

# Boris Hejblum

*Associate Professor in Biostatistics*  
*(Maître de Conférences)*

Université de Bordeaux - ISPED  
146 rue Léo Saignat  
33076 Bordeaux cedex  
FRANCE

✉ [boris.hejblum@u-bordeaux.fr](mailto:boris.hejblum@u-bordeaux.fr)  
French, 31 years old

## Research experience

- 2016–present **Associate Professor (*Maître de Conférences*)**, Bordeaux University.  
Inserm Bordeaux Public Health U1219, Inria Bordeaux Sud-Ouest, *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 - 2019 **Associate Professor**, Bordeaux University, France.
- Ph.D. courses:
- R for development & performance (graduate class, 12h per year)
- Master in Public Health Data Science and Master in Biostatistics courses:
- likelihood estimation and multivariate regression (graduate class, 30h per year)
  - factor methods for multivariate data analysis (graduate class, 30h per year)
  - Bayesian analysis and sampling methods (graduate class, 30h per year)
  - Genomics data analysis (graduate class, 20h per year)
  - sparse Partial Least Squares methods (graduate class, 7h per year)
  - ANOVA regression (graduate class, 7.5h)
  - hypothesis testing (graduate class, 30h)
  - R software for data science (undergraduate and graduate class, 30h per year)
- 2019 **International Ph.D. course**, University of Copenhagen (Graduate School of Health and Medical Sciences), Denmark.
- Bayesian analysis for medical studies
- 2012 - 2014 **Teaching Assistant**, Bordeaux University, France.
- Master in Public Health and Master in Biostatistics courses:
- MCMC methods for Bayesian analysis (graduate class, 12h)
  - sparse Partial Least Squares methods (graduate class, 5h)
  - basic statistics (undergraduate class 16h)
  - logistic regression (undergraduate class, 12h)
  - R software (undergraduate class 9h)
- 2018 - ... **Ph.D. student supervisor**.
- Marine Gauthier (2018 - ... 50%)
  - Soufiane Ajana (2017 - 2019 15%)
- 2014 - 2018 **Intern supervisor**, master thesis.
- Anthony Devaux (2019 100%)
  - Thomas Ferte (2019 100%)
  - Marine Gauthier (2018 100%)
  - Roxane Coueron (2018 50%)
  - Paul Tautzia (2017 50%)
  - Chariff Alkhassim (2015 50%)
  - Damien Chimits (2014 50%)

---

## Grants

- 2018-2020 Principal Investigator of the Inria associate-team "Statistical Workforce for Advanced Genomics using RNA-seq" (SWAGR : 36K€ over 3 years – renewable)
- 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

**Artificial Intelligence for health:** I have developed various artificial intelligence approaches to solve biomedical data analysis bottlenecks. In particular, I am working on machine learning approaches to automate the processing of flow and mass cytometry measurements, and also on automated medical diagnosis from both structured data and free text medical notes in English, French and Chinese through language agnostic algorithms.

**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.












**Electronic Health Records:** I am currently developing models to perform probabilistic record linkage to match electronic health records without using identifier variables, and to predict disease phenotype from electronic health record data, with application in infection and rheumatoid arthritis.

**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.

**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.

---

## Software development & maintenance

- 2020 **dearseq:** an  package to perform Differential Expression Analysis for RNA-seq data through a robust variance component test. Submitted to . Development version available on GitHub. *Creator & maintainer.*
- 2019 **vici:** an interactive  Shiny application to perform for accurate estimation of vaccine induced cellular immunogenicity with bivariate linear modeling. Available online or locally from the CRAN. Development version available on GitHub. *Creator & maintainer.*
- 2019 **phenotypr:** an  package for probabilistic phenotyping patients from electronic health records using both diagnosis codes and natural language processed medical notes using diagnosis codes. Development version available on GitHub. *Creator & maintainer.*
- 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 & maintainer.*
- 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. *Creator & maintainer.*

---

## Active international research collaborations

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

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

**François Caron**, *University of Oxford, Department of Statistics*, Oxford (United-Kingdom), Associate Professor.

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

**Paul Kirk**, *MRC-Biostatistics Unit, University of Cambridge*, Cambridge (United-Kingdom), Group leader.

---

## Research visits abroad

- 2018-2019  
(2×3 weeks) **MRC Biostatistics Unit, Cambridge University**, Cambridge (United-Kingdom)  
invited by Sylvia Richardson, Professor.
- 2018 (1 week) **Rand Corporation, Statistics group**, Santa Monica (CA, USA)  
invited by Denis Agniel, Associate Statistician.
- 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 international peer-reviewed scientific journals

*Annals of Applied Statistics, BioData Mining, Bioinformatics, Biometrics, Cancer Reports, Computational Statistics Data Analysis, Journal of Statistical Computation and Simulation, PLOS Computational Biology, Scientific Reports, Statistical Applications in Genetics and Molecular Biology*

---

## Academic responsibilities

- 2019 **Webmaster for the French Biometric Society** (*Société Française de Biométrie*).
- 2019 **Co-organizer of the Bordeaux Statistics Seminar** (quarterly).
- 2017-2019 **Organizer of the Inserm BPH Biostatistics** (biweekly) **seminar**.
- 2018 **Co-organizer of the workshop in honor of Daniel Commenges' 70<sup>th</sup> birthday**.
- 2012-2014 **Founder of the ISPED Ph.D. students** (weekly) **seminar**.
- 2009-2010 **President** (formerly Secretary General) **of the ENSAI Business Networking Forum**.  
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

### ▷ Books:

- L Desquilbet, S Granger, BP Hejblum, A Legrand, P Pernot, NP Rougier, *Vers une recherche reproductible : Faire évoluer ses pratiques*, 2019.  
ISBN : 979-10-97595-05-0

### ▷ Preprints:

- M Gauthier, D Agniel, R Thiébaut, BP Hejblum, dearseq: a variance component score test for RNA-Seq differential analysis that effectively controls the false discovery rate, *bioRxiv* 635714, 2019. DOI: 10.1101/635714

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

- BP Hejblum, C Alkhassim, R Gottardo, F Caron, R Thiébaut, Sequential Dirichlet process mixture of skew t-distributions for model-based clustering of flow cytometry data, *Annals of Applied Statistics*, 13(1):638-660, 2019.  
DOI: 10.1214/18-AOAS1209
- 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, *Scientific Data*, 6:180298, 2019.  
DOI: 10.1038/sdata.2018.298
- E Lhomme, BP Hejblum, C Lacabartz, A Wiedemann, JD Lelièvre, Y Lévy, R Thiébaut, L Richert. Analyzing cellular immunogenicity in vaccine clinical trials: a new statistical method including non-specific responses for accurate estimation of vaccine effect, *Journal of Immunological Methods*, (in press), 2019.  
DOI: 10.1016/j.jim.2019.112711
- 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, *Statistical Methods for Medical Research*, (in press), 2019.  
DOI: 10.1177/0962280219837676
- S Ajana, A Niyazi, L Bretillon, BP Hejblum, H Jacqmin-Gadda, C Delcourt, Benefits of dimension reduction in penalized regression methods for high dimensional grouped data: a case study in low sample size, *Bioinformatics*, btz135, (in press), 2019.  
DOI: 10.1093/bioinformatics/btz135
- R Thiébaut, BP Hejblum, H Hocini, H Bonnabau, J Skinner, M Montes, C Lacabartz, L Richert, K Palucka, J Banchereau, Y Levy, Gene expression signatures associated with immune and virological responses to therapeutic vaccination with Dendritic Cells in HIV-infected individuals, *Frontiers in Immunology*, 10:874, 2019.  
DOI: 10.3389/fimmu.2019.00874
- D Commenges, C Alkhassim, R Gottardo, BP Hejblum, R Thiébaut, cytometree: a Binary Tree Algorithm for Automatic Gating in Cytometry Analysis, *Cytometry: Part A*, 93(11):1132–1140, 2018.  
DOI: 10.1002/cyto.a.23601
- JA Sinnott, F Cai, S Yu, BP Hejblum, C Huong, IS Kohane, KP Liao, PheProb: probabilistic phenotyping using diagnosis codes to improve power for genetic association studies, *Journal of the American Medical Informatics Association*, 25(10):1359–1365, 2018.  
DOI: 10.1093/jamia/ocy056

- 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*, , 70:1113–1117, 2018.  
DOI: 10.1002/acr.23444
- 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*, 27(4):1099–1114, 2018.  
DOI: 10.1177/0962280216653427
- 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*, 90(22):e1979–e1988, 2018.  
DOI: 10.1212/WNL.0000000000005607
- 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 Rechtién, 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 Thiébaut, 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
- B Lique, P Lafaye de Micheaux, BP Hejblum, R Thiébaut, 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 Thiébaut, 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 Thiébaut, 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 Thiébaut, 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.

## Selected communications

▷ Oral communications: (\* indicates invited talks)

- UK 2019 \* B Hejblum; PDW Kirk, Scaling up nonparametric Bayesian clustering with MCMC for big data applications, *12<sup>th</sup> International Conference of the ERCIM WG on Computational and Methodological Statistics*, London.
- France 2019 B Hejblum, E Lhomme, R Thiébaut, L Richert, VICI: a Shiny app for accurate estimation of Vaccine Induced Cellular Immunogenicity with bivariate modeling, *UseR! 2019*, Toulouse.
- Taiwan 2019 \* B Hejblum, M Gauthier, R Thiébaut, D Agniel, A variance component score test applied to RNA-Seq differential analysis, *3<sup>rd</sup> EcoSta Conference*, Taichung.
- France 2019 \* B Hejblum, M Gauthier, R Thiébaut, D Agniel, Controlling Type-I error in RNA-seq differential analyses through a variance component score test with an application to tuberculosis infection, *Séminaire de l'équipe de Statistique de l'Institut de Recherche Mathématique de Rennes (IRMAR)*, Rennes.
- France 2018 \* B Hejblum, M Gauthier, R Thiébaut, D Agniel, Controlling type-I error and false discoveries in RNA-seq differential analyses through a variance component score test, *Bioinfo-Biostat GenoToul Annual Day*, Toulouse.
- Spain 2018 B Hejblum, D Agniel, A Variance Component Score Test for RNA-Seq Differential Analysis in Vaccine Trials, *IBC 2018, 29<sup>th</sup> International Biometric Conference*, Barcelona.
- 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, 38<sup>th</sup> Annual Conference of the International Society for Clinical Biostatistics*, Vigo.
- 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 Thiébaut, 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 Thiébaut, 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 \* B Hejblum, F Caron, R Thiébaut, Bayesian nonparametric modeling of flow cytometry data with Dirichlet process mixtures, *Ph.D. students working group of the LSTA (Laboratoire de Statistique Théorique et Appliquée) in Paris 6 University*.
- Spain 2013 R Thiébaut, 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, AIDS Research and Human Retroviruses*, 29 (11), A5-A6, Barcelona.
- Norway 2012 B Hejblum, J Skinner, R Thiébaut, 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.

▷ Written communications:

- 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 Thiébaut, 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*.