Boris Hejblum

Associate Professor in Biostatistics

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

✉ *boris.hejblum@u-bordeaux.fr*
French, 29 years old

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 Thiébaut (`Rodolphe.Thiebaut@u-bordeaux.fr`), François Caron (`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 per year)
- Bayesian analysis and sampling methods (graduate class, 24h per year)
- sparse Partial Least Squares methods (graduate class, 7h per year)
- ANOVA regression (gratuate class, 7.5h)
- hypothesis testing (graduate class, 30h)
- R software (undergraduate and graduate class, 15h per year)

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 (undegratuate 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 (undegratuate class, 12h)

2014-2018   **Intern supervisor**, master thesis.
- Marine Gauthier (2018 – 100%)
- Roxane Coueron (2018 – 50%)
- Paul Tauzia (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 1st.

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 developping 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 rheumatoïd arthritis.

Software development

2018   **phenotypr**: an ℝ package for probabilitic phenotyping patients from electonic health records using both diagnosis codes and natural language processed medical notesusing diagnosis codes. Development version available on GitHub. *Creator & maintainer*.

2017   **ludic**: an ℝ package for probabilistic record linkage using diagnosis codes. Available on CRAN, development version on GitHub. *Co-creator & maintainer*.

2017   **cytometree**: an ℝ package for automatic gating and annotation of flow-cytometry data. Available on CRAN, development version on GitHub. *Co-creator*.

2017   **sslcov**: an ℝ package for covariance semi-supervised learning. Available on GitHub. *Co-creator*.

2016   **tcgsaseq**: an ℝ 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 ℝ 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 ℝ 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 ℝ package to analyze longitudinal gene-expression data at the gene set level. Available on CRAN, development version on GitHub. *Co-creator & maintainer*.

Active international 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.

Research visits abroad

2018 (1 week)   **Rand Corporation, Statistical 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

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

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)

- JA Sinnott, F Cai, S Yu, <u>BP Hejblum</u>, 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, (in press), 2018. DOI: 10.1093/jamia/ocy056

- S Lefèvre-Arbogast, D Gaudout, J Bensalem, L Letenneur, JF Dartigues, <u>BP Hejblum</u>, C Féart, C Delcourt, C Samieri, Pattern of polyphenol intake and the long-term risk of dementia in older persons, *Neurology* (in press), 2018. DOI: 10.1212/WNL.0000000000005607

- <u>BP Hejblum</u>, 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, <u>BP Hejblum</u>, 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 Rechtien, L Richert, H Lorenzo, G Martrus, <u>BP Hejblum</u>, 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*, <u>BP Hejblum</u>, 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, <u>BP Hejblum</u>, 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, <u>BP Hejblum</u>, 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`

- <u>BP Hejblum</u>, 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*, <u>BP Hejblum</u>*, 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, <u>B Hejblum</u>, 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 & <u>BP Hejblum</u>, 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 Alkhassim, R Gottardo, <u>BP Hejblum</u>, R Thiébaut, cytometree: a Binary Tree Algorithm for Automatic Gating in Cytometry Analysis (in revision).

- <u>BP Hejblum</u>, 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).

- <u>BP Hejblum</u>, C Alkhassim, R Gottardo, F Caron, R Thiébaut, Sequential Dirichlet process mixture of skew t-distributions for model-based clustering of flow cytometry data (in revision).

- S Chan, <u>BP Hejblum</u>, A Chakrabortty, T Cai, Semi-Supervised Estimation of Covariance with Application to Phenome-wide Association Studies with Electronic Medical Records Data (in revision).

- <u>BP Hejblum</u>, 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).

- <u>BP Hejblum</u>, S Chan, T Cai, Adjusted PheWAS with Semi-Supervised Estimation of Conditional Covariance in Presence of Surrogates Variables and Non-Random Sampling (in preparation).

Selected communications

▷ Oral communications: (* indicates invited talks)

Spain 2018 <u>B Hejblum</u>, D Agniel, Type I error and False discovery rate control in RNA-seq differential analyses through a variance component score test, *ISCB 2017, 29th International Biometric Conference, Barcelona, 2018*.

UK 2017 * <u>B Hejblum</u>, 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 Thiébaut, Bayesian analysis of time-course flow cytometry data with Dirichlet process mixture modeling, *27th 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, *8th 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 Thiébaut, Bayesian nonparametric modeling of flow cytometry data with Dirichlet process mixtures.
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, Barcelona 2013, AIDS Research and Human Retroviruses 29 (11), A5-A6*.
Norway 2012	B Hejblum, J Skinner, R Thiébaut, Application of Gene Set Analysis of Time-Course gene expression in a HIV vaccine trial, *33rd Annual conference of the International Society for Clinical Biostatistics, Bergen 2012*.

▷ Written communications:

France 2017	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, *BNP11 Meeting: 11th 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 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*.
Netherlands 2013	B Hejblum, J Skinner, R Thiébaut, 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, *13th International Symposium for Veterinary Epidemiology and Economics ISVEE13, Maastricht 2012*.