



Biostatistics/Bioinformatics Unit



In the last decade, a number of technologies that generate vast amounts of data have been popularised. For instance, microarrays measure mRNA expression levels for tens of thousands of genes simultaneously, tiling arrays assess enrichment in millions of chromosomal locations, and next generation sequencing technologies deliver hundreds of millions of genomic sequences in a single experiment. Nowadays researchers face not only the challenge of obtaining scientifically relevant data, but also of extracting as much valuable information from them as possible. Statistics is the science that transforms data into information. It provides a disciplined and scientifically sound framework to test scientific hypotheses and to learn about the systems and processes that generate biomedical data. Also, the experimental design theory guides researchers as to the best way to conduct experiments in order to reach their goals. We offer scientists support in the following areas: (i) experimental design (sample size calculation, study design, planning of statistical methodology); (ii) data analysis (clinical or biomedical databases, high-throughput data, *eg*, genomics, proteomics); (iii) statistical methodology; and (iv) software (help in using statistical software, development of software to meet special data analysis or study design needs).

The Biostatistics and Bioinformatics Unit was created in January 2008. By the end of the year the facility was staffed by its manager, David Rossell, and a research officer, Evarist Planet. During 2008, we have been involved in 25 collaborative research projects that have arisen from thirteen groups at IRB Barcelona. In addition, we have provided technical guidance in a number of projects focused on fields such as gene regulation, developmental biology, oncology, bioinformatics and molecular medicine.

Our mission is to offer the IRB Barcelona research community a competitive advantage by increasing both the quality and speed of its research. Quality has been furthered by making available cutting-edge methodology and tailored solutions to specific problems while speed has been increased by developing software tools to facilitate the generation and interpretation of experimental results.

In terms of methodological research, we have developed the GaGa model for differential expression analysis, which has contributed to proving that several chromatin-regulating transcription factors share a common regulatory programme. As another example, we have derived a framework for Bayesian Gene Set Enrichment Analysis, which has facilitated assessment of the biological relevance of findings from gene expression studies. Most of this research has been either published

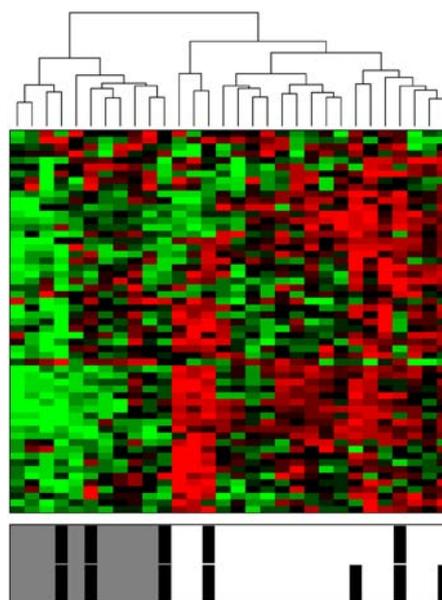


Figure 1. Hierarchical clustering analysis of gene expression data reveals associations with clinical outcomes.

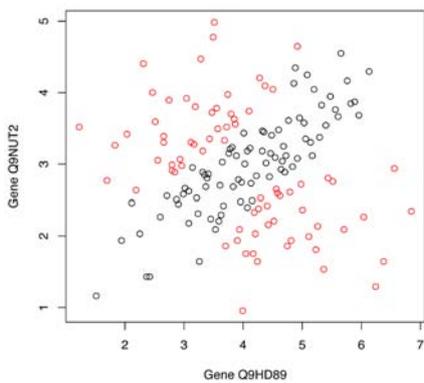
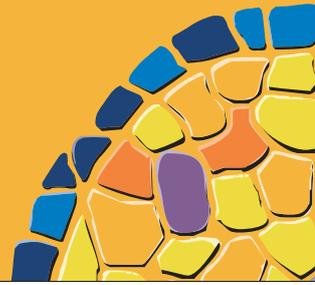


Figure 2. Expectation-maximisation algorithm reveals correlation between two genes in the presence of noise. Black circles indicate correlated observations, red circles indicate observations arising from noise.

or submitted for publication in scientific journals, thereby contributing to consolidating IRB Barcelona as a cutting-edge research institution.

In terms of software, we have developed routines to automatically produce reports with hyper-links to a number of on-line databases and resources. This has allowed researchers to obtain, for instance, additional information about specific genes or gene networks with a single click on their computer.

In collaboration with the IT Department, we have also provided a web browser interface which allows researchers to access their results moments after we have produced them. This development circumvents the inherent delay caused by copying large files with results on compact discs and sending them to researchers.

We have collaborated with IRB Barcelona groups in a number of research projects on developmental biology, structural and computational biology, molecular medicine and oncology.

Publications

Font-Burgada J, Rossell D, Auer H and Azorín F. *Drosophila* HP1c isoform interacts with the zinc-finger proteins WOC and relative-of-WOC (ROW) to regulate gene expression. *Genes Dev*, 22(21), 3007-23 (2008)

Rossell D, Baladandayuthapani V and Johnson VE. Bayes factors based on test statistics under order restrictions. In *Bayesian Evaluation of Informative Hypotheses in Psychology* (H Hoijtink, I Klugkist, P Boelen, ed.), Springer (2008)

Rossell D, Guerra R and Scott C. Semi-parametric differential expression analysis via partial mixture estimation. *Stat Appl Genet Mol Biol*, 7(1), 15 (2008)

Paper on differential expression analysis

Rudy Guerra, Rice University (Houston, USA) and Clayton Scott, University of Michigan (Ann Harbor, USA)

Paper on sequential design for high-throughput experiments

Peter Müller, MD Anderson Cancer Center (Houston, USA)

Collaborations

Paper on Bayes factors

Valen Johnson and Veerabhadran Baladandayuthapani, MD Anderson Cancer Center (Houston, USA), Herbert Hoijtink, Irene Klugkist and Paul A Boelen, Utrecht University (Utrecht, The Netherlands)