

# Boris Hejblum

*Associate Professor in Biostatistics*

## Research experience

- 2016–present **Associate Professor (*Maître de Conférences*)**, Bordeaux University.  
Inria BSO, Inserm Bordeaux Public Health U1219, *SISTM* team.
- 2016 **Research Associate**, Department of Biostatistics, *Harvard School of Public Health*.
- 2015–2016 **Postdoctoral Research Fellow**, Department of Biostatistics, *Harvard School of Public Health*.
- 2011–2015 **Research Assistant** (Ph.D. student), Department of Biostatistics, *ISPED Bordeaux School of Public Health, Bordeaux University*.
- Apr.–Sept. 2011 **Research Assistant** (intern), Inserm U897, *Biostatistics team*.  
Development of dynamic statistical models applied to the epidemiology of myocardial infarction.
- May–Jul. 2010 **Statistician Assistant**, *AltraBio* (start-up in biotechnologies), Lyon, France.  
Analysis of transcriptomics data of preclinical trials (internship).

## Education

- 2011–2015 **Ph.D. in Biostatistics**, Bordeaux University.  
Integrative analysis of high-dimensional data applied to vaccine research.  
Advisors: Pr. Rodolphe Thiébaud ([Rodolphe.Thiebaut@u-bordeaux.fr](mailto:Rodolphe.Thiebaut@u-bordeaux.fr)),  
François Caron ([caron@stats.ox.ac.uk](mailto:caron@stats.ox.ac.uk))
- 2008–2011 **Master of Science (M.Sc.) in Statistics** (equivalent: *diplôme d'ingénieur*), ENSAI, National School for Statistics and Information Analysis (*École Nationale de la Statistique et de l'Analyse de l'Information*), Rennes, France.  
Specialization in biostatistics, with high honors
- 2011 **Master of Science (M.Sc.) in Statistics and Econometrics** (*Master de statistique et économétrie*), Department of Mathematics, University of Rennes 1.  
Dual degree partnership in conjunction with studies at ENSAI (additional education focused on scientific research).
- 2009 **Bachelor of Science (B.Sc.) in Mathematics** (equivalent: *licence de mathématiques*), Pierre and Marie Curie University – Paris 6 (UPMC), Paris, France.  
In conjunction with studies at ENSAI (dual curriculum, by correspondence).
- 2006–2008 **Post-Secondary Preparatory Classes** (Classes Préparatoires aux Grandes Écoles – CPGE), Lycée Hoche, Versailles, France.  
University-level courses required in preparation for competitive exams into top universities, engineering, and graduate schools (France's "Grandes Écoles"). Major in Mathematics and Physics.
- 2006 **High school diploma**, Lycée Richelieu, Rueil-Malmaison, France.  
with high honors

---

## Teaching experience

- 2016–2018 **Associate Professor**, Bordeaux University, France.  
Master in Epidemiology and Master in Biostatistics level courses:
- Factor methods for multivariate data analysis (graduate class, 30h)
  - Bayesian analysis and sampling methods (graduate class, 24h)
  - sparse Partial Least Squares methods (graduate class, 7h)
  - ANOVA regression (graduate class, 7.5h)
- 2013–2014 **Teaching Assistant**, Bordeaux University, France.  
Master in Epidemiology and Master in Biostatistics level courses:
- MCMC methods for bayesian analysis (graduate class, 12h)
  - sparse Partial Least Squares methods (graduate class, 5h)
  - basic statistics (undergraduate class 16h)
  - logistic regression (undegraduate class, 12h)
  - R software (undergraduate class 9h)
- 2012-2013 **Teaching Assistant**, Bordeaux University, France.  
Master in Epidemiology and Master in Biostatistics level courses:
- sparse Partial Least Squares methods (graduate class, 5h)
  - logistic regression (undegraduate class, 12h)
- 2014-2018 **Intern supervisor**, master thesis.
- Marine Gauthier (2018 – 100%)
  - Roxane Coueron (2018 – 50%)
  - Paul Tautzia (2017 – 50%)
  - Chariff Alkhassim (2015 – 50%)
  - Damien Chimits (2014 – 50%)

---

## Grants

- 2018 Principal Investigator of the Inria associate team SWAGR
- 2016 Travel grant from the Harvard Program in Quantitative Genomics (PQG) to attend the ENAR conference.
- 2011 Ph.D. grant from the EHESP (*École des Hautes Études en Santé Publique*, Rennes, France) – ranked 1<sup>st</sup>.

---

## Research expertise

**Evidence synthesis causal analysis:** I studied stochastic modeling of life-course health data. The developed idea was to explore potential causal factors of myocardial infarction by relating the drift of a degradation process with metadata from the literature.








**Statistical genomics & high dimensional data:** I have a strong interest in models for high dimensional data. I am familiar with the multiple testing issue and potential strategies to face it. I have worked on sparse Partial Least Squares methods, and with other dimension reduction approaches such as the random forests or the LASSO. I have analyzed gene expression data in a clinical trial context and I am familiar with the specificities of this kind of data, such as preprocessing.

**Bayesian nonparametric models:** I am interested in statistical learning methods such as nonparametric Bayesian mixture of skew distributions for the clustering of large cell populations.

**Electronic Health Records:** I am currently developing models to perform probabilistic record linkage to match electronic health records without using identifiable variables.

---

## Software development

- 2017 **ludic**: an  package for probabilistic record linkage using diagnosis codes. Available on CRAN, development version on GitHub. *Co-creator & maintainer.*
- 2017 **cytometree**: an  package for automatic gating and annotation of flow-cytometry data. Available on CRAN, development version on GitHub. *Co-creator.*
- 2017 **sslcov**: an  package for covariance semi-supervised learning. Available on GitHub. *Co-creator.*
- 2016 **tcgsaseq**: an  package to analyze (longitudinal) RNA-seq data (at the gene set level). Available on CRAN, development version on GitHub. *Co-creator & maintainer.*
- 2016 **kernscr**: an  package to perform survival analysis by gene sets in presence of competing risks. Available on CRAN, development version on GitHub. *Co-creator & maintainer.*
- 2015 **NPflow**: an  package to perform clustering of large cell populations with Dirichlet process mixture of skew-Normal and skew-t distributions. Available on CRAN, development version on GitHub. Uses C++ code to speed up computation. *Co-creator & maintainer.*
- 2014 **TcGSA**: an  package to analyze longitudinal gene-expression data at the gene set level. Available on CRAN, development version on GitHub. *Co-creator & maintainer.*

---

## Research collaborations

**Denis Agniel**, *Rand Corporation, Statistical group, Santa Monica (CA, USA), Associate Statistician.*

**Katherine P. Liao**, *Brigham and Women's Hospital - Harvard Medical School, Rheumatology, Boston (MA, USA), Assistant Professor.*

**Tianxi Cai**, *Harvard TH Chan School of Public Health, Department of Biostatistics, Boston (MA, USA), Professor.*

**Pierre E. Jacob**, *Harvard University, Department of Statistics, Cambridge (MA, USA), Assistant Professor.*

**François Caron**, *University of Oxford, Department of Statistics, Oxford (United-Kingdom), Research Fellow.*

**David Furman**, *Stanford University, Stanford School of Medicine, Microbiology and Immunology Department, Stanford (California, USA), Postdoctoral Fellow.*

**Jason Skinner**, *Baylor Institute for Immunology Research, Dallas (Texas, USA), Senior Postdoctoral Fellow.*

---

## Research visits abroad

- 2016-2017  
(2 × 1 week) **Harvard University, Department of Biostatistics**, Cambridge (MA, USA)  
invited by Tianxi Cai, Professor.
- 2013-2014  
(3 × 1 week) **University of Oxford, Department of Statistics**, Oxford (United-Kingdom)  
invited by François Caron, Research Fellow.
- 2012 (1 month) **Benaroya Research Institute**, Chaussabel Laboratory, Seattle (WA, USA)  
invited by Damien Chaussabel, Director of Systems Immunology.
- 2011 (1 month) **Baylor Institute for Immunology Research**, Dallas (TX, USA).

---

## Reviewer for scientific journals

*Biometrics, Annals of Applied Statistics, BioData Mining, Statistical Applications in Genetics and Molecular Biology*

---

## Academic Associations

- 2012–2014 **Coordinator of the ISPED Ph.D. students** (weekly) seminar.
- 2009–2010 **President** (formerly Secretary General) **of the ENSAI Business Networking Forum Team.**  
Responsible for organizing the yearly networking event between companies and ENSAI students
- 2009 **Vice President of the ENSAI Student Council.**  
Organize and coordinate associative activities and social life at the school

---

## Publications

▷ **Published/in press:** (\* indicates equal contribution)

- S Lefèvre-Arbogast, D Gaudout, J Bensalem, L Letenneur, JF Dartigues, BP Hejblum, C Féart, C Delcourt, C Samieri, Pattern of polyphenol intake and the long-term risk of dementia in older persons, *Neurology* (in press), 2018.
- BP Hejblum, J Cui, L Lahey, A Cagan, JA Sparks, S Shaw, J Sokolove, T Cai, KP Liao, Association of specific anti-citrullinated peptide antibodies with coronary artery disease in rheumatoid arthritis, *Arthritis and Care Research*, (in press), 2017.  
DOI: 10.1002/acr.23444
- D Agniel, BP Hejblum, Variance component score test for time-course gene set analysis of longitudinal RNA-seq data, *Biostatistics*, 18(4):589–604, 2017.  
DOI: 10.1093/biostatistics/kxx005
- A Rechten, L Richert, H Lorenzo, G Martrus, BP Hejblum, C Dahlke, R Kasonta, M Zinser, H Stubbe, U Matschl, A Lohse, V Krähling, M Eickmann, S Becker, VEBCON Consortium, R Thiébaud, M Altfeld, and M Addo, Systems Vaccinology Identifies an Early Innate Immune Signature as a Correlate of Antibody Responses to the Ebola Vaccine rVSV-ZEBOV, *Cell Reports*, 20(9):2251–2261, 2017.  
DOI: 10.1016/j.celrep.2017.08.023
- KP Liao\*, JA Sparks\*, BP Hejblum, IH Kuo, J Cui, LJ Lahey, A Cagan, VS Gainer, W Liu, TT Cai, J Sokolove, T Cai, Phenome-wide association study of autoantibodies to citrullinated and non-citrullinated epitopes in rheumatoid arthritis, *Arthritis & Rheumatology*, 69:742–749, 2017.  
DOI: 10.1002/art.39974
- M Neykov, BP Hejblum, JA Sinnott, Kernel Machine Score test for pathway analysis in the presence of semi-competing Risks, *Statistical Methods for Medical Research*, (sous presse), 2016.  
DOI: 10.1177/0962280216653427
- B Liquet, P Lafaye de Micheaux, BP Hejblum, R Thiébaud, Group and sparse group Partial Least Square approaches applied in genomics context, *Bioinformatics*, 32 (1): 35-42, 2016.  
DOI: 10.1093/bioinformatics/btv535

- BP Hejblum, J Skinner, R Thiébaud, TcGSA: a gene set approach for longitudinal gene expression data analysis, *PLOS Computational Biology*, 11 (6):e1004310, 2015. DOI: 10.1371/journal.pcbi.1004310
- D Furman\*, BP Hejblum\*, N Simon, V Jojic, CL Dekker, R Thiébaud, RJ Tibshirani, MM Davis, A systems analysis of sex differences reveals an immunosuppressive role for testosterone in the response to influenza vaccination, *Proceedings of the National Academy of Sciences of the United States of America*, 111(2):869–874, 2014. DOI: 10.1073/pnas.1321060111.
- R Thiébaud, B Hejblum, L Richert, L'analyse des "Big Data" en recherche clinique, *Revue d'Épidémiologie et de Santé Publique*, 62(1):1–4, 2014. DOI: 10.1016/j.respe.2013.12.021.
- D Commenges & BP Hejblum, Evidence synthesis through a degradation model applied to myocardial infarction, *Lifetime data analysis*, 19(1):1–18, 2013. DOI: 10.1007/s10985-012-9227-3.

▷ In revision/under review/submitted/in preparation:

- D Commenges, C Alkassim, R Gottardo, BP Hejblum, R Thiébaud, cytomtree: a Binary Tree Algorithm for Automatic Gating in Cytometry Analysis (in revision).
- BP Hejblum, G Weber, KP Liao, N Palmer, S Churchill, P Szolovits, S Murphy, I Kohane, T Cai, Probabilistic Record Linkage of De-Identified Research Datasets Using Diagnosis Codes (in revision).
- BP Hejblum, C Alkassim, R Gottardo, F Caron, R Thiébaud, Sequential Dirichlet process mixture of skew t-distributions for model-based clustering of flow cytometry data (in revision).
- S Chan, BP Hejblum, A Chakraborty, T Cai, Semi-Supervised Estimation of Covariance with Application to Phenome-wide Association Studies with Electronic Medical Records Data (in revision).
- BP Hejblum, G Weber, KP Liao, N Palmer, S Churchill, P Szolovits, S Murphy, I Kohane, T Cai, Genotype Association Testing through Probabilistically Linked De-identified Electronic Health Records (in preparation).
- BP Hejblum, S Chan, T Cai, Adjusted PheWAS with Semi-Supervised Estimation of Conditional Covariance in Presence of Surrogate Variables and Non-Random Sampling (in preparation).

## Selected communications

▷ Oral communications:

- UK 2017 B Hejblum, A Bayesian model-based approach to finding cell-type level associations in heterogeneous methylation samples, *BSU invited Seminar*.
- Spain 2017 B Hejblum, D Agniel, Type I error and False discovery rate control in RNA-seq differential analyses through a variance component score test, *ISCB 2017, 38th Annual Conference of the International Society for Clinical Biostatistics, Vigo, 2017*.
- USA 2016 B Hejblum, D Agniel, Time-course Gene Set Analysis of longitudinal RNA-seq data, *ENAR 2016 Spring Meeting, Austin (TX)*.
- Italy 2014 B Hejblum, F Caron, R Thiébaud, Bayesian analysis of time-course flow cytometry data with Dirichlet process mixture modeling, *27<sup>th</sup> International Biometric Conference, Florence 2014*.

- France 2014 B Hejblum, R Genuer, R Thiébaud, Variable selection in high-dimensional dataset: comparison of sPLS with other approaches in an HIV vaccine trial, *8<sup>th</sup> International Conference on Partial Least Squares and Related Methods, Paris 2014*.
- France 2014 Invited speaker at the Ph.D. students working group of the LSTA (*Laboratoire de Statistique Théorique et Appliquée*) in Paris 6 University, B Hejblum, F Caron, R Thiébaud, Bayesian nonparametric modeling of flow cytometry data with Dirichlet process mixtures.
- Spain 2013 R Thiébaud, B Hejblum, J Skinner, M Montes, G Chene, K Palucka, J Banchereau, Y. Levy, Integrative Analysis of Responses to Dendritic-Cell Vaccination Identifies Signatures Correlated with Control of HIV Replication: The DALIA Trial, *AIDS Vaccine 2013, Barcelona 2013, AIDS Research and Human Retroviruses 29 (11), A5-A6*.
- Norway 2012 B Hejblum, J Skinner, R Thiébaud, Application of Gene Set Analysis of Time-Course gene expression in a HIV vaccine trial, *33<sup>rd</sup> Annual conference of the International Society for Clinical Biostatistics, Bergen 2012*.

▷ **Written communications:**

- France 2017 BP Hejblum, C Alkhassim, R Gottardo, F Caron, R Thiébaud, Sequential Dirichlet process mixture of skew t-distributions for model-based clustering of flow cytometry data, *BNP11 Meeting: 11<sup>th</sup> Conference on Bayesian Nonparametrics, Paris, 2017*.
- USA 2015 B Hejblum, T Cai, G Weber, PIC-SURE Patient Linkage Working Group, Probabilistic Patient Linkage Algorithms for PIC-SURE, *BD2K all Hands Meeting 2015, Bethesda, MA*.
- United Kingdom 2014 B Hejblum, F Caron, R Thiébaud, Hierarchical analysis of time-course flow cytometry data with Dirichlet process mixture modeling, *Medical Research Council Conference on Biostatistics in celebration of the MRC Biostatistics Unit's centenary year, Cambridge 2014*.
- Netherlands 2013 B Hejblum, J Skinner, R Thiébaud, Time-course Gene Set Analysis applied in a HIV vaccine trial, *Statistical Methods for (post)-Genomics Data 2013, Amsterdam 2013*.
- Netherlands 2012 C Chauvin, C Fablet, V Bal, B Hejblum, P-O Goffard, P Sanders, E Jouy, Identification of the best antimicrobial usage indicator for resistance trend analysis: an application to time series in poultry production, *13<sup>th</sup> International Symposium for Veterinary Epidemiology and Economics ISVEE13, Maastricht 2012*.