data structure
   c. Type of variables
   d. SNP and gene expression data
   e. Univariate data analysis
      i. Frequency tables
      ii. Summary statistics
      iii. Plots
   f. Bivariate data analysis
      i. Measures of correlation
      ii. Contingency tables, measures of associations, measures of risk (RR and OR)

2. Important distributions
   a. Binomial
   b. Normal
   c. T (Student dist.)
   d. Chi-squared
   e. F (Fisher-Snedecor dist.)
   f. q-q plots

3. Principles of statistical inference
   a. Inference about populations parameters,
   b. Maximum Likelihood Estimation,
   c. Bias, variance and mean squared error
   d. Statistical tests. Type I error, Power and sample size computation
   e. Wald, likelihood-ratio and score test

4. Important statistical tests
   a. One sample t-test for the mean
   b. Two-sample t-test for equality of means with unequal variances
   c. Two sample t-test for equality of two means with equal variances
   d. F-test on equal variances
   e. Wilcoxon rank test for the equality of two means
   f. Binomial test
   g. Chi-squared test for independence of two factors in contingency tables
   h. Fisher test for 2x2 tables
   i. Correlation test
   j. Normality tests
   k. Outliers test
5. Multiple testing
   a. Application of tests to a whole set of variables
   b. Distribution of p-values under the null
   c. Family-Wise Error Rate (FWER) and False Discovery Rate (FDR)
   d. Methods for multiple testing correction

6. Resampling methods for inference
   a. Bootstrap estimates and confidence intervals
   b. Permutation tests

7. Regression models
   a. Linear regression
   b. Measures of performance: Explained variation ($R^2$)
   c. One-way analysis of variance
   d. Two-way analysis of variance
   e. Logistic regression
   f. Penalized regression: LASSO (Regression with variable selection)

8. Resampling methods for model selection and validation
   a. Apparent, internal and external validation
   b. Bootstrap validation
   c. Cross-validation
   d. Simultaneous model selection and validation

9. Models for survival analysis
   a. Nonparametric estimators of the survival and cumulative hazard functions
   b. Semiparametric Cox’s proportional hazards model

10. Unsupervised methods: Cluster analysis and PCA
    a. Distance
    b. Linkage cluster analysis
    c. K-means cluster analysis
    d. Dimension reduction: Principal Component Analysis

11. Supervised data-mining methods for classification
    a. Classification
    b. Measures of classification accuracy: classification error, sensitivity and specificity, ROC curve, AUC
    c. Classification and Regression trees (CART)
    d. Random Forest (RF)
    e. Support Vector Machine
    f. Neural Networks

**Lecturers and trainers:**
Malu Calle, Systems Biology Dept, EPS, Uvic
Jordi Solé, Dept. of Information and Digital Technologies, EPS, Uvic
Genome Bioinformatics

The aim of this course is to introduce the most important methods and tools for sequence analysis and sequence alignment in the context of comparative genomics and functional genomics.

Program:
- Concepts of genomics. Functional elements of the genome
- Biological Databases
- Probabilistic models for sequence alignment
- Algorithms for pair-wise sequence alignment
- Multiple sequence alignment
- Bioinformatic tools for sequence analysis
- Sequence analysis with R and Bioconductor
- Methods in comparative genomics
- HMM for prediction of conserved motifs
- Methods for functional sites prediction

Lecturers and trainers:
Enrique Blanco, Genetics Department/IBUB, UB
Mireia Olivella, Systems Biology Dept, EPS, Uvic
Josep M. Serrat, Systems Biology Dept, EPS, Uvic
David Torrents, Joint IRB-BSC program on Computational Biology, BSC, ICREA
Analysis of complex disease association studies

The aim of this course is to introduce the most important methodologies for the analysis of the genetic component of complex diseases. It is a practical course that combines lectures with practical sessions using R for illustration of the different methodologies.

Program:
- Variation in the Human Genome
- Population Genetics and Linkage Disequilibrium
- The International HapMap Project
- SNP prioritization and Tag SNP selection
- Genotyping platforms and Next Generation sequencing
- Association studies: Candidate Gene Studies, Candidate Region Studies, GWA Studies
- Data Quality Control: Population Stratification, Hardy-Weinberg Equilibrium
- Single-locus Tests of Association Studies: Chi-square test and logistic regression
- Haplotype analysis in Association Studies
- Confounding and Population Stratification
- Genome-Wide Associations Studies
- Genotype imputation methods
- Copy Number Variant Association Studies
- Analysis of Gene-environment and gene-gene interactions
- Follow-up studies. Survival analysis. Predictive and prognostic models
- Measures of biomarker predictive accuracy

Lecturers and trainers:
Marinona Bustamante, CREAL: Biostatistics Program. PRBB
Malu Calle, Systems Biology Dept, EPS, Uvic
Juan Ramón González, CREAL: Biostatistics Program. PRBB
Transcriptomics: Analysis of Microarray gene expression data

The main objective of this course is to introduce the most important methods of processing (preprocessing) and analyzing microarray expression data analysis. It aims to find out the main problems that can be studied with microarrays and how to design, process and analyze the corresponding experiments. Appropriate software to carry out each stage of the process will be introduced.

Program:
- Concepts of gene regulation
- Gene expression measurement
- Gene expression databases
- Experiments with DNA microarrays. Design and execution
- Data preprocessing: Exploration, normalization, filtering
- Detection of differentially expressed genes and related statistical problems (power, multiple comparisons, etc.)
- Classification and prediction from microarray data
- Functional analysis and biological interpretation

Lecturers and trainers:
Josep L. Mosquera, Statistics and Bioinformatics Unit. IRVall d'Hebron
Àlex Sànchez, Statistical Dept, UB and Statistics and Bioinformatics Unit. IRVall d'Hebron
Josep M. Serrat, Systems Biology Dept, EPS, Uvic
Next Generation Sequencing analysis

This is a hands-on training course with the aim of introducing the most important methodologies for NGS data analysis. The course starts with a briefly introduction to NGS technologies and covers data analysis of RNA-Seq, ChIP-Seq and whole-exome sequencing experiments for variants detection.

Program:

- Next Generation Sequencing Technologies
- Bioconductor for high-throughput sequence analysis
- Short read formats
- Alignment of reads to a reference genome
- Alignment formats
- Summarization
- Exon sequencing
- DNA-seq experiments for variant calling: SNP and rare variant detection
- RNA-seq experiments for differential gene expression
- Gene set enrichment for RNA-seq differential expression results
- ChIP-Seq analysis of DNA regions of interest
- Annotating ChIP peaks

Lecturers and trainers:
Juan Ramón González, CREAL: Biostatistics Program. PRBB
Lorena Pantano, Ascidea
Benjamín Rodríguez, qGenomics

Epigenomics

This course provides an overview of the epigenetic mechanisms and their link to gene regulation. The aim of the course is to introduce the most important methods for genomics data analysis.

Program:

- Epigenetic mechanisms of gene regulation
- DNA methylation
- Histone modifications
- Epigenetics databases
- Analysis of epigenetic profiles
- NGS epigenetic data analysis

Lecturers and trainers:
Antonio Gómez, Cancer Epigenetics, Idibell
Holger Heyn, Cancer Epigenetics, Idibell
Sergi Sayols, Cancer Epigenetics, Idibell
Sebastián Morán, Cancer Epigenetics, Idibell
Introduction to Metagenomics

This course will introduce the field of metagenomics as well as the areas where it can be applied. The aim is to understand what questions can be explored with this technique and get an idea of the existing tools and approaches for data analyses. The course will also include hands-on sessions. The main focus will be microbial metagenomics, a field that has grown dramatically during the last five years.

Program:

- Introduction to the field of metagenomics and its progress along with High-Throughput Sequencing and High-Throughput Computing. Microbial metagenomics: community composition vs. community metabolism. Contribution of microbial metagenomics to diverse fields such as oceanic community ecology and ecosystem functioning, human health, environmental monitoring and agriculture. General guidelines for microbial metagenomic projects.

- Analysis of microbial taxonomic diversity based on PCR-amplified and non-amplified rDNA/RNA metagenomic fragments (Tags) using mostly QIIME and MOTHUR.
  - Quality control of 454 and Illumina tags. Denosing of 454 tags and merging of Illumina paired-end tags
  - Chimera detection
  - De-novo generation of Operational Taxonomic Units (OTUs); discussion on clustering thresholds and approaches. Assignation of metagenome-derived tags (mtags) to reference OTUs.
  - Taxonomic assignation
  - Basic community ecology analyses using R and Vegan

- Analysis of microbial community metabolism:
  - Basic quality control of Illumina reads
  - Analyses using MEGAN and BLAST
  - Introduction to other popular tools like MG-RAST and IMG
  - Reference databases and approaches for annotation
  - Assembly of metagenomes using Ray and subsequent functional analyses
  - Implications of the results for ecosystem functioning

Lecturers and trainers:
Ramiro Logares, Institut de Ciències del Mar, CSIC
Proteomics

The aim of this course is to introduce the necessary concepts and the most important computational methods for proteomics data analysis. The course combines lectures with hands-on sessions using R for illustration of the different methodologies.

Program:

Day 1 (Dra. Cristina Chiva)

Introduction to proteomics (theoretical session)
- Introduction to proteomics (compared to other -omics)
- Mass spectrometry. Definitions.
- Elements of a mass spectrometer
- Concepts of ion, isotope, resolution
- Tandem MS
- Basis of peptide fragmentation
- PMF based protein identification
- MSMS based identification
- Database search

MSMS interpretation and database searches (practical session)
- Manual de novo sequencing of one MSMS peptide spectrum
- Data formats and conversion (.raw to .mgf or .mzXML)
- Data base search

Day 2 (Dra. Cristina Chiva)

Protein identification (theoretical session)
- Peptide-protein inference, protein grouping and FDR/FNR
- Overview of proteomics databases.
- Unique peptides
- PTM's and PTM's site localization tools (Score, phosphoRS)

Protein identification (practical session)
- Results visualization with Proteome Discoverer
- Practical exercises with real data
- Scripts to correlate data obtained from .msf files (eg: z to MW)

Day 3 (Dra. Eva Borràs)

Protein quantification (theoretical session)
- Experimental design
- Label free quantification (Spectral count, top3, iBAQ, peptide areas)
- Isotopic labelling (MS1: SILAC; MS2: TMT)
• Normalization of data
• Log scale. How to calculate a ratio
• Missing values
• Targeted proteomics

Protein quantification (practical session)

• Quantitative analysis of label-free proteomics data
• Quantitative analysis of labelled proteomics data

Day 4 (Dra. Eva Borràs)

Applications of bioinformatics to the proteomics field (theoretical session)

The idea here is to show the students some nice example on how a bioinformatician can contribute into the proteomics research

• Repositories
• Statistics
• Example of combination of unique transitions
• Decision tree
• MacCoss window overlap

Statistics in proteomics (practical session)

• R packages
• MSStats & SRMStats
• OpenMS: overview of already developed tools

Lecturers and trainers:

Eva Borràs, Proteomics Facility. Department of Experimental Health Sciences, UPF

Cristina Chiva, Proteomics Facility. Department of Experimental Health Sciences, UPF
Metabolomics

The aim of this course is to introduce the most important concepts and computational methods for metabolomics data analysis. The course combines lectures with hands-on sessions using R for illustration of the different methodologies.

Program:
- Experimental design
- Sample analysis & metabolite identification by mass spectrometry (MS) and nuclear magnetic resonance (NMR)
- Data processing for NMR, LC/MS and GC/MS
- Statistical and chemometric analysis
- Pathway analysis

Lecturers and trainers:
Oscar Yanes, Centre for Omic Sciences, URV
Maria Vinaixa, Centre for Omic Sciences, URV
Miguel Ángel Rodríguez, Centre for Omic Sciences, URV
Interactomics: Systems Biology

This course, with a strong practical focus, is intended to provide students with the capacity to use data from databases of molecular interactions to build and analyze biological networks. The aim of the course is the analysis of network topology, the modeling of motif dynamics, and to establish relationship between topology and biological function.

Program:
- Biological interactions: classes and biological relevance.
- Methods for the determination of molecular interactions.
- Molecular interaction databases.
- Parsing interaction databases.
- Building graph objects.
- Topological analysis of graphs.
- Genetic circuits dynamics
- Noise in genetic circuits
- Multicellular genetic circuits
- Integrative omic analyses
- Interactome properties and analyses
- Systems-level view of cancer

Lecturers and trainers:
Jordi Planas, Systems Biology Dept, EPS, UVic
Jordi Serra Musach, Idibell
Integrative Genomics

The aim of this course is to introduce the most important methodologies for Integrative Genomics. This includes visualization of multidimensional genomic data, enrichment analysis and bioinformatic tools for functional annotation and construction of networks. It is a practical course that combines lectures with hands-on sessions for illustration of the different methodologies.

Program:
- Introduction to Integrative Cancer Genomics
- Visualizing multidimensional cancer genomics data
- Cancer genomics data resources
- Introduction to enrichment analysis
- Sample level enrichment analysis (SLEA)

Lecturers and trainers:
Núria López-Bigas, Biomedical Genomics, Research Unit on Biomedical Informatics (GRIB) UPF-PRBB, Barcelona