David Rossell

Biostatistics / Bioinformatics Unit

Modern biomedical research has fostered the development of novel technologies capable of generating vast amounts of data. High-throughput genomic or proteomic experiments, for instance, produce data from as few as hundreds up to millions of genes or proteins. Nowadays researchers face not only the challenge of obtaining relevant scientific data, but also of extracting valuable information from it. Statistics is the science that transforms data into information. Founded on probability theory, it provides a disciplined and scientifically sound framework to test hypotheses and to learn about the systems and processes that generate biomedical data. The experimental design theory also guides researchers to conduct experiments in such a way that the subsequent data analysis can provide the information they need, that is, that the experimental goals are met.

Our collaboration with IRB Barcelona researchers focuses on the design and analysis of high-throughput experiments: next-generation sequencing, microarrays, tiling arrays and mass spectrometry. Regarding next-generation sequencing, we have participated in ChIP-Seq studies to characterise genome-wide transcription factor binding sites and histone methylation, and in RNA-Seq studies to characterise and quantify alternative splicing patterns. With respect to microarrays and tiling arrays, we have worked not only with datasets produced in-house but also with an increasing number of publicly available datasets, assessing the extent to which findings obtained in animal systems can be extrapolated to human patients. Regarding mass spectrometry, we have helped to design and analyse iTRAQ and to label free quantification experiments to find differentially expressed proteins that can serve as biomarkers. In addition to these study-specific collaborations, we placed emphasis on developing tools and methodology that benefits the IRB Barcelona community at large. We have developed statistical methods to infer alternative splicing based on RNA-Seq studies (Figure 1), and a novel Bayesian framework which offers important advantages in a wide range of hypothesis-testing problems.

In terms of informatics, we have structured a number of public gene expression datasets so that their contents are easily accessible. Furthermore, we have developed statistical data analysis software to screen them for patterns.

We have developed computer programmes to reduce the time needed to perform routine tasks and to implement novel data analysis methodology. We have made all our software available and have developed interfaces to facilitate its use.

Other activities during 2009 include providing assistance with statistical methodology and organising workshops to train IRB Barcelona researchers in the use of several data analysis software packages.
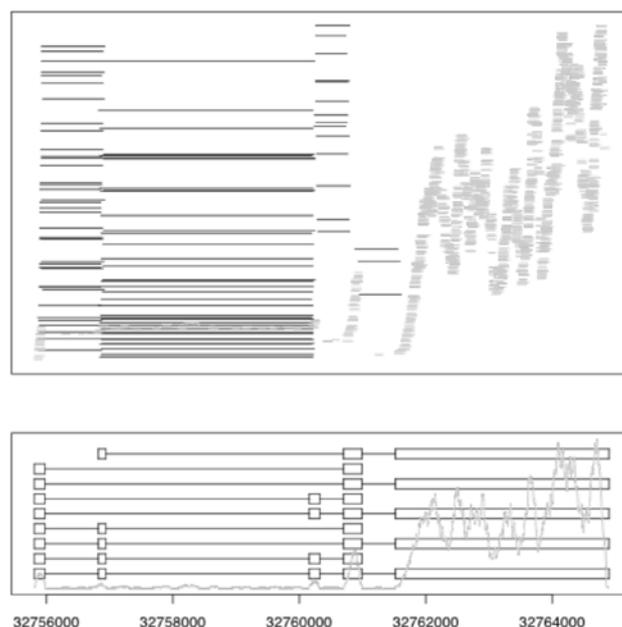

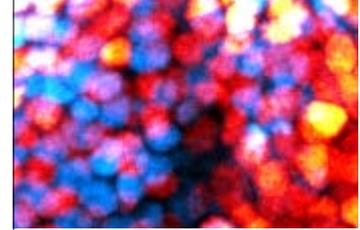*Figure 1. Inferring alternative splicing from paired end RNA-Seq data.*

Services for IRB Barcelona researchers
Experimental design
Sample size, study design and planning of statistical methodology

Data analysis
Clinical or biomedical databases, and genomics data

Assistance with statistical methodology in own or others' research

Software
Help in using software, development of software to meet special data analysis or study design needs

Research Group Members

Unit Manager:
David Rossell

Research Officer:
Evarist Planet

Scientific output

Publications

Geslain R, Cubells L, Bori-Sanz T, Alvarez-Medina R, Rossell D, Martí E and de Pouplana LR. Chimeric tRNAs as tools to induce proteome damage and identify components of stress responses. *Nucleic Acids Res*, Epub Dec 8 (2009)

Rossell D. GaGa: a parsimonious and flexible model for differential expression analysis. *Ann Appl Statist*, **3**(3), 1035-51 (2009)

Collaborations

Evaluation of CSF immunodepletion and fractionation strategies for MS-based biomarker discovery
Jacques Borg and Joan Guinovart, IRB Barcelona (Barcelona, Spain)

Non-local priors for high-dimensional variable selection
Valen E Johnson, MD Anderson Cancer Center (Houston, USA); Donatello Telesca, University of California (Los Angeles, USA)

Prior densities for default Bayesian hypothesis tests
Valen E Johnson, MD Anderson Cancer Center (Houston, USA)

Sequential sample sizes for high-throughput experiments
Peter Müller, MD Anderson Cancer Center (Houston, USA)