

Florian Rohart

PhD in Applied Statistics

Brisbane
QLD, Australia
✉ f.rohart@uq.edu.au
<http://florian.rohart.free.fr>
28/11/1985 – Single

Professional experience

- 2013 – 2014 **Postdoctoral Research Fellow**, *Working on the development of multivariate statistical approaches for cross platform analyses*, AIBN, Brisbane, Australia.
- 2012 – 2013 **Research/Teaching Assistant**, INSA Toulouse.
- 2009 – 2012 **PhD**, *Phenotypic prediction and variable selection in high dimensional linear and linear mixed models*.
Supervision of : Béatrice Laurent-Bonneau, INSA Toulouse & Magali San-Cristobal, INRA Toulouse, France
- March–July 2009 **Internship, Master 2 Research**, *Confidence region & sensitivity analysis*, ACTIA-Toulouse, France, the internship was about failure detection from automotive data.

Education

- 2008 – 2009 **Master 2 Research in Applied Mathematics**, *with distinction*, Learning of several stochastic tools and statistics tools: ACP, SVM, Rice formula, Delta method, M- & Z-estimators,
Université Paul Sabatier, Toulouse, France
- 2006 – 2008 **Master 1 and Licence of fundamental mathematics**, *with distinction (10^{ème}/61 and 9^{ème}/76)*, Université Paul Sabatier, Toulouse, France.
- 2003 – 2006 **Two-year highly selective classes to prepare for the competitive exams to the french “Grandes Ecoles”, specializing in mathematics and physics (MPSI/MP)**.
Lycée Georges Clémenceau, Reims, France
- Juin 2003 **Baccalauréat Scientifique**, *with distinction*, Lycée Gay-Lussac, Chauny, France.

First Author Publications

- 2014 **YuGene: A simple approach to scale gene expression data derived from different platforms for integrated analyses**, K-A Lê Cao, F.Rohart, L. McHugh, O. Korn, C.A. Wells, *Genomics* 10.1016/j.ygeno.2014.03.001.
Fixed Effects Selection in High Dimensional Linear Mixed Models, Rohart F., B. Laurent and M. SanCristobal, *Computational Statistics and Data Analysis*. 10.1016/j.csda.2014.06.022.
- 2012 **Phenotypic Prediction Based on Metabolomic Data on the Growing Pig from three main European Breeds.**, Rohart F., Paris A., Laurent B., Canlet C., Molina J., Mercat M.J., Tribout T., Muller N., Iannucelli N., Villa-Vialaneix N., Liaubet L., Milan D. and San Cristobal M., *J Anim Sci.* 2012 Dec;90(13):4729-40.
- 2011 **Multiple Hypotheses Testing For Variable Selection.**, Rohart F., *Submitted*.

Communications

Rohart F. A Molecular Classification of Mesenchymal Stromal Cells (MSCs). *Genomics of Development and Disease*, May 2014, Brisbane, QLD, Australia

Rohart F. A stable statistical tool to find a signature of Mesenchymal Stem Cells (MSCs). *The International Biometric Society, Biometrics by the canal*, December 1-5th, 2013, Mandurah, WA, Australia

Rohart F. Finding Mesenchymal Stem Cell (MSC) gene expression signature regardless of the tissue of origin. Poster communication at *The Australasian Microarray and Associated Technologies Association (AMATA) conference*, October 13-16th, 2013, Surfers Paradise, QLD, Australia

Rohart F. Variable Selection in High Dimensional Linear Mixed Model through ℓ^1 Penalization. Tenth colloquium of "*Jeunes Probabilistes et Statisticiens*", April 16-20th, 2012, CIRM Marseille, France

Rohart F. Multiple Hypotheses Testing For Variable Selection. Presentation given at "*Statistical Methods for Post-Genomic Data*" January 26-27th, 2012 in Lyon; at "*Statistiques Mathématiques et Applications*" August 28th - September 2nd, 2011 in Fréjus; and at "*Séminaire des Doctorants du Département de Génétique Animale de l'INRA*", April 5-6th, 2011 in Limoges, France

Rohart F., C. Canlet, N. Villa-Vialaneix, J. Molina, D. Milan, B. Laurent, A. Paris and M. SanCristobal (2011) Feature selection in 1H NMR-based metabolomic profiles enables the prediction of growth phenotypes in various pig breeds and highlight few metabolites involved in this complex trait. *5èmes Journées Scientifiques du Réseau Français de Métabolomique et Fluxomique*, May 23-25th, 2011, Jussieu, France

Rohart F., N. Villa-Vialaneix, A. Paris, C. Canlet, J. Molina, D. Milan, B. Laurent and M. SanCristobal (2010) Phenotypic prediction based on metabolomic data: lasso vs Bolasso, primary data vs wavelet data. *9th World Congress on Genetics Applied to Livestock Production*, August 1st-6th, 2010, Leipzig, Germany.

Organisations

Member of the organising committee of "Journées Statistiques du Sud", June 20-22th, 2012, Toulouse, France.

Member of the organising committee of "quinzième Séminaire des Doctorants du Département de Génétique Animale", March 26-27th, 2012, Ile d'Oléron, France.

Supervision-Teaching

- 2014 **Guest lecturer**, *Bioinfo Masters*, University of Queensland, Brisbane, Australia.
- 2012 **Speaker in "Statistiques pour l'analyse de données post-génomiques haut-débit"**, *Génotoul/Interbio - Génopole Toulouse Midi-Pyrénées*, Professional training for statistics users that do not possess special skills in mathematics but wish to acquire autonomy to analyze their own datasets.
- 2011 **Co-supervisor of a 3-months project and a 3-months internship**, *Transcriptomic analysis*, INSA Toulouse, INRA Toulouse, France.
- 2009–2012 **Teacher of Mathematics from Licence 1 to Licence 3**, *Statistics modelling, Probabilities & Statistics, Differential calculus and Integral calculus, . . .*, INSA Toulouse, France .

Skills

	Languages		Computer
French	native	System	Windows, Linux, Mac
English	fluent (written and spoken)	Software	R, Matlab, Maple, L ^A T _E X, ...
Spanish :	scholar	Computer Language	Notion of C/C++, learning java

Personal interests

Sport	Football, tennis, hiking (Mare e Monti, Corse, Sept 2012)
Photo	Amateur photographer http://florian.rohart.free.fr/Photos
Cultural activities	Cinema -old movies such as Metropolis, Forbidden Planet-, theatre, opera, reading of thriller and fantasy books -The Farseer Trilogy- in english or french language
Music	Learning guitar
Other	Scout leader for three years

Summary of research work

Most of my research work has been done in the context of the PhD entitled “phenotypic prediction and variable selection in high dimensional linear and linear mixed models”.

Recent technologies have provided scientists with genomics and post-genomics high-dimensional data; there are always more variables that are measured than the number of individuals. In this high-dimensional context **I worked on a real dataset** which comes from the pig species and **high-throughput biotechnologies**. Metabolomic data has been measured with NMR spectroscopy and phenotypic data has been mainly obtained post-mortem. There were two objectives. On one hand, we aimed at obtaining good prediction for the production phenotypes and on the other hand we wanted to pinpoint metabolomic data that explain the phenotype under study. Thanks to the Lasso method applied in a linear model (Tibshirani, 1996) -which is a ℓ^1 -penalization of the least squares criterion-, **we showed that metabolomic data has a real prediction power for some important phenotypes for livestock production**, such as a lean meat percentage and the daily food consumption, cf. (Rohart et al., 2012).

The second objective was a problem of variable selection. Classic statistical tools such as the Lasso method or the FDR procedure were investigated and I developed new powerful methods. **I have developed one variable selection method for ordered selection and one for non ordered selection**. The methods are sequential multiple testing procedures inspired by Baraud et al. (2003). The procedures are designed to **perform in high dimensional linear models and non asymptotic results** are given under conditions on the signal (Rohart, 2011). Both methods showed really good simulation results, the study was performed with the *mht* R-package that I have developed (<http://cran.r-project.org/>). The non ordered variable selection procedure was also used to construct networks and the simulations showed good results.

Since supplemental data were available on the real dataset such as the batch or the family relationships between the animals, linear mixed models were considered. **I have developed a new method that performs fixed effects selection in high dimensional linear mixed models**. The method relies on an multicycle ECM algorithm which turned out to be faster than the usual methods such as the lmmLasso of Schelldorfer et al. (2011). Thanks to its structure, it can be combined with any variable selection methods built for linear models. However, the convergence property of this algorithm depends on the method that is used. The multiple hypotheses testing procedure shows good empirical results, the simulation has been done with the *MMS* R-package (<http://cran.r-project.org/>), cf. (Rohart et al.,

2014).

My postdoctoral position at the Australian Institute of Bioengineering and Nanotechnologies (AIBN) consisted in developing new multivariate approaches for cross platform analyses.

A lot of biological studies are publicly available online, sometimes answering the same biological question. Instead of looking at a phenomenon on a per dataset basis, performing a meta-analysis of similar studies could unravel new interesting things (e.g. biomarkers) because of a higher statistical power. Unfortunately, each dataset comes from a particular experiment that was performed in one lab, on one commercial platform; simply combining these different datasets mainly highlights the technical differences instead of the underlying biology. **We developed a new simple and easy method called YuGene to integrate microarrays studies** (Lê Cao et al., 2014).

References

- Baraud, Y., Huet, S., and Laurent, B. (2003). Adaptive test of linear hypotheses by model selection. *The Annals of Statistics*, 31(1):225–251.
- Lê Cao, K.-A., Rohart, F., McHugh, L., Korn, O., and Wells, C. (2014). Yugene: A simple approach to scale gene expression data derived from different platforms for integrated analyses. *Genomics*. 10.1016/j.ygeno.2014.03.001.
- Rohart, F. (2011). Multiple hypotheses testing for variable selection.
- Rohart, F., Laurent, B., and SanCristobal, M. (2014). Selection of fixed effects in high dimensional linear mixed models using a multicycle ecm algorithm. *Computational Statistics and Data Analysis*. 10.1016/j.csda.2014.06.022.
- Rohart, F., Paris, A., Laurent, B., Canlet, C., Molina, J., Mercat, M., Tribout, T., Muller, N., Iannucelli, N., Villa-Vialaneix, N., Liaubet, L., Milan, D., and San Cristobal, M. (2012). Phenotypic prediction based on metabolomic data on the growing pig from three main european breeds. *Journal of Animal Science*, 90(13):4729–4740.
- Schelldorfer, J., Bühlmann, P., and van de Geer, S. (2011). Estimation for high-dimensional linear mixed-effects models using ℓ_1 -penalization. *Scandinavian Journal of Statistics*, 38:197–214.
- Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society*, B 58(1):267–288.