Activity Report 2019

Project-Team SISTM

Statistics In System biology and Translational Medicine

Table of contents

Project-Team SISTM

Creation of the Team: 2013 April 02, updated into Project-Team: 2015 January 01

Keywords:

Computer Science and Digital Science:

A3.3.2. - Data mining
A3.3.3. - Big data analysis
A3.4.1. - Supervised learning
A3.4.2. - Unsupervised learning
A3.4.4. - Optimization and learning
A3.4.5. - Bayesian methods
A6.1.1. - Continuous Modeling (PDE, ODE)
A6.2.4. - Statistical methods
A6.3.1. - Inverse problems
A6.3.4. - Model reduction
A6.4.2. - Stochastic control
A9.2. - Machine learning

Other Research Topics and Application Domains:

B1.1. - Biology
B1.1.5. - Immunology
B1.1.7. - Bioinformatics
B1.1.10. - Systems and synthetic biology
B2.2.4. - Infectious diseases, Virology
B2.2.5. - Immune system diseases
B2.3. - Epidemiology
B2.4.1. - Pharmaco kinetics and dynamics
B2.4.2. - Drug resistance

1. Team, Visitors, External Collaborators

Research Scientists
Daniel Commenges [Inserm, Senior Researcher, HDR]
Melanie Prague [Inria, Researcher]

Faculty Members
Rodolphe Thiébaut [Team leader, Univ de Bordeaux, Professor, HDR]
Marta Avalos Fernandez [Univ de Bordeaux, Associate Professor, HDR]
Robin Genuer [Univ de Bordeaux, Associate Professor]
Boris Hejblum [Univ de Bordeaux, Associate Professor]
Laura Richert [Univ de Bordeaux, Associate Professor, HDR]

Post-Doctoral Fellows
Quentin Clairon [Inserm, Post-Doctoral Fellow, from Mar 2019]
Irène Balelli [Inserm, Post-Doctoral Fellow, until Jul 2019]
Laura Villain [Inria, Post-Doctoral Fellow]
Binbin Xu [Univ de Bordeaux, Post-Doctoral Fellow, until Jun 2019]

PhD Students

    Marie Alexandre [Inria, PhD Student]
    Louis Capitaine [Univ de Bordeaux, PhD Student]
    Marine Gauthier [Univ de Bordeaux, PhD Student]
    Edouard Lhomme [Univ de Bordeaux, PhD Student]
    Hadrien Lorenzo [Inserm, PhD Student]
    Perrine Soret [Univ de Bordeaux, PhD Student]

Technical staff

    Fabien Barbier [Inserm, Engineer, until Aug 2019]
    Solenne Delahaye [Inserm, Engineer, until Dec 2019]
    Melany Durand [Inserm, Engineer]
    Miriam Krueger [Inserm, Engineer, from Sep 2019]
    Clement Nerestan [Inria, Engineer, from Oct 2019]
    Van Hung Tran [Inserm, Engineer, until Sep 2019]

Interns and Apprentices

    Thibaud Alin [Inria, from Mar 2019 until Aug 2019]
    Anthony Devaux [Inria, from Feb 2019 until Jul 2019]
    Victor Gasque [Inserm, from Jun 2019]
    Mohamed Harith Ibrahim [Inria, from Jun 2019 until Jul 2019]
    Clément Lemoigne [Inria, from Jul 2019 until Sep 2019]
    Guillaume Sotton [Inria, from Feb 2019 until Aug 2019]
    Alice Herteau [Inserm, from Feb to Oct 2019]
    Thomas Ferte [Inserm, from Feb to Oct 2019]
    Valentin Kochegarov [Inserm, from May to Jul 2019]

Administrative Assistants

    Cécile Boutros [Inria, Administrative Assistant, until Apr 2019]
    Audrey Plaza [Inria, Administrative Assistant, from Jul 2019]
    Sandrine Darmigny [Inserm, Administrative Assistant]
    Anton Ottavi [Inserm, European Projects coordinator]
    Destandau Eugénie [Inserm, Communication, until Dec 2019]

2. Overall Objectives

2.1. Overall Theoretical Objectives

The overall objective of SISTM is to develop statistical methods for the integrative analysis of health data, especially those related to clinical immunology and vaccinology to answer specific questions risen in the application field. To reach this objective we are developing statistical methods belonging to two main research areas:

- Statistical and mechanistic modeling, especially based on ordinary differential equation systems, fitted to population and sparse data
- Statistical learning methods in the context of high-dimensional data

These two approaches are used for addressing different types of questions. Statistical learning methods are developed and applied to deal with the high-dimensional characteristics of the data. The outcome of this research leads to hypotheses linked to a restricted number of markers. Mechanistic models are then developed and used for modeling the dynamics of a few markers. For example, regularized methods can be used to select relevant genes among 20000 measured with microarray/RNA-seq technologies, whereas differential equations can be used to capture the dynamics and relationship between several genes followed over time by a q-PCR assay or RNA-seq. We apply the methods developed by the team in data science approaches to vaccine trials in order to elucidate the effects and mechanisms of action of vaccines and immunotherapies and to accelerate their clinical development.

2.2. Overall Applied Objectives

Data are generated in clinical trials or biological experimentations. Our main application of interest is the immune response to vaccines or other immune interventions (such as exogenous cytokines), in the context of HIV or Ebola infection. The methods developed in this context can be applied in other circumstances but the focus of the team on immunology and vaccinology is important for the relevance of the results and their translation into practice, thanks to a longstanding collaboration with several immunologists and the implication of the team in the Labex Vaccine Research Institute (http://vaccine-research-institute.fr). Examples of objectives related to this application field are:

- To understand how immune responses are generated with immune interventions (vaccines or immunotherapies)
- To predict what would be the immune responses to a given immune intervention in order to design next studies and to adapt interventions to individual persons or to specific populations

3. Research Program

3.1. Mechanistic learning

When studying the dynamics of a given marker, say the HIV concentration in the blood (HIV viral load), one can for instance use descriptive models summarizing the dynamics over time in term of slopes of the trajectories [56]. These slopes can be compared between treatment groups or according to patients' characteristics. Another way for analyzing these data is to define a mathematical model based on the biological knowledge of what drives HIV dynamics. In this case, it is mainly the availability of target cells (the CD4+ T lymphocytes), the production and death rates of infected cells and the clearance of the viral particles that impact the dynamics. Then, a mathematical model most often based on ordinary differential equations (ODE) can be written [48]. Estimating the parameters of this model to fit observed HIV viral load gave a crucial insight in HIV pathogenesis as it revealed the very short half-life of the virions and infected cells and therefore a very high turnover of the virus, making mutations a very frequent event [47].

Having a good mechanistic model in a biomedical context such as HIV infection opens doors to various applications beyond a good understanding of the data. Global and individual predictions can be excellent because of the external validity of a model based on main biological mechanisms. Control theory may serve for defining optimal interventions or optimal designs to evaluate new interventions [40]. Finally, these models can capture explicitly the complex relationship between several processes that change over time and may therefore challenge other proposed approaches such as marginal structural models to deal with causal associations in epidemiology [39].

Therefore, we postulate that this type of model could be very useful in the context of our research that is in complex biological systems. The definition of the model needs to identify the parameter values that fit the data. In clinical research this is challenging because data are sparse, and often unbalanced, coming from populations of subjects. A substantial inter-individual variability is always present and needs to be accounted as this is the main source of information. Although many approaches have been developed to estimate the parameters of non-linear mixed models [51], [59], [43], [49], [44], [58], the difficulty associated with the complexity of ODE models and the sparsity of the data leading to identifiability issues need further research.

Furthermore, the availability of data for each individual (see below) leads to a new challenge in this area. The structural model can easily be much more complex and the observation model may need to integrate much more markers.

3.2. High-dimensional statistical learning

With the availability of omics data such as genomics (DNA), transcriptomics (RNA) or proteomics (proteins), but also other types of data, such as those arising from the combination of large observational databases (e.g. in pharmacoepidemiology or environmental epidemiology), high-dimensional data have became increasingly common. Use of molecular biological technics such as Polymerase Chain Reaction (PCR) allows for amplification of DNA or RNA sequences. Nowadays, microarray and Next Generation Sequencing (NGS) techniques give the possibility to explore very large portions of the genome. Furthermore, other assays have also evolved, and traditional measures such as cytometry or imaging have became new sources of big data. Therefore, in the context of HIV research, the dimension of the datasets has much grown in term of number of variables per individual than in term of number of included patients although this latter is also growing thanks to the multi-cohort collaborations such as CASCADE or COHERE organized in the EuroCoord network [1]. As an example, in a phase 1/2 clinical trial evaluating the safety and the immunological response to a dendritic cell-based HIV vaccine, 19 infected patients were included. Bringing together data on cell count, cytokine production, gene expression and viral genome change led to a 20 Go database [55]. This is far from big databases faced in other areas but constitutes a revolution in clinical research where clinical trials of hundred of patients sized few hundred of Ko at most. Therefore, more than the storage and calculation capacities, the challenge is the comprehensive analysis of these data-sets.

The objective is either to select the relevant information or to summarize it for understanding or prediction purposes. When dealing with high-dimensional data, the methodological challenge arises from the fact that data-sets typically contain many variables, much more than observations. Hence, multiple testing is an obvious issue that needs to be taken into account [52]. Furthermore, conventional methods, such as linear models, are inefficient and most of the time even inapplicable. Specific methods have been developed, often derived from the machine learning field, such as regularization methods [57]. The integrative analysis of large data-sets is challenging. For instance, one may want to look at the correlation between two large scale matrices composed by the transcriptome in the one hand and the proteome on the other hand [45]. The comprehensive analysis of these large data-sets concerning several levels from molecular pathways to clinical response of a population of patients needs specific approaches and a very close collaboration with the providers of data that is the immunologists, the virologists, the clinicians...

4. Application Domains

4.1. Systems Biology and Translational medicine

Biological and clinical researches have dramatically changed because of the technological advances, leading to the possibility of measuring much more biological quantities than previously. Clinical research studies can include now traditional measurements such as clinical status, but also thousands of cell populations, peptides, gene expressions for a given patient. This has facilitated the transfer of knowledge from basic to clinical science (from "bench side to bedside") and vice versa, a process often called "Translational medicine". However, the analysis of these large amounts of data needs specific methods, especially when one wants to have a global understanding of the information inherent to complex systems through an "integrative analysis". These systems like the immune system are complex because of many interactions within and between many levels (inside cells, between cells, in different tissues, in various species). This has led to a new field called "Systems biology" rapidly adapted to specific topics such as "Systems Immunology" [53], "Systems vaccinology" [50], "Systems medicine" [42]. From the data scientist point of view, two main challenges appear: i) to deal with the massive amount of data ii) to find relevant models capturing observed behaviors.

---

[1]see online at http://www.eurocoord.net

4.2. HIV immunotherapies

The management of HIV infected patients and the control of the epidemics have been revolutionized by the availability of highly active antiretroviral therapies. Patients treated by these combinations of antiretrovirals have most often undetectable viral loads with an immune reconstitution leading to a survival which is nearly the same to uninfected individuals [46]. Hence, it has been demonstrated that early start of antiretroviral treatments may be good for individual patients as well as for the control of the HIV epidemics (by reducing the transmission from infected people) [41]. However, the implementation of such strategy is difficult especially in developing countries. Some HIV infected individuals do not tolerate antiretroviral regimen or did not reconstitute their immune system. Therefore, vaccine and other immune interventions are required. Many vaccine candidates as well as other immune interventions (IL7, IL15) are currently evaluated. The challenges here are multiple because the effects of these interventions on the immune system are not fully understood, there are no good surrogate markers although the number of measured markers has exponentially increased. Hence, HIV clinical epidemiology has also entered in the era of Big Data because of the very deep evaluation at individual level leading to a huge amount of complex data, repeated over time, even in clinical trials that includes a small number of subjects.

4.3. Translational vaccinology

Vaccines are one of the most efficient tools to prevent and control infectious diseases, and there is a need to increase the number of safe and efficacious vaccines against various pathogens. However, clinical development of vaccines - and of any other investigational product - is a lengthy and costly process. Considering the public health benefits of vaccines, their development needs to be supported and accelerated. During early phase clinical vaccine development (phase I, II trials, translational trials), the number of possible candidate vaccine strategies against a given pathogen that needs to be down-selected in early clinical development is potentially very large. Moreover, during early clinical development there are most often no validated surrogate endpoints to predict the clinical efficacy of a vaccine strategy based on immunogenicity results that could be used as a consensus immunogenicity endpoint and down-selection criterion. This implies considerable uncertainty about the interpretation of immunogenicity results and about the potential value of a vaccine strategy as it transits through early clinical development. Given the complexity of the immune system and the many unknowns in the generation of a protective immune response, early vaccine clinical development nowadays thus takes advantage of high throughput (or "omics") methods allowing to simultaneously assess a large number of response markers at different levels ("multi-omics") of the immune system. This has induced a paradigm shift towards early-stage and translational vaccine clinical trials including fewer participants but with thousands of data points collected on every single individual. This is expected to contribute to acceleration of vaccine development thanks to a broader search for immunogenicity signals and a better understanding of the mechanisms induced by each vaccine strategy. However, this remains a difficult research field, both from the immunological as well as from the statistical perspective. Extracting meaningful information from these multi-omics data and transferring it towards an acceleration of vaccine development requires adequate statistical methods, state-of-the art immunological technologies and expertise, and thoughtful interpretation of the results. It thus constitutes research at the interface between disciplines: data science, immunology and vaccinology. Our main current areas of application here are early phase trials of HIV and Ebola vaccine strategies, in which we participate from the initial trial design to the final data analyses.

5. Highlights of the Year

5.1. Highlights of the Year

5.1.1. Team structure

The SISTM team was re-structured into three research axes (formerly two) in 2019:
- Axis "High Dimensional Statistical Learning" (coordinator Boris Hejblum)
- Axis "Mechanistic Learning" (coordinator Mélanie Prague)
- Axis "Translational Vaccinology" (coordinator Laura Richert)

The third research axis on "Translational vaccinology" was created in order to formalize research activities already performed previously in a less structure way. This axis is dedicated to applied research questions in early stage clinical vaccine trials, with two objectives:

– to elucidate the potential effects and mechanisms of action of vaccines and immunotherapies in integrative statistical analyses of the induced responses at various levels of the immune system

– to better inform future trial designs and statistical analysis methods by means of modelling and methodological developments.

The three axes collaborate closely with each other.

In fact, the third axis gives the motivating examples leading to the methodological work done in the two other axes. The first axis deals with the raw high-dimensional data generated in clinical epidemiology or biological studies and aims at reducing the dimension of the problem or better annotate the data available (e.g. automatic gating of cytometry data). The second axis aims at building mechanistic model to understand and predict the biological phenomenons by using the available information. The idea then is that the results of this modelling part feed the third axis to define the next strategies to be evaluated in clinical studies and the design of these studies.

5.1.2. Team composition

The SISTM team core has changed in December 2019: Daniel Commenges, DRE Inserm, HDR (emeritus from September 2014) has retired from research activities. Daniel Commenges founded the "Biostatistic" team of Bordeaux within an Inserm unit in the 1990s. The research team was officially labelled by Inserm in the early 2000s and was lead by Daniel Commenges until 2013. The team split in 2014 into two teams: the Inserm "Biostatistic" team (led by Hélène Jacqmin-Gadda), and the "SISTM" team (led by Rodolphe Thiébaut) that joined the Inria BSO center.

5.1.3. Funded projects

- Launch of the Graduate's School Digital Public Health (PI: R Thiébaut) including the Master of Public Health Data Sciences that started with its first cohort of 9 international students in Sep 2019.

- Positive response for funding of the H2020 IP-Cure-B project (Immune profiling to guide host-directed interventions to cure hepatitis B infections, project coordinator: Pr. F. Zoulim, Inserm U1052 CRCL), in which the work package "Data Science" is led by the SISTM team. The project will be launched in January 2020.

- Kick-off of the EDTCP-2 funded project PREVAC-UP (partnership for research on Ebola vaccinations – extended follow-up and clinical research capacity build-up, project coordinator: Pr Y. Yazdanpanah, Inserm), in which the work package "Systems vaccinology" is led by the SISTM team.

- A new collaboration has started with the pharmaceutical company Ipsen on the integration of "omics" data into in-silico modelling of early-stage clinical trials in cancer. This project will be conducted with a "*CIFRE*" (*Conventions Industrielles de Formation par la REcherche*) PhD contract starting in January 2020.

- *Action de Développement technologique* VASI: Visualization and Analytics Solution for Immunologists.

- The Ebovac2 IMI project on Ebola vaccine development has been extended to 11/2020 (no cost extension).

- Associate Team DYNAMHIC: Dynamical Modeling of HIV Cure in Collaboration With Harvard Program for evolutionary dynamics.

5.1.4. Advancements in projects

- A translational phase I clinical trial of an experimental placental malaria vaccine, conducted by an interdisciplinary consortium including members of the SISTM team (Primalvac trial), has reached its publication, with a manuscript accepted for publication in the Lancet Infectious Diseases

- Two HIV clinical vaccine trials have reached their final stage with all resultats available, including integrative data analyses of the immune responses, and the corresponding manuscripts are in preparation (ANRS VRI01 trial and ANRS 149 Light trial).
- The two phase II Ebola vaccine trials conducted by the IMI-2 EBOVAC2 consortium that is coordinated by Rodolphe Thiébaut are terminated. Results have been presented at international conferences and the manuscripts with the primary results are either submitted (EBL2001 study, submitted to the Lancet) or in preparation (EBL2002 study). Systems vaccinology analyses of the data from these trials are ongoing in the SISTM team.
- Robin Genuer co-authored a book with Jean-Michel Poggi on random forests entitled *Les forêts aléatoires avec R* in *Presses Universitaires de Rennes*, Rennes, France.

5.1.5. Awards

- Award for Doctoral Supervision and Research Activity (PEDR) attributed by the University of Bordeaux to Marta Avalos and Robin Genuer

6. New Software and Platforms

6.1. marqLevAlg

KEYWORDS: Optimization - Biostatistics

FUNCTIONAL DESCRIPTION: An R package for function optimization. Available on CRAN, this package performs a minimization of function based on the Marquardt-Levenberg algorithm. This package is really useful when the surface to optimize is non-strictly convex or far from a quadratic function. A new convergence criterion, the relative distance to maximum (RDM), allows the user to have a better confidence in the stopping points, other than basic algorithm stabilization.

RELEASE FUNCTIONAL DESCRIPTION: This package has been updated so that the optimization is done in parallel and goes faster. This update has been done in collaboration Viviane Philipps and Cecile Proust-Lima from Inserm BPH

- Partner: INSERM
- Contact: Melanie Prague
- URL: https://cran.r-project.org/web/packages/marqLevAlg/index.html

6.2. VSURF

Variable Selection Using Random Forests

KEYWORDS: Classification - Statistics - Machine learning - Regression

FUNCTIONAL DESCRIPTION: An R package for Variable Selection Using Random Forests. Available on CRAN, this package performs an automatic (meaning completely data-driven) variable selection procedure. Originally designed to deal with high dimensional data, it can also be applied to standard datasets.

RELEASE FUNCTIONAL DESCRIPTION: * add RFimplem parameter which allows to choose between randomForest, ranger and Rborist to compute random forests predictors. This can be a vector of length 3 to chose a different implementation for each step of VSURF() * update of the parallel and clusterType parameters to also give the possibility to choose which step to perform in parallel with a clusterType per step * add progress bars and information of the progress of the algorithm, and also an estimated computational time for each step

- Contact: Robin Genuer
- URL: https://github.com/robingenuer/VSURF

6.3. NPflow

Bayesian Nonparametrics for Automatic Gating of Flow-Cytometry Data

KEYWORDS: Bayesian estimation - Bioinformatics - Biostatistics

FUNCTIONAL DESCRIPTION: Dirichlet process mixture of multivariate normal, skew normal or skew t-distributions modeling oriented towards flow-cytometry data pre-processing applications.

- Contact: Boris Hejblum
- URL: https://CRAN.R-project.org/package=NPflow

6.4. COVVSURF

Combination of Clustering Of Variables and Variable Selection Using Random Forests

KEYWORDS: Classification - Statistics - Cluster - Machine learning - Regression

- Contact: Robin Genuer
- URL: https://github.com/robingenuer/CoVVSURF

6.5. clogitLasso

KEYWORDS: Biostatistics - Bioinformatics - Machine learning - Regression

FUNCTIONAL DESCRIPTION: R package to fit a sequence of conditional logistic regression models with lasso, for small to large sized samples.

RELEASE FUNCTIONAL DESCRIPTION: Optimisation

- Partner: DRUGS-SAFE
- Contact: Marta Avalos Fernandez
- URL: https://cran.r-project.org/web/packages/clogitLasso/index.html

6.6. TcGSA

Time-course Gene Set Analysis

KEYWORDS: Bioinformatics - Genomics

FUNCTIONAL DESCRIPTION: An R package for the gene set analysis of longitudinal gene expression data sets. This package implements a Time-course Gene Set Analysis method and provides useful plotting functions facilitating the interpretation of the results.

- Contact: Boris Hejblum
- URL: https://CRAN.R-project.org/package=TcGSA

6.7. NIMROD

Normal approximation Inference in Models with Random effects based on Ordinary Differential equations

KEYWORDS: Ordinary differential equations - Statistical modeling

FUNCTIONAL DESCRIPTION: We have written a specific program called NIMROD for estimating parameter of ODE based population models.

- Contact: Melanie Prague
- URL: http://etudes.isped.u-bordeaux2.fr/BIOSTATISTIQUE/NIMROD/documentation/html/index.html

6.8. tcgsaseq

Time-Course Gene Set Analysis for RNA-Seq Data

KEYWORDS: Genomics - Biostatistics - Statistical modeling - RNA-seq - Gene Set Analysis

FUNCTIONAL DESCRIPTION: Gene set analysis of longitudinal RNA-seq data with variance component score test accounting for data heteroscedasticity through precision weights.

- Contact: Boris Hejblum
- URL: https://CRAN.R-project.org/package=tcgsaseq

6.9. cytometree

KEYWORDS: Clustering - Biostatistics - Bioinformatics

FUNCTIONAL DESCRIPTION: Given the hypothesis of a bimodal distribution of cells for each marker, the algorithm constructs a binary tree, the nodes of which are subpopulations of cells. At each node, observed cells and markers are modeled by both a family of normal distributions and a family of bimodal normal mixture distributions. Splitting is done according to a normalized difference of AIC between the two families.

- Contact: Boris Hejblum
- URL: https://CRAN.R-project.org/package=cytometree

6.10. CRTgeeDR

KEYWORDS: Missing data - Statistics - Regression

FUNCTIONAL DESCRIPTION: The CRTgeeDR package allows you to estimates parameters in a regression model (with possibly a link function). It allows treatment augmentation and IPW for missing outcome. It is particularly of use when the goal is to estimate the intervention effect of a prevention strategy agains epidemics in cluster randomised trials.

- Contact: Melanie Prague
- URL: https://cran.r-project.org/web/packages/CRTgeeDR/index.html

6.11. ludic

KEYWORDS: Probability - Biostatistics

FUNCTIONAL DESCRIPTION: An R package to perform probabilistic record Linkage Using only DIagnosis Codes without direct identifiers, using C++ code to speed up computations. Available on CRAN, development version on github.

- Contact: Boris Hejblum
- URL: https://CRAN.R-project.org/package=ludic

6.12. CoDaPCA

KEYWORDS: Unsupervised learning - PCA

FUNCTIONAL DESCRIPTION: R functions associated to the article Avalos et al. Representation Learning of Compositional Data. NeurIPS 2018 http://papers.nips.cc/paper/7902-representation-learning-of-compositional-data

- Contact: Marta Avalos Fernandez
- URL: https://github.com/sistm/CoDa-PCA

6.13. Left-censored Lasso

KEYWORDS: Biostatistics - Machine learning

FUNCTIONAL DESCRIPTION: R function associatied to the article Soret et al. Lasso regularization for left-censored Gaussian outcome and high-dimensional predictors. BMC Medical Research Methodology (2018) 18:159 https://doi.org/10.1186/s12874-018-0609-4

RELEASE FUNCTIONAL DESCRIPTION: https://github.com/psBiostat/left-censored-Lasso

- Contact: Marta Avalos Fernandez

6.14. dd-sPLS

Data-Driven Sparse PLS

KEYWORDS: Marker selection - Classification - Regression - Missing data - Multi-Block - High Dimensional Data - PLS - SVD

SCIENTIFIC DESCRIPTION: Allows to build Multi-Data-Driven Sparse PLS models. Multi-blocks with high-dimensional settings are particularly sensible to this. Whatsmore it deals with missing samples (entire lines missing per block) thanks to the Koh-Lanta algorithm. SVD decompositions permit to offer a fast and controlled method.

FUNCTIONAL DESCRIPTION: That software solves the missing samples problem selecting interesting variables under multi-block supervised settings.

- Contact: Hadrien Lorenzo
- URL: https://hadrienlorenzo.netlify.com/projects/ddspls/

6.15. kernscr

KEYWORDS: Genomics - Biostatistics

FUNCTIONAL DESCRIPTION: An R package to perform KERNel machine score test for pathway analysis in the presence of Semi-Competing Risks

- Contact: Boris Hejblum
- URL: https://CRAN.R-project.org/package=kernscr

6.16. phenotypr

KEYWORDS: Phenotyping - Automatic labelling - Automatic Learning

FUNCTIONAL DESCRIPTION: Machine learning prediction algorithm for predicting a clinical phenotype from structured diagnostic data and CUI occurrence data collected from medical reports, previously processed by NLP approaches.

- Contact: Boris Hejblum
- URL: https://github.com/borishejblum/phenotypr

6.17. R2GUESS

Graphical processing Unit Evolutionary Stochastic Search

FUNCTIONAL DESCRIPTION: R2GUESS package is a wrapper of the GUESS (Graphical processing Unit Evolutionary Stochastic Search) program. GUESS is a computationally optimised C++ implementation of a fully Bayesian variable selection approach that can analyse, in a genome-wide context, single and multiple responses in an integrated way. The program uses packages from the GNU Scientific Library (GSL) and offers the possibility to re-route computationally intensive linear algebra operations towards the Graphical Processing Unit (GPU) through the use of proprietary CULA-dense library.

- Contact: Rodolphe Thiebaut

7. New Results

7.1. Mechanistic learning

7.1.1. Ebola models

New models have been developed for the response to the Ebola vaccine. The first one has been fitted to Phase 1 trials and has given interesting predictions of the long term duration of the response that are confirmed with the new data coming from phase 2 trials. These results have been published in Journal of Virology. Then, a new model including the B cell memory response has been defined and its mathematical proprieties have been studied. A manuscript has been submitted to Journal of Theoretical Biology. The next step is to estimate model parameters using EBL2001 clinical trial data.

New publication: Pasin C et al. Dynamics of the Humoral Immune Response to a Prime-Boost Ebola Vaccine: Quantification and Sources of Variation. J Virol. 2019 Aug 28;93(18). pii: e00579-19. doi: 10.1128/JVI.00579-19. Print 2019 Sep 15.

7.1.2. Estimation method

A new approach is currently under development by Quentin Clairon to estimate model parameters using a regularization method based on the control theory. The estimation method used an approximation of the original ODE solution for each subject. The expected advantages of this approach are i) to mitigate the effect of model misspecification on estimation accuracy ii) to regularize the estimation problem in presence of poorly identifiable parameters, iii) to avoid estimation of initial conditions. The method is still under development but preliminary results have been presented at the Viral dynamics conference in October 2019.

7.2. High-dimensional and statistical learning

7.2.1. Automatic analysis of cell populations

New publication:

Hejblum BP, Alkhassim C, Gottardo R, Caron F, Thiébaut R, 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.

7.2.2. High-dimensional compositional data analysis

Perrine Soret (PhD student in the axis "High-dimensional and statistical learning", supervised by M. Avalos) has applied our expertise in high dimensional data analysis to human microbiome field of research:

Soret P, Vandenborght LE, Francis F, Coron N, Enaud R, The Mucofong Investigation Group, Avalos M, Schaeverbeke T, Berger P, Fayon M, Thiébaut R and Delhaes L. Respiratory mycobiome and suggestion of inter-kingdom network during acute pulmonary exacerbation in cystic fibrosis. To appear in *Scientific Reports*.

7.2.3. Missing Value Treatment in Longitudinal High Dimensional Supervised Problems

Poor blood sample quality introduces a large number of missing values in the context of sequencing data production. Furthermore, strong technical biases may force the analyst to remove the considered sequenced samples. Then entire day dependent data are then missing. Hadrien Lorenzo (PhD student in the axis "High-dimensional and statistical learning", supervised by J. Saracco and R. Thiébaut) has developed a multi-block approach: the dd-sPLS method. dd-sPLS has been applied to high dimensional data analysis of different fields of research:

Lorenzo, H., Misbah, R., Odeber, J., Morange, P. E., Saracco, J., Trégouët, D. A., and Thiébaut, R. High-dimensional multi-block analysis of factors associated with thrombin generation potential. In 2019 IEEE 32nd International Symposium on Computer-Based Medical Systems (CBMS) (pp. 453-458). IEEE. https://hal.archives-ouvertes.fr/hal-02429302

Ellies-Oury, M. P., Lorenzo, H., Denoyelle, C., Saracco, J., and Picard, B. An Original Methodology for the Selection of Biomarkers of Tenderness in Five Different Muscles. Foods, 8(6), 206 (2019). https://hal.archives-ouvertes.fr/hal-02164157

7.3. Translational vaccinology

7.3.1. HIV vaccine development

We have finalized the data science analyses of two HIV vaccine clinical trials: 1) ANRS VRI01, a randomized phase I/II trial evaluating for different prime boost vaccine strategies in healthy volunteers; 2) ANRS 149 LIGHT, a randomized phase II trial comparing a prime-boost therapeutic HIV vaccine strategy to placebo in HIV-infected patients undergoing antiretroviral treatment interruption. This included integrative statistical analyses using sPLS methods (as developed by the team) to relate markers from different high-dimensional immunogenicity or gene expression assays or virological assays to each other. In the ANRS VRI01 data set this allowed to disentangle the immune responses induced by the different vaccines used in the prime-boost strategies, showing specific effects of one of the vaccines (MVA HIV-B). We further identified a gene expression signature that correlates with later functional T-cell responses across the three different prime-boost association in which the MVA HIV-B vaccine was used. The corresponding manuscripts are currently in preparation.

Other HIV vaccine trials are currently being set-up by the French VRI (Vaccine Research Institute) and the European consortium EHVA, with strong contributions of SISTM team members to the trial designs.

7.3.2. Ebola vaccine development

The main results of the two randomized phase II Ebola vaccine trials conducted by the IMI-2 EBOVAC2 consortium (coordinated by Rodophe Thiébaut from the SISTM team) were finalized in 2019. and presented at international conferences. The results showed that the tested vaccine strategy (two-dose heterologous Ad26.ZEBOV and MVA-BN®-Filo Ebola vaccine regimen, developed by Janssen) was safe and immunogenic in both European and African volunteers (EBL2001 and EBL2002 trials). Deeper analyses of the induced immune responses are currently ongoing, and systems vaccinology analyses will soon start in the SISTM team.

The SISTM team is also a partner in the related IMI-2 EBOVAC1 and EBOVAC3 consortia (assesing the same vaccine regimen in other trial populations and/or trial phases), in which the main contributions of the team are related to mechanistic modeling of the immune responses (ongoing).

7.3.3. Vaccine development against other pathogens

Two other phase I vaccine trials (one testing a placental malaria vaccine, Primalvac trial; and one testing a nasal Pertussis vaccine, BPZE-1 trial), in which members of the SISTM team were strongly involved, have shown promising results. Results of the Primalvac trial have been accepted for publication in the Lancet Infectious Disease journal, and results of the BPZE-1 trial have been submitted for publication.

7.3.4. Methodological developments for vaccine trials

At the interface between the axis on "Mechanistic learning" and the axis "Translational vaccinology", modelling done within a PhD project (M. Alexandre, supervised by R. Thiébaut and M. Prague) has informed the definition of the primary endpoint and statistical analysis method to be used in two therapeutic HIV vaccine trials with antiretroviral treatment interruption (EHVA T02 and ANRS DALIA-2). The methodological choices and their rationale have been presented to the governance bodies of the research consortia, patient associations and have been submitted for ethics and regulatory approvals. This will be subject to a specific methodology publication.

Edouard Lhomme (PhD student in the axis "Translational Vaccinology, supervised by L. Richert) has developed a statistical method for functional T-cell assay data from vaccine trials (in particular the intracellular cytokine staining assay) that takes into account non-specific immune responses. We propose using a bivariate linear model for the analysis of the cellular immune responses to obtain accurate estimations of the vaccine effect. We benchmarked the performance of the model in terms of both bias and control of type-I and -II errors, and applied it to simulated data as well as real pre- and post-vaccination data from two recent HIV vaccine trials (ANRS VRI01 and ANRS 149 LIGHT in HIV-infected participants). This method has been published in the Journal of Immunological Methods and is now used in the SISTM team as the standard method for analyses of functional cellular data with non-stimulated control conditions, for instance in the currently ongoing analysis of cellular proliferation data from the EBOVAC2 EBL2001 trial. We have also established an online interface based on R Shiny to make this analysis method available for use by immunologists without specific training in statistical modelling (https://shiny-vici.apps.math.cnrs.fr/).

7.3.5. *Prediction of the survival of patients based on RNA-seq data*

In collaboration with the Inria MONC team, with the Inserm Angiogenesis and Tumor micro-environment team, and with clinicians from Milan and Bergen, the project GLIOMA-PRD aims to improve the prediction of the evolution of the lower grade glioma, a primary brain tumor, based on clinical, imaging and genomic data. In the SISTM team, we first had to determine the sufficient sample size for determining a predictive signature based on RNA-seq data for the survival of the patients. We concluded that 50 patients were a good enough sample size for this aim. We then explored the potential methods to analyze these data, with a particular focus on the methods grouping the genes by pathways, as the pilot data (The Cancer Genome Atlas Research Network, NEJM, 2015) showed a high correlation structure. We particularly compared two methods, Generalized Berk-Jones (GBJ), proposed by [54], and tcgsaseq, proposed by [1]. The first method could be applied to the survival context and thus be appropriate for our data.

Once the RNA-seq data were available, we could observe a high batch effect since the data were sequenced in two different lanes. One of these two batch included only patients that exhibited a particular tumor at the PET-scan, called "COLD", while the second batch included both patient labelled as "COLD", but also the other type of tumor, called "DIFFUSE". As this experiment structure might leads to confusing the difference due to the batch effect with the one due to the biological different, we had to explore methodologies that could remove this batch effect.

Once this batch effect removed, we will then analyze the RNA-seq data in order to identify the genes, or the group of genes, that could be predictive of the survival of the patients.

8. Bilateral Contracts and Grants with Industry

8.1. Contracts and Grants with Industry

Implication in research for the development of Ebola vaccine has lead to several indirect contracts with industry:

- The EBOVAC1, EBOVAC2 and EBOVAC3 project, collaboration with Janssen from Johnson et Johnson.
- The Prevac trial vaccine trial (legal sponsors: Inserm, NIH, London School of Hygiene and Tropical Medicine) involves collaborations with Merck and Janssen. The purpose of this study is to evaluate the safety and immunogenicity of three vaccine strategies that may prevent Ebola virus disease (EVD) events in children and adults. Participants will receive either the Ad26.ZEBOV (rHAd26) vaccine with a MVA-BN-Filo (MVA) boost, or the rVSVΔG-ZEBOV-GP (rVSV) vaccine with or without boosting, or placebo. The EDCTP-2 funded Prevac-UP project is set as a continuation of Prevac trial in the same framework.

A new collaboration has started with the pharma company Ipsen on the integration of OMICS data into an in-silico trials pipeline (Cifre Phd to start in January 2020)

9. Partnerships and Cooperations

9.1. Regional Initiatives

The team have strong links with :

- Research teams of the research center Inserm U1219 : "Injury Epidemiology, Transport, Occupation" (IETO), "Biostatistics", "Pharmacoepidemiology and population impact of drugs", "Multimorbidity and public health in patients with HIV or Hepatitis" (MORPH3Eus), "Computer research applied to health" (ERIAS) emerging research team.
- Bordeaux CHU ("Centre Hospitalier Universitaire").
- Institut Bergonié, Univ Bordeaux through the Euclid F-CRIN Clinical Trials platform and CIC-EC (CIC1401)
- Inria Project-team MONC, M3DISIM and CQFD

The project team members are involved in:

- EUCLID/F-CRIN clinical trials platform (Laura Richert)
- The Clinical Epidemiology module of the Clinical Investigations Center (CIC1401) (Laura Richert)
- The research project "Self-management of injury risk and decision support systems based on predictive computer modelling. Development, implementation and evaluation in the MAVIE cohort study" funded by the Nouvelle-Aquitaine regional council (Marta Avalos).
- Phenotyping from Electronic Health Records pilot project in cooperation with with the ERIAS Inserm emerging team in Bordeaux and the Rheumatology service from the Bordeaux Hospital (Boris Hejblum)
- A cancer research project (GLIOMA-PRD) in collaboration with Inria MONC team and with the the Inserm Angiogenesis and Tumor micro-environment team on glioblastoma

9.2. National Initiatives

- Labex Vaccine Research Institute (VRI) There are strong collaborations with immunologists involved in the Labex Vaccine Research Institute (VRI) as Rodolphe Thiébaut and Laura Richert are leading the Data science division (previously Biostatistics/Bioinformatics) http://vaccine-research-institute.fr.
- Collaboration with Inserm PRC (pôle Recherche clinique).
- Collaboration with Inserm Reacting (REsearch and ACTion targeting emerging infectious diseases) network
- Collaboration with Inserm RECap (Recherche en Epidémiologie Clinique et en Santé Publique) network

9.2.1. Expert Appraisals

- Rodolphe Thiébaut is a member of the CNU 46.04 (Biostatistiques, informatique médicale et technologies de communication).
- Rodolphe Thiébaut is a member of the Scientific Council of Inserm.
- Mélanie Prague is an expert for ANRS (France Recherche Nord&Sud Sida-HIV Hépatites) in the CSS 3 (Recherches cliniques et physiopathologiques dans l'infection à VIH) and AC 47 (Dynamique et contrôle des épidémies VIH et hépatites).
- Laura Richert is an expert for the PHRC (Programme hospitalier de recherche Clinique).
- Marta Avalos is an expert for the ANSM (Agence nationale de sécurité du médicament et des produits de santé)

9.2.2. *Various Partnership*

The project team members are involved in:

- DRUGS-SAFE platform funded by ANSM (Marta Avalos). Initiated in 2015-2018. Renewed for 2019.
- F-CRIN (French clinical research infrastructure network), initiated in 2012 by ANR under "Programme des Investissements d'avenir". (Laura Richert)
- INCA (Institut National du Cancer) funded the project *Evaluation de l'efficacité d'un traitement sur l'évolution de la taille tumorale et autres critères de survie : développement de modèles conjoints.* (Principal PI Virginie Rondeau Inserm U1219, Mélanie Prague is responsible of Work package 4 "mechanistic modeling of cancer: 5800 euros").
- Contrat Initiation ANRS MoDeL-CI: Modeling the HIV epidemic in Ivory Coast (Principal PI Eric Ouattara Inserm U1219 in collaboration with University College London, Mélanie Prague is listed as a collaborator).

9.3. European Initiatives

9.3.1. *FP7 & H2020 Projects*

The member of SISTM Team are involved in EHVA (European HIV Vaccine Alliance):

EHVA: European HIV Vaccine Alliance: a EU platform for the discovery and evaluation of novel prophylactic and therapeutic vaccine candidates

Coordinator: Inserm/University of Lausanne. Other partners: EHVA consortium gathers 41 partners. Duration: 60 months. 01 /01 /2016 - 31 /12 /2020

With 37 million people living with HIV worldwide, and over 2 millions new infections diagnosed each year, an effective vaccine is regarded as the most potent public health strategy for addressing the pandemic. Despite the many advances in the under-standing, treatment and prevention of HIV made over the past 30 years, the development of broadly-effective HIV vaccine has remained unachievable. The EHVA international alliance, which includes academic and industrial research partners from all over Europe, as well as sub-Saharan Africa and North America, will work to discover and progress novel vaccine candidates through the clinic. EHVA fosters a multidisciplinary approach to the challenge of developing broadly effective HIV vaccines. EHVA's program primary goals are:

- To develop a Multidisciplinary Vaccine Platform (MVP) for prophylactic and therapeutic HIV vaccines
- To move at least two novel prophylactic vaccine candidates to clinical development
- To identify immune correlates associated with control of HIV replication following immunological intervention
- To establish a strong scientific basis for further development of EHVA vaccine candidates in larger clinical trials

  To this purpose, EHVA bring to the field 4 multidisciplinary research platforms representative of the latest advances in clinical trials and preclinical vaccine development. These four platforms cover all aspects of vaccine development from early-stage discovery to clinical trials.

- The Discovery Platform will work to disclose promising vaccine candidates based on the induction of T-cell and antibody responses (ie, neutralizing antibody and non-neutralizing antibody).
- The Immune-Profiling Platform will advance assays to predict the immunogenicity of potential vaccine candidates. The ability to generate a profile of a potential vaccine candidate, using models that emulate the immune system's response, will assist with benchmarking novel and existing vaccine candidates.

- The Data Management/Integration and Down-Selection Platform is developed around the WP10 led by Rodolphe Thiébaut. SISTM provides here state-of-the-art statistical tools for the analysis and interpretation of complex data and algorithms for the efficient selection of vaccines.

- The Clinical Trial Platform includes pharmaceutical industry expertise for late stage development, a network of top European clinical centers for conducting large cohort studies, as well as relationships with leading scientists based in Africa. Future testing of EHVA vaccine in Sub-Saharan Africa is a research priority because it is the area of the world with the greatest number of people infected with HIV.

IP-CURE-B: Immune profiling to guide host-directed interventions to cure HBV infections. Coordinated by Inserm (France), the project includes a total of 13 Beneficiaries: Centre Hospitalier Universitaire Vaudois (Switzerland), Karolinska Institutet (Sweden), Institut Pasteur (France), Universita degli studi di Parma (Italy), Fondazione IRCCS CA' Granda – Ospedale maggiore policlinico (Italy), Universitaetsklinikum Freiburg (Germany), Ethniko Kai Kapodistriako Panepistimio Athinon (Greece), Fundacio Hospital Universitari vall d'Hebron (Spain), Gilead Sciences Inc. (USA), Spring Bank Pharmaceuticals, Inc (USA), European Liver Patients Association (Belgium), Inserm Transfert SA (France). Duration: 60 months. $01/01/2020 - 31/12/2024$

HBV infections, are a major global public health threat with over 257 million people worldwide chronically infected and over 887,000 deaths per year. 4.7 million people live with HBV in the European Union (EU) and European Economic Area (EEA). W.H.O. estimates that HBV causes almost 40% of the cases of hepatocellular carcinoma (HCC), which is the 2nd leading cause of cancer-related mortality worldwide. HBV kills nearly 900,0000 people around the world each year. The current prophylactic vaccine has no impact on established chronic infection.

The objective of the IP-CURE-B project is to develop novel curative concepts for chronic hepatitis B (CHB). Specific aims will be to: 1) improve the rate of functional cure of CHB by boosting innate immunity with immune modulators and stimulating adaptive immune responses with a novel therapeutic vaccine; ii) characterize immune and viral biomarker signatures for patient stratification and treatment response monitoring; iii) integrate biological and clinical data to model the best combination treatment for future trials; iv) model the effectiveness of novel curative therapies with respect to disease spectrum, patient heterogeneity, and constraints of National Health Systems.

The project organization combines: i) a Proof of Concept clinical trial of a combination of 2 novel compounds stimulating innate immunity; ii) a preclinical immune therapy platform in humanized mice combining immune-modulatory strategies to stimulate innate immunity, rescue exhausted HBV-specific T cells and generate anti-HBV adaptive responses; iii) extensive virologic and immune profiling to identify correlates of cure in patients, iv) the integration of large biological and clinical data-sets, v) a cost-effectiveness modelling of new therapeutic interventions, vi) project management, vii) results exploitation and dissemination.

In the IP-CURE-B project, SISTM coordinates WP6 Data science platform for data integration and statistical modeling which will provide powerful data management and statistical tools for the analysis and interpretation of the complex heterogeneous and high-dimensional data generated in the other WPs. For data management and data sharing, SISTM will leverage on a data warehouse system, based on Lab-key Server, the primary structure already established within the EU funded H2020 EHVA project. SISTM will develop and apply statistical methods for integrating data from several assay platforms to better describe and understand the mechanisms of the experimental products and to define predictive signatures of viral control and functional cure. Indeed, the immune system forms a sophisticated network of tissues, cells and molecules that interact in order to achieve viral control. Understanding how this complex network responds to interventions aimed at HBV functional cure requires the use and integration of data from multiple assay technologies. Two main strategies will be used: 1) statistical approaches to relate and down-select several high-dimensional data from the various assays in humanized mice and humans; 2) a modelling approach, taking into account

biological knowledge and the results from the first step, to better capture and understand the non-linear relationships between the components of the immune system, viral control and their dynamics over time. Statistical and mechanistic models will be used, based on ordinary differential equation systems or other approaches. At the end of the process, if an adequate model is identified, this can be used to down-select immunomodulatory and vaccine regimens and make in silico predictions about optimized strategies or stratified treatment approaches. These approaches have been successfully applied in HIV immunotherapy trials and in vaccine trials by SISTM.

9.3.2. *Collaborations in European Programs, Except FP7 & H2020*

The members of SISTM are also involved in Innovative Medicine Initiative 2 (IMI2) projects which are all under the IMI Ebola+ program that was launched in response to the Ebola virus disease outbreak of 2014. SISTM is active in 3 projects which are all in collaboration with Janssen Vaccines & Preventions B.V. The overall aim of the EBOVAC program is to assess the safety, immunogenicity and efficacy of a novel 2-dose Ad26 + MVA prophylactic vaccine regimen against Ebola Virus Disease. In this context, the 3 projects develop as follows:

EBOVAC1: Development of a Prophylactic Ebola Vaccine Using an Heterologous Prime-Boost Regimen.

Coordinated by London School of Hygiene & Tropical Medicine (United Kingdom). Other beneficiaries: Janssen a Pharmaceutical Companies of Johnson & Johnson, The Chancellor, Masters and Scholars of the University of Oxford (United Kingdom), Inserm (France), University of Sierra Leone (Sierra Leone). Duration: 84 months. 01 /12 /2014 - 30 /11 /2021.

EBOVAC1 is dedicated to the Phase I and III development of prime-boost vaccine based on Ad26.ZEBOV and MVA-BN-Filo. Phase I was conducted in the US, the UK and in Africa (Sierra Leone, Uganda, Kenya and Tanzania) for a total of 231 volunteers enrolled. Phase III was conducted in Sierra Leone in several phases leading to the successful enrolment of more than 2800 volunteers including around 500 children aged 1-17 years. In EBOVAC1, SISTM is modelling the immune response to the Ad26.ZEBOV and MVA-BN-Filo, using the data obtained in the project.

EBOVAC2: Development of a Prophylactic Ebola Vaccine Using a 2-Dose Heterologous Vaccination Regimen: Phase 2.

Coordinated by Rodolphe Thiébaut with the following partners: Inserm (France), Labex VRI (France), Janssen Pharmaceutical Companies of Johnson & Johnson, London School of Hygiene & Tropical Medicine (United Kingdom), The Chancellor, Masters and Scholars of the University of Oxford (United Kingdom), Le Centre Muraz (Burkina Faso), Inserm Transfert (France). Duration: 72 months. 01 /12 /2014 - 30 /11 /2020.

EBOVAC2 main objective is to provide extensive and robust data on the safety and immunogenicity of the Ad26.ZEBOV and MVA-BN-Filo vaccine. This was designed by: 1. Carrying out translational studies to link vaccine elicited immune responses in humans to protection from Ebola in vaccinated non-human primates 2. Carrying out Phase II trials in African and European volunteers in approximately 6 countries, four in Africa and two in the EU with an overall target enrolment of approximately 1,500 subjects. Given the compressed nature of this development program, the Phase II studies were conducted in parallel with the planned Phase III study (EBOVAC1). The rationale for inclusion of European volunteers in Phase 2, in addition to the trials in Africa, is to allow for higher sensitivity in safety signal detection in populations with low incidence of febrile illnesses, to generate negative control specimens for assay development, to allow for inclusion of health care workers or military personnel that may be deployed to Ebola-endemic regions. 3. Evaluating the vaccine response in special population groups, such as children (ages 1-17 years), the elderly (ages 50-65) and individuals infected with HIV, to confirm safety and immunogenicity. The Phase II trials started as soon as preliminary safety data were available from Phase I trials. 4. Monitoring and characterizing immune response to the proposed vaccine through different set of analysis of the humoral and cellular response with different approaches (ICS, luminex, gene expression analysis, T and B

cell activation assays, Virus neutralization assays...) leading to a unique set of data. In EBOVAC2, in addition to the coordination of the whole project, SISTM is involved in the statistical analysis of the results obtained by the VRI lab responsible for an important part of the exploratory work, but also in the integrative data analysis of these high dimension and complex data. A Labkey environment was established in SISTM for EBOVAC2 to facilitate the exchange and following treatment of the project data.

EBOVAC3: Bringing a prophylactic Ebola vaccine to licensure.

Coordinated by the London School of Hygiene & Tropical Medicine (United Kingdom). Other beneficiaries: Janssen a Pharmaceutical Companies of Johnson & Johnson, Inserm (France), The University of Antwerpen (Belgium), University of Sierra Leone (Sierra Leone). Duration: 60 months. 01 /06 /2018 - 30 /05 /2023.

EBOVAC3 aims at supporting an essential part of the remaining clinical and manufacturing activities required for licensure in the European Union (EU) and the United States (US) for the candidate heterologous Ad26.ZEBOV and MVA-BN-Filo prophylactic vaccine regimen against Ebola virus disease. As a follow-up project, the IMI2 funded EBOVAC3 project, has started in June 2018. In this project, the vaccine strategy is further evaluated in specific populations in Africa (infants in Guinea and Sierra Leone; and front line workers in RDC). The project includes a work package on modelling, which is led by Rodolphe Thiébaut. Three workshop have been organized in Bordeaux (October 29th-30th, 2018), Arcachon (May 2nd-3rd, 2019) and Leiden (November 20th, 2019) to discuss and collaborate with the EBOVAC3 partners on the planned modelling work.

PREVAC-UP: The Partnership for Research on Ebola VACcinations-extended follow-UP and clinical research capacity build-UP.

SISTM is also involved in PREVAC-UP, an EDCTP2 project in direct link with the research carried out on the Ebola vaccines.

Coordinated by Inserm (France). Other beneficiaries: CNFRSR (Guinea), CERFIG (Guinea), LSHTM (UK), COMAHS (Sierra-Leone), NIAID (USA), NPHIL (Liberia), USTTB (Mali), Centre pour le Développement des Vaccins (Mali), Inserm Transfert SA (France). Duration: 60 months. 01 /01 /2019 - 31 /12 /2023.

Human-to-human transmission of Ebola virus in West Africa was interrupted in 2016 but the risk of reemergence of the disease is real. Thus, efforts to develop a safe and effective vaccine against Ebola virus disease with a durable prophylactic effect in communities must continue. The PREVAC-UP project is built around the PREVAC consortium. The Partnership for Research on Ebola Vaccinations (PREVAC) is an international consortium including the French Institute of Health and Medical Research, the London School of Hygiene & Tropical Medicine, the US National Institutes of Health, health authorities and scientists from Guinea, Liberia, Mali and Sierra Leone, a non-governmental organization (Alliance for International Medical Action), and Merck, Johnson & Johnson and Bavarian Nordic companies. The PREVAC trial is a phase IIB, randomized, placebo controlled, multicentre trial evaluating the safety and immunogenicity over 12 months of three vaccine strategies in children and adults. Participants are randomized to one of five groups: (i) vaccination with Ad26.ZEBOV prime and MVA-BN-Filo boost, (ii) vaccination with rVSVδG-ZEBOV-GP prime and a boost of the same vaccine, (iii) vaccination with rVSVδG-ZEBOV-GP vaccine without boost, (iv) placebo group 1 and (v) placebo group 2. Preliminary phases started in Liberia and Guinea in March 2017; the main phase of the trial evaluating the five regimens will begin in Liberia, Guinea Sierra Leone and Mali in April 2018 with an enrolment targets of 1,400 adults and 1,400 children.

PREVAC-UP two primary objectives are to determine (i) the long-term immunogenicity and safety and (ii) durability of humoral and cellular immune responses of Ebola vaccine regimes over 60 months. We will also evaluate the effect of co-infections, such as malaria and helminths on the immune response to vaccination. An integrative statistical analysis of the immune response will be used under the coordination of SISTM to explore the mechanism of action of the vaccines

and to identify early correlates of durable antibody induction. PREVAC-UP will also build on the extensive community mobilization efforts previously generated through PREVAC to provide a trans-national platform for social and health science research and training. Finally, this research proposal will expand and sustain capacity building and training of scientists in the four participant African countries. This program is expected to significantly impact Ebola prevention and control in adults and children in Africa. PREVAC-UP will also strengthen capacity for science relevant to the development and evaluation of new vaccines in sub-Saharian Africa.

In PREVAC-UP, SISTM leads the WP4 Utilisation of a system vaccinology approach using integrative statistical analyses and mechanistic modelling of the immune response to explore the interrelationship of immune response to Ebola vaccines. System vaccinology approach helps in better understanding and predicting the response to vaccines as demonstrated in the context of yellow fever, flu and many other vaccines. The idea is to integrate the massive data generated by high-throughput technologies (transcriptomics, flow cytometry, multiplex data) and population characteristics (sociodemographics and coinfections) to isolate the main markers/signatures associated to the vaccine response. Then, a mechanistic model of the response can be built and hopefully predict the individual long-term response. The PREVAC trial is a unique opportunity for setting up such an approach and apply it to the most advanced vaccine platforms against Ebola. The Inserm-SISTM team has produced several publications highlighting how within-host mechanistic models could play an important role in predicting vaccine efficacy and in improving treatment regimens, notably in HIV. The team has started to work on modelling the response to the Ad26.ZEBOV/MVA platform. In PREVAC-UP, it is expected that signatures and the mechanistic model itself will be different according to the type of vaccine as, specifically, the rVSV is a replicative vector. Two main outcomes are expected. One is a better understanding of the individual variability of the immune response and another is the prediction of the response with two specific aspects: after a new boost and on the long-term (5 years) for a new vaccinees. Identification and validation of an early correlate of later antibody responses would allow early prediction of whether an individual, or group of individuals is likely to be a poor responder and then to recommend subsequent interventions to test in this subset (such as change in vaccination strategy or additional boosts). Heterogeneity in antibody responses is expected within each group as it has been observed in former studies. In PREVAC-UP, information will be collected to inform the reason of this variability. Specific aspects will be explored such as the impact of malaria and various infectious agents on the immune response. Integrating such information in a mechanistic model of the immune response may help understanding the pathway leading to blunted response in vaccines and also to generate new hypotheses that could be biologically validated later on. Another important aspect of the modelling approach is the quantification of the impact of each potential factor helping to order the relative importance of various factors. In conclusion, this work is definitely at the confluence of the other work packages, integrating and ordering all the available information to understand and predict the effects of the promising vaccine strategy evaluated in the PREVAC trial.

9.3.3. Collaborations with Major European Organizations

University of Oxford;

London School of Hygiene and Tropical Medicine;

University Hospital Hambourg (UKE);

Heinrich Pette Institute for Experimental Virology, Hambourg;

MRC, University College London;

MRC Biostatistics Unit, University of Cambridge;

The University of Antwerpen;

University of Milan;

University of Bergen.

9.4. International Initiatives

9.4.1. Inria International Labs

Inria@EastCoast
Associate Team involved in the International Lab:

9.4.1.1. DYNAMHIC

Title: DYNAMical modeling of HIV Cures

International Partner (Institution - Laboratory - Researcher):

Harvard University (United States) - Harvard Program for Evolutionary Dynamics - Alison HILL

Start year: 2019

See also: https://team.inria.fr/dynamhic/

The aim of the DYNAMHIC Associate Team is to bring together a mathematical biology team at Harvard and the Inria team SISTM of applied statisticians at Bordeaux Sud-ouest. This collaboration will allow the analysis of unique pre-clinical non human primates data of HIV cure interventions. In particular, we will focus on immunotherapy and therapeutic vaccine, which are very promising in term of efficacy and are at the leading edge of pre-clinical research in the area. The novelty of the approach is to propose an integrative project studying complex biological processes with novel mathematical statistical models, which has the potential to yield predictive computational tools to assist in the design of both therapeutic products and clinical trials for HIV cure

Finally, the associate team is the opportunity to provide the research group with an official administrative framework. And, to continue to develop a promising research topic connected but different from those funded up to now.

Inria@SiliconValley
Associate Team involved in the International Lab:

9.4.1.2. SWAGR

Title: Statistical Workforce for Advanced Genomics using RNAseq

International Partner (Institution - Laboratory - Researcher):

RAND Corporation (United States) - Statistics group - Denis Agniel

Start year: 2018

See also: https://team.inria.fr/swagr/

The SWAGR Associate Team aims at bringing together a statistical workforce for advanced genomics using RNAseq. SWAGR combines the biostatistics experience of the SISTM team from Inria BSO with the mathematical expertise of the statistics group at the RAND Corporation in an effort to improve RNAseq data analysis methods by developing a flexible, robust, and mathematically principled framework for detecting differential gene expression. Gene expression, measured through the RNAseq technology, has the potential of revealing deep and complex biological mechanisms underlying human health. However, there is currently a critical limitation in widely adopted approaches for the analysis of such data, as edgeR, DESeq2 and limma-voom can all be shown to fail to control the type-I error, leading to an inflation of false positives in analysis results. False positives are an important issue in all of science. In particular in biomedical research when costly studies are failing to reproduce earlier results, this is a pressing issue. SWAGR propose to develop a rigorous statistical framework modeling complex transcriptomic studies using RNAseq by leveraging the synergies between the works of B. Hejblum and D. Agniel. The new method will be implemented in open-source software as a Bioconductor R package, and a user friendly web-application will be made available to help dissemination. The new method will be applied to clinical studies to yield significant biological results, in particular in vaccine trials through existing SISTM partnerships. The developed method is anticipated to become a new standard for the analysis of RNAseq data, which are rapidly becoming common in biomedical studies, and has therefore the potential for a large impact.

9.5. International Research Visitors

9.5.1. Visits of International Scientists

9.5.1.1. Internships

- Eva Reiner (Germany), intern in the Translational Vaccinology axis (March-July 2019)
- Aaron Sonabend, PhD student from Harvard University, collaborator in the High-dimensional statistical learning axis (June-August 2019) funded by the Harvard Rose Fellowship.

9.5.2. Visits to International Teams

9.5.2.1. Research Stays Abroad

- Boris Hejblum did a research stay at the Biostatistics Unit of The Medical Research Council at the University of Cambridge (Cambridge, UK) for a cumulative period of 1.5 month in 2019. This stay was devoted to collaborative work with Paul DW Kirk on scalalble bayesian computational methods.
- Boris Hejblum did a research stay at the Rand Corporation (offices in both Santa Monica CA and Boston MA) and at the Harvard Medical School (Boston MA, USA) for a cumulative period of 2 weeks in 2019. This stay was devoted to collaborative work with Denis Agniel in the context of the SWAGR Associate Team and with Tianxi Cai on high-dimensional statistical inference.
- Mélanie Prague did a research stay abroad in Harvard.

10. Dissemination

10.1. Promoting Scientific Activities

10.1.1. Scientific Societies Involvement

- Boris Hejblum is a member of the chairing committee of the *Société Française de Biométrie*, the French Chapter of the International Biometric Society
- Boris Hejblum is a board member of the "MAchine Learning et Intelligence Artificielle" (MALIA) group of *Société Française de Statistique* (SFdS)
- Mélanie Prague is a board member of the "Jeunes Statisticien.ne.s" group of SFdS
- Marta Avalos is a board member of the "Stat&Sport" group of SFdS

10.1.2. Scientific Events: Organisation

10.1.2.1. General Chair, Scientific Chair

Mélanie Prague co-organized with Jérémy Guedj from Inserm the 4th Workshop on Virus Dynamics https://viraldynamics.sciencesconf.org, 21-23 October in Paris, France.

10.1.3. Scientific Events: Selection

10.1.3.1. Chair of Conference Program Committees

- Robin Genuer Co-organizes a reading group called "Smiling in Bordeaux" (http://www.math.u-bordeaux.fr/~mchave100p/smiling/)
- Boris Hejblum organizes the Biostatistics Seminar Series at the Bordeaux Population Health Inserm Research Center
- Boris Hejblum co-organizes the Bordeaux Statistics Seminar Series (https://indico.math.cnrs.fr/category/367/)

10.1.3.2. Member of the Conference Program Committees

- Mélanie Prague was a member of the 51th Days of Statistics (JdS), the most important annual scientific event in the French-speaking statistical community, organized by the French Statistical Society (SFdS), June 2019, Nancy, France
- Rodolphe Thiébaut was a member of the scientific committee of the national conference on clinical research (EPICLIN)
- Rodolphe Thiébaut is a member of the scientific committee of the IWHOD International Workshop on HIV Observational Databases since 2013 (http://iwhod.org/Committee) and chair in 2020.

10.1.3.3. Reviewer

- NeurIPS (Robin Genuer)
- Workshop at NeurIPS 2019 "ML4H: Machine Learning for Health" (Marta Avalos)

10.1.4. Journal

10.1.4.1. Member of the Editorial Boards

Lifetime Data Analysis (Daniel Commenges)

Statistics Surveys (Daniel Commenges)

Associate editor of International journal of Biostatistics (Melanie Prague)

IMIA Yearb Med Inform (Rodolphe Thiébaut)

10.1.4.2. Reviewer - Reviewing Activities

- AIDS (Rodolphe Thiébaut)
- Am J Public Health (Mélanie Prague)
- Biometrics (Mélanie Prague, Boris Hejblum)
- Bioinformatics (Boris Hejblum)
- IMIA Yearb Med Inform (Marta Avalos)
- International Journal of Epidemiology (Daniel Commenges)
- Journal of Computation Statistics and Data Analysis (Boris Hejblum)
- Journal of the American Statistical Association (Robin Genuer)
- Journal of the Royal Statistical Society: Interaction (Mélanie Prague)
- JRSS-B (Mélanie Prague)
- Operations research (Robin Genuer)
- PLOS Computational Biology (Boris Hejblum)
- Society of clinical trial (Mélanie Prague)
- Statistical Methods in Medical Research (Mélanie Prague)
- Statistical science (Mélanie Prague)
- Trials (Laura Richert)

10.1.5. Invited Talks

- 3rd EcoSta Conference – 25-27th June 2019. Taïchung, Taïwan. A variance component score testapplied to RNA-Seq differential analysis. [B Hejblum]
- 12th International Conference of the ERCIM Working Group on Computational and Methodological Statistics – 14-16th December 2019, London, UK. Scaling up nonparametric Bayesian clustering with MCMC for big data applications. [B Hejblum]
- The 9th International Digital Public Health Conference – 20-23 November 2019, Marseille, The Challenges of Implementing Healthcare Technology and Innovation across Europe and Beyond [R Thiebaut]

- Conférence du centre de recherche du CHU de Sainte Justine – 17 October 2019, Montreal, Canada - *Science des données en épidémiologie clinique: exemples en vaccinologie*. [R Thiebaut]
- 4th Workshop on viral dynamics – 21-23rd October 2019, Paris - Modeling to optimize vaccine development against Ebola [R Thiebaut]
- INFECTION, IMMUNITY AND INFLAMMATION PROGRAMME Mini-symposium Mathematical Immunology in memoriam Robin Callard - 3rd October 2019, Institute of Child Health, London - Modelling in Immunology from 2007 to tomorrow, an overview of the work done with and inspired by Robin Callard [R Thiebaut]
- Workshop ITUN CESI, CHU Nantes, 3rd September. Understanding and predicting the effect exogeneous IL-7 in HIV infected patients through mathematical modelling: toward personnalised medicine. [R Thiebaut]
- 14th Colloquium of ADEA *La transformation du système de santé : l'hôpital de demain*, Cambo les Bains, September. L'intelligence artificielle [R Thiebaut]
- Robin's Callard symposium – Institute of Child Health, London – 4 October [R Thiebaut]

10.1.6. Oral contributions in conferences and colloquium

- Several members contributed to the 51th Days of Statistics (JdS), the most important annual scientific event in the French-speaking statistical community, organized by the French Statistical Society (SFdS), June 2019, Nancy, France [R Genuer, M Prague, M Avalos]
- UseR! 2019, Toulouse, France – July 2019. VICI: a Shiny app for accurate esti-mation of Vaccine Induced Cellular Immunogenicity with bivariate modeling. [B Hejblum]
- Joint Statistical Meeting 2019, Denver, USA – July 2019. Can you trust differential expression methods for RNA-seq data analysis ? [B Hejblum]
- 6th *Colloque francophone international sur l'enseignement de la statistique* 2019, Nancy, France– June 2019. *Enseigner la science des données en santé publique* [B Hejblum]
- Medinfo 2019, Lyon, 25-30 August 2019 [R Thiebaut]
- Several Talks at Virus Dynamics, October, Paris [R. Thiebaut, M. Alexandre, Q. Clairon].
- 32nd IEEE International Symposium on Computer-Based Medical Systems (CBMS) [H. Lorenzo].
- GDR Statistics and Health, Paris, France, October 2019 [L. Villain].
- StatOmique day - November 2019 [M. Gauthier]

10.1.7. Poster contributions in conferences and colloquium

- The French Health Data System applied to health research Meeting - June 2019, Rennes, France [M Avalos]
- PAGE June Stocholm [M. Prague, M. Alexandre]
- StatOmique day - Novembrer 2019, Paris [L. Villain]

10.1.8. Scientific Expertise

- Rodolphe Thiébaut is an expert for INCA (Institut National du Cancer) for the PHRC (Programme hospitalier de recherche Clinique en cancérologie) and for the PRME (Programme de recherche médico-économique en cancérologie).
- Rodolphe Thiébaut is a member of the CNU 46.04 (Biostatistiques, informatique médicale et technologies de communication).
- Rodolphe Thiébaut is a member of the Scientific Council of Inserm.
- Rodolphe Thiébaut is a member of the commitee "Biologie des Systèmes et Cancer (Plan Cancer)", a member of the Scientific Advisory Board of the "Institut Pierre Louis d'Epidémiologie et de Santé Publique" (UPMC, Dir : Dominique Costagliola), a member of the independent committee of international trials ODYSSEY and SMILE, a member of the scientific council of Muraz's Center (Bobo-Dioulasso, Burkina Faso)

- Mélanie Prague is an expert for ANRS (France Recherche Nord&Sud Sida-HIV Hépatites) in the CCS13 (Recherches cliniques et physiopathologiques dans l'infection à VIH) and AC 47.
- Laura Richert is an expert for the PHRC (Programme hospitalier de recherche Clinique).
- Marta Avalos is an expert for the ANSM (Agence nationale de sécurité du médicament et des produits de santé)

10.1.9. Research Administration

Daniel Commenges is the director of the Biostat-Info axis in the Inserm BPH (Bordeaux Public Health) institute.

Rodolphe Thiébaut is the director of the department of Public Health in University of Bordeaux and a member of the Inserm Scientific Council

Laura Richert is coordinator of the Clinical epidemiology module of the Clinical Investigations Center (CIC1401 Bordeaux)

10.2. Teaching - Supervision - Juries

10.2.1. Teaching

In class teaching

Master: Rodolphe Thiébaut is head of the Digital Public Health graduate program, University of Bordeaux.

Master : All the permanent members and several PhD students teach in the Master of Public Health (M1 Santé publique, M2 Biostatistique and/or M2 Epidemiology and/or M2 Public Health Data Science) and the Digital Public Health graduate program, University of Bordeaux.

Master : Marta Avalos, Robin Genuer, Louis Capitaine and Marine Gautier teach in the Master of Applied Mathematics and Statistics (1st and/or 2nd year), University of Bordeaux.

Master : Marta Avalos teaches in the 2nd year of the Master of "Management international : Développement pharmaceutique, Production et Qualité opérationnelle", University of Bordeaux.

Bachelor: Laura Richert and Edouard Lhomme teach in PACES and DFASM1-3 for Medical degree at Univ. Bordeaux

Master: Hadrien Lorenzo teach in the third year engineering school ENSAI, Rennes and in the "Ecole Santé Science" Program (dual degree in Health and Science), University of Bordeaux.

Master: Laura Richert teaches in the Master of Vaccinology from basic immunology to social sciences of health (University Paris-Est Créteil, UPEC)

Teaching unit coordination: Laura Richert, Rodolphe Thiébaut, Robin Genuer, Boris Hejblum and Marta Avalos coordinate several teaching units of Master in Public Health (Biostatistics, Epidemiology, Public Health), M2 Public Health Data Science. Laura Richert coordinates the teaching unit "Experimental Designs" (M2 Epidemiology) and teaching unit "Omics Data" (M2 Public Health Data Science), University of Bordeaux

Internation summer school: Laura Richert coordinates the summer school course "Big data in immunology" University of Bordeaux. All permemanent team members teach in this course.

H Lorenzo participated to the University Course *Big data et statistique pour l'ingénieur (BDSI)*, Ecole Nationale Supérieure de Cognitique (ENSC), Bordeaux.

E-learning

Master: Marta Avalos is head of the first year of the e-learning program of the Master of Public Health, University of Bordeaux.

Master: Marta Avalos teaches in the e-learning program of the Master of Public Health (1st and 2nd year)

ODL University Course: Robin Genuer is head of the Diplôme universitaire "Méthodes statistiques en santé.

ODL University Course: Mélanie Prague teaches in the Diplôme universitaire "Méthodes statistiques de régression en épidémiologie".

ODL University Course: Laura Richert co-coordinates and teaches in the Diplôme universitaire "Recherche Clinique".

ODL University project: Robin Genuer participated to the IdEx Bordeaux University "Défi numérique" project "BeginR" (http://beginr.u-bordeaux.fr).

10.2.2. Supervision

HdR: Laura Richert, *Statistical approaches and methodological developments to optimize clinical vaccine research*, Université de Bordeaux, defence date 17th Dec 2019

PhD: Edouard Lhomme, *Analyse des déterminants de la réponse immunitaire post-vaccination dans des stratégies vaccinales expérimentales*, defence date 25th Nov, directed by Laura Richert

PhD: Hadrien Lorenzo, *Analyse supervisée multibloc en grande dimension*, defence date 27th Nov, co-directed by Rodolphe Thiébaut and Jérôme Saracco (CQFD team)

PhD: Perrine Soret, "Régression pénalisée de type Lasso pour l'analyse de données biologiques de grande dimension : application à la charge virale du VIH censurée par une limite de quantification et aux données compositionnelles du microbiote", Université de Bordeaux, defence date 28th Nov, supervisor: Marta Avalos

PhD: Soufiane Ajana "Comparison of linear and non-linear machine learning approaches to predict Age-related Macular Disease (AMD) risk in a survival framework", defense's date 4th Nov, co-supervised by Boris Hejblum and Hélène Jacqmin-Gadda (Inserm) and Cécile Delcourt (Inserm), from Sept 2016.

PhD in progress: Marie Alexandre "Mechanistic modeling and optimization of vaccine response in HIV and Ebola", co-directed by Mélanie Prague and Rodolphe Thiébaut, from Oct 2018.

PhD in progress: Marine Gauthier "Methods for bulk and single-cell RNA-seq data analysis in vaccine research", co-directed by Boris Hejblum and Rodolphe Thiébaut, from Sept 2018.

PhD in progress: Louis Capitaine, *Random forests for high-dimensional longitudinal data*, from Oct 2017, co-directed by Robin Genuer and Rodolphe Thiébaut.

PhD in progress: Madelyn Rojas *Self-management of injury risk and decision support systems based on predictive computer modelling. Development, implementation and evaluation in the MAVIE cohort study*, from Oct 2017, (Injury Epidemiology team, Inserm U1219, ED SP2) co-directed by Emmanuel Lagarde (Inserm) and Marta Avalos.

Several team's members supervised Master 2 (T. Alin, A. Devaux, G. Sotton, A. Herteau, T. Ferte) and Master 1 (V. Gasque, M.H. Ibrahim, C. Lemoigne, V. Kochegarov) internship students.

10.2.3. Juries

Boris Hejblum was involved in the PhD defence jury of Soufiane Ajana (university of Bordeaux).

Marta Avalos was involved in the PhD defence jury of Perrine Soret (University of Bordeaux).

Laura Richert was involved in the PhD defence jury of Edouard Lhomme (University of Bordeaux).

Robin Genuer was reviewer of the thesis of Antonio Sutera (University of Liège).

Rodolphe Thiébaut was involved in the PhD defence jury of Edouard Lhomme, Hadrien Lorenzo, Perrine Soret, Sophie Lefèvre-Arbogast, Corentin Segalas (University of Bordeaux), Simon Bussy, Hugo Arlegui, Jeanne Tamarelle (Paris), and Alban Caporossi (Lyon). He was also a reviewer of the HDR of Encarnita Mariotti-Ferrandiz (Sorbonne University, Paris) and took part in her defence jury.

Mélanie Prague was involved in the PhD defence jury of Chiara Nicolo (University of Bordeaux) and Steven Sanche (University of Montreal, CA)

Marta Avalos is a member of the follow-up dissertation committee of 2 PhD students: Alexandre Conanec (Statistics, IMB, ED MI) and Madelyn Rojas (Epidemiology, ED SP2).

Mélanie Prague is a member of the follow-up dissertation committee of 2 PhD students: Jonas Beal (Institut Curie) and Marie astrid Metten (University of Rennes).

Laura Richert, Rodolphe Thiébaut, Robin Genuer, Boris Hejblum, Edouard Lhomme and Marta Avalos participated to the juries of Master in Public Health (Biostatistics, Epidemiology, Public Health)

Robin Genuer and Marta Avalos participated to the juries of Master of Applied Mathematics and Statistics (2nd year), University of Bordeaux.

Edouard Lhomme participated to the juries of medical thesis defenses, Medical School of Bordeaux University

10.3. Popularization

10.3.1. Internal or external Inria responsibilities

Mélanie prague is part of two committees at Inria: "ADT - Aide au développement technologique" and "CER - Comission Emploi recherche".

10.3.2. Articles and contents

Mélanie Prague participated to the video reportage produced by the Inria's communication service

10.3.3. Interventions

Mélanie Prague participated at the "Fête de la Science" for high school students on October.

10.3.4. Internal action

R Thiébaut gave the talk "VIRUS EBOLA: Comment la modélisation permet d'aider au développement des vaccins ?" in November at "Unithé ou café", Inria BSO, Talence.

H Lorenzo presented a poster *Apprentissage supervisé pour données massives multi-blocs incomplètes* in the anual workshop (*JED*) of the Doctoral School of Societies, Politics and Public Health (ED SP2), University of Bordeaux.

11. Bibliography

Major publications by the team in recent years

[1] D. AGNIEL, B. P. HEJBLUM. *Variance component score test for time-course gene set analysis of longitudinal RNA-seq data*, in "Biostatistics", 2017, vol. 18, n° 4, pp. 589-604, https://hal.inria.fr/hal-01579077

[2] D. COMMENGES, C. ALKHASSIM, R. GOTTARDO, B. P. HEJBLUM, R. THIÉBAUT. *cytometree: A binary tree algorithm for automatic gating in cytometry analysis*, in "Cytometry Part A", 2018, vol. 93, n° 11, pp. 1132-1140, https://hal.inria.fr/hal-01887966

[3] D. COMMENGES, H. JACQMIN-GADDA. *Dynamical Biostatistical Models*, Chapman and Hall/CRC, 2015, https://hal.inria.fr/hal-01580149

[4] A. JARNE, D. COMMENGES, M. PRAGUE, Y. LEVY, R. THIÉBAUT. *Modeling CD4 + T cells dynamics in HIV-infected patients receiving repeated cycles of exogenous Interleukin 7*, in "Annals of Applied Statistics", 2017, https://hal.inria.fr/hal-01579008

[5] C. PASIN, F. DUFOUR, L. VILLAIN, H. ZHANG, R. THIÉBAUT. *Controlling IL-7 injections in HIV-infected patients*, in "Bulletin of Mathematical Biology", 2018

[6] M. PRAGUE, D. COMMENGES, J. M. GRAN, B. LEDERGERBER, J. YOUNG, H. FURRER, R. THIÉBAUT. *Dynamic models for estimating the effect of HAART on CD4 in observational studies: Application to the Aquitaine Cohort and the Swiss HIV Cohort Study*, in "Biometrics", 2017, https://hal.inria.fr/hal-01406614

[7] A. RECHTIEN, L. RICHERT, H. LORENZO, G. MARTRUS, B. P. HEJBLUM, C. DAHLKE, R. KASONTA, M. ZINSER, H. STUBBE, U. MATSCHL, A. LOHSE, V. KRÄHLING, M. EICKMANN, S. BECKER, R. THIÉBAUT, M. ALTFELD, M. ADDO. *Systems Vaccinology Identifies an Early Innate Immune Signature as a Correlate of Antibody Responses to the Ebola Vaccine rVSV-ZEBOV*, in "Cell Reports ", 2017, vol. 20, n° 9, pp. 2251 - 2261, https://hal.inria.fr/hal-01579246

[8] L. VILLAIN, D. COMMENGES, C. PASIN, M. PRAGUE, R. THIÉBAUT. *Adaptive protocols based on predictions from a mechanistic model of the effect of IL7 on CD4 counts*, in "Statistics in Medicine", 2018, vol. 38, n° 2, pp. 221-235

Publications of the year

Doctoral Dissertations and Habilitation Theses

[9] E. LHOMME. *Analysis of the determinants and modelisation of the post- vaccination immune response in experimental vaccine strategies*, Université de bordeaux, November 2019, https://hal.inria.fr/tel-02430642

[10] H. LORENZO. *Supervised analysis of high dimensional multiblock data*, Université de bordeaux, November 2019, https://hal.inria.fr/tel-02433612

[11] L. RICHERT. *Statistical approaches and methodological developments to optimize clinical vaccine research*, Université de Bordeaux, December 2019, Habilitation à diriger des recherches, https://hal.inria.fr/tel-02426909

[12] P. SORET. *Penalized Lasso regression for the analysis of high-dimensional biological data: application to HIV viral load censored by a limit of quantification and to microbiota compositional data*, Université de bordeaux, November 2019, https://hal.inria.fr/tel-02425113

Articles in International Peer-Reviewed Journals

[13] S. AJANA, N. N. ACAR, L. BRETILLON, B. P. 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*, in "Bioinformatics", 2019 [*DOI :* 10.1093/BIOINFORMATICS/BTZ135], https://hal.inria.fr/hal-02425449

[14] M. CHAVENT, R. GENUER, J. SARACCO. *Combining clustering of variables and feature selection using random forests*, in "Communications in Statistics - Simulation and Computation", February 2019, pp. 1-20 [*DOI :* 10.1080/03610918.2018.1563145], https://hal.archives-ouvertes.fr/hal-02013631

[15] M. COPPRY, C. LEROYER, M. SALY, A. G. VENIER, B. MEGARBANE, C. SLEKOVEC, X. BERTRAND, S. PARER, S. ALFANDARI, E. CAMBAU, C. LAWRENCE, B. CLAIR, A. LEPAPE, P. CASSIER, D. TRIVIER, A. BOYER, J. ASSELINEAU, V. DUBOIS, R. THIÉBAUT, A. M. ROGUES. *Exogenous acquisition of Pseudomonas aeruginosa in intensive care units: a prospective multi-centre study (DYNAPYO study)*, in "Journal of Hospital Infection", August 2019 [*DOI :* 10.1016/J.JHIN.2019.08.008], https://hal.archives-ouvertes.fr/hal-02315665

[16] S. DOERKEN, M. F. AVALOS, E. LAGARDE, M. SCHUMACHER. *Penalized logistic regression with low prevalence exposures beyond high dimensional settings*, in "PLoS ONE", May 2019, vol. 14, n° 5, e0217057 p. [*DOI :* 10.1371/JOURNAL.PONE.0217057], https://hal.inria.fr/hal-02140472

[17] L. GRAS, M. MAY, L. P. RYDER, A. TRICKEY, M. HELLEBERG, N. OBEL, R. THIÉBAUT, J. GUEST, J. GILL, H. CRANE, V. DIAS LIMA, A. D'ARMINIO MONFORTE, T. STERLING, J. MIRÓ, S. MORENO, C. STEPHAN, C. SMITH, J. TATE, L. SHEPHERD, M. SAAG, A. RIEGER, D. GILLOR, M. CAVASSINI, M. MONTERO, S. INGLE, P. REISS, D. COSTAGLIOLA, F. W. WIT, J. A. C. STERNE, F. DE WOLF, R. B. GESKUS. *Determinants of Restoration of CD4 and CD8 Cell Counts and Their Ratio in HIV-1-Positive Individuals With Sustained Virological Suppression on Antiretroviral Therapy*, in "Journal of Acquired Immune Deficiency Syndromes", 2019, vol. 80, n° 3, pp. 292-300 [*DOI :* 10.1097/QAI.0000000000001913], https://hal.sorbonne-universite.fr/hal-02279469

[18] B. P. HEJBLUM, C. ALKHASSIM, R. GOTTARDO, F. CARON, R. THIÉBAUT. *Sequential Dirichlet Process Mixtures of Multivariate Skew t-distributions for Model-based Clustering of Flow Cytometry Data*, in "Annals of Applied Statistics", 2019, vol. 13, n° 1, pp. 638-660, https://arxiv.org/abs/1702.04407 - 39 pages, 11 figures [*DOI :* 10.1214/18-AOAS1209], https://hal.inria.fr/hal-01579063

[19] B. P. HEJBLUM, G. M. WEBER, K. P. LIAO, N. P. PALMER, S. CHURCHILL, N. SHADICK, P. SZOLOVITS, S. N. MURPHY, I. KOHANE, T. CAI. *Probabilistic record linkage of de-identified research datasets with discrepancies using diagnosis codes*, in "Scientific Data ", 2019, vol. 6, 180298 p. [*DOI :* 10.1038/SDATA.2018.298], https://hal.inria.fr/hal-01942279

[20] P. HELLARD, M. F. AVALOS, G. LEFORT, R. PLA, I. MUJIKA, J.-F. TOUSSAINT, D. B. PYNE. *Elite Swimmers' Training Patterns in the 25 Weeks Prior to Their Season's Best Performance: Insights Into Periodization From a 20-Years Cohort*, in "Frontiers in Physiology", April 2019, vol. 10 [*DOI :* 10.3389/FPHYS.2019.00363], https://hal.inria.fr/hal-02095063

[21] É. LHOMME, B. P. HEJBLUM, C. LACABARATZ, A. WIEDEMANN, J.-D. LELIÈVRE, Y. LEVY, 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*, in "Journal of Immunological Methods", December 2019, 112711 p. [*DOI :* 10.1016/J.JIM.2019.112711], https://hal.inria.fr/hal-02425379

[22] É. LHOMME, C. MODET, A. AUGIER, S. FAYE, T. S. DABAKUYO-YONLI, C. LÉVY-MARCHAL, E. D'ORTENZIO, Y. YAZDANPANAH, G. CHENE, A. H. BEAVOGUI, L. RICHERT. *Enrolling study personnel in Ebola vaccine trials: from guidelines to practice in a non-epidemic context*, in "Trials", December 2019, vol. 20, n° 1, 422 p. [*DOI :* 10.1186/S13063-019-3487-0], https://hal.inria.fr/hal-02426902

[23] D. Low, S. Lefèvre-Arbogast, R. Dominguez-Gonzalez, M. Urpi Sarda, P. Micheau, M. Pétéra, D. Centeno, S. Durand, E. Pujos-Guillot, A. Korosi, P. J. Lucassen, L. Aigner, C. Proust-Lima, C. Manach, B. P. Hejblum, C. Helmer, C. Andres-Lacueva, S. Thuret, C. Samieri. *Diet-Related Metabolites Associated with Cognitive Decline Revealed by Untargeted Metabolomics in a Prospective Cohort*, in "Molecular Nutrition and Food Research", 2019, e1900177 p. , Epub ahead of print [*DOI :* 10.1002/mnfr.201900177], https://hal.archives-ouvertes.fr/hal-02167524

[24] O. Lucar, M. Sadjo Diallo, C. Bayard, A. Samri, N. Tarantino, P. Debré, R. Thiébaut, F. Brun-Vézinet, S. Matheron, R. Cheynier, V. Vieillard. *B7-H6-mediated downregulation of NKp30 in natural killer cells contributes to HIV-2 immune escape*, in "AIDS", 2019, vol. 33, n° 1, pp. 23-32 [*DOI :* 10.1097/QAD.0000000000002061], https://hal.archives-ouvertes.fr/hal-02307134

[25] C. Moore, W. Stöhr, A. Crook, L. Richert, J.-D. Lelièvre, G. Pantaleo, F. García, S. Vella, Y. Lévy, R. Thiébaut, S. McCormack. *Multi-arm, multi-stage randomised controlled trials for evaluating therapeutic HIV cure interventions*, in "Lancet HIV", May 2019, vol. 6, n° 5, pp. e334-e340 [*DOI :* 10.1016/S2352-3018(19)30082-7], https://hal.inria.fr/hal-02426904

[26] A. Niehrs, W. Garcia-Beltran, P. Norman, G. Watson, A. Hölzemer, A. Chapel, L. Richert, A. Pommerening-Röser, C. Körner, M. Ozawa, G. Martrus, J. Rossjohn, J.-H. Lee, R. Berry, M. Carrington, M. Altfeld. *A subset of HLA-DP molecules serve as ligands for the natural cytotoxicity receptor NKp44*, in "Nature Immunology", September 2019, vol. 20, n° 9, pp. 1129-1137 [*DOI :* 10.1038/s41590-019-0448-4], https://hal.inria.fr/hal-02426901

[27] C. Pasin, I. Balelli, T. Van Effelterre, V. Bockstal, L. Solforosi, M. Prague, M. Douoguih, R. Thiébaut. *Dynamics of the Humoral Immune Response to a Prime-Boost Ebola Vaccine: Quantification and Sources of Variation*, in "Journal of Virology", September 2019, vol. 93, n° 18 [*DOI :* 10.1128/JVI.00579-19], https://hal.inria.fr/hal-02280792

[28] A. Samri, C. Charpentier, M. S. Diallo, M. Bertine, S. Even, V. Morin, A. Oudin, C. Parizot, G. Collin, A. Hosmalin, R. Cheynier, R. Thiébaut, S. Matheron, F. Collin, R. Zoorob, F. Brun-Vézinet, B. Autran. *Limited HIV-2 reservoirs in central-memory CD4 T-cells associated to CXCR6 co-receptor expression in attenuated HIV-2 infection*, in "PLoS Pathogens", 2019, vol. 15, n° 5, e1007758 p. [*DOI :* 10.1371/journal.ppat.1007758], https://hal.sorbonne-universite.fr/hal-02165899

[29] B. Silenou, M. F. Avalos, C. Helmer, C. Berr, a. Pariente, H. Jacqmin-Gadda. *Health administrative data enrichment using cohort information: Comparative evaluation of methods by simulation and application to real data*, in "PLoS ONE", January 2019, vol. 14, n° 1, e0211118 p. [*DOI :* 10.1371/journal.pone.0211118], https://hal.inria.fr/hal-02008499

[30] R. Thiébaut, B. P. Hejblum, H. Hocini, H. Bonnabau, J. Skinner, M. Montes, C. Lacabaratz, L. Richert, K. Palucka, J. Banchereau, Y. Lévy. *Gene Expression Signatures Associated With Immune and Virological Responses to Therapeutic Vaccination With Dendritic Cells in HIV-Infected Individuals*, in "Frontiers in Immunology", April 2019, vol. 10, 874 p. [*DOI :* 10.3389/fimmu.2019.00874], https://hal.archives-ouvertes.fr/hal-02267014

[31] T. van Stigt Thans, J. Akko, A. Niehrs, W. Garcia-Beltran, L. Richert, C. M. Stürzel, C. Ford, H. Li, C. Ochsenbauer, J. Kappes, B. H. Hahn, F. Kirchhoff, G. Martrus, D. Sauter, M. Altfeld, A. Hölzemer. *Primary HIV-1 Strains Use Nef To Downmodulate HLA-E Surface Expression*, in "Journal of Virology", October 2019, vol. 93, n° 20 [*DOI :* 10.1128/JVI.00719-19], https://hal.inria.fr/hal-02426899

Conferences without Proceedings

[32] H. LORENZO, R. MISBAH, J. ODEBER, P.-E. MORANGE, J. SARACCO, D.-A. TRÉGOUËT, R. THIÉBAUT. *High-Dimensional Multi-Block Analysis of Factors Associated with Thrombin Generation Potential*, in "CBMS 2019 - 32nd IEEE International Symposium on Computer-Based Medical Systems (CBMS)", Cordoba, Spain, IEEE, June 2019, pp. 453-458 [*DOI :* 10.1109/CBMS.2019.00094], https://hal.archives-ouvertes.fr/hal-02429302

Scientific Books (or Scientific Book chapters)

[33] L. DESQUILBET, S. GRANGER, B. P. HEJBLUM, A. LEGRAND, P. PERNOT, N. P. ROUGIER, E. DE CASTRO GUERRA, M. COURBIN-COULAUD, L. DUVAUX, P. GRAVIER, G. LE CAMPION, S. ROUX, F. SANTOS. , U. RÉGIONALE DE FORMATION À L'INFORMATION SCIENTIFIQUE ET TECHNIQUE DE BORDEAUX (editor) *Towards reproducible research : Evolve your practices*, Unité régionale de formation à l'information scientifique et technique de Bordeaux, May 2019, pp. 1-161, https://hal.archives-ouvertes.fr/hal-02144142

[34] R. GENUER, J.-M. POGGI. *Les forêts aléatoires avec R*, Presses Universitaires de Rennes, 2019, https://hal.archives-ouvertes.fr/hal-02409535

Other Publications

[35] L. CAPITAINE, R. GENUER, R. THIÉBAUT. *Fréchet random forests*, December 2019, https://arxiv.org/abs/1906.01741 - working paper or preprint, https://hal.archives-ouvertes.fr/hal-02409430

[36] L. CAPITAINE, R. GENUER, R. THIÉBAUT. *Random forests for high-dimensional longitudinal data*, February 2019, https://arxiv.org/abs/1901.11279 - working paper or preprint, https://hal.archives-ouvertes.fr/hal-02002979

[37] M. GAUTHIER, D. AGNIEL, R. THIÉBAUT, B. P. HEJBLUM. *dearseq: a variance component score test for RNA-Seq differential analysis that effectively controls the false discovery rate*, May 2019, working paper or preprint [*DOI :* 10.1101/635714], https://hal.inria.fr/hal-02138664

[38] B. XU, C. GIL-JARDINÉ, F. THIESSARD, É. TELLIER, M. F. AVALOS, E. LAGARDE. *Neural Language Model for Automated Classification of Electronic Medical Records at the Emergency Room. The Significant Benefit of Unsupervised Generative Pre-training*, December 2019, https://arxiv.org/abs/1909.01136 - working paper or preprint, https://hal.inria.fr/hal-02425097

References in notes

[39] O. AALEN, K. KJETIL ROYSLAND, J. GRAN, B. LEDERGERBER. *Causality, mediation and time: a dynamic viewpoint*, in "Journal of the Royal Statistical Society: Series A (Statistics in Society)", 2007, vol. 175, n⁰ 4, pp. 831–861

[40] F. CASTIGLIONE, B. PICCOLI. *Cancer immunotherapy, mathematical modeling and optimal control*, in "Biometrical Journal", 2007, vol. 247, n⁰ 4, pp. 723-32

[41] R. M. GRANICH, C. F. GILKS, C. DYE, K. M. DE COCK, B. G. WILLIAMS. *Universal voluntary HIV testing with immediate antiretroviral therapy as a strategy for elimination of HIV transmission: a mathematical model*, in "Lancet", 2009, vol. 373, n⁰ 9657, pp. 48-57, 0140 6736 English

[42] L. HOOD, Q. TIAN. *Systems approaches to biology and disease enable translational systems medicine*, in "Genomics Proteomics Bioinformatics", 2012, vol. 10, n⁰ 4, pp. 181–5

[43] Y. HUANG, D. LIU, H. WU. *Hierarchical Bayesian methods for estimation of parameters in a longitudinal HIV dynamic system*, in "Biometrics", 2006, vol. 62, n⁰ 2, pp. 413–423

[44] E. KUHN, M. LAVIELLE. *Maximum likelihood estimation in nonlinear mixed effects models*, in "Computational Statistics & Data Analysis", 2005, vol. 49, n⁰ 4, pp. 1020–1038

[45] K.-A. LE CAO, P. MARTIN, C. ROBERT-GRANIÉ, P. BESSE. *Sparse canonical methods for biological data integration: application to a cross-platform study*, in "BMC bioinformatics", 2009, vol. 10, 34 p.

[46] C. LEWDEN, D. SALMON, P. MORLAT, S. BEVILACQUA, E. JOUGLA, F. BONNET, L. HERIPRET, D. COSTAGLIOLA, T. MAY, G. CHÊNE. *Causes of death among human immunodeficiency virus (HIV)-infected adults in the era of potent antiretroviral therapy: emerging role of hepatitis and cancers, persistent role of AIDS*, in "International Journal of Epidemiology", 2005, vol. 34, n⁰ 1, pp. 121-130, 0300 5771 English

[47] A. S. PERELSON, A. U. NEUMANN, M. MARKOWITZ, J. M. LEONARD, D. D. HO. *HIV-1 dynamics in vivo: virion clearance rate, infected cell life-span, and viral generation time*, in "Science", 1996, vol. 271, n⁰ 5255, pp. 1582-6

[48] A. S. PERELSON. *Modelling viral and immune system dynamics*, in "Nature Reviews Immunology", 2002, vol. 2, n⁰ 1, pp. 28-36

[49] J. PINHEIRO, D. BATES. *Approximations to the log-likelihood function in the nonlinear mixed-effects model*, in "Journal of Computational and Graphical Statistics", 1995, vol. 4, n⁰ 1, pp. 12–35

[50] B. PULENDRAN. *Learning immunology from the yellow fever vaccine: innate immunity to systems vaccinology*, in "Nature Reviews Immunology", 2009, vol. 9, n⁰ 10, pp. 741-7

[51] H. PUTTER, S. HEISTERKAMP, J. LANGE, F. DE WOLF. *A Bayesian approach to parameter estimation in HIV dynamical models*, in "Statistics in Medicine", 2002, vol. 21, n⁰ 15, pp. 2199–2214

[52] A. REINER, D. YEKUTIELI, Y. BENJAMINI. *Identifying differentially expressed genes using false discovery rate controlling procedures*, in "Bioinformatics", 2003, vol. 19, n⁰ 3, pp. 368–375

[53] C. SCHUBERT. *Systems immunology: complexity captured*, in "Nature", 2011, vol. 473, n⁰ 7345, pp. 113-4

[54] R. SUN, S. HUI, G. D. BADER, X. LIN, P. KRAFT. *Powerful gene set analysis in GWAS with the Generalized Berk-Jones statistic*, in "PLoS genetics", 2019, vol. 15, n⁰ 3, e1007530 p.

[55] R. THIÉBAUT, B. P. HEJBLUM, L. RICHERT. *[The analysis of "Big Data" in clinical research.]*, in "Epidemiology and Public Health / Revue d'Epidémiologie et de Santé Publique", January 2014, vol. 62, n⁰ 1, pp. 1–4 [*DOI :* 10.1016/J.RESPE.2013.12.021], http://www.hal.inserm.fr/inserm-00933691

[56] R. THIÉBAUT, H. JACQMIN-GADDA, A. BABIKER, D. COMMENGES. *Joint modelling of bivariate longitudinal data with informative dropout and left-censoring, with application to the evolution of CD4+cell count*

and HIV RNA viral load in response to treatment of HIV infection, in "Statistics in Medicine", 2005, vol. 24, n° 1, pp. 65-82

[57] R. TIBSHIRANI. *Regression shrinkage and selection via the lasso*, in "Journal of the Royal Statistical Society: Series B (Statistical Methodology)", 1996, vol. 58, pp. 267–288

[58] Y. WANG. *Derivation of various NONMEM estimation methods*, in "Journal of Pharmacokinetics and pharmacodynamics", 2007, vol. 34, n° 5, pp. 575–593

[59] H. WU. *Statistical methods for HIV dynamic studies in AIDS clinical trials*, in "Statistical Methods in Medical Research", 2005, vol. 14, n° 2, pp. 171–192