David **Montaner**

R · Python · Machine Learning · Statistics

St Albans - AL1 4DL - U.K.

0775 497 22 35 | david.montaner@gmail.com | www.dmontaner.com | dmontaner | dmontaner

"More than 15 years playing with data."

Professional Work Experience

Data Scientist

London - U.K.

Genomics England Ltd. - Bioinformatics Department — *2015 - present*

- Genomics England Ltd. is the leading partner of the largest genomic consortium in the world: the 100,000 Genomes Project.
- My main duty is to model clinical and genomic data for 100,000 individuals and millions of variables (genetic variants) using R and Python. I use supervised classification methods (SVM, KNN, LDA and Random forests) to predict clinical outcomes from genomic variables. The aim is for instance to discriminate cancer and normal samples, classify different tumors or predict disease stages from the patient genetic information. I also used logistic regression models to evaluate disease risk, and Cox proportional hazards models for survival analysis.
- I parsed and collected clinical information from hundreds of files in many different formats: XML, JSON, xls, csv. I organized this information in our internal LabKey data management system (SQL). I created an R library which calls LabKey web services to automatize the data intake. I then collected genomic information stored in MongoDB, combined it with the clinical information, and used our own Python libraries to serve it to our clients trough a REST API.
- I computed descriptive statistics and using R and Python, and created plots and visual representations of the data using the ggplot2 library. I used unsupervised classification methods (KNN and hierarchical clustering) and principal component analysis to find out outliers and biases in our data. I inserted those in automatically generated reports using knitr, pandoc markdown and LaTeX, I derived a pipeline to export the reports to PDF, HTML, doc, text and wiki formats. I also created some Jupyter Notebooks and some R Shiny websites for data visualization.

Chief Data Scientist

Valencia - Spain

Genometra S.L. — *2010 - 2015*

- I co-founded the company. I set it up as an spin-off company of the CIPF research center where I was working. I carried out all the foundations administrative formalities and legal agreements.
- I supervised a team of 3 technical staff (computer scientist, biologist and statistician) and one commercial. I did train them at the beginning and keep them up to date with the latest technologies in Bioinformatics. We also had several students and apprentices which I supervised.
- At the beginning the company was marketing some of the software we had developed in the research center for the preprocessing of raw genomic data. Later, we developed our own internal pipelines, software and methodologies and we provided consulting services, training, and software and data base customization.
- I developed several R packages, some of them contributed to public repositories as CRAN or Bioconductor. I also developed several pipelines using Python and Linux shell script. All this software was used to preprocess raw genomic data, query databases or web services for extra biological information, carrying out hypothesis testing or class predictions and automatizing the creation of reports.

Head of the Biostatistics Unit

Valencia - Spain

Principe Felipe Research Center (CIPF) - Computational Genomics Department — *2005 - 2010*

- I was head of the Biostatistics unit within the Bioinformatics department. I supervised a team of other 3 statisticians.
- In the Bioinformatics department we developed a suit of web tools for the analysis of genomic data. Those tools have been now running form more than 15 years and can be found at www.babelomics.org. Babelomics is the most used site in the world for functional genomics analysis. I developed all the statistical back end of the application. I did use R and Bioconductor libraries and also developed my own ones, now contributed to Bioconductor. I did use Python, WEKA, AWK and many other open source code.
- I developed or adapted several statistical methodologies to the analysis of gene expression data, Sometimes for our tools or pipelines, sometimes for particular data analysis or collaborations: linear and logistic regression, resampling and permutation methodologies, naive Bayes approaches, multiple hypothesis testing, nested models, mixture models, sample size computation.
- I wrote more than 50 scientific (pair reviewed) publications and a couple of book chapters. I presented or organized more than 30 congresses and international workshops.
- I have supervised several MSc students and one PhD.

Par Time Lecturer

Valencia - Spain

University of Valencia - Mathematics and Computer Science Departments — *2009 - 2015*

- Teaching Probability and Statistics to Undergraduate Mathematics students.
- Teaching Statistics and Experimental Design to Biology and Economics students.
- Teaching databases, R, Python, and Bash programming MSc students.

Epidemiologist

Bristol - U.K.

University of Bristol - Social Medicine School — *2002 - 2004*

- Preprocessing and statistical analysis of large cohort studies. I did use STATA, SAS and SPSS for all kind of descriptive statistics and plots, survival analysis, regression analysis and hypothesis testing.

Education

PhD in Bioinformatics and Functional Genomics

Valencia - Spain

Universitat de Valencia - Biotechnology — *2009 - 2013*

- I explored different statistical methodologies to push forward the knowledge in functional genomics. The DNA sequence of most genes is known, but their function, how they act in the body, is generally undescribed. The aim was to be able to predict gene functionality form genetic experimental data.
- I collected information from hundreds of different experiments publicly available. I did preprocess them and standardize signals to get a data matrix of 25,000 genes and 30,000 experiments. I did combine this experimental information with that of many databases describing gene functionality. Then I tested several approaches to infer and analyze genetic function.
- My final proposal used Multidimensional Logistic Regression to infer gene activity and functionality. I used a weighed model which incorporated prior experimental information and corrected for gene colinearity and correlation. My approach was also novel because it allowed for the interpretation of several genomic characteristics at the same time (current methods work just with a single genetic dimension or measure set but there are many available)
- I implemented the Methodology in an R package called mdgsa which is available in the Bioconductor repository. The methodology was also used as a back end of the Babelomics web suit.

MSc in Probability & Statistics

Madrid - Spain

Universidad Complutense de Madrid - Mathematics Department — *2002 - 2004*

- I tackled the missing data imputation problem. I explored the EM algorithm in the context of multiple imputation, I used bootstrap and some other resampling techniques for incomplete data analysis.

MSc in Statistics with Applications in Medicine

Southampton - U.K.

University of Southampton - Statistics Department — *2001 - 2002*

- Experimental design, stratification, sample size calculation, demographic analysis, clinical trials, spatial analysis.

Degree in Mathematics

Madrid - Spain

Universidad Complutense de Madrid — *1995 - 2001*

- Statistics and Operations Research specialty.

Skills

Programming R (dplyr, ggplot2, shiny, caret), Python (numpy, scipy, pandas), Linux Bash, AWK, SQL, MongoDB, Spark

Statistics Data processing. Data modeling. Machine learning. Clustering. Network analysis, Plots and data visualization.

Languages Spanish, English & French.

Selected Publications

1. *Integrated gene set analysis for microRNA studies.* Bioinformatics (Oxford, England). 2016;
2. *Family-based genome-wide association study in Patagonia confirms the association of the DMD locus and cleft lip and palate.* European journal of oral sciences. 2015;
3. *Babelomics 5.0: functional interpretation for new generations of genomic data.* Nucleic acids research. 2015; 43(W1):W117-21.
4. *Pathway network inference from gene expression data.* BMC systems biology. 2014; 8 Suppl 2:S7.
5. *Four new loci associations discovered by pathway-based and network analyses of the genome-wide variability profile of Hirschsprung's disease.* Orphanet journal of rare diseases. 2012; 7:103.
6. *Large-scale transcriptional profiling and functional assays reveal important roles for Rho-GTPase signalling and SCL during haematopoietic differentiation of human embryonic stem cells.* Human molecular genetics. 2011; 20(24):4932-46.

7. *The MicroArray Quality Control (MAQC)-II study of common practices for the development and validation of microarray-based predictive models.* Nature biotechnology. 2010; 28(8):827-38. NIHMSID: NIHMS235927

8. *Hypoxia promotes efficient differentiation of human embryonic stem cells to functional endothelium.* Stem cells (Dayton, Ohio). 2010; 28(3):407-18.

9. *Initial genomics of the human nucleolus.* PLoS genetics. 2010; 6(3):e1000889.

10. *Multidimensional gene set analysis of genomic data.* PloS one. 2010; 5(4):e10348.

11. *Functional genomics of 5- to 8-cell stage human embryos by blastomere single-cell cDNA analysis.* PloS one. 2010; 5(10):e13615.

12. *DNA methylation epigenotypes in breast cancer molecular subtypes.* Breast cancer research : BCR. 2010; 12(5):R77.

13. *Gene set internal coherence in the context of functional profiling.* BMC genomics. 2009; 10:197.

    SNOW, a web-based tool for the statistical analysis of protein-protein interaction networks. Nucleic acids research. 2009; 37(Web Server issue):W109-14.

14. *CLEAR-test: combining inference for differential expression and variability in microarray data analysis.* Journal of biomedical informatics. 2008; 41(1):33-45.

15. *GEPAS, a web-based tool for microarray data analysis and interpretation.* Nucleic acids research. 2008; 36(Web Server issue):W308-14.

16. *Direct functional assessment of the composite phenotype through multivariate projection strategies.* Genomics. 2008; 92(6):373-83.

17. *Functional profiling of microarray experiments using text-mining derived bioentities.* Bioinformatics (Oxford, England). 2007; 23(22):3098-9.

18. *Prophet, a web-based tool for class prediction using microarray data.* Bioinformatics (Oxford, England). 2007; 23(3):390-1.

19. *From genes to functional classes in the study of biological systems.* BMC bioinformatics. 2007; 8:114.

20. *The detection, treatment and control of high blood pressure in older British adults: cross-sectional findings from the British Women's Heart and Health Study and the British Regional Heart Study.* Journal of human hypertension. 2006; 20(10):733-41.

21. *Next station in microarray data analysis: GEPAS.* Nucleic acids research. 2006; 34(Web Server issue):W486-91.

22. *Clustering of risk factors and social class in childhood and adulthood in British women's heart and health study: cross sectional analysis.* Ebrahim S, **Montaner D**, Lawlor DA. BMJ (Clinical research ed.). 2004; 328(7444):861.