MID-TERM EVALUATION REPORT

July 2018
The C3BI (https://c3bi.pasteur.fr) was launched in 2015 with strong support from the Institut Pasteur, as an integral part of its 2014-2018 Strategic Plan (see Appendix 1.1). The importance of computational approaches for the development of modern biology and medicine was clearly understood. The ambition was to build a leading national and international centre in bioinformatics, related disciplines and their applications in biology and health. Since then, our main mission has been to raise the knowledge and skills of the whole Institut Pasteur in these domains. Interaction with the 33 Institut Pasteur around the world, notably in developing countries, is also a major objective. Since 2015, we have been rapidly growing, through the appointment of about 10 engineers in bioinformatics and biostatistics per year, and the creation of 4 new research Units over that same period. One of these Units (InBio, started in 2017) is a common project with Inria, a French national research institution focusing on computer science and applied mathematics. In 2016, we became a joint “Unité Mixte de Service et de Recherche” (USR) with the CNRS, affiliated not only to the biology department (INSB), but also to the computer science (INS2I) and environment & ecology (INEE) departments. Our goal was to establish strong connections with formal sciences (computer science, mathematics and statistics) and to develop interdisciplinary research projects. By mid-2016, we applied to the French "Plan Investissement Avenir" (PIA)-Convergence Call, and got awarded with a grant of 12 million € over 9 years to set up the INCEPTION programme on integrative and multidisciplinary approaches for research about the emergence of pathologies in individuals and populations. This funding made it possible to develop ambitious projects at the interface between biology, data sciences and social sciences.

The C3BI includes two main components:

- The “Espace Recherche” (research area) is an umbrella structure for affiliated research entities: 9 Units and 2 junior groups (or G5). The Units that existed before 2015 are still primarily affiliated to their scientific department, while the new Units or G5 are generally affiliated to the C3BI.
- The “Hub de Bioinformatique et Biostatistique”, or simply Hub, is responsible for providing services to experimental research Units and platforms, for performing data analyses, for developing applications and data processing pipelines, and for providing training. These services are provided to all the Units on the Paris campus and also to the Institut Pasteur International Network. The Hub involves 50 research engineers in bioinformatics and biostatistics, which are organised in 6 groups with complementary expertise.

In the following report, we will first introduce the C3BI, its missions, and its trajectory since 2015, as well as the first results obtained and the new perspectives of the C3BI for the future. In a second part, we will introduce the research Units, starting with the 3 research Units of primary C3BI affiliation. These Units will be evaluated during the October 2018 session. The 8 other Units with secondary C3BI affiliation are not evaluated, but their activities in bioinformatics and biostatistics are briefly described in this document, as they actively participate to the C3BI via their publications, teaching, seminars, annual retreats, grant applications, etc. In fact, the C3BI is a privileged place to exchange on all methodological aspects of their research. Lastly, we will expose the activities and results of the Hub and its 6 expert groups.
# Abbreviation list

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Full Form</th>
</tr>
</thead>
<tbody>
<tr>
<td>APHP</td>
<td>Assistance Publique Des Hôpitaux de Paris (Public Hospital Network of Paris)</td>
</tr>
<tr>
<td>CL</td>
<td>Conseil de laboratoire (Laboratory council)</td>
</tr>
<tr>
<td>CNRS</td>
<td>Centre National de Recherche Scientifique</td>
</tr>
<tr>
<td>CODIR</td>
<td>Comité de direction (Executive committee)</td>
</tr>
<tr>
<td>COPIL</td>
<td>Comité de pilotage (Steering committee)</td>
</tr>
<tr>
<td>CR</td>
<td>Chargé de Recherche (Research associate)</td>
</tr>
<tr>
<td>CRIT</td>
<td>Centre de Ressources et Innovation Technologiques</td>
</tr>
<tr>
<td>DR</td>
<td>Directeur de recherche (Director of research)</td>
</tr>
<tr>
<td>HTS</td>
<td>High Throughput Sequencing</td>
</tr>
<tr>
<td>HR</td>
<td>Human Resources Department</td>
</tr>
<tr>
<td>INEE</td>
<td>Institut National Écologie et Environnement</td>
</tr>
<tr>
<td>INRIA</td>
<td>Institut National de Recherche en Informatique et en Automatique</td>
</tr>
<tr>
<td>INS2I</td>
<td>Institut National d'Informatique et ses Interactions du CNRS</td>
</tr>
<tr>
<td>INSB</td>
<td>Institut National des Sciences Biologiques du CNRS</td>
</tr>
<tr>
<td>IP</td>
<td>Institut Pasteur</td>
</tr>
<tr>
<td>NGS</td>
<td>Next Generation Sequencing</td>
</tr>
<tr>
<td>IPIN</td>
<td>Institut Pasteur International Network</td>
</tr>
<tr>
<td>PI</td>
<td>Principal Investigator</td>
</tr>
</tbody>
</table>
The eight following research Units have a secondary affiliation to the C3BI, and will not be evaluated in October 2018. Their activities in bioinformatics and biostatistics are quickly described in this document, as they actively participate to the C3BI.

| 2.1 | Evolutionary Bioinformatics ................................................................. 17 |
| 2.2 | Statistical Genetics ............................................................................. 21 |
| 2.3 | InBio: Experimental and Computational Methods for Modeling Cellular Processes .................................................................................................................. 26 |
| 2.4 | Systems Biology ..................................................................................... 30 |
| 2.5 | Human Genetics and Cognitive Functions .................................................. 35 |
| 2.6 | Mathematical Modelling of Infectious Diseases Unit .................................. Error! Bookmark not defined. |
| 2.7 | Imaging and Modeling ........................................................................... 37 |
| 2.8 | Structural Bioinformatics ...................................................................... 38 |
| 2.9 | Spatial Regulation of genomes ............................................................... 39 |
| 2.10 | Microbial Evolutionary Genomics ........................................................... 40 |
| 2.11 | Human Evolutionary Genetics .................................................................. 41 |
| 2.12 | Decision and Bayesian computation ....................................................... 42 |

3 BIOINFORMATICS AND BIOSTATISTICS HUB ................................................................. 43

| 3.1 | Introduction to the Bioinformatics and Biostatistics Hub .......................... 44 |
| 3.2 | Hub Expert Group SABER: Statistics Applied to Biology and Experimental Results ................................................................. 56 |
| 3.3 | Hub Expert Group TEG: Transcriptome and Epigenome ........................... 60 |
| 3.4 | Hub Expert Group WINTER: Web Integration ............................................ 65 |
| 3.5 | Hub Expert Group ALPS: Algorithmics and programming in science ........ 69 |
| 3.6 | Hub Expert Group FUNGEN: Functional Genomics ................................ 73 |
| 3.7 | Hub Expert Group GIPhy: Genome Informatics and Phylogenetics ............ 78 |
1 C3BI: MISSIONS, ORGANISATION, RESULTS AND PERSPECTIVES
1.1 Missions and development of the C3BI since March 2015

Biology is undergoing a profound and irreversible change. We now have access to huge amounts of data, including genomic data, which contains essential information and can be subject to multiple applications in the field of health in particular. Similar to other fields previously, such as astronomy, biology is becoming a computational science, increasingly using mathematical modelling and computational resources. Beyond the scientific aspects, the processing of new biological data will certainly play an important and increasing role in medicine, drug discovery, diagnosis and personalized treatment. The processing of this data, which has increased in a “super-exponential” manner, much faster than Moore’s Law, and the multiplicity of issues and objectives, raise new problems and challenges for computational and mathematical sciences such as complex stochastic models, fast algorithms, storage and sharing of data, etc. Bioinformatics, biostatistics and biomathematics are areas evolving rapidly, through the development of numerous training programmes, recruitment of large numbers of scientists and engineers both in institutional and private companies, as well as the creation of large centres. This growth converges to gather sufficient critical mass focused on answering multiple biological questions that meets the growing needs in these fields. Traditionally, two types of activities are distinguished in these areas: services to biologists, which provides them with help to process and analyse their data; and research, which focuses on developing new models and new methods of analysis. In reality, there is a continuum between these two types of activities; for example, more and more “dry biologists”, whose main working tool is the computer (rather than the “wet” bench), generally do not develop new methods, but use available data and computer programs to answer a wide variety of biological questions. Adapting methods that have been developed in other contexts to biology is also an essential activity.

In 2014, Institut Pasteur perfectly understood this radical evolution of biology. Important means were leveraged in the 2014-2018 Strategic Plan, to follow this evolution and to offer the Institut Pasteur community opportunities to benefit from this recent evolution. The ambition was to set up a transversal centre within the institute, with a strong critical mass, which counts at the national and international levels in bioinformatics, its related disciplines, and their applications to biology and health. The focus was particularly placed on infectious diseases and pathogens, for which the Institut Pasteur already benefits from a solid and visible reputation. Significant resources were allocated, including the recruitment of 40 bioinformatics engineers between 2014 and 2017, the setup of 3 new research Units with strong methodological component, and the renovation of a building, with a capacity of 135 people. The main mission of the “Centre de Bioinformatique, Biostatistique et Biologie Intégrative” (Bioinformatics, Biostatistics and Integrative Biology Centre) or “C3BI” has always been to raise the institute’s overall level in these domains, developing advanced methodological research, assisting the experimental labs in their data analyses, organizing trainings and setting up common software programs and pipelines to analyse the data produced by the institute and beyond. Interactions with the Institut Pasteur International Network (IPIN, 33 institutes worldwide, particularly in Africa and South-East Asia) were also a clear objective; the demand for bioinformatics being particularly strong in developing countries.

In March 2015, the C3BI was officially created after a wide-world consultation of a team of senior scientists from Institut Pasteur led by Didier Mazel. The initial team included 10 collaborators: 7 bioinformatics engineers recruited by the end of 2014, Marie-Agnès Dillies and Christophe Malabat, established research engineers at Institut Pasteur, in charge of the “Hub de Bioinformatique et Biostatistique”, and Olivier Gascuel, CNRS Research Director and head of the C3BI. In addition, the contribution of several deputy directors, Eduardo Rocha, Marie-Agnès Dillies, Michael Pressigout and Magnus Fontes were instrumental in the early phase of development of the C3BI at the organisational and operational level. Eduardo Rocha and Magnus Fontes left to pursue other missions from 2017 onwards.

IGDA (International Group for Data Analysis) was created in 2014 to cater for the IPIN capacity building needs in bioinformatics. In 2017, IGDA’s activities were included in the C3BI and Antonio Borderia continues his mission for 30% of his time as part of the C3BI.

The tasks of the C3BI in the launching phase of the C3BI have been significant, taking on many actions simultaneously:

- Recruitment of 12 high-level engineers for the Hub, via an international call (370 applicants, 45 candidates interviewed by 3 juries: scientific, technical and human resources).
- Establishment of a rigorous procedure to receive and process the projects submitted to the Hub by the experimental Units and platforms.
- Call for new C3BI-affiliated junior groups (G5) and Units (50 applicants, international interviewing committee, 3 candidates selected by the Scientific Council of Institut Pasteur, 2 eventually recruited: Hugues Aschard and Gregory Batt, who is Research Director at Inria; his Unit is a joint project with this institution.
- Application to become a joint Unit with the CNRS. The objectives were: to federate the Institut Pasteur research Units with marked activities in bioinformatics, biostatistics and biomathematics; to establish strong links with the CNRS, not only on
the biological level, but also in computer science; to allow the successful development of C3BI, on the service side, already well supported, but also on the research and methodological development side.

In 2016, our CNRS application was accepted. We have thus become an “Unité mixte de Service et de Recherche”, the USR 3756 (Institut Pasteur & CNRS), with two primary and balanced affiliations to INSB (biological sciences) and INSA2I (computer science), and a secondary affiliation to INEE (environment & ecology). The implementation of Hub services continued, as well as our efforts on our training mission within the institute and beyond. C3BI teaching activities in the IPIN started, with an international school organised in Senegal on next generation sequencing data (NGS). We held well-attended bioinformatics and biostatistics seminars, on a bi-monthly basis, with prestigious speakers such as Chris Sander on Structural Bioinformatics or Ludovic Orlando on Ancient DNA (see Appendix 1.6). The Unit of Olivier Gascuel (Evolutionary Bioinformatics) started in early 2016, and those of Hugues Aschard (Statistical Genetics) and Gregory Batt (InBio: Experimental and Computational Methods for Modeling Cellular Processes) at the end of the year. The two main achievements of 2016 were:

- The continuation of recruitment of bioinformatics and biostatistics engineers, with targeted profiles, different from those in 2015 (220 candidates, 40 interviewed by 3 juries, 10 recruitments).
- The development of the INCEPTION programme, submitted to the French call “Convergence” of the “Investissements d’Avenir”, on the integrative and interdisciplinary approaches to understand and fight the emergence of diseases in populations and individuals. The C3BI played a central role in this proposal, with eight PIs involved among a dozen in total, and a strong data science component (see below and Appendix 1.4 and 1.5). The grant was awarded in August 2016 (among 5 for about 40 applications, in all scientific fields), with an endowment of 12 million € over 9 years. It was gradually put in place at the end of the year for a launch in January 2017. INCEPTION is led by Olivier Gascuel and Thomas Bourgeron. The success of this project allowed us to set-up a strong incentive policy toward multidisciplinary research.

In 2017, the Hub reached a critical mass sufficient to ensure a broad spectrum of skills. It brought together nearly 40 engineers, recruited from 2014 to 2016, but also a dozen of previously recruited engineers from the campus, who had been previously isolated in their Units and services. To adapt to this change of scale, the Hub was re-organized and structured in small groups (3 to 9 collaborators) bringing together expertise on the major bioinformatics domains (transcriptomics, epigenetics, web applications…). The primary-affiliated research Units participated in the recruitment and reached or approached their maximum headcount. Courses were launched, notably a compulsory course on biological data analysis, for first-year Ph.D. students. This course was requested by the Institut Pasteur scientific board and was funded by INCEPTION, just as all C3BI courses currently taught on the campus. INCEPTION activities started (Ph.D. students’ recruitment, interdisciplinary projects, scientific events, etc.) supported by the recruitment of a Project Manager dedicated to INCEPTION: Marina Caillet (see Appendix 1.4 and 1.5). A final wave of recruitments for the Hub was launched at the end of the year, with targeted profiles following the demands of the departments and platforms, channelled through our Steering Committee. We received 190 applications resulting in 11 new recruitments. Angèle Bénard joined us as Scientific and Administrative Deputy Director of the C3BI in February 2018.

In 2018, the Hub growth phase reached the capacity envisioned in the 2014-2018 strategic plan. It includes today 50 research engineers. A third of them are detached in research Units and platforms and spend 20% of their time in the Hub to develop common projects (e.g. teaching, pipelines…). We moved to the new Alexandre Yersin building, renovated for the C3BI, with 135 seats. Four research Units are currently grouped together in this building, in addition to the Hub. A call was launched in 2017-2018 by INCEPTION to recruit new junior groups (G5) in computational biology, to round out our strength. We received more than 50 applications, which were evaluated by an international committee and then the institute’s Scientific Council. A new G5 is expected to start by the beginning of 2019, its research area deals with the algorithmic of genomes and metagenomes, an essential topic for the institute in the coming years. Eventually more than 50 articles were published between the beginning of the year and the 1st of July 2018, some in prestigious journals such as Nature and P.N.A.S. (see Appendix 2.1 and 3.4).
1.2 Organisation and governance

As of July 2018, the C3BI is composed of about 80 members with primary affiliation, and about 120 with secondary affiliation. It is broken down into two large components: the “Espace Recherche” (research area), which brings together the research Units, and the “Hub de Bioinformatique et Biostatistique” (Bioinformatics and Biostatistics Hub). The Systems Biology Laboratory headed by Benno Schwikowski has an intermediate status, between research and services/collaborations with the experimental research Units of the campus (see Appendix 1.3 for all entities and members with primary affiliations).

The “Espace Recherche” includes 11 research Units, most of which, 7 in fact, existed prior to the C3BI with secondary affiliation. The 3 research Units with primary affiliation have been created since the launch of the C3BI and have a strong orientation towards computer science, and mathematical and statistical methods. The C3BI allows for the harmonious development of these activities, which could only be partially supported in a department that would focus on biological questions. The C3BI makes it possible to build synergies around methodological aspects; in this regard, INCEPTION programme stands out as the best example. Another outstanding illustration of this fruitful environment is the organisation of the C3BI seminars, which takes pride in its excellent reputation and attendance by many researchers from the institute but also from organisations outside the institute (see Appendix 1.6). The level of investment of secondary Units in method development is variable, but all of them have recognized activities in this area, particularly among young researchers, as evidenced by the last C3BI retreat which brought together about 90 participants and enabled many fruitful interactions (see Appendix 1.8). The 11 research Units are as follows:

- **Experimental and Computational Methods for Modelling Cellular Processes** - Gregory Batt  
  Start February 2017 - Joint project with Inria - Primary Affiliation
- **Statistical Genetics** (G5) - Hugues Aschard  
  Start September 2016 - Primary Affiliation
- **Evolutionary Bioinformatics** - Olivier Gascuel  
  Start January 2016 - Primary Affiliation
- **Decision and Bayesian Computation** (G5) - Jean-Baptiste Masson  
  Start January 2017 - Secondary Affiliation
- **Human Evolutionary Genetics** - Lluis Quintana Murci  
  Secondary Affiliation
- **Human Genetics and Cognitive Functions** - Thomas Bourgeron  
  Secondary Affiliation
- **Imaging and Modelling** - Christophe Zimmer  
  Secondary Affiliation
- **Mathematical Modelling Of Infectious Diseases** - Simon Cauchemez  
  Secondary Affiliation
- **Microbial Evolutionary Genomics** - Eduardo Rocha  
  Secondary Affiliation
- **Spatial Regulation of Genomes** - Romain Koszul  
  Secondary Affiliation
- **Structural Bioinformatics** - Michael Nilges  
  Secondary Affiliation

The Bioinformatics and Biostatistics Hub comprises 50 engineers. A third of them are detached in the research Units and platforms, another third is embedded in the experimental Units to work on long-term projects, the last third constitutes the Hub core, in charge of maintenance tasks (e.g. Galaxy Pasteur), the implementation of pipelines and software programmes, and the processing of short and medium-term projects submitted to the Hub. The detached engineers are working 80% of their time within the host entities, and 20% within the Hub for common activities (e.g. teaching). They are detached for long periods, from 2 to 5 years (renewable) for the platforms and research Units, respectively. Embedded and Hub core engineers (moving from one status to another is fluid depending on the projects) remain affiliated to the Hub, with varying degree of involvement in the projects submitted to the Hub. They are organized in 6 expert groups, dealing with different fields and under the responsibility of senior research engineers:

- **Genome informatics and phylogenetics** - Alexis Criscuolo
- **Web integration** - Hervé Ménager
- **Functional genomics** - Natalia Pietrosemoli
- **Algorithmics and programming in science** - Nicolas Maillet
- **Transcriptome and epigenome** - Claudia Chica
- **Statistics applied to biology and experimental results** - Gaël Millot
The governance of C3BI is ensured by the “Comité de Direction du C3BI” (C3BI-CoDir, Executive Committee), which meets every two weeks to deal with all day-to-day, budgetary, organizational, strategic and other issues. Reports of these meetings are sent to the entire C3BI. As of today, the C3BI-CoDir consists of:

- Olivier Gascuel – Director
- Angèle Bénard – Scientific and Administrative Deputy Director
- Marie-Agnès Dillies – Service Deputy Director, Co-Head of the Hub
- Christophe Malabat – Co-Head of the Hub
- Antonio Borderia – International Activities and Teaching
- Mélanie Ridel – Administrative Assistant

The “Conseil de Laboratoire” (CL, Laboratory Council) is made up of members of various personnel categories, including young researchers (Ph.D. students, Post-doctoral scientists, CR ...) and research engineers from the Hub. It was recently decided to complete the CL by the PIs of the research Units with primary affiliation to the C3BI (i.e. G. Batt and H. Aschard), and some PIs with secondary affiliation (e.g., E. Rocha). The CL played a very active role in the organization and the programme of our annual retreats, and in the scientific outreach and activities of the C3BI in general (see Appendix 1.8). It has met 4 times since the creation of the USR (January 2016), always together with the C3BI-CoDir, to foster the dialogue within the USR and between the various categories of personnel. The CL is continuously evolving to ensure its strong involvement in all aspects of the C3BI.

The last governing body of C3BI is the “Comité de Pilotage” (COPIL, Steering Committee). Actually, the COPIL does not manage the whole C3BI, but only the Hub and the services, rather than the research, which is carried out autonomously as is the norm at the institute and elsewhere. The Hub distributes significant human resources to the campus and the IPIN. It is essential that the distribution of these resources and the monitoring of the Hub’s activities remain, through the COPIL supervision, under the control of all the departments and bodies of the Institute. The COPIL is also involved in the strategic directions, in particular the recruitment profiles of engineers, in order to favour the major needs of the campus (e.g. bio-imaging in 2017, epigenetics in 2016), and the contents of our training programmes (e.g. mandatory courses in data analysis for Ph.D. students). The COPIL consists of one representative from each department, of representatives from the main platforms, and of representatives from the scientific and technical boards of the institute. The COPIL reviews and selects the long-term projects submitted to the Hub, to ensure impartiality as well as scientific and strategic relevance of these projects. The entire COPIL meets 1 to 2 times a year. The “bureau” (core) of the COPIL meets 3 times a year to arbitrate 4 to 6 long-term project requests. Details of the members, the organisation of the COPIL (steering committee), and the meetings’ agenda are shown in Appendix 1.2.
1.3 Achievements in Research for Three Years: Highlights

We summarise here some of the most visible achievements obtained since the start of the C3BI, in March 2015. All details, publications, projects, etc. are given in the research Units sections, pages 17-44.

The C3BI’s Units are recognized for their expertise and regularly publish papers in top international scientific journals. As of the 1st of July 2018, and since 2015, the C3BI Units published about 80 articles. More than 30 papers were published since the beginning of 2018 up to July 1st in Nature (2), P.N.A.S., Elife, Nature Communications (2), Nature Immunology (2), Nature Biotechnology (2), Nature Ecology and Evolution, American Journal of Human Genetics, Journal of Infectious Diseases, Nucleic Acid Research, Bioinformatics (2), PLOS Computational Biology...

C3BI has proven to be quite attractive and succeeded in recruiting top scientists during that period, from the PI level (H. Aschard, G. Batt, J.B. Masson) to the researcher level (Chargé de Recherche/Research Associate: G. Dumas (IP), M. Rotival (IP), H. Salje (IP), T. Rolland (CNRS), O. Rendueles Garcia (CNRS), C. Vestergaard (CNRS), J. Ruess (Inria)). Moreover, several of C3BI researchers received awards and ERC grants (C. Zimmer (Thérèse Lebrasseur Prize 2016), R. Toro (Open Science Prize 2016), Romain Koszul (ERC grant 2017, Pasteur Vallery-Radot Prize 2017), O. Gascuel (Inria–French Academy of Sciences Grand Prize 2017)).

The C3BI, its researchers and PIs played a key role in the success of the INCEPTION project on integrative and multidisciplinary approaches for research into the emergence of pathologies through individuals and populations (12 M€ / 9 years, held by O. Gascuel and T. Bourgeron, funded by the French "Projet Investissement Avenir" - Convergence Call 2016). Marina Caillat has been the project manager of INCEPTION since May 2017. Three of the four axes of INCEPTION are headed by C3BI PIs (Method for Integrative Biology, M. Nîlges, C. Zimmer, O. Gascuel; Emergence of Diseases in Populations, S. Cauchemez, E. Rocha, A. Fontanet; Emergence of Diseases in Individuals, T. Bourgeron, L. Quintana Murci, M.-L. Gougeon). INCEPTION involves all the Institut Pasteur Units, via calls for proposals, training and scientific activities, working in synergy with our partners (CNRS, Inserm, INRA, CEA, AP-HP, Paris Sciences et Lettres, Paris-Diderot University and the “Frontières du Vivant” – FdV – doctoral school). In 2017, progress was made on the project in several areas: recruitment of multidisciplinary Ph.D. students in epidemiology and human genetics; integrative projects on zoonoses in Africa and the resurgence of whooping cough in Europe; training on biological data analysis for Ph.D. students, and other bioinformatics and biostatistics courses, seminars and events; a rapid response Unit to intervene as quickly as possible in the event of a major pandemic outbreak (such as plague in Madagascar early 2018). Most importantly, INCEPTION has a yearly call to recruit new young research groups (G5) on its themes. Etienne Simon-Lorière (Evolutionary genomics of RNA viruses) was recruited in January 2018, and we expect to recruit a new G5 on the algorithmic of genomes and metagenomes, after the 2017-2018 call (>50 applicants, see Appendix 1.5).

1.4 Achievements in Services for Three Years: Highlights

Here, we summarise the main achievements in C3BI bioinformatics and biostatistics services since its start in March 2015. All recruitments, projects, publications, satisfaction surveys, etc. are described in the Hub section, page 44-76.

Since its creation, the Hub has hired 40 permanent research engineers in bioinformatics and biostatistics, an area where recruitment is notoriously difficult. Moreover, our calls attracted hundreds of applicants (370 in 2015, 220 in 2016, 190 in 2017), which enabled us to achieve very high-level recruitments. Most Hub staff members have a Ph.D., and one or several postdoctoral year(s) of experience, as well as a number of publications. They all have solid experience in bioinformatics/biostatistics. A number of them are also well experienced in teaching.

Since its start, the Hub has worked on more than 280 projects. It is involved in 22 long-term projects approved by the C3BI COPIL, such as: the PIbNet project, which provides the tools required to analyse microbial genomes aiming at monitoring diseases and disease emergence. A recent satisfaction survey disseminated throughout the institute showed that experimental research Units and platforms are highly satisfied by the services provided by the Hub, with an average grade of 4.3 on a scale of 1 to 5 ([https://c3bi.pasteur.fr/c3bi-surveys/c3bi-satisfaction-survey-2018-results/](https://c3bi.pasteur.fr/c3bi-surveys/c3bi-satisfaction-survey-2018-results/) and Appendix 1.9).

Hub input on these projects is recognised in about 90 publications co-authored by Hub staff members. As expected there was a strong increase recently, with about 27 publications since the beginning of 2018 until the 1st of July. The Hub has also developed services to help scientists on the campus on a daily basis. Weekly open desks and on-line, instantaneous answers to short questions contribute to this daily support.

The Hub is strongly involved in teaching and training activities, as described in section 1.6. Since 2015, the Hub has designed 10 new courses on programming, statistics and NGS data analysis, closely mixing theory and practice. These courses are open to the campus but also, for some of them, to the scientific community outside of Institut Pasteur and represent about 300 hours of courses offered to the students and researchers of the IP Paris Campus and of the IPIN annually.
1.5 General overview of the bi-monthly Seminars of the C3BI

https://c3bi.pasteur.fr/seminars/

Organising team members:
Olaya Rendueles García (MEG), Maxime Rotival (HEG), Christophe Becavin (Hub), Pascal Campagne (Hub), Angèle Bénard (C3BI)

Overview:
The objective of the C3BI seminars is to build a scientific community within the Institut Pasteur and beyond around the research topics and expertise of the C3BI namely, bioinformatics, biostatistics and integrative biology. We are seeking to reach out to scientists who already have an interest in these topics but also to potential collaborators who might need to integrate expertise related to C3BI topics in their research project. Over the past 3 years, the C3BI has organised high-level scientific seminars with 61 eminent external speakers coming from major research organisations all over the world. These 1-hour bi-monthly seminars have been attended in average by around 40 participants from Institut Pasteur but also by scientists coming from other external research institutions. Some seminars have attracted up to 100 scientists, including up to 30 external guest attendees from different institutes around Paris (e.g. Curie, Institut Imagine, Gustave Roussy, IBENS, Ligue contre le cancer...). The quality of our speakers has been regularly praised by the audience. The organising team is committed to include a high degree of diversity in the guest speakers list. As an example, 17 women have been among the speakers since the beginning of the C3BI seminars series, in a field where the representation of women is notably scarce. Moreover, we thrive to invite junior group leaders and promising early career scientists from worldwide research institutions. The topics covered in these seminars range broadly from “Big Data” analysis to Bayesian statistics, bioimage informatics, biological networks, genomics or mathematical modelling, overall covering all fields and research topics of the C3BI. We offer the invited speaker to meet and discuss with 5 to 6 research scientists during the day and around lunch, furthermore scientific discussion is encouraged right after the seminars during a more informal collation. Alternatively, we also offer the opportunity to research engineers, PhD and post-doctoral students as well as researchers of the C3BI to expose their latest research results once a month during 30-minute sessions. The objective of these internal seminars is to share knowledge, disseminate C3BI research expertise, foster discussions and identify opportunities for collaborations in the C3BI themes. Overall, the appreciation for C3BI seminars is high with a mean of 4/5 in a satisfaction survey conducted recently internally (https://c3bi.pasteur.fr/c3bi-surveys/c3bi-satisfaction-survey-2018-results/ and see Appendix 1.9). The comments gathered from this survey range from the recognition of high scientific level of the seminars to the appreciation of the quality of the speakers and the great diversity of topics.

Speakers highlights: (see Appendix 1.6 for a complete list of C3BI seminars)

- **Prof. Lazlo-Albert Barbasi**, Northeastern University College of Science, “Network Medicine: From Cellular Networks to the Human Disease”, Tuesday 12th of June 2018, **100 participants**
- **Prof. Amos Bairoch**, from University of Geneva Medical school, “Some lessons learned during 30 years of biocuration activities”, Thursday 8th of March 2018, **70 participants**
- **Prof. Roderic Guiguo**, Group leader, Computational Biology of RNA Processing, Centre for Genomics Regulation, Barcelona “The molecular anatomy of the human body”, 11th of January 2018, **60 participants**
- **Dr. Ewan Birney**, European Institute for Bioinformatics, “Big data in biology: challenges and opportunities”, Thursday 5th of October 2017, **90 participants**
- **Prof. Gene Meyers**, Director, Centre for Systems Biology Dresden, Max Planck Institute for Molecular Cell Biology and Genetics, “Computer Vision & Computational Optics for Bioimage Informatics”, 21th September 2017, **75 participants**

**Perspectives:**
We will pursue our efforts in scientific outreach by inviting high-quality scientists and by including higher degree of diversity in our speakers list. We will thrive to use the opportunities of the C3BI seminars to foster new research collaborations. Furthermore, we are exploring the possibility to video-record the C3BI seminars in order to make them accessible to a wider audience, particularly to the Institut Pasteur International Network of all over the world. Eventually we will ensure that communication about the C3BI seminars reaches out to all interested inside and outside of the Institut Pasteur campus.
1.6 C3BI Teaching & Training programme at Paris Campus

https://c3bi.pasteur.fr/trainings/

Teaching and training coordinators:
Marie-Agnès Dillies (Hub), Stéphane Descorps-Declère (Hub), Angèle Bénard (C3BI)

Outlook and highlights:
Teaching & training is a critical mission of the C3BI and many C3BI scientists take an active role in a variety of courses at Institut Pasteur. Since September 2016, the C3BI has organised, co-organised or coordinated 21 courses involving more than 350 participants in total and representing close to 750 hours of teaching and training. These courses included around 20 people per course in average as a majority of them involved practical sessions, proven to be more efficient within small groups of students, but up to 120 participants for some theoretical sessions. In fact, most course formats alternate theoretical knowledge and practical “hands-on” exercises. These courses aim at endowing students with a solid scientific methodological background and at providing them with a prototypic data analysis workflow. At the end of the course, the participants have become familiar with the most commonly used software for scientific data analyses and visualisation. The pedagogical objectives of the courses are to provide basic knowledge and practical handling of bioinformatic tools as well as fundamentals in experimental planning and biostatistics. The range of course topics offered is varied, 10 in total, and mainly revolves around next generation sequencing data analysis, biostatistics, programming and phylogenetics; skills that are in high demand at the Institut Pasteur Paris campus. The most frequent topics taught at the C3BI so far deal with introduction to high-throughput sequencing data analysis and biostatistics. These courses target mainly Ph.D. and post-doctoral student audiences but some of them can be attended by collaborators from technician to principal investigator level. Some of the C3BI courses are integrated in several education programmes such as the internal Institut Pasteur continuing education offer, the doctoral course programme or the Pasteur courses (https://www.pasteur.fr/en/education/programs-and-courses/pasteur-courses) open to external participants or the course on data analysis within the Biology Master Programme at École Normale Supérieure (https://www.enseignement.biologie.ens.fr/spip.php?article91). Close to 40 research scientists from the C3BI, mainly from the Hub, have been involved in a course over the past three years and all C3BI members are encouraged to offer and design new course programmes on a regular basis.

Courses highlights:
- “Introduction to Data Analysis”, 3 sessions: Jan 16th to Feb 17th, 2017; Oct 16th to Nov 17th, 2017; Jan 8th to Feb 2nd 2018, 35 students/session, https://c3bi.pasteur.fr/training-introduction-to-data-analysis-201718/
- “Bioinformatics of protein–protein interactions for wet lab scientists”, 3rd to 7th of April 2017, participants: 24 during hands-on sessions, 21 during lectures, https://c3bi.pasteur.fr/training-bioinformatics-of-protein%C2%ADprotein-interactions-for-wet-lab-scientists/

Perspectives and strategy
A campus wide survey on the C3BI rated the relevance and satisfaction on the trainings delivered by the C3BI both at 4.4/5 (https://c3bi.pasteur.fr/c3bi-surveys/c3bi-satisfaction-survey-2018-results/ see Appendix 1.9). The feedback received at the end of each of the courses is incorporated into the next version of the course in order to improve and adapt the content and pedagogical objective of the courses on a yearly basis. Several new courses will be offered for the academic year 2018-2019, more specifically a programming peer-learning based course and a course targeting specifically a medical audience are in preparation.
1.7 C3BI Teaching in the Institut Pasteur International Network

Teaching and training coordinators:
Antonio Borderia (C3BI), Angèle Bénard (C3BI)

Outlook and highlights

One critical mission of the C3BI is to build capacity in bioinformatics and biostatistics at the Paris campus but also within all the Institut Pasteur International Network (IPIN). Several bioinformatics and biostatistics courses are organised each year to fulfil this mission. The main objectives of these courses are to build capacity in bioinformatics and biostatistics as well as to foster the development of a bioinformatic community across the IPIN.

The C3BI organises bioinformatics and biostatistics courses throughout the IPIN several times a year. Up to recently, those courses were jointly organised with the International Group for Data Analysis (IGDA), which was incorporated into the C3BI. This collective effort engages scientists from the IPIN either through travel grants incentive to participants from the IPIN, or through mixed training teams always including local teachers and teachers from other IP of the IPIN. The courses span a high variety of trending topics at the cutting edge of science such as metagenomics, precision medicine or GWAS. They have endowed more than 400 scientists with basic tools and concepts in bioinformatics and biostatistics that allow them to analyse their own data independently and to expand their knowledge in the field of bioinformatics and biostatistics. The originality of some of the courses coordinated by the C3BI within IPIN lies within the combination of high quality theoretical sessions backed up by hands-on practical sessions where participants receive individualized support to work on their own data. Eventually each and every one of the C3BI courses is carefully designed to adapt to the local needs of the research community as well as to fit with the regional global health priorities taking in account the epidemiological relevance of pathogens studied for example. Eventually, trainings are also carefully tailored to the level of knowledge and skill range of the participants, which present a major but exciting challenge.

Courses highlights:

• **Leishield Training Course on Next Generation Sequencing.** November 2015, 1 week, Paris (France) 15 participants, [https://c3bi.pasteur.fr/training-leishield-training-course-on-next-generation-sequencing/](https://c3bi.pasteur.fr/training-leishield-training-course-on-next-generation-sequencing/)

• **INDA Hands-on NGS-Statistics course.** October 2016, 2 weeks, M’Bour (Senegal) 27 participants - [https://c3bi.pasteur.fr/training-inda-handson-ngsstatistics-course-senegal-2016](https://c3bi.pasteur.fr/training-inda-handson-ngsstatistics-course-senegal-2016)

• **C3BI Hands-on NGS course.** November 2016, 2 weeks, Paris (France) 23 participants - [https://c3bi.pasteur.fr/training-handson-ngs-course-paris/](https://c3bi.pasteur.fr/training-handson-ngs-course-paris/)


• **C3BI Hands-on NGS course,** November 2017, 2 weeks, Hong Kong (China), 80 participants, [https://c3bi.pasteur.fr/training-c3bi-handson-ngs-course-hongkong-2017/](https://c3bi.pasteur.fr/training-c3bi-handson-ngs-course-hongkong-2017/)
Introduction:
Since its creation, the common efforts of the C3BI and IGDA have resulted in the organisation of 14 courses within 7 different countries across 4 continents, totalling up to 424 participants coming from more than 24 different countries.

Outcome:
Most of those courses have been co-designed and co-organised with the hosting IP of the IPIN, often based on a proposal originating from an IP of the IPIN. A participative project and active continuous discussion ensured the involvement of local stakeholders, including scientists from the local IP and from other IP of the IPIN, research engineers from the Hub, senior lecturers of IP Paris as well as administrative staff from all organisations involved. All actors collectively decided on a syllabus based on their respective area of expertise, the local scientific and public health priorities, the research and skill level of participants and the pedagogical objectives. This approach resulted in courses highly tailored to regional needs. Eventually, the coordinating team from C3BI/IGDA assisted by the local IP, ensured the fair selection of speakers, teaching assistants and participants, the scientific outreach and marketing of the course, the logistics and eventually the smooth running of the teaching and training course activities.

At least 8 different topics have been taught within C3BI/IGDA courses. These topics range widely from basic Next-Generation Sequencing analysis and biostatistics to cutting edge mathematical modelling or GWAS. Some topics were chosen to respond to a request from the IPIN, such as phylogenetics but also to reflect recent rapid development in a specific field such as metagenomics for example. Furthermore, Open Science and multidisciplinary approaches were the topic of three courses that involved the IPIN.

The originality of the C3BI/IGDA courses lies within the dual theoretical and hands-on feature of these courses. Participants acquired state of the art knowledge on a highly specialised topic, discussed with experts bioinformaticians, and analysed and interpreted their own data directly with the tools and methodologies taught during the theoretical sessions. This promoted a high level of commitment and motivation from the participants. Two small thematic courses for the Leishield Consortium intended to provide participants with basic knowledge on DNA-sequencing analysis without working directly on data. In addition, a “Train the Trainer” course aiming at providing teaching skills to local researchers of the IPIN was organised in 2015 on High Performance Computing (HPC) and Galaxy and was part of a wider strategy of long term capacity building.

Eventually three open science and multidisciplinary workshops were held in Uruguay, Denmark and France to support open, collaborative and multidisciplinary approaches around projects such as Labex Milieu Intérieur or arbovirus evolution. Satisfaction surveys at the end of each course highlight the high level of satisfaction with scores overall ranging from 3 to 4.5 (5 maximum score) (see Appendix 1.9). The categories that score the highest included usefulness of the course or availability of the teachers for example.

In order to support participating coming from other IPIN, IGDA/C3BI provided travel grants that covered flight and accommodation costs for about one third of the total participants and most of non-local participants (200 distributed over 4 courses). IGDA/C3BI also sponsored expenses related to travel of teachers invited from IP Paris and the IPIN to these courses. Eventually the local hosting IP provides in-kind support in terms of logistics for guest teachers and participants.

The sustainability of C3BI teaching and training activities is guaranteed by a dedicated budget of the C3BI, endowed by Institut Pasteur and can be supplemented in some cases through moderate participation fees. Occasionally, application to travel grants such as Calmette-Yersin provides additional funding opportunities for participants.

National & International Collaborations:
In order to foster the development of a bioinformatics community within the IPIN, we favour and support exchange of scientists from the IPIN by inviting teachers from one IP to actively participate in the pedagogical team and design of courses organised in another IP of the IPIN.

In addition, the organisation of courses on the Paris Campus has helped facilitating collaborations and competence transfer between bioinformaticians from the IPIN and Institut Pasteur Paris.

Throughout the past 3 years, we have been engaging in multiple collaborations through the co-organisation of courses with various local universities and other actors contributing to strengthening bioinformatics capacity globally. As an example, the first “Hands-on” NGS-GWAS course organised in Brazil was led in collaboration with Oswaldo Cruz Foundation and the University of Sao Paulo. The “Hands-on” course organised on GWAS-NGS in Senegal in 2015 was also organised in partnership with University Gaston Berger. Eventually the course in M'Bour in 2016 involved H3Abionet (https://h3abionet.org), a Pan African Bioinformatics network that aims at supporting project developing bioinformatics capacity in the Africa and the African Institute for Mathematical Sciences (https://www.aims-senegal.org). Our future strategy includes the further strengthening and active involvement in multiple other collaborations, namely the European Bioinformatics Institute, the Wellcome Trust Overseas Course programme and the Doherty Institute through co-design of courses, strategic goal harmonisation and sharing of resources.
Scientific outreach:

In 2018, we are organising a steering committee meeting that will be held around a mini-symposium on the theme: “Bioinformatics across the IPIN”. The aim of this event is to facilitate interactions between bioinformaticians across the IPIN, and between the IPIN and the Paris campus in order to support the development of a bioinformatics community within the IPIN with strong links and support from the Paris Campus. We have identified and invited 2 bioinformaticians from each IP of the IPIN, gathering around 60 participants in total and we are planning to organise short talks, scientific round table discussions as well as poster sessions.

Webinars & MOOC:

- **High Performance Computing Cluster** is a Small Private Online Course designed to teach new users, the basic rules of access and use of the resources provided by the High-Performance Computing Cluster of IP Paris that was co-designed by the Direction des Services Scientifiques and Antonio Borderia. It is an ongoing course that has trained a total 547 students registered up to today. [https://moocs.pasteur.fr/courses/Institut_Pasteur/DSI_01/1/about](https://moocs.pasteur.fr/courses/Institut_Pasteur/DSI_01/1/about)

- **Principles and Trends in Genomics and Computational Biology** was a MOOC designed to introduce students to basic techniques when dealing with molecular and computational biology. This MOOC was financed by the Institut Pasteur International Network Association and done in collaboration with the Oswaldo Cruz Foundation and Institut Pasteur. 1437 registered users followed the training and 198 received a certificate of completion. [https://moocs.pasteur.fr/courses/course-v1:Fiocruz_Institut_Pasteur+01+02/about](https://moocs.pasteur.fr/courses/course-v1:Fiocruz_Institut_Pasteur+01+02/about)

Awards and Grants:

- C3BI/IGDA will participate to the PHINDaccess, a “Twinning” project funded by an H2020 call from EU aiming at strengthening the research institutes in less privileged countries in a specific area of research. The goal is to transfer competences in order to increase research capacities around data generation, management, sharing and analysis. This grant involves other European Institutes such as the Max-Planck Institute, Robert Koch Institute and the Centre for Genomic Regulation (CRG).

Perspectives and strategy:

We are currently planning the organisation of two new courses during 2018, one on “Introduction to Molecular Phylogenetics” that will take place in Hong Kong in October and another one on “Microbiome & Health” that will take place in December in Montevideo.

In addition, we are currently planning to organise missions of visiting scientists between IP Paris and 5 other IP of the IPIN. The missions will consist in a collaboration involving a visit of several weeks of one scientist from IP Paris and one researcher from one of the 5 selected IP of the IPIN. This exchange will offer the opportunity to both scientists to work concomitantly on a common project that will promote peer learning and coaching in bioinformatics. It will also provide all participants with new skills and a fresh perspective in their area of expertise.

Among the future course options that we are foreseeing for 2019, we are interested in exploring the possibilities to organise a course on the theme of “Long reads”, especially nanopore sequencing. This is based on the request from the IPIN and the rapid development of these new technologies that provides interesting tools adapted to the challenges encountered in various research areas of the IPIN.

Eventually we are currently in discussion for partnering with the European Bioinformatics Institute and possibly the Wellcome Trust Sanger institute, in order to organise larger “Train the Trainer” courses and to capitalise on the opportunity to integrate the wider CABANA network in South America. We are interested in contributing and co-designing a collaborative strategic and successful approach to strengthening bioinformatics capacity in Africa and more specifically to reducing the inequalities in access to bioinformatics capacity across the continent, potentially through the organisation of French-English bilingual courses.
1.8 Perspectives and strategic plan in computational biology

With the increasing scale of Omics data, biology becomes a computational science. In addition to the traditional approach where experiments are designed to test specific hypotheses, it is no longer possible to ignore the reverse approach that allows hypotheses to be generated and verified from large data collections. With this in mind, the C3BI needs to focus on a strong computational component and a broad spectrum of skills. The Institut Pasteur made significant efforts in the 2014-2018 strategic plan to meet this challenge by recruiting 40 high-level bioinformaticians and biostatisticians, as well as several research Units with a strong methodological component. Efforts in strengthening and structuring interactions with the rest of the campus, in scientific outreach, in software development coordination as well as in increasing visibility at the national and international levels will be critical elements for achieving the main objective set for C3BI; in other words, to raise the level of the institute in bioinformatics, biostatistics and in related disciplines.

It is essential to persevere in our efforts in order to establish long-term high-level capacities in these fields at the Institut Pasteur; these will be crucial for the future of biology and medicine. In order to achieve this, we must act on several levels:

- By maintaining the recruitment of methodological Units to ensure sufficient critical mass and broaden the skills spectrum. The priority areas are: algorithmics, machine learning, sequences, genomics and metagenomics, biological systems modelling, and biological networks. Beyond these relatively targeted profiles, there are more general questions about large-scale data processing, which involves cross-cutting and multidisciplinary methodological approaches as well as privileged access to IP-generated data in order to validate these approaches and to promote collaborations.

- By ensuring the continuity between experimental and methodological Units, through the recruitment of young multidisciplinary researchers, typically cross-curriculum. Research engineers with a targeted profile should be recruited for the major cross-cutting projects already in the pipeline (drug resistance, infectious diseases, aging, etc.) and for the new methodological Units. More generally, the recruitment of methodological profiles should be reinforced at all levels, especially among Ph.D. students.

- By training all scientific staff in all C3BI disciplines, from PhD students to PIs. In particular, this includes: analysis of biological data, software usage, data management and sharing, and high-level specialised courses, as seen at CNRS or INSERM thematic schools.

- By strengthening the interactions within the C3BI teams and with the campus Units. The aim is to bring closer together the members of the Hub and the C3BI research teams in order to create the synergies that are essential for the emergence of large, high-impact projects. In addition, we will work toward ensuring balanced interactions between C3BI and other Institut Pasteur research groups. Eventually, we want to ensure that the C3BI responds adequately and effectively to the analytical and methodological needs of the researchers.

- By federating the C3BI teams around a collaborative strategy with the IPIN, through several offers such as diverse and high-quality trainings, a range of services and some collaborative research projects.

- By strengthening our mission of cross-cutting research and scientific outreach within the Pasteur Campus through C3BI seminars, annual conferences and a reinforced presence at major international conferences.

- By developing interactions with computational research institutions, in particular CNRS-INS2I and Inria, as well as fostering privileged partnerships with European institutions specializing in the C3BI topics (EBI-EMBL, CRG, SIB, etc.). In particular, we will be leveraging on the ELIXIR network, and on existing and new collaborative projects.

- By engaging in technology transfer and software development. This will lead to the development of relevant analytical tools and promising industrial applications in public health and biology. The contribution of experimental biologists will be essential in the evaluation and improvement of these tools. Antonio Borderia, who shares his activities with the technology transfer office (DARRI), will help identifying opportunities and liaising with the DARRI on this matter.

- By strengthening links between the C3BI, the CRIT (Centre de Ressources et Innovation Technologiques) and its various components, especially on image analysis or flow cytometry, coordinating methodological and software development efforts to maximize the effectiveness and impact of the tools developed.

- By maintaining our efforts in scientific computing and data management. This must be done through regular consulting between the Information Technology services and the researchers. A competitive research centre such as the Institut Pasteur must (1) be at the cutting edge of available computing technologies and (2) produce results as quickly as possible in a highly competitive context. Beyond computing power, improvements in the areas of data management and sharing will be essential. In fact, future projects of the Institut Pasteur will certainly require those at an exponential rate.
2  RESEARCH UNITS AND LABORATORIES
2.1 Evolutionary Bioinformatics

Olivier Gascuel, DR (CNRS & IP), olivier.gascuel@pasteur.fr, https://research.pasteur.fr/en/team/evolutionary-bioinformatics/

Team members

Research engineers: F. Lemoine, A. Zhukova
Ph.D. students: M. Morel, J. Voznica (starting Sept. 2018)
Post-doc students: M. Davila Felipe, M. Moslonka
Visiting scientists: S. Duchene (AUS), K. Ocaña (BR)
Master interns: T. Dot, L. Blassel

Keywords: Phylogenetics, Phylodynamics, Molecular Epidemiology, HIV, Modeling, Algorithmics

Outlook and highlights

Our Unit started in 2016, with two people (FL, OG). Our work was centred on statistical questions related to phylogenetic inference. We revisited the famous phylogenetic bootstrap of Felsenstein (>35,000 citations) and designed a new version tailored to large data sets, recently published in Nature. In 2017 the Unit grew, and we started tackling more applied questions related to viral evolution and epidemics. We are working on a new maximum-likelihood method to infer ancestral character states in very large trees, with applications to phylogeography and drug resistance mutations of HIV. We develop approximate Bayesian methods (ABC) to estimate the parameters of phylodynamics models of outbreaks (e.g. Ebola), and investigate the impact of non-random sampling (e.g. induced by partner notification) in this context. We collaborate with teams from the UK, South-Africa, Cuba, Brazil and Australia. We are involved in an H2020 project: Virogenesis, on large scale analyses of virus datasets. We played (and still play) a major part in the INCEPTION programme (12 M€ / 9 years) on multidisciplinary approaches to study the emergence of diseases in individuals and populations. We teach phylogenetics and its applications to molecular epidemiology on the Paris campus and in the Institut Pasteur International Network. In 2017, Olivier Gascuel received the “Grand Prix Inria - Académie des Sciences” for numerical sciences.

Publication highlights


Web servers:
- https://ngphylogeny.fr/
- https://pastml.pasteur.fr/

Comparison of Felsenstein’s (FBP) and transfer (TBE) supports on a phylogeny of 9,147 HIV-AIDS strains. (Nature 2018) The nine subtypes (A, B, C, D, F, G, H, J, and K) of HIV-1-M are clearly separated and their branching is consistent with that of previous studies. These nine subtypes have a high transfer (TBE) support, often close to 100%. Conversely, the Felsenstein’s (FBP) support is weaker, in particular for the strongly prevalent subtypes (A, B, C, D, and G). For example, FBP does not “see” the cluster of subtype B while this one is very clearly identified and has a support of 99% with TBE. See text for explanations.
Introduction
The huge amount of molecular data available nowadays can help addressing new and essential questions in evolution. However, reconstructing evolution requires models, algorithms, and statistical and computational methods of ever increasing complexity. Our Unit aims at developing new methods and algorithms that are able to tackle efficiently the ever-increasing amount of sequence data, in the fields of phylogenetics and molecular epidemiology. Most of our phylogenetic methods and software applies to a broad spectrum of questions, species and sequences, from mammals to viruses, DNA to proteins, and genes to complete genomes. This results in a strong impact of some of our papers, e.g. PhyML (2003 & 2010, ~20,000 citations), but also of recent tools, e.g. SMS (Lefort et al. 2017) for model selection with >90 citations. Our second aim is to apply these tools (and others) to pathogens, mostly viruses, and especially HIV. The goals are multiple: understand their evolution (e.g. the emergence and transmission of drug resistance mutations (Zhukova et al. 2017), decipher their genome (e.g. to confirm the existence of the 10th gene of HIV, Cassan et al. 2017), design surveillance tools (e.g. to control outbreaks). Most of the current methods to tackle these questions are based on Bayesian approaches, which are computationally heavy and not able to process the large datasets available nowadays. Developing methods that scale with the “deluge” of data is a real challenge, in terms of algorithmics but also modeling. Our LSD software using quadratic programming for dating molecular phylogenies, is a good example of such an approach. LSD is not only very fast but also fairly accurate (To et al. 2015).

Our Unit started in 2016. We were only two for the first year (FL, OG), concentrating on statistical issues related to phylogenetic inference. In 2017 the Unit grew and the questions being tackled became more diverse, as well as the background of the team members, ranging from probability and statistics (MDF) to epidemiology (MM), with a strong computer science component (AZ, DC) and experienced people in evolutionary biology (SI, TF). All of our projects are collaborative with a mix of these various disciplines. In the following, we present two recent achievements representative of this approach.

Projects and results (two examples)
Renewing Felsenstein’s phylogenetic bootstrap. The phylogenetic bootstrap was proposed by Joseph Felsenstein more than 30 years ago. This method, based on resampling and replications, is used extensively to assess the robustness of phylogenetic inferences. Its usefulness, simplicity and interpretability made it extremely popular in evolutionary studies, to the point that it is generally required for publication of phylogenies. Felsenstein’s article has been cited more than 35,000 times and is ranked in the top 100 of the most cited scientific papers of all time. In 2017, it was cited more than 2,000 times. However, it is commonly acknowledged that Felsenstein’s bootstrap is not appropriate for large datasets containing hundreds or thousands of taxa, which are now common thanks to high-throughput sequencing technologies. While such datasets generally contain a lot of phylogenetic information, the Felsenstein’s bootstrap proportions (FBP) tend to be low, especially when the tree is inferred from a single gene, or only a few genes. The reason for such degradation is explained by the core methodology of Felsenstein’s bootstrap. A bootstrap branch must match exactly a branch in the original tree estimate, to be accounted for in the bootstrap support of that branch. A difference of just one taxon is sufficient for the bootstrap branch to be counted absent, while it is nearly identical to the original branch. The standard approach is to remove “rogue” (phylogenetically unstable) taxa and relaunch the analysis, but this is statistically questionable and computationally expensive. Moreover, with large trees inferred branches are likely to have errors and a large fraction of taxa may be unstable.

We recently published in *Nature* an article (Lemoine et al. 2018) proposing a new version of phylogenetic bootstrap, in which the presence of original branches in bootstrap trees is measured using a gradual “transfer” distance, as opposed to the original version using a binary presence/absence index. This distance is normalized in the [0, 1] range and averaged over all bootstrap trees. We so obtain the “transfer bootstrap expectation” (TBE), which replaces the branch presence frequency of FBP (i.e. the expectation of a 0/1 function), by the expectation of a nearly continuous function. By construction, TBE supports are necessarily higher than FBP’s and the difference is substantial for deep branches. When combined with consistent tree estimation, TBE rarely supports poor branches. Our results with mammal, HIV and simulated data sets, clearly demonstrate its usefulness, especially with deep branches and large trees, where branches known to be essentially correct are supported by TBE but not by FBP. In previous Figure we compare both supports using a large HIV1-M data set. Although the deep branching of the subtypes is poorly supported by FBP, it becomes apparent with TBE. For example, the subtype B clade (3,559 taxa) has a support of only 3% using FBP, but a support of 99% using TBE. This clade contains all subtype B sequences, plus two taxa detected as B recombinant; this means that both supports are likely to be correct insofar as they state that this clade is incorrect (FBP) or nearly correct (TBE). However, FBP fails to detect any phylogenetic signal, whereas TBE reveals that this signal is very strong. Importantly, TBE supports are easily interpreted as fractions of unstable taxa, and the ability of TBE to identify the most unstable taxa (e.g. recombinant HIV sequences) makes it possible to study them further, understand why they are phylogenetically unstable, and revise the branch supports. TBE computation and other phylogenetic tools are available from http://booster.c3bi.pasteur.fr. Currently, we are still working on the subject, to elucidate the mathematical bases of the transfer distance and investigate other branch-support approaches.
Reconstructing ancestral scenarios; applications to molecular epidemiology. The reconstruction of ancestral scenarios is widely used to study the evolution of characters along a phylogenetic tree, for example to infer ancestral molecular characters and their changes in time, or in phylogeography to trace back the geographical locations and moves of species. Standard methods are based on parsimony and likelihood. In the likelihood framework one assumes a probabilistic evolutionary model and commonly uses the marginal posterior probabilities of the character states, and the joint reconstruction of the most likely scenario. Both approaches are somewhat unsatisfactory. Marginal reconstructions provide users with state probabilities, but these are difficult to interpret and use, while joint reconstructions select a unique state for every tree node and thus do not reflect the uncertainty of inferences.

We proposed (Ishikawa 2018) a fast and simple approach, which is in between these two extremes. We use decision-theory concepts and the Brier criterion to associate each node in the tree to a set of likely states. In the tree regions where the uncertainty is low, a unique state is predicted for the nodes. In the uncertain parts, typically around the tree root, several states are predicted. The algorithm has linear time complexity and applies to very large trees. To visualize the results, we cluster the neighbouring nodes associated to the same states and use graph visualization tools. This method is implemented in PastML software, which is freely available from https://pastml.pasteur.fr/.

Our results on simulated data consistently show the accuracy and robustness of the approach. We apply the method to a dataset of more than 3,000 sequences from HIV-1M subtype C sampled worldwide, to study their phylogeography and drug resistance mutations. Results are very convincing: we quickly retrieve and visualize the main transmission routes of HIV-1C; we demonstrate that drug resistance mutations mostly emerge independently under the treatment pressure, but some resistance clusters are found, likely corresponding to transmissions among untreated patients. The clusters found in the past are still present today with higher prevalence. This indicates a high potential for predicting the emergence of problematic, resistant sub-epidemics to be surveyed with special care.

National & International Collaborations

- Analysing HIV/SIV progression in controller vs. non-controller individuals. Bioinformatics analyses, paper in preparation with our collaborators from APHP, Hôpital Cochin, Laboratoire de Virologie, Paris.
- Determining patterns of migration and HIV incidence within rural KwaZulu-Natal. Application of PastML software, paper in preparation with our collaborators from the Africa Health Research Institute (AHRI, ZA), and LSHTM, UK.
- Phylodynamics analysis of the Cuban HIV epidemic. Grant submitted to “Partnership Hubert Curien – Carlos J. Finlay 2019”, with our collaborators from the Tropical Medicine Institute “Pedro Kouri”, Havana, Cuba (project accepted in August 2018, for 2 years).
- Modeling drug resistance emergence and transmission in HIV-1. Analysis of very large datasets from the UK cohort (>40,000 pol sequences) and Pangea Consortium (>10,000 complete genomes). Ongoing project, with UK collaborators.

Teaching and Training Activities

- Introduction to Phylogeny. This course has been taught in 2015, 2016, 2017 on the Parisian campus, in Vietnam (Hanoi 2016, Quy Nhon 2017) and in Morocco (Casablanca 2016). A new edition is planned in Hong-Kong in October 2018.

Software and tools

- http://booster.c3bi.pasteur.fr: Booster web server and free software, implementing the transfer bootstrap
- https://ngphylogeny.fr/: Reimplementation and update of Phylogeny.fr (~3,000 citations; Dereeper 2018)

Grants and awards

- NG-Phylogeny.fr: Institut Français de Bioinformatique, 2-year salary: D. Correia, Pls. S. Cohen (Orsay U.), O. Gascuel
- VIROGENESIS: European H2020 project, ~400 K€, O. Gascuel PI for the CNRS, large scale analyses of virus data
- INCEPTION: 12 M€ / 9 years, Pls O. Gascuel and T. Bourgeron, see Appendix 1.4.

Perspectives and strategy

We develop phylodynamics models and methods to study viral epidemics. This new, trendy domain is in between usual mathematical epidemiology, typically based on compartment models, and phylogenetics. The models used in phylodynamics share the common assumption that new cases are detected randomly. However, a key counter-example is given by partner notification (PN), one of the most widely used control measure against sexually transmitted infections. Our goal is to assess and model the public health impact of PN on HIV epidemics, and study possible observation biases induced by PN. Phylodynamics models are usually complex and writing down the likelihood of the data is often impossible. Thus, we develop approximate Bayesian computation (ABC) methods, which we apply to several data sets and questions (e.g. Ebola).
We will continue investigating statistical issues in phylogenetics (e.g. branch supports, bootstrapping, model selection, tree comparison), develop fast algorithms (e.g. dating with relaxed clock models, phylogenetic placement), and explore new approaches (e.g. deep learning and its application to HIV data, notably to predict resistances to drugs).

**Publications**


2.2 Statistical Genetics

Hugues Aschard; DR, haschard@pasteur.fr, https://research.pasteur.fr/en/team/statistical-genetics/

Team members:
Research engineers: Hanna Julienne; Apolline Gallois
PhD students: Florian Privé
Post-doctoral students: Vincent Laville; Amaury Vaysse

Keywords: Statistics, Genetics, Computational Biology, Epidemiology

Outlook and highlights:
The enormous amount of genetic and genomic data generated in the last decade shows great promise in our ability to understand better human diseases and improving public health. Yet, the genetic architecture of complex human phenotypes remains elusive, and important questions are still unanswered. Our research addresses methodological issues related to the analysis of large multidimensional data in genetics and genomics. It focuses in particular on the development and application of innovative and computationally efficient methods that aim at i) improving association mapping in large genomics datasets where multiple correlated variables are measured across multiple biological levels; ii) allowing for the robust evaluation of causal models that include both genetic, genomic, clinical and environmental data; and iii) identifying and targeting discoveries that have the highest potential clinical utility.

Publication highlights:

Variance components of adjusted variables. (From Aschard et al, 2017 PMID:29038595.) (a–d) Illustrations of the components of the variance of outcome Y before and after adjusting for other variables. The predictor of interest X is displayed in red. In a, the adjusting variables (U1 and U2) are true causal factors that have direct effects on Y; therefore, adjusting Y for U1 and U2 (thus yielding Yadj) decreases the variance of Y. In b, the true factors are not measured, but a variable C influenced by U1 and U2, is measured. Adjusting Y for C decreases the residual variance of Y but also introduces a component of the variance specific to C. In c, the covariate shares factors with Y but is also influenced by X. When the effect of X on C is concordant with the effect of X on Y, a power loss may be induced. In d, Y is not associated with the predictor, and adjusting for C can induce a false-association signal by introducing the effect of X into the residual of Y.
**Introduction:**

My work during the past year and a half aimed at:

1. Building my research group, which included identifying and recruiting two post-doctoral scientists, 2 research engineers, and several interns; setting up hardware and computational resources; getting familiar with all Pasteur administrative processes, in particular those related to scientific projects (e.g. IRB and grant application).

2. Identifying potential collaborators within the Institut Pasteur. Several projects are pending additional discussion and/or the identification of source of funding. As of today, two projects have been effectively initiated. These projects aim at i) exploring causal relationship between DNA, gene expression and methylation in human data (Lluis Quintana-Murci). A manuscript is in its final stage and will be submitted soon. ii) Performing a genome-wide association study in a Listeria sequencing database (Marc Lecuit). Analyses are still ongoing.

3. Applying for funding. I first set up sub-contracts between the Institut Pasteur and both my previous institution and University of Washington. Both are now effective. I also started applying for European (EMBO) and French (ANR) grants. Results for the latter application are pending.

4. Starting projects defined in my research goal as part of my new position at the Institute Pasteur. Works have focused on three projects: i) extending and applying an advanced method for the analysis of multi-phenotype individual-level data, ii) developing and implementing an approach for the combined analysis of GWAS (genome-wide association study) summary statistics data, and iii) developing strategies to estimate the contribution of gene-environment interaction in human to phenotypic variance. I expect the results from these analyses to be submitted before the end of the year. Details of these projects are provided in the next section.

**Results since the creation of the Unit:**

Altogether, finalization of previous work, new studies initiated since the creation of my research group at Pasteur, and collaborative works, have led to the publication of approximately 20 papers between September 2016 and April 2018. Among the various studies we’ve been involved in, three are central to the research group:

a. *The CMS approach*

Large-scale cohort programs have initiated the collection of thousands of phenotypes and omics data across millions of individuals in order to improve the quality and efficacy of health care. One of the primary objectives of these initiatives is to map genetic variants of complex disease phenotypes in the context of large-scale high dimensional data. Future success in the analysis of such data relies on the development of new methodological approaches. We have developed CMS (Covariates for Multi-phenotype Studies), an innovative and computationally efficient approach geared toward the analysis of rich phenotypic data. Our approach keeps the univariate properties of determining association between a single outcome and a single predictor, but as in multivariate approaches, leverages other available variables producing increases in power equivalent to a two or even three folds increase in sample size. The manuscript describing the method has been published in October 2018 in Nature Genetics. Meanwhile, I initiated collaborations with other research groups interested in the application of the method for molecular phenotypes. In particular our group is now working with a team at UCLA (University California Los Angeles) to apply CMS in the Finnish Metabolic Syndrome in Men (METSIM) cohort, performing systematic association screening between 600K common genetic variants and more than 100 metabolites in 10,000 individuals. Preliminary results from this analysis have been presented at the American Society of Human Genetics (ASHG) annual meeting, and a manuscript is under submission.

b. *The JASS approach*

Genome-wide association studies (GWAS) have proven successful in identifying thousands of significant genetic associations for multiple traits and diseases. However, GWAS meta-analyses across different diseases and traits have received limited attention, even though multivariate analysis has the potential to improve the detection of genetic variants. We developed JASS (Joint Analysis of Summary Statistics), a computationally efficient framework for the joint analysis of multiple phenotypes based on GWAS summary statistics. Our framework solves several practical and methodological issues which have been overlooked in previous studies. We have finalized all methodological aspects of the approach during the past months and started a large-scale application of the approach using 45 publicly available GWAS summary statistics. A manuscript describing the application is now under submission. Besides methodological work and application, our group also led, through collaboration with the C3BI Hub, the implementation of a web-interface allowing researcher to perform disease-oriented application of JASS (http://jass.pasteur.fr/index.html). The interface will lead to a second publication, which is also under submission.

c. *Contribution of gene-environment interaction to phenotypic variance*

Several genome-wide Gene-by-Environment (GxE) interaction in human traits and diseases have been published with the aim of identifying single genetic variants interacting with environmental exposure in human traits and diseases. However, little has been done to estimate the overall contribution of GxE interactions to trait heritability. We are developing methods for estimating the contribution of GxE to phenotypic variances of multifactorial diseases, and exploring whether these results will
provide new insight on disease prevalence. Part of this work is done in collaboration with the Gene-Lifestyle Interactions in Cardiovascular Traits from the CHARGE consortium and initiated by colleagues from the University of Washington. This is a long-term project and results from the primary study will likely take one additional year. However, this work already has led to the publication of a method directly related to the topic (Laville et al. 2018).

National & International Collaborations
- Harvard School of Public Health, Boston, USA: Fine mapping analysis in breast cancer data from the BCAC consortium
- Harvard Medical School, Boston, USA: Gene-environment interaction on Glaucoma from the NEIGHBORHOOD consortium
- University of Washington, Saint-Louis, USA: gene-environment analysis in the CHARGE consortium
- University of California Los Angeles, Los Angeles, USA: application of the CMS approach in the METSIM study
- University of California San Francisco, San Francisco, USA: multivariate phenotype analysis in the SICCA study

Teaching and Training Activities
- Guest lecture M2 Ecole Pasteur/Cnam de santé publique,
- Guest lecture M2 Génétique épidémiologique et biomarqueurs at University Paris-Saclay,
- BST47 course on Advanced statistical genetics at the Harvard School of public health

Software and tools
- Joint analysis of GWAS summary statistics
  https://gitlab.pasteur.fr/statistical-genetics/jass
- Estimation of variance explained by a set of SNP for model including interaction effects, from summary statistics
  https://gitlab.pasteur.fr/statistical-genetics/VarExp
- Method for selecting covariates in multiple linear regression
  https://github.com/haschard/CMS

Grants and awards
- Sub-contract with University of Washington for an NIH-R01 grant ($20,000) - ongoing
- Sub-contract with Harvard School of Public Health for an NIH-R03 grant ($120,000) - ongoing
- ANR JCJC (335,000eur) - pending final review of the scientific committee
- Sub-contract with University of Washington for an NIH-R01 grant ($80,000) - submitted

Perspectives and strategy
I focused the past year and a half on building a team and consolidating the network of collaborators. At the same time, the main projects of the group have seen major progress. The primary mid-term objective of the group is to finalize these ongoing projects. However, we have also recently initiated several new projects, including new method development and application of these methods through collaborations. The preliminary stage of these projects makes it difficult to predict accurately the extent of their potential outcome, nevertheless some are promising. Two main axes have been identified and will be central to our future research:

1) We aim first at improving the computational efficiency of several methods we have developed. Indeed, both multivariate approaches and gene-environment interaction methods we recently published perform well in cohort of moderate sample size (e.g. N < 20K) but will be intractable for the analysis of very large-scale datasets. We recently sent a proposal to access the UK Biobank cohort, which includes approximately 500K individuals with complete genotype data (over 90 million genetic variants) and extensive phenotypic and environmental data (>1000 phenotypes measured). This dataset will be used as a benchmark during development, and for large-scale application.

2) Second, we aim at developing and applying new innovative approaches for the analysis of multivariate molecular phenotype data. In particular, we are interested in the identification of genetic variants and environmental factors associated with difference between molecular phenotypes at the network level. Most existing methods have used machine learning approaches. While these approaches have some advantages and offer some flexibility, validation and replication of results is often challenging. Here, we will consider formal statistical frameworks allowing for comparison between covariance matrices and other form of pairwise correlations.
Publications

With C3BI affiliation


Without C3BI affiliation


2.3 InBio: Experimental and Computational Methods for Modeling Cellular Processes

Gregory Batt; DR, gregory.batt@{inria.fr, pasteur.fr}


Team members: (As of July 14, 2018)

**Permanent researchers:** Gregory Batt, Jakob Ruess

**Research engineers:** Andjela Davidovic, François Bertaux, Steven Fletcher

**PhD students:** Chetan Aditya, Sebastian Sosa, Virgile Andréani

**Post-doctoral students:** Matthieu Pichené

**Keywords:** Quantitative Systems And Synthetic Biology; Single Cell And Cell Population Dynamics; Stochastic Models; Optimal Experimental Design; Active Learning And Optimization

**Outlook and highlights:**

InBio is a Unit of the C3BI, jointly created by Institut Pasteur and Inria, a French national research institution focusing on computer science and applied mathematics. Within InBio, we develop experimental and computational methods to better perturb, observe, characterize, and ultimately understand and control, biological systems. Our main objectives are the development of iterative approaches for the identification of predictive models of cellular processes, and their use for optimization and control. We therefore combine wet and dry biology. Our core expertise is the analysis and control of dynamical systems in quantitative systems and synthetic biology.

The Unit has been created in February 2017 and the wet lab opened in June 2017. Since that time, we built the team, hiring people of complementary expertise. In particular, two permanent staff joined the group: Jakob Ruess, an Inria research scientist hired in October 2016, and Francois Bertaux, a Pasteur engineer, hired in April 2018. Since 2016, we published in PLoS computational biology, Bioinformatics, The Journal of Chemical Physics, Interface, and Nature Communications.

**Publications highlights:**


**Balancing a genetic toggle switch by real-time feedback control.** Using an *in silico* feedback control loop, we demonstrate that a bistable genetic toggle switch can be dynamically maintained near its unstable equilibrium position for extended periods of time. The environment of the cell is modified in real-time according to the control strategy so that a genetic toggle switch implemented in the cell remains in a balanced state.

**Balancing a genetic toggle switch by real-time feedback control.** Using an *in silico* feedback control loop, we demonstrate that a bistable genetic toggle switch can be dynamically maintained near its unstable equilibrium position for extended periods of time. The environment of the cell is modified in real-time according to the control strategy so that a genetic toggle switch implemented in the cell remains in a balanced state. The environment of the cell is modified in real-time according to the control strategy so that a genetic toggle switch implemented in the cell remains in a balanced state. The environment of the cell is modified in real-time according to the control strategy so that a genetic toggle switch implemented in the cell remains in a balanced state.
Introduction

InBio is an interdisciplinary research group, performing wet and dry biology in the same lab. We combine systems and synthetic biology approaches with control and active learning methods, and stochastic and statistical modeling frameworks.

Our main long-term goal is to develop a comprehensive methodological framework supporting the development of a quantitative understanding of cellular processes. Given a process of interest and current knowledge on the system, the problem is to decide iteratively which strain to construct and which experiment to run to characterize the process in an optimal manner. More generally, we are interested in understanding, controlling and optimizing cellular processes from the single cell to the cell population levels.

On the methodological side, we employ single cell models to represent the biological processes and develop methods for model reduction, sensitivity analysis, inference of model parameters, experimental design, and control, based on techniques such as global optimization, stochastic simulation, and moment closure. In addition, we develop software to support these novel methods, video microscopy image analysis and microscopy automation.

Results since the creation of the Unit

In this section, we will on the one hand report on recent scientific activities, notably dealing with starting novel research projects and directions in the context of the new group, and on the other hand report on results published by group members during the period July 2016 - July 2018.

Two important additions to the team have been made since December 2015, when the decision to create InBio was taken. The first is the recruitment of a mathematician, Jakob Ruess, as an Inria research scientist (chargé de recherche) in October 2016. The second is the recruitment of FranÇois Bertaux as a Pasteur engineer (permanent staff, ingénieur de recherche confirmé), in charge of experimental platform development. Virgile Andréani joined the group as a PhD student in Oct 2017, and since we moved from Palaiseau to the Parisian Pasteur campus, we hired 4 PhD student, two started in Feb 2018, Sebastian Sosa Carrillo and Chetan Aditya, and two starting in October 2018, Elise Weill and Arthur Carcano, and one postdoctoral researcher, Matthieu Pichené. Additionally, Andjela Davidovic, an engineer from the C3BI Hub, joined our group a few months ago and might hopefully continue her work with us for a year on a part time affiliation (40%).

A significant part of our time for this period was dedicated to building the lab (works ended in April 2017), equipping it and effectively starting wet lab activities and the construction of experimental platforms for real-time experiments. Our arrival on the campus was the opportunity to interact with many scientists at Institut Pasteur. One can notably mention A. Jacquier and C. Saveanu, S. van Teeffelen, F. Schweisguth, S. Shorte and N. Aulner, S. Novault, M.-A. Dillies and C. Malabat, S. Cauchemez, C. Zimmer, B. Arcangioli, D. Bikard, S. Goba, A. Imbert, R. Levayer, J.-B. Masson, F. Tangy, and Y. Jakob. A scientific collaboration has started with Philippe Glaser in the context of the Ph.D. work of Virgile Andréani, and additional collaborations with Jacques Bellalou and Thomas Gregor might be initiated in the context of the works of Sebastian Sosa Carrillo and Andjela Davidovic, respectively. We also planned the research that will be done in the context of 4 projects, all starting in 2017: a European Fet-Open project, Cosy-Bio, two ANR projects, Memip and Cogex, and an Inria project, IPL Cosy. On the theory side, we made progress on method developments for parameter inference using single cell data, real-time control of heterogeneous cell populations, and optimal experimental design. Lastly, reporting was a non-negligible effort since 7 activity reports were asked by our two institutions in the last 12 months.

In the last two years, we have published 9 papers. These publications are listed below. In Ruess et al (2017), we considered stochastic reaction networks in which the parameters are randomly distributed over the population and propose a new sensitivity index that captures the robustness of system outputs upon changes in the characteristics of the parameter distribution. This notion provides a more complete picture of a system’s robustness than the standard ones since one does not assume that all cells in the population have the same parameters and are perturbed in the same way. A Monte Carlo estimator of this sensitivity index is constructed by using the Girsanov’s likelihood ratio method in combination with estimation algorithms that make use of a marginalization of the path distribution of stochastic reaction networks and lead to Rao-Blackwellized estimators with reduced variance. In Llamosi et al. (2016), we proposed a quantitative modeling framework that attributes specific parameter values to single cells for a standard model of gene expression. In the presence of significant cell-to-cell heterogeneity, parameters of models of intracellular processes should indeed be fitted to individual cells, yielding a population of models of non-identical individuals. We combined single-cell measurements of the response of yeast cells to repeated hyperosmotic shocks and state-of-the-art statistical inference approaches for mixed-effects models to infer multidimensional parameter distributions describing the population. We then derived specific parameters for individual cells. Their analysis showed that single-cell identity was, at least partially, captured by the parameter values of the gene expression model. For this research, we needed high quality single cell traces extracted from video microscopy movies. To do so, we developed CellStar, an image analysis tool for brightfield yeast images (Versari et al., 2017). The two above-mentioned articles deal with phenotypic heterogeneity. Either to quantify it in models or to measure the impact of its change on system’s performance. Another
direction of the group is to develop and demonstrate methods for real-time control of intracellular processes. In Lugagne et al. (2017), we investigated the control of multistable gene regulatory networks, which are ubiquitously found in nature and play critical roles in cell differentiation and decision-making. Using an in silico feedback control loop, we demonstrated that a bistable genetic toggle switch can be dynamically maintained near its unstable equilibrium position for extended periods of time. Importantly, we showed that a direct method based on dual periodic forcing is sufficient to simultaneously maintain many cells in this undecided state. In Chait et al. (2017), we demonstrated simultaneous control of many individual cells over days. Our automated and programmable platform broadly enables experiments that bridge individual and population behaviours. We demonstrated: (i) population structuring by independent closed-loop control of gene expression in many individual cells, (ii) cell–cell variation control during antibiotic perturbation, (iii) hybrid bio-digital circuits in single cells, and freely specifiable digital communication between individual bacteria. These two recent papers, in addition to a previous one (Uhlendorf et al, PNAS 2012), placed us among the leaders of a nascent field called cybergenetics, that could be defined as the confluence of modern genetic manipulation techniques, powerful measurement technologies, and advanced analysis methods, enabling a new area of research in which systems, communications, and control theory notions are used for synthetically regulating cellular processes at the gene level (M. Khammash). Our visibility in this field can be appreciated through invitations to join European projects (Cosy-Bio) or to participate to selective workshops (Control of Cellular and Molecular Systems at MBI, Ohio; Design-Build-Test in Synthetic Biology: Closing the Loop at Univ. of Edinburgh; Cybergenetics invited session at CCTA Copenhagen; Cybergenetics – at the interface between living and non-living regulatory systems at Symposium of German Association for Microbiology (VAAM)).

National & International Collaborations
At the national and international levels, we collaborate with Eugenio Cinquemani (Inria Grenoble), Dirk Drasdo (Inria Paris), Calin Guet (IST Austria), Pascal Hersen (MSC lab, CNRS and Paris Diderot), Gasper Tkacik (IST Austria), Lingchong You (Duke University), and Christoph Zechn (Max-Planck Institute for Molecular Cell Biology & Genetics).

Teaching and Training Activities
- Computational biology, Master Approches Interdisciplinaires du Vivant (AIV - Center for Research and Interdisciplinary), Paris-Diderot and Paris-Descartes Universities, G. Batt (coordinator and teacher 48h) and J. Ruess (teacher 24h)
- Modeling and engineering of biological systems, Institut de Technologie et d’Innovation de Paris Sciences et Lettres (PSL-ITI), G. Batt (co-coordinator 80h and teacher 8h) and J. Ruess (8h)

Software and tools
- MicroMator: automating microscopy experiments (under development). S Fletcher (InBio)

Technology transfer
It is often said that model-based approaches are beneficial to guide the engineering of biological systems. Our main objective is to develop methods to streamline this process. Although our objective is clearly to make a proof of concept, the methods that we will develop could potentially lead to intellectual property and technology transfer. We also patented novel recombinases in the context of past work with the Weiss lab at MIT (US Patent 15/410,875).

Grants and awards
- Control engineering of biological systems for reliable synthetic biology applications (Cosy-Bio, 3M€), H2020 FET-Open (2017-2020), with D. di Bernardo (coord, Tigem), F. Menolascina (Edinburgh U), M. di Bernardo (Naples U), P. Hersen (Paris Diderot U), M. Khammash (ETHZ), G.-B. Stan (Imperial College), and L. Marucci (Bristol U).
- Real-time control of synthetic microbial communities (IPL Cosy, 700k€), Inria Project Lab (2017-2020, ~700k€), with E. Cinquemani (coord., Inria), H. Geiselmann (LiPhy, CNRS/Grenoble Univ.), J.-L. Gouzé (Inria), J.-P. Richard (Inria), F. Bonnans (Inria), B. Laroche (INRA Jouy-en-Josas), and H. Youk (Delft Univ.).
- Mixed-effect models for cellular processes: methods, tools and applications (Memip, 500k€), ANR (2017-2020), G. Batt coordinator, with P. Hersen (CNRS/Paris Diderot), E. Cinquemani (Inria) and M. Lavieille (Inria/Polytechnique)
- Computer-aided control of gene expression (Cogex, 440k€), ANR (2017-2019), with P. Hersen (Coord., CNRS/Paris Diderot) and G. Truan (LISBP, CNRS/INSA)
- Stochastic models: Scalable model checking (StochMC, 360k€), ANR (2014-2018), with B. Genest (Coord., Inria), W. Zielonka (LIASA), and H. Gimbert (LaBRI)
Perspectives and strategy

• Develop technological platforms, mathematical methods and biological systems to reach the ambitious objectives set in the multiple starting research projects in the 3 - 4 years to come
• Consolidate our position as central actors on cybergenetics at the European and world levels.
• Develop collaborations with other Pasteur teams, notably on antibiotic treatments and therapeutic protein production
• Find smooth operation modes between the two institutions (few administrative hurdles remain)

Publications

JB Lugagne, SS Carrillo, M Kirch, A Köhler, G Batt*, P Hersen* (2017), Balancing a genetic toggle switch by real-time feedback control and periodic forcing, Nature Communications, 8 (1), 1671

R Chait*, J Ruess*, T Bergmiller, G Tkačik, CC Guet (2017), Shaping bacterial population behavior through computer-interfaced control of individual cells, Nature Communications, 8 (1), 1535


2.4 Systems Biology

Benno Schwikowski; DR, benno@pasteur.fr, http://www.systemsbiology.fr

Team members:

**Permanent researchers:** Benno. Schwikowski

**PhD students:** Pierre Bost, Irina Nikolayeva (until February 2018)

**Interns (alumni since 2015):** Urszula Czerwinska, Sharad Goulam Abas, Floriane Garnier, Pierre Bost, Aurélie Nerin.

**Keywords:** Systems biology, systems medicine, Data integration, Statistics, Network biology

**Outlook and highlights:**

We develop and validate computational methods for the integrative data-driven study of complex diseases. Most of our projects are collaborations with biological and clinical researchers, and technology specialists.

Since 2016, we have developed a novel underlying machine learning model that allowed us to discover the first biomarker for severe dengue that could be replicated in independent datasets, across blood types and measurement technologies (Nikolayeva et al., 2018). Furthermore, we developed and validated an automated method that allowed immunologists from Institut Pasteur working on the LabEx project Milieu Intérieur to automatically analyse flow cytometry experiments on a very large scale (Chen et al., 2015). In addition, we designed a large-scale pipeline for the identification of small peptides from DNA sequences (Friedman et al., 2017). LEAN, the first truly efficient method to identify 'hot spots' in networks was also developed in our team (Gwinner et al., 2017). Eventually, the Cytoscape-GRAVITY platform, developed and validated within our team, provides tools to visualize and analyse complex disease data in the context of autism and cancer (http://gravity.pasteur.fr). It is noteworthy that our Cytoscape platform is now cited more than four times per day (Web of Science).

**Publications highlights:**


Introduction

Our group develops, applies, and validates computational modelling approaches to decipher complex diseases in the context of iterated cycles of data collection/experimentation and analysis. We embed the development of novel computational methods into collaborations around specific biological and clinical research questions. Immunity is a core theme, and diseases of interest include dengue infection, Sjögren’s syndrome, cancer, and autism. We strive to obtain robust and interpretable results by (i) directing our models towards the discovery of relevant associations of ‘omics data with biological/clinical phenotypes, (ii) exploiting a maximum of different types of relevant data, (iii) developing machine learning models whose structure is carefully calibrated to the disease and available data. We have particular expertise and experience with molecular network models.

Results since de creation of the Group (2016)

Theoretical foundations, method and tool development

Finding subnetwork ‘hot spots’ is a computationally hard problem, for which only heuristic approaches exist. We uncovered an issue in the most common statistical formulation of this problem and outlined potential solutions and best practices (Nikolayeva et al., 2017).

Detecting genetic interactions without a potentially lossy pre-filtering step has been computationally infeasible so far. Developing a highly efficient, filter-free implementation allowed us, in collaboration with the team of Anavaj Sakuntabhai, to analyse and detect novel genetic interactions in Psoriasis across around 5,300 genomes.

We assisted the same team in the parallelization of a novel method to detect a class of genetic associations (multi-way epistasis) in families (Loucoubar et al., 2016).

As of this writing, the seminal paper for our Cytoscape network analysis and visualization platform has been cited more than 6x per day (Shannon et al., 2003). We continue to implicate ourselves in the organization of a global community around Cytoscape.

We contributed to formulating a global vision on the future of ‘exposome’ research (Escher et al., 2016).

Method development and application to complex disease

Severe dengue infection accounts for high global health burden, but it remains difficult to detect in the early phase of the disease. Until now, data-driven approaches have not yielded reproduced biomarkers. In collaboration with the group of Anavaj Sakuntabhai, we developed a novel (and general) nonlinear machine learning approach based on monotonic functions. This allowed us to identify the first biomarker for severe dengue to replicate in independent datasets, across blood types and measurement technologies (Nikolayeva et al., 2018).

The analysis of flow cytometry data requires domain-specific knowledge, and remains tedious, preventing the systematic evaluation of large datasets from large cohorts. To help with the analysis of flow cytometry experiments across the 1000-donor, 8-panel data from the Labex Milieu Intérieur project, we designed, implemented, and validated a largely automatic analysis approach (Chen et al., 2015).

In collaboration with the lab of Jost Enninga, we studied Shigella infection of host cells at the single-cell level. Using adapted statistics and network analysis tools, we unveiled a new bacterial strategy to subvert the host response (Lippman et al., 2015).

Small peptides are hard to discover in both sequence data and experimentally. In a collaboration with the C3BI Hub and mass spectrometrists, we designed and validated a novel method to detect small peptides across many prokaryotes on a large scale. Our finding is that small peptides are surprisingly frequent (Friedman et al., 2017).

Following up on the insights from our theoretical study on subnetwork identification (Nikolayeva et al., 2017), we developed the first efficient, parameter-free computational statistical model that can be used to analyse any genome-wide set of gene p-values in a network context. Applying this in a collaboration with medical geneticist Elisabeth Tournier-Lasserve, we detected, and validated, in a mouse model, a novel physiological effect occurring in the Cerebral Cavernous Malformation disease that could not be detected by gene-by-gene, or pathway analysis. (Gwinner et al., 2017).

National & International Collaborations

• Olivia Doppelt-Azroual, Anna Zhukova (IP, C3BI Hub)
• Anavaj Sakuntabhai (IP, Functional Genetics of Infectious Disease)
• Darragh Duffy (IP, Functional Genetics of Infectious Disease)
• Rachel Golub (IP, Lymphopoiesis group)
• Jacques-Eric Gottenberg (CHRU Strasbourg, Natl. Centre for systemic autoimmune diseases)
• Ido Amit (Weizmann Institute Israel, Immuno-genomics lab)
• Sampsa Hautaniemi (U. Helsinki, Systems Biology of Drug Resistance in Cancer group/assoc. clinicians)
Teaching and Training Activities

- Mini-training on single-cell data analysis (IP, multi-week training course for two interns)
- Technology updates to the Hub single-cell working group
- Lectures on Systems Biology (Paris VII/Institut Curie)
- Exosome workshop series (Imperial College London April/June 2018)
- IP: during 2015–2018, 3 interns/3 PhD students/3 postdocs trained

Software and tools

- Prediction server for small peptides (http://disco-bac.web.pasteur.fr)
- LEANR implementation LEAN of genomic analysis (https://tinyurl.com/leanrpage)
- Cytoscape platform (http://cytoscape.org; currently >6 citations/day, Google Scholar)

Technology transfer

- Invention disclosure for dengue biomarkers (with Anavaj Sakuntabhai)

Grants and awards

- HERCULES H2020 Systems Medicine/ovarian cancer (~2020, tbn)
- NECESSITY: EU-IMI2: Biomarker discovery/Sjögren’s syndrome (2019–, tbn)
- ENS PhD scholarship (Pierre Bost MSc)

Perspectives and strategy

Just like more established areas of bioinformatics/computational biology (e.g., sequence alignment), in ‘systems’ or ‘integrative’ biology, the interface between biology, technology, and computation evolves rapidly. New technologies (e.g., single cell technology) generate qualitatively new data on different biological levels to address new questions for which biological concepts are, in part missing and technical data characteristics (noise etc.) are unknown. In the context of medical questions, clinical data of heterogeneous quality represents a particularly interesting new layer of complexity and challenge.

Developing computational models to deal with this data thus requires intense collaboration with technologists and biologists towards a joint project vision, driven by high levels of investment and trust, and the willingness to require interdisciplinary expertise.

Our new collaboration with technology pioneer Ido Amit (since last fall co-advisor of our joint PhD student Pierre Bost, first manuscript submitted) lets us co-develop and apply leading-edge single-cell methodology, which we then bring to the campus, for example, through our new collaboration with IP immunologist Rachel Golub (work in progress with Rachel’s students Maxime Petit and Florian Specque).

In the context of our work on Sjögren’s disease, we have been applying extremely sensitive technology to measure interferons, developed by Darragh Duffy (IP/Milieu Intérieur) to validate our data-driven stratification of patients in the French ASSESS Sjögren’s cohort (manuscript in preparation).

A technically new component for our research is the prediction of drug effects in two newly funded projects (HERCULES and NECESSITY projects on cancer and Sjögren’s syndrome, 2 postdocs to be hired). Here, we expect to apply and develop models that take (potentially patient-specific) gene networks into account.

Publications:

(* = shared first/last authorship)


The eight following research Units have a secondary affiliation to the C3BI, and will not be evaluated in October 2018. Their activities in bioinformatics and biostatistics are quickly described in this document, as they actively participate to the C3BI.
2.5 Human Genetics and Cognitive Functions

Thomas Bourgeron, Prof. DR (University Paris Diderot), thomasb@pasteur.fr, https://research.pasteur.fr/en/team/human-genetics-and-cognitive-functions/

Team members involved in the C3BI

<table>
<thead>
<tr>
<th>Category</th>
<th>Team Members</th>
</tr>
</thead>
<tbody>
<tr>
<td>Permanent researchers</td>
<td>G. Dumas, E. Ey, C. Leblond, T. Rolland, R. Toro</td>
</tr>
<tr>
<td>PhD students</td>
<td>C. Carton</td>
</tr>
<tr>
<td>Research engineers</td>
<td>F. Cliquet, A. Mathieu</td>
</tr>
</tbody>
</table>

Keywords: Autism Human Genetics Psychiatry Synapse Neuroscience

Outlook and highlights:

Our laboratory gathers psychiatrists, neuroscientists and geneticists to understand the causes of autism. We previously identified the first mutations in the NLGN-NRXN-SHANK pathway associated with autism. This pathway is known for playing a role in synapse formation and in the balance of excitation and inhibition within the brain. We are leading the genetic work package of EU-AIMS/AIMS2-TRIALS, the largest European project on the research on autism and the French genetic project on autism. Both projects are focused on deep genotyping and deep-phenotyping of individuals with autism and controls including whole genome sequence, brain imaging (EEG and MRI) and a battery of cognitive tests. Our group is also developing new methods for analysing whole genome and brain imaging data as well as new paradigms for characterizing mouse social and vocal behaviours. Thomas Bourgeron is co-leading with Olivier Gascuel the INCEPTION project (12 M€ / 9 years) that applies integrative biology to understand the emergence of diseases in populations and individuals. He is a member of the French Academy of Science and a recipient of the chair of excellence of the Bettencourt-Schueller Foundation. He received the IPSEN prize for Neuronal Plasticity in 2015.

Publications highlights:


The genes related to autism: This figure illustrates the findings for 203 genes with de novo mutations identified in individuals with ASD. A preliminary functional annotation clustering analysis using GeneMANIA (http://www.genemania.org) and DAVID (http://david.abcc.ncifcrf.gov) indicates that 36% of the proteins have at least one interaction (direct or indirect) with another protein, 61% are expressed in the brain, and 14% are known to be involved in synaptic function.
2.6 Mathematical Modelling of Infectious Diseases

*Simon Cauchemez, DR; simon.cauchemez@pasteur.fr,*
*https://research.pasteur.fr/fr/team/mathematical-modelling-of-infectious-diseases/

**Team members**

**Permanent researchers:** Henrik Salje

**Research engineers:** Alessio Andronico

**PhD students:** Catherine Eisenhauer, Noémie Courtejoie

**Post-doctoral scientists:** Birgit Nikolay, Juliette Paireau, Quirine ten Bosch, Nathanaël Hoze, Anthony Cousien

**Keywords:** Mathematical Modelling, Statistics, Epidemic Dynamics, Transmission, Outbreak Control

**Outlook and highlights**

The main objective of the Unit is to develop state-of-the-art statistical and mathematical methods to address the many challenges epidemiologists are confronted with during epidemics, with the aim to increase the understanding of how pathogens spread in populations as well as the impact of interventions, to support policy making and optimize control strategies. Our approach is therefore highly multidisciplinary, looking at infectious diseases through multiple perspectives (statistics, modelling, epidemiology, surveillance, Public Health, policy making, microbiology), multiple scales and multiple data streams. Since the creation of the Unit in 2013, we have been involved in response activities to a number of major outbreaks such as Ebola in West Africa (analyses of chains of transmission), Zika in the Americas (first estimates of the risk of microcephaly associated with Zika, real-time modelling to support health care planning), MERS-CoV in the Middle East (understanding the complex transmission dynamics).

**Publications highlights**


**Transmission trees of Ebola virus:** The date of symptom onset of cases having infected more than 3 cases is indicated in each circle. The size of the circles is proportional to the number of people infected by each case. © Institut Pasteur
2.7 Imaging and Modeling

Christophe ZIMMER, DR; czimmer@pasteur.fr, https://research.pasteur.fr/en/team/imaging-and-modeling/

Team members

Permanent researchers: F. Mueller  
Research engineers: M. Lelek, C. Weber, B. Lelandais  
PhD students: W. Ouyang, M. Woringer  

Keywords: Advanced Imaging, Chromatin Architecture, Host-Pathogen Interactions, Computational Modeling, Deep Learning

Outlook and highlights:
The Imaging and Modeling Unit is an interdisciplinary team created in 2013 that combines expertise in physics, applied mathematics, optics and cell biology. The lab develops computational and experimental approaches to characterize and quantitatively predict selected cellular processes. Current projects concentrate on (i) developing advanced imaging methods, especially single molecule based 3D super-resolution microscopy, by a combination of optical instrumentation and tailored algorithms, (ii) investigating the spatial architecture of the genome and its functional implications in yeast and human cells using live cell microscopy or high resolution imaging and molecular dynamics simulations of polymers, (iii) quantifying gene expression in single cells using RNA-FISH and dedicated image analysis, (iv) using high-resolution imaging to study host-pathogen interactions, with an emphasis on the early HIV replication cycle, (v) using deep learning methods to analyse complex images from cellular microscopy and medical imaging.

Publications highlights:


Image and modelling approaches developed in the lab: a-c) 3D super-resolution microscopy: a fluorescence microscopy system featuring a deformable lens (a) is used to image single photoswitching molecules (b). Computational analysis using a dedicated algorithm (ZOLA-3D) allows reconstruction of 3D super-resolution images of nuclear pores in a whole HeLa cell (c, colour indicates depth). d) Imaging of a single chromosome in a human cell with standard widefield microscopy (left) and ZOLA-3D (right). e) A molecular dynamics simulation of chromosomes in an entire human nucleus. Predictions of this polymer model are compared to Hi-C data and imaging experiments to understand the mechanisms of 3D chromatin organization.
2.8 Structural Bioinformatics

Michael Nilges, DR; michael.nilges@pasteur.fr, https://research.pasteur.fr/en/team/structural-bioinformatics/

Team members

Permanent researchers: B. Bardiaux, G. Bouvier, A. Blondel, P.L. Chau, F. Cordier, N. Izadi-Pruneyre, T. Malliavin, R. Pellarin

Research engineers: N. Duclet-Savatier, T. Huynh

PhD students: I. Pitard, D. Monet, L. Ortega Varga

Post-doctoral students: F. Allain, B. Worley

Keywords: Integrative Structural Biology, Computational Biophysics, Molecular dynamics, Protein-ligand interactions, Computational Drug Discovery

Outlook and highlights

Our Unit (created in 2001) uses computational and experimental methods (NMR, EM) to understand the three-dimensional structures of biological macromolecules and their dynamics, and to interfere with their dynamics or assembly by computational drug discovery. One research focus is the development of computational methods for data integration for integrative structural biology methods. We introduced the first fully Bayesian modeling method for NMR (Inferential Structure Determination, Rieping et al., Science 2005), which we extend to other experimental data (cross-linking mass spectrometry, EM). Key examples of systems studied are bacterial secretion systems (type 2, 4 and 7). M. Nilges coordinates the "Investissement d’Avenir" projects bip:bip (2010-2017) and CACSICE (2011-2019), which aim at developing computational methods and to provide experimental infrastructure for integrative structural biology (e.g., the Titan Krios EM at Institut Pasteur), respectively.

Publications highlights

2.9 Spatial Regulation of genomes

Koszul Romain, DR (CNRS & IP); romain.koszul@pasteur.fr

Team members
Permanent researchers: M. Marbouty, A. Piazza
Research engineers: A. Thierry, A. Cournac
PhD students: T. Foutel-Rodier, B. Conin, L. Baudry
Post-doctoral scientists: V. Scolari, R. Montagne, C. Cockram

Keywords: Genome organization, chromatin, segregation, metagenomics, synthetic chromosome

Outlook and highlights:
The team Spatial Regulation of Genomes (RSG), in the Department of Genomes and Genetics is focused on the field of chromosome organization in eukaryotic and prokaryotic microorganisms. The team was created in 2011 as a junior group and was transformed into a full Unit in 2017 (with the support of ERC starting and consolidator grants). For instance, the team has recently described the dramatic reorganization of the entire budding yeast genome over the mitotic and meiotic cell cycles, focusing on the roles of cohesin and condensins in the folding of the genomes at these stages. While studying the biology of genome folding with a combination of experimental, synthetic and computational approaches, the laboratory has also developed techniques aiming at improving genomic/metagenomic analyses, which have opened up novel areas of research, such as bridging viral and episomal molecules to bacterial chromosomes in complex microbiomes. The group encompasses a wide range of expertise, including physicists, computer scientists and biologists working together on interdisciplinary studies.

Publications highlights:

Two recent works on the genome folding of bacteria E. coli and yeast S. cerevisiae. A) We revealed the multiscale, intertwined organization of the model bacteria E. coli and unveiled new roles for nucleoid associated proteins in the maintenance of the higher-order architecture. B) We provided direct evidence for the existence of arrays of stable chromatin loops along budding yeast chromosomes during meiotic prophase. We used synthetic chromosomes to track homolog pairing and meiotic recombination.
2.10 Microbial Evolutionary Genomics

Eduardo PC ROCHA, DR (CNRS), erocha@pasteur.fr
https://research.pasteur.fr/en/team/microbial-evolutionary-genomics/

Team members
Researchers: E. Rocha, M. Touchon, O. Rendueles (CNRS)
Research engineers: A. Perrin
PhD students: C. d’Humières, R. Denise
Master interns: M. Haudiquet
Visiting scientists: V. de Crecy (U Florida)

Keywords: Comparative genomics, Bioinformatics, Evolution, Population genetics.

Outlook and highlights
The Unit is composed of researchers from different backgrounds including pharmacists, computing scientists, and biologists. It has a large experience on the use of comparative genomics and population genetics approaches to study microbial evolution. In the last few years the lab has shown how interactions between bacteria and eukaryotes affect the evolution of antagonistic and mutualistic interactions, and how they can evolve fast by hyper-mutagenesis during experimental evolution. The lab has also demonstrated the impact of mobile genetic elements, both conjugative elements and phages, in the evolution of bacterial genomes, in the creation of novel functions, and in shaping bacterial defence systems. Recently, we have developed an individual-based model to study the interactions between bacteria and phages and their consequences for community dynamics and bacterial evolution.

The laboratory produces and distributes open source software, often in collaboration with the C3BI Hub. We collaborate with teams from France, UK, Spain, Denmark, and USA. Since 2016 we have been involved in many projects including five ANRs and an ERC (ended December 2017), and published 39 scientific articles. We played an important role in setting up the C3BI and the INCEPTION programme.

Publication highlights

First study showing clear parallels between experimental and natural evolution of symbioses. [Clerissi, Nat Comm, 2018]
2.11 Human Evolutionary Genetics

Lluis QUINTANA-MURCI, DR (CNRS & Professor IP); quintana@pasteur.fr
https://research.pasteur.fr/en/team/human-evolutionary-genetics/

Team members

Research scientists: E. Patin, G. Laval, M. Rotival

Research engineers: H. Quach

Technician: C. Harmant

Post-doc scientists: M. O’Neill, L. Rubio, J. Mendoza

Master interns: S. Cuadros

PhD students: M. Lopez, J. Choin, M. Silvert, L. Husquin

Keywords: Population genetics, genomics, demography, selection, immunity, gene expression, epigenetics

Outlook and highlights

Our Unit, which was created in 2007, is focused on understanding how natural selection, human demography and lifestyle have shaped the patterns of diversity of the human genome, to understand how this may impact phenotype variation and disease. Our current projects are focused on (i) the genetic architecture of human populations, migrations patterns and admixture events; (ii) the occurrence of positive selection in the human genome and the relationship between population demography and the burden of deleterious alleles; and (iii) the relationship between genetic diversity, epigenetic patterns (in particular DNA methylation) and changes in lifestyle and habitat of human populations. Our research is also focused on the study of the genetic and non-genetic factors driving variation in human immune responses, as this helps to lay the foundations of precision medicine related to infectious and immune-related disorders. To do so, our laboratory combines population genetics and cellular genomic approaches, with computational modelling and development of new statistical frameworks, often working closely to theoretical population geneticists, immunologists, epidemiological geneticists as well as anthropologists. Lluis Quintana-Murci coordinates the Labex Milieu Intérieur (http://www.milieuinterieur.fr/), which aims at dissecting the interplay between genetics and environment and their impact on the human immune system.

Publication highlights

2.12 Decision and Bayesian computation

Jean-Baptiste Masson, CR (IP), jbmasson@pasteur.fr, https://goo.gl/aQotqj

Team members

Research engineers: F. Laurent (permanent), S. Doureligne, M. El Beheiry (50% Curie)
PhD students: C. Guerinot (with PM Lledo)
Alumni: S. Clais

Keywords: Bayesian Inference, Random Walk, Behaviour, Bayesian Induction, Connectome, Virtual reality, “Human in the loop” algorithms

Outlook and highlights

Our Unit started in February 2017. The lab joins theoretical physicists and applied mathematicians to explore decision-making at various scales in problems where physics, statistics and Bayesian Inference have to be combined to both understand the biological dynamics and to infer the models from experimental data. The lab tackles 3 subjects. 1) We combine Random Walk theory, multiple models of physical environments of the cell and Bayesian inference to extract information from single biomolecule dynamics. We have/are developing a software platform TRamWAY that performs full Bayesian approach to biomolecule random walks from single molecule images to model selection. Recent advances have allowed us to map the evolution of synaptic environment with phosphorylation of glycine receptors and measure the spatio-temporal evolution of Gag during the HIV-1 Virion assembly. 2) We are developing hybrid approaches to understand the link between behaviour and neural architecture in the Drosophila larva. We leverage genetic advances to optogenetically activate single neuron cell lines (>1500 lines with 5 000 000 larvae) to screen for relations between single neuron activity and behaviour. We combine semi-supervised learning to Bayesian induction to both create neural-relevant behaviour dictionaries and study their dynamics. 3) Our last project is motivated by creating a new approach to 3D image analysis by joining human cognition, data exploration and small data learning. We have designed a software platform, DIVA, that automatically includes any data in Virtual reality environment and allows both interaction with the data and data treatment. We are using this technology for both pre-clinical and clinical applications. Finally, the DIVA platform has been deployed in the image analysis hub of the Institut Pasteur and in the electron microscopy platform of the Curie Institute. The start-up associated to the project will be officially open in September 2018 with Mohamed El Beheiry as CEO.

Tools and Software

• TRamWAY https://github.com/DecBayComp/TRamWAY
• Escale: https://github.com/DecBayComp/escale

Publications


3 BIOINFORMATICS AND BIOSTATISTICS HUB
3.1 Introduction to the Bioinformatics and Biostatistics Hub

Introduction

The Hub was created in early 2015 as part of the Centre of Bioinformatics, Biostatistics and Integrative Biology (C3BI). Its main mission is to provide a strong support to the campus in bioinformatics and (bio-)statistics. This support can have very different forms depending on the needs of the Institut Pasteur researchers, from simple advice to trainings and collaborative projects on the long term. The driving force behind the Hub conception was to integrate bioinformaticians and statisticians working in experimental research Units to create a synergy among them and to favour creativity and development of novel ideas. Recruitments started in 2014 and the Hub reached its expected size early 2018. It is now composed of fifty-one research engineers, among which forty engineers were recruited between 2014 and 2018.

Building process of the Hub: four recruitment sessions

The five-year strategic plan of Institut Pasteur (2014-2018) consisted in recruiting ten bioinformaticians and statisticians per year during four years. The process started in 2014 and three other recruitment sessions were organised in 2015, 2016 and 2017. The first session followed a specific scheme due to the fact that it was prior to the C3BI creation. Indeed, three one-day sessions (June, September and December 2014) were organized to select three or four candidates per session. Ten candidates pre-selected by the Human Resource (HR) department were interviewed by four evaluation committees (scientific, technical, HR and soft skills). Decisions were made at the end of each day.

The last three sessions were conducted in four phases:

- Candidates submitted their application on a dedicated website developed by Christophe Malabat to manage process, files, reviews and interviews.
- The C3BI director, voluntary scientists from the Hub, the C3BI research Units, the HR and IT departments graded the applications according to the education level, expertise in bioinformatics, motivation, teaching experience and reference letters. Each application was reviewed by at least three persons.
- Applications were sorted according to those grades and selected by an evaluation committee composed by the C3BI director and scientific deputy director, the Hub co-heads, two members of the HR department, one engineer from the Hub and one or two senior researchers from the C3BI.
- Selected applicants were interviewed by three evaluation boards (scientific, technical and HR) during two and a half days. Each interview lasted twenty minutes, with ten minutes of presentation from the applicant and ten minutes for questions. Decisions were made at the end of each day, just after the last interviews.

Two short additional sessions were organised in 2016 to compensate for the two offer declinations among selected candidates.

Table 1 summarises the characteristics of each session (number of applicants, number of interviews, number of selected applicants).

<table>
<thead>
<tr>
<th>Year</th>
<th>Number of applications</th>
<th>Number of selected candidates</th>
<th>Number of interviews</th>
<th>Number of selected applicants</th>
</tr>
</thead>
<tbody>
<tr>
<td>2014</td>
<td>Unknown (HR)</td>
<td>24</td>
<td>96</td>
<td>7</td>
</tr>
<tr>
<td>2015</td>
<td>371</td>
<td>45</td>
<td>135</td>
<td>12</td>
</tr>
<tr>
<td>2016</td>
<td>217</td>
<td>36</td>
<td>108</td>
<td>9</td>
</tr>
<tr>
<td>2017</td>
<td>188</td>
<td>32</td>
<td>96</td>
<td>12</td>
</tr>
</tbody>
</table>

Table 1: Characteristics of recruitment sessions
Figure 1: Evolution of the Hub staff members since its creation

Junior engineers (with at least 3 years of experience) as well as experienced scientists (post-graduates with various postdoctoral experience) were selected to cover a broad range of bioinformatics and statistics expertise. Upon request of the management of Institut Pasteur, profiles with dedicated expertise were privileged in order to match the bioinformatic needs of recent research Units.

Additionally, eleven people joined the Hub through internal mobility. Indeed, Christophe Malabat left the Genetics of Molecular Interactions Unit headed by Alain Jaquier, where he had been working as a bioinformatician for ten years. Marie-Agnès Dillies left the Transcriptome and Epigenome Platform headed by Jean-Yves Coppée, where she had been working as a statistician for twelve years. In addition, Pierre Lechat and Stéphane Descorps-Decère joined the Hub in 2015 after the Bioinformatics platform headed by Ivan Moszer closed in 2014. In 2016, Olivia Doppelt-Azeroual, Fabien Mareuil, Hervé Ménager, Bertrand Néron, Corinne Mauvais and Louis Jones from the Centre of Informatics for Biology (CIB) headed by the Director of Information System joined the Hub. Bernd Jagla joined the Hub in January 2016, after working at the Transcriptome and Epigenome platform since 2009, and was then detached to the Biomarker platform headed by Milena Hasan.

In August 2015, one of the early recruits resigned for personal reasons and was replaced by one candidate on the waiting list of the 2015 session. In 2017, Louis Jones retired but was not replaced. Olivia Doppelt-Azeroual resigned in April 2018 and a recruitment will be organised in autumn to select her replacement.

Current structure of the Hub

Hub staff members can have three different status: they can either be part of the Hub core, be embedded in research Units, or be detached in platforms or research Units for several years.

Hub core:
People in the Hub core work mainly on short and medium projects submitted by the IP campus. They are physically located in the Yersin building and are under the HR responsibility of the Hub.

Embedded engineers:
Embedded Hub engineers work part or full time on long projects, that is, projects that last more than three months of cumulated workload. This is a very flexible status, since all but detached staff members can be embedded part time. They are physically located in the Yersin building but move to the research Unit(s) in which they are working as often as necessary, from one day to four days a week for the duration of the project. They have the same HR status as Hub core staff members. An “Embedding agreement” (see Appendix 3.1) describes the relationship between the Hub and the host Unit.

Detached engineers:
Detached engineers are seconded into a research Unit or a platform for a fixed number of years (five years in the case of a research Unit and two years in the case of a platform). They are under the HR responsibility of the head of the Unit or the platform in which they are detached.

The decision to detach an engineer in a platform or a research Unit is taken by the Direction Générale of Institut Pasteur (General management). Detachments in platforms are examined every two years and extension decisions are made by the steering committee (COPIL).

A “Detachment agreement” (see Appendix 3.2) describes the relationship between the Hub and the host Unit or platform.
Recent status development:
A new type of detachment has recently emerged following a request from large scale projects that include a strong bioinformatics component. These projects can spend their grant to hire a bioinformatician for one-year full time or more. Three of them turned to the Hub to hire one of its bioinformaticians rather than recruiting an external candidate, with the advantage of recruiting an engineer with known strong competencies. The full salary of the engineer is then taken in charge by the project. Three Hub staff members have been detached to granted projects so far, namely Emna Achouri (Marco Vignuzzi’s Unit), Violaine Saint André (Labex, Milieu Intérieur) and Freddy Cliquet (Thomas Bourgeron’s Unit).

Regardless of their status, all Hub staff members dedicate 20% of their time to common tasks, including trainings and development of internal collaborative projects such as pipelines or software development but also participating to open desks, answering short questions or attending Hub meetings.

Expert groups:
The Hub is composed of six expert groups that are headed by a senior member, as shown on figure 2. These groups gather embedded and Hub core engineers according to their scientific expertise. Detached engineers are in close contact with one of the groups but are not formally affected to it. These groups reflect important fields of expertise well represented within the Hub, namely Genomics, Transcriptomics and Epigenomics, Phylogenetics, Statistics, Algorithmics, and Web development. Their size has been defined as a compromise between homogeneity of the members’ scientific expertise and the need to keep them small enough to favour exchanges and discussions. Some important topics are not represented within these groups, such as metagenomics, which is carried by too few Hub staff members. The possibility for Hub engineers to be affected to two different groups has been suggested, as the profile of some of them are at the interface between different areas of bioinformatics.

Groups have a fundamental role of scientific outreach activity around their area of expertise. Meetings are organised about once or twice a month and are an opportunity for group members to present an on-going or finished project or discuss a particular scientific subject. They are open to all Hub staff members and are an important opportunity to exchange around specific and sharp questions. Internal group meetings are also organised and are an opportunity to carry out common projects within the group, such as the definition of a common reference or a list of FAQs (Frequently Asked Questions) to be posted on the group web page, the development of a tool to be shared by all Hub staff members, or any other scientific activity carried out in common within the group.

Group leaders also have a strong role in the decision process on submitted projects through the project committee, as they are the best-positioned to evaluate and match the resources available in terms of expertise and time in their group with the requirement of a project. Finally, short meetings are organised once a month between each group leader and the Hub co-heads to review the group’s progress, address possible scientific or human issues encountered in specific projects or any other topic that needs to be discussed.

The group structure is expected to evolve according to the development of the expertise of Hub staff members and changes in their status (Hub core, embedded or detached) with time and project opportunities, as well as possible leavings and
recruitments. The Hub is and should remain a flexible structure in order to adapt to its development, the constant and fast pace in the field of bioinformatics and the need for exchanges and discussions between bioinformaticians and statisticians.

**Hub missions**

Hub missions are diverse and include providing advice, engaging in collaborative projects, performing regular literature surveillance, benchmarking newly developed methods, training, as well as participating to national and European bioinformatics organisations (IFB, ELIXIR, France Génomique, StatOmique). Here, we first introduce our project management tool and then we describe in more detail the various missions of the team. Finally, we present the results of the last C3BI satisfaction survey conducted on the campus between March and May 2018 (see Appendix 1.9).

**BISTRO:**

BISTRO stands for Beautiful Interface System To Report and Organise. It has been developed by Damien Mornico, Rachel Torchet and Christophe Malabat. BISTRO is our management system designed to accurately track the real-time progress of all Hub actions, from the shortest (short questions, open desks) to the most extensive (collaborative projects). It includes public as well as private pages, which can be viewed by the Hub staff members only. It is a versatile tool that has been constantly adapted to the growing needs of the group and has become crucial for the Hub’s daily work. It allows to process any type of request instantaneously with the same level of priority. BISTRO also guarantees that all Hub staff members have equal access to the upcoming projects and questions. Only trainings are organised outside BISTRO, through the C3BI web site.

**Short questions:**

Anyone on the campus can ask a question on our web page. So-called “short questions” are automatically forwarded as an email to all Hub staff members. Short questions get an answer within forty-eight hours and may lead to an appointment. The Hub has answered 157 short questions since we started recording them early 2015.

**Open desks:**

Open desks are organised every Tuesday morning from 10:00 to noon. They take place in the Hub meeting room but will soon be transferred to the cafeteria at the common space in the basement of the Yersin building. Anyone on the campus can come with their question, data or piece of code in order to discuss with engineers and find a solution to their problem. The discussion may lead to the submission of a project. Indeed, very often open desks are their first contact with the Hub prior to the establishment of a short/long term collaboration.

Hub engineers are divided in twelve teams of four people with different backgrounds. Each team is on duty every twelve weeks. However, all other Hub staff members must be close by during the open desk sessions in order to provide additional support to their colleagues in case of need. All open desk discussions are tracked in BISTRO. So far, 165 open desk questions have been recorded, with a mean satisfaction score of 4.23/5.

**Figure 3: Open-desk and short question satisfaction survey.**

**Collaborative projects:**

Collaborative projects are submitted to the Hub through BISTRO, using a dedicated web form. Scientists of the campus can submit a project without prior interaction with any Hub member, but for most of the cases, project submission comes after a
first interaction with a Hub member, either after an “open desk” visit or through an informal contact. The submission form includes several fields in order to describe the project and the needs as much as possible. Every submitted project is forwarded to all Hub staff members and anyone can share their opinion on it and choose to work the lead on it. Every two weeks a Hub internal committee, named the Project Committee, gauges the feasibility of the submitted projects by Hub members. The Project Committee is composed of all expert group leaders and Hub managers. The role of this committee is:

- to estimate if the project can be done by someone from the Hub
- to estimate the project workload/duration
- to determine if the availability and expertise of Hub staff members can match the project needs

Until recently, projects were classified as short, medium or long depending on the associated workload. Short projects (less than two weeks of effective working day) were directly accepted and assigned by the Hub managers and the project committee. Medium projects (between two weeks and three months) needed to be reviewed by the C3BI direction. Long projects (more than three months) were reviewed by an external committee called the core C3BI steering committee.

The core steering committee is composed of Claudine Médigue from the CNRS, and Didier Mazel, Marc Delarue, Patrick Trieu-Cuot, Olivier Schwarz and Jessica Quintin from Institut Pasteur (see 1.1 Missions and development of the C3BI and Appendix 1.2). It meets two to three times a year and examines the applications for long projects as well as requests for extension. Engineers are then embedded in the host team for a long period of time, usually one or two years. The committee can accept the project but can also ask for more information and postpone its decision if the project is poorly described (see Appendix 3.12). They can also decide to terminate a project on the basis of poor progression, delay or they can decline projects if there are no human resources available for its completion.

The project classification and the decision process have been recently revised to adapt to the real duration and progress of submitted projects. Indeed, many short projects last a little more than two weeks but are not re-qualified as medium projects to avoid additional delays in their completion. Therefore, the core steering committee, upon suggestion of Hub co-heads, has decided to consider only two types of projects, namely short (less than one month) and long (more than one month). The decision process for short projects or one-year long projects remains unchanged. In contrast, the decision process for long projects of a few months has been modified. Indeed, the description of submitted long projects will be sent to the steering committee by email once a month. The committee will then decide to either accept, reject, ask for more information or ask for the project to be formally examined at the following committee meeting.

At each step of the process projects can be rejected upon different reasons:

- Absence of the required expertise to complete the project. This is a concerted decision between the applicant and the Hub.
- Incompatible time frame between the project’s timeline as defined by the applicant and the availability of Hub staff members. This is a concerted decision between the applicant and the Hub.
- Un-exploitable data. This decision is always supported by a written report from the Hub member who has examined data quality.
- Insufficient scientific content. This decision is taken by the steering committee and supported by a written report. Poorly described projects or projects whose topic is outside the boundaries of the scientific strategy of the Institut Pasteur are rejected. These projects can be re-submitted after modification by the applicant.
- No respect of the general policy on long projects. This decision is taken by the steering committee. Long projects are supported by the Hub for at most two years without any financial compensation. For an extension beyond two years, applicants must cover the full salary of the engineer dedicated to their projects.

The inability to take on submitted projects is very rare and always explained to the applicant. Twenty-five of the 280 projects have been rejected in concertation with the Hub and the project applicant. Two of the sixteen long projects have been rejected by the steering committee.

Statistics of submitted projects

Since the creation of the Hub in 2015, 280 projects have been submitted so far. Figure 5 summarises the distribution across the different project types and status. Most of the projects are short (183, 65%) and closed. 58 (20 %) are medium and 39 (15%) are long. The recent evolution of the projects classification into only short and long will results into 219 short projects and 61 long projects.
Projects are submitted by research Units from all scientific departments of Institut Pasteur. The department with the highest number of projects is the Genomes and Genetics department, as shown on Figure 5a.

The Hub has collaborated with more than 70% of Institut Pasteur Units, so far 105 out of 141 Units. Figure 5b shows the distribution of the project counts per Unit, that is the number of projects submitted per research Unit. Thus, one half of the research Units with which we collaborate submitted only one project. Interestingly, the other half submitted more than one project, suggesting that they were satisfied with the first one and wanted to submit further projects. The figure shows that, the Molecular Parasitology and Signaling Unit headed by Gerald Spaeth, submitted up to twelve projects.

**Figure 4:** Project type and status distribution

**Figure 5:** Distribution of projects across departments. The Hub works with all departments and centres of Institut Pasteur. Some of them like Genome and Genetics have important requirement of bioinformatics support;
Project count by research Unit. Half of the Units have submitted 1 project to the Hub, while the other half have submitted 2 or 3 projects. Very few research Units have submitted more than 7 projects. Overall, this shows a critical need for bioinformatics support that can potentially require the implication in one long project of the detachment of the Hub member in that unit.

**Project progression:**

Once an engineer has been assigned to a submitted project, he becomes the project manager (PM) of this project. BISTRO sends an automatic notification via email to the applicant with the name of the assigned PM. The PM is the reference for the applicant for the whole duration of the project and is responsible for the progression of the work. Other Hub staff members may also be involved as operators. All the relevant information regarding the project (names of PM and operators, goals and methodological steps, progress status) are available in BISTRO. Only the applicant and Hub staff members can have access to this detailed information through a login and a password defined at the time of the first submission.

**Satisfaction questionnaire:**

For each project, a satisfaction questionnaire is sent when the project is closed or waiting for publication. This survey is the opportunity for the researcher to provide his/her feedback on the global development of the project, the collaboration with the Hub, the project manager and operators, the results obtained and the time of completion. A general comment can also be added at the end of the survey. Among the 151 closed or awaiting publication projects, we have received 73 responses (48%). The median of the satisfaction score is 5, with a mean around 4.7. Time of completion is the only criteria that receives slightly lower grades.

**Figure 6: Projects evaluation.** A satisfaction questionnaire is sent at the end of each project and Hub users are asked to rate 4 aspects of the project from 1 (poor) to 5 (high). The 4 questions are about global appreciation, exchanges with the Hub, results and time of completion. The feedback received is very positive with a mean rate of 4.7 for 3 of the 4 indicators. For the time of completion mean is around 4.3. Comments associated to the survey are listed in the Appendix 3.3.
Training

Since its creation, the C3BI has been entrusted with the mission of organising trainings in bioinformatics for biologists. Thus, with about fifty junior and senior scientists and a wide variety of scientific skills, the Hub has been responsible for coordinating the implementation of several training courses.

Having the opportunity to develop our trainings de novo, we have chosen to implement an educational model guided by the principle of "learning by doing", and to focus our efforts more particularly on bio-analysis tools.

The Institut Pasteur Boards asked the C3BI to develop a mandatory training program dedicated to all first year PhD students entering the Institute. This request together with the bioanalysis needs in the campus led us to identify three strong axes of course development: First, there is a strong demand for statistics methodology for daily use. Second, machine learning approaches are gaining ground in all disciplines and artificial intelligence methods are key to future research. Finally, sequence analysis remains fundamental and essential with Omics data. These three axes gave rise to three major courses:

- **Introduction to Data Analysis**: this fifty-hour course covers all the mandatory steps to build state of art analysis workflows (design, collection, curation, hypothesis testing and data mining) and produce good quality scientific graphic visualisations.
- **Scientific Programming in Python**: this course gives participants basic knowledge in python and state-of-the-art machine learning methods to analyse their own data sets.
- **Sequence Analysis**: the objective of this course is to allow autonomous and critical use of some of the major sequence analysis software by combining theoretical presentations and practical applications.

For each course, we have selected tools that are considered as a reference for the community (R, Python, Galaxy ...) and we are committed to transmit good practices such as proper experimental design, literal programming, source control, notebook, etc.

We wish to further develop our offer in the coming years with courses that reflect the strong expertise gathered by Hub experts, in particular in Next Generation Sequencing data analysis. We intend to develop courses that differentiate from the broad offer that already exists in the field, in France and more generally in Europe, by taking advantage of the specificity and long history of Institut Pasteur in infectious diseases.

Bioinformatics national and international organisations

The Hub is involved in various national and international organisations, either upon request from the C3BI or upon the initiative of its staff members, as shown in table 2.

<table>
<thead>
<tr>
<th>Organization</th>
<th>Hub member(s) involved</th>
<th>Contribution and involvement</th>
</tr>
</thead>
<tbody>
<tr>
<td>SFBI</td>
<td>Hub co-heads and 23 Hub staff members</td>
<td>The Hub is supporter as a legal entity and Hub engineers are individual members</td>
</tr>
<tr>
<td>IFB</td>
<td>Bertrand Néron</td>
<td>Correspondent for full life-cycle cost</td>
</tr>
<tr>
<td></td>
<td>Fabien Mareuil</td>
<td>Member of infrastructure working group (GRISBI). Involved in CRISPRCasFinder project funded by IFB.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Member of National Network Cluster Resources working group</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Involved in New Generation Phylogeny.fr: Refactoring Phylogeny.fr for innovative phylogenetic services project funded by IFB and in collaboration with O. Gascuel’s Unit</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Involved in Facilitate Galaxy analysis environment construction for research communities project funded by IFB</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Involved in ARIA for hybrid structure determination on the cloud project funded by IFB in collaboration with M. Nilges’ Unit</td>
</tr>
<tr>
<td></td>
<td>Hervé Ménager</td>
<td>Co-leader of the IFB “Catalogue of French resources in bioinformatics” action</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Involved in the “Catalysing interoperability and integration between bioinformatics resources” action</td>
</tr>
<tr>
<td></td>
<td>Laurent Bouri</td>
<td>18 months contract on the following project</td>
</tr>
<tr>
<td></td>
<td></td>
<td>&quot;Development of the Catalogue of French Resources in Bioinformatics&quot;</td>
</tr>
<tr>
<td>ELIXIR</td>
<td>Hervé Ménager</td>
<td>“Workflows and Workbenches” pillar of the tools platform</td>
</tr>
</tbody>
</table>
Interactions with the Hub finally reached a critical mass. The next immediate goal is to consolidate its structure as well as the processes that drive the activities. Publication led by the interoperability platform, the “Enabling the reuse, extension, scaling, and reproducibility of scientific workflows” Implementation Study:

More recently, we hired another Hub staff representative, as the responsible for the tasks of the C3BI and the Hub have been in contact with the SFBi board since 2017 for hosting JOBIM conference in Paris in 2021. If the application of the C3BI is accepted, the Hub intends to be strongly involved in the organisation of this event.

The Hub is a member of the French Institute for Bioinformatics (IFB) which gathers all bioinformatics platforms in France. We are currently hosting a short-term contract IFB engineer, Laurent Bouri, who will develop the Catalogue of French resources in bioinformatics.

The Hub is also involved in the European consortium for bioinformatics (ELIXIR). In that context, Hervé Ménager is the IP representative, as the responsible for the “Workflows and Workbenches” pillar of the tools platform, for the ELIXIR infrastructure. It has participated in various activities within the IFB, and has been funded (12 months contract) by the ELIXIR/EXCELERATE program to participate to the development of the bio.tools registry of bioinformatics tools and services.

Most recently, we hired another research engineer, Jérôme Raciazek, for a ten-month contract to work on an implementation study led by the interoperability platform, the “Enabling the reuse, extension, scaling, and reproducibility of scientific workflows”.

Publications and tool development

Since the creation of the team, Hub staff members work has resulted in many publications and tools:

- 91 publications. The table below describe the distribution between the different status of Hub staff members and the authorship positions

<table>
<thead>
<tr>
<th>Author position</th>
<th>Hub core</th>
<th>Embedded</th>
<th>Detached</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>First</td>
<td>6</td>
<td>3</td>
<td>5</td>
<td>14</td>
</tr>
<tr>
<td>Second</td>
<td>6</td>
<td>8</td>
<td>4</td>
<td>18</td>
</tr>
<tr>
<td>Third</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>15</td>
</tr>
<tr>
<td>More than third</td>
<td>20</td>
<td>13</td>
<td>10</td>
<td>43</td>
</tr>
<tr>
<td>Last</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Total</td>
<td>37</td>
<td>29</td>
<td>25</td>
<td>91</td>
</tr>
</tbody>
</table>

- 66 talks (39 by Hub core, 8 by embedded staff and 19 by detached staff)
- 62 posters (39 by Hub core, 6 by embedded staff, 17 by detached staff)
- 48 distributed tools (19 by Hub core, 19 by embedded staff and 10 by detached staff; https://c3bi.pasteur.fr/hub/tools/)

The publications of the detached staff are listed at the end of this section. Publications of the other Hub staff members will be listed at the end of the section of their respective expert group.

Future directions:

After the last recruited bioinformatics and biostatistics experts joined Institut Pasteur in late 2017 and early 2018, the Hub has finally reached a critical mass. The next immediate goal is to consolidate its structure as well as the processes that drive its interactions with the Institut Pasteur scientific community. This includes:

- Optimise the protocol followed by the project committee in order to accelerate the process regarding the candidate projects, as well as to streamline the decision making on Hub procedures. The committee is intended to meet every two weeks but these meetings might need to take place on a more regular basis.
- Renew our efforts to outreach the scientific community of Institut Pasteur and make them aware of the kind of support and services we are ready to provide/undertake. To this end we will continue advertising our different activities in the BIP (the IP weekly newsletter) and through our web site. Aside from this, and following our recent move to the Yersin building, we will organise an “open day” to present our domains of expertise and work dynamics to the IP campus. This kind of initiative may be repeated twice a year and can be coupled with a presentation during the regular Welcome Day where newcomers are introduced to the IP community.
Apart from improving the daily operation of the Hub, there are a few axes that we wish to develop or extend:

- Definition of an individual training plan for all Hub staff members, in order to adapt to the constantly evolving bioinformatics landscape. This would allow us to fulfill the needs of IP’s research community without having necessarily to rely on new recruitments. To this end, on the immediate future, we plan to develop expertise in metagenomics, long reads, single cell and flux cytometry data analysis, and machine learning approaches.
- Foster collaborations to be associated with grant proposal together with IP research Units.
- Promote the development of technological solutions in collaboration with the Biomics pole.
- Creation of a Metagenomics group, associated with the development of this expertise within the Hub.
- Development of bioinformatics methods and tools with IP partners, eventually leading to the submission of grant proposals.

One important issue that the Hub will face in the coming years will be to keep the motivation of Hub engineers. Most of them have a high level of education and thus a high level of expectation with respect to the kind of tasks and projects they will be involved in. Thus, it is crucial to be able to propose ambitious projects on the long term in collaboration with research Units or technological platforms. This trend is already in place as it can be seen from the proportion of embedded, detached and Hub core engineers (28%, 40% and 32%). It is also important to favor the progression of the careers of Hub staff members. Group leaders and senior engineers will be encouraged to apply for a promotion through the COMESP internal evaluation procedure, in order to become expert engineers. The involvement of junior engineers in long and ambitious projects submitted by the campus will be favored to allow them to improve and increase their area of expertise and autonomy.

**Publications of the 15 detached staff members of the Hub:**


### 3.2 Hub Expert Group SABER: Statistics Applied to Biology and Experimental Results

**Gael Millot, gael.millot@pasteur.fr**  

#### General Overview

**Team members:**
- 6 Research engineers

**Hub core (5):**
- Emeline Perthame
- Gael Millot
- Steven Volant
- Vincent Guillemot
- Pascal Campagne

**Embedded in research Units (1):**
- Andjela Davidovic; Experimental and Computational Methods for Modeling Cellular Processes

**Keywords:** Biostatistics, Applied Mathematics, Modeling, Experimental Design, Computing

#### Outlook and highlights:

The SABER group consists of 6 research engineers, most of us recruited in 2016 or after. Our main objective is to support the analysis and interpretation of biological results using statistic and mathematic approaches. Our skills include applied mathematics, biostatistics, modeling, machine learning, transcriptomics, and computing. To date, we collaborated to 47 projects, submitted by 36 Pasteur Units from 10 different departments. This currently led to 10 publications as well as 4 publicly available tools, covering various fields, like SHAMAN (metagenomics, [http://shaman.c3bi.pasteur.fr/](http://shaman.c3bi.pasteur.fr/)), MEMHDX (mass spectrometry, [http://memhdx.c3bi.pasteur.fr/](http://memhdx.c3bi.pasteur.fr/)), and JASS (statistical genetics, [http://jass.pasteur.fr/](http://jass.pasteur.fr/)). Beside scientific research, an important part of our activity is dedicated to teaching, notably with the creation and follow-up of the mandatory course “Introduction to Data Analysis” for Pasteur PhD students. Globally, we were involved in 5 different training programs, which represent 186 hours of teaching in 2016 - 2017 and 195 hours in 2017 - 2018.

#### Main Projects:

1. **Host microbiota modification by the pathogen Listeria monocytogenes [SV, 2 months]**
   
   PI: Javier Pizarro-Cerda, Yersinia Unit, Microbiology dept.
   
   Published in PNAS 2016 (10.1073/pnas.1523899113), SHAMAN tool available at [http://shaman.c3bi.pasteur.fr/](http://shaman.c3bi.pasteur.fr/)

2. **Comparing V-J gene usage for specific antibody generation [GM, 2 months]**
   
   PI: Pierre Bruhns, Antibodies in Therapy and Pathology Unit, Immunology dept.
   
   Submitted, Comat tool available at [https://gitlab.pasteur.fr/gmillot/Comat](https://gitlab.pasteur.fr/gmillot/Comat)

3. **Implementation of a statistical tool for HDX-MS data analyses [SV, 5 months]**
   
   PI: Sébastien Brier, Mass Spectrometry for Biology, Chemistry and Structural Biology dept.
   

4. **Infection of *Ixodes ricinus* by *Borrelia burgdorferi* sensu lato by in peri-urban forests of France [EP, 2 weeks]**
   
   PI: Valérie Choumet, Environment and Infectious Risks Unit, Infection & Epidemiology dept.
   
   Published in PLoS One 2017 (10.1371/journal.pone.0183543)

---

**Figure 1:** the 26 achieved projects of SABER. Projects submitted by a Pasteur department are represented on the left side of the panel. Specific projects, depending on C3BI, Institut Pasteur International Network (RIIP) or other centres (like Pasteur Platforms), are on the right. "Achieved projects" include closed projects (with publication, manuscript submitted, manuscript in preparation or no publication expected) and running project with publications.

Number of publications is indicated above bars. In addition to these 10 publications, VG published one paper when he was a SABER member embedded in the Hugues Aschard Pasteur Unit (see the publication list below).
Detailed description

Introduction:

The SABER group was created with the objective to support the research work of Pasteur Units when statistic or mathematic approaches are required. Our group is composed of 6 research engineers. SV joined the Hub when it was created in 2014. The 5 other members were recruited in 2016 or after (Figure 2). VG was embedded for one year in the Hugues Aschard Pasteur Unit before being fully integrated into the SABER group. Our backbone expertise includes applied mathematics, statistics, modeling, machine learning, experimental design, differential analysis, omics, clustering and computing (R, Python or bash), which is in agreement with the main demands from the research Units for the last two years. The group also benefits from particular expertise, such as long experience as researcher (post-doc or associate professor), wet-laboratory skills, or diverse educational backgrounds (math, stat, physics or biology), which facilitates the communication between the SABER group and biologists requesting stat/math analysis. To date, we collaborated to 47 projects submitted by 36 Pasteur Units from 11 different departments (see Appendix 3.11). Twenty-six were achieved (Figure 1), leading to 10 publications and 13 projects waiting for a publication. In parallel, we are keeping up to date with the last developments in our field, through trainings, active membership in societies, like "Statomique" a group supported by the Société Française de BioInformatique (SFBI) and by CNRS (GdR Molecular Bioinformatics) and with the visit of prestigious scientists in the Hub (Prof. Hervé Abd, University of Texas, Dallas, USA, in May-June 2018).

Results since the conception of the expert group, two illustrative examples:

Host microbiota modification by the pathogen *Listeria monocytogenes* [SV, 2 months]

**PI:** Javier Pizarro-Cerdà, Yersinia Unit, Microbiology dept.

The *Listeria monocytogenes* (Lm) bacteria is responsible for gastroenteritis in healthy individuals and for a severe invasive disease in immunocompromised patients. The virulence mechanisms of this bacterium are poorly known but seem to depend on epidemic strains. To investigate this, mice were orally infected with Lm strains from different outbreaks. We developed a web-based interface, named SHAMAN (SHiny Application for Metagenomics ANalysis), to estimate the relative abundance of different bacteria genus in mouse microbiota. SHAMAN results from a stat (SV - SABER group) - bioinfo (A.Ghozlane - Biomics) collaboration. It encompasses all the steps of data analysis (including data normalization) coupled to high levels of statistical analyses (i.e., generalized linear models) with the strong advantage that this is transparent for non-statistician users, thanks to the user-friendly Shiny interface. This allowed to show that Lm strain type I secretes a bacteriocin specifically in the gut that alters host intestinal microbiota, allowing Lm colonization of the intestine and, consequently, invasion of deeper organs. Therefore, our work showed that epidemic listeriosis implicates not only interactions between Lm and host cells, but also interactions between Lm and the host intestinal microbiota that are critical for the establishment of infection.

Published in PNAS 2016 (10.1073/pnas.1523899113), Web of Science citations: 22.

SHAMAN tool available at http://shaman.c3bi.pasteur.fr/ (google analytics: 100 users/month, cited in 5 papers).

Comparing V-J gene usage for specific antibody generation [GM, 2 months]

**PI:** Pierre Bruhns, Antibodies in Therapy and Pathology Unit, Immunology dept.

Immunoglobulin G (IgG) antibodies secreted by mature B cells in mammal bodies results from a complex process. Each B cell secretes a single antibody protein generated by the assembly of one V and one J DNA cassette, among the several available on the chromosome, by intrachromosomal DNA rearrangement. The different combination of V-J cassette usage, referred to as "V-J gene usage", is such that it should cover a large possibility of antigen detection. However, the number of different circulating antibodies that recognize a *same* antigen, also called the "antigen repertoire of an individual", is not known. The aim of the project was to compare the IgG repertoire of different mice immunized with the Tetanus Toxoid (TT) antigen, a component of the Tetanus Vaccine. To achieve this, the CelliGO technology was developed. This technology sorts the mature B cells that specifically secrete the anti-TT IgG and then single-cell sequences the mRNA IgG domain responsible for the antigen recognition. In other words, CelliGO counts, for each analysed mouse, the number of times each V and J cassette is used to create the anti-TT IgG. The results are recapitulated in a contingency matrix of V-J gene usage. For this project, we developed a R toolbox, called "Comat", that provides quality control reports, matrix distances, permutation test, corrected p values and graphics for the 2 by 2 comparison of contingency matrices of same dimension in batches. Using Comat, we showed that mouse
B cells use a high diversity of V-J cassettes to form IgG that recognize the TT antigen. The analysis of mice immunized with another antigen (rabbit Glucose-6-Phosphate Isomerase) indicated that the frequencies of V-J cassette usage depend on the antigen used for immunization, suggesting an antigen repertoire restriction.

Manuscript submitted in Nature Biotechnology

Comat tool available at https://gitlab.pasteur.fr/gmillot/Comat

For a complete list of the projects of the SABER expert group see Appendix 3.5.

**Inter-group Activities:**

Collaboration with the other hub expertise groups or platforms is a backbone of the SABER group activity, as it represents 39% (18/47) of the projects taken in charge by SABER members. In details, collaborations represent 2% (1/47) with ALPS, 4% (2/47) with TEG, 6% (3/47) with FunGen, 8% (4/47) with WINTER and 17% (8/47) with Biomics. The SHAMAN and JASS web interfaces result from such inter group collaborations (see below).

**Teaching and Training Activities:**

An important part of our activity is dedicated to teaching. This notably involved the creation and the running of the mandatory course "Introduction to Data Analysis" for Pasteur PhD students. For the last 2 academic years, we totalised 381 hours of teaching in 5 different training programs (detailed in the Appendix 1.7). Our teaching activity led to the publication of a book chapter (Abdi et al, Canonical Correlation Analysis, Springer ed. 2017, DOI 10.1007/978-1-4614-7163-9_110191-1) and a book (Millot, Comprendre et réaliser les tests statistiques à l'aide de R, 4rd edition increased. De Boeck editions, 2018, https://c3bi.pasteur.fr/gael-millot-livres/).

**Software and tools:**

- SHAMAN: web interface for quantitative metagenomic analysis [SV in coll with Biomics]
- MEMHDX: web interface for Hydrogen/deuterium exchange (HDX) mass spectrometry (MS) analysis [SV].
  [http://memhdx.c3bi.pasteur.fr/], initially published in Hourdel et al, Bioinformatics 2016. Two papers totally, google analytics: 50 users/month
- Comat: tool for the 2 by 2 comparison of contingency matrices of same dimension in batches [GM].
  [https://gitlab.pasteur.fr/gmillot/Comat], manuscript submitted.
- JASS: Web interface for joint analysis of GWAS summary statistics [VG in coll with ALPS, WINTER & Aschard].
  [http://jass.pasteur.fr], still in development.

**Technology transfer:**

SV, in collaboration with the Unit of Julia Chamot-Rooke (Mass spectrometry for Biology), patented the commercial use of MEMHDX.

**Perspectives and strategy:**

Beside our current responsibilities in (1) management of projects submitted by Pasteur research Units, (2) rapid support (open-desk and short questions via the Hub webpage) and (3) knowledge transmission, we are maintaining our effort in active technological and innovation watch, notably those in agreement with demands from Pasteur research Units. To date, single cell "omic", metabolomic and flow cytometry analyses are at the top priority list of the competences that need to be acquired or improved by SABER members.

**Publications of SABER expert group:**


3.3 Hub Expert Group TEG: Transcriptome and EpiGenome

Claudia Chica; claudia.chica@pasteur.fr, https://research.pasteur.fr/en/team/expertise-group-teg-transcriptome-and-epigenome/

General Overview

Team members:
9 Research engineers

Hub core (1):
- Christophe Bécavin

Embedded in research Units (5):
- Anne Biton; Human genetics and cognitive functions
- Giovanni Bussotti; Molecular parasitology and signalling
- Claudia Chica; Nuclear organization and oncogenesis
- Blaise Li; Mechanisms of epigenetic inheritance
- Adrien Pain; Genetics and genomics of insect vectors

Detached in the following Units & platforms (3):
- Violaine Saint-André; Human evolutionary genetics & Immunobiology of dendritic cells
- Rachel Legendre and Hugo Varet; BIOMICS: Transcriptome & Epigenome platform

Intern M.Sc. (1):
- Maëlle Daunesse (2-year program)

Keywords: Expression quantification, Gene regulation, Comparative and differential omics, Multidimensional data integration

Outlook and highlights:
The group consists of 9 research engineers, most of us recruited during the second half of 2015. Our common interest is the study of biological function through the lens of the information transactions that lead from the genotype to the phenotype. In particular, we deal with the multiple kinds of sequence information measured by high throughput technologies. We have collaborated with 39 research Units from 10 different departments/platforms in the framework of 62 projects that have already produced 12 publications and will most likely be followed by 4 more in the coming 6 months. Specific developments required by some of these projects have become independent and publicly available tools such as Listeriomics, a web platform for “-omics” data visualization; Sequana/Sequanix, a set of standardised pipelines for sequencing analysis with a dynamic graphical interface. As an additional activity of outreach to the scientific community of Institut Pasteur (IP) and beyond, we have participated as organizers and/or trainers in 15 of the courses and workshops offered by the C3BI and in some of the IFB/AVIESAN/Elixir teaching initiatives.

Main Projects:
- **Listeriomics** - Development of a web platform for visualization and analysis of Listeria omics data. Pascale Cossart; Bacteria-Cell Interactions; Cell biology & infection [CB; 2.5 years; 2 publications; 1 tool]
- **Bioinformatics support for the LeiSHield project.** Gerald Spaeth; Molecular parasitology and signalling; Parasites and insect vectors [GB; 2.5 years; 2 publications]
- **Insect Vector Genomics.** Kenneth Vernick; Genetics and genomics of insect vectors; Parasites and insect vectors [AP 2 years; 2 publications]

Figure 1: The 62 achieved projects by TEG.
Number of projects submitted by the 11 departments of the Institut Pasteur, C3BI, Institut Pasteur International Network (IPIN) or other centres (such as Institut Pasteur Platforms) are represented on the y axis. Number of publications is indicated above bars. In addition, 6 articles were published as the result of collaborations established by group members before their integration to the Hub.
Detailed description

Introduction:
The majority of the TEG members joined the Hub between August and December 2015 with a couple of more recent additions in 2016 and 2018. Four of these members were already working in different research Units at IP; this has eased their integration into the Hub and their knowledge of the IP research community has facilitated our expert group mission. Some group members have come to fill in an expertise gap in the analysis of the sequencing techniques for chromatin profiling and quantification of the non-coding genome. They are involved in projects that deal with the role of the epigenome in maintaining a specific transcriptional status or in tuning the transition between different biological conditions. Other members have applied their analytical and statistical expertise to the dissection of high dimensional data, for the description of complex phenotypes such as disease or pathogenicity. Other group members have instead focused on comparative studies where the effect of genomic variation in establishing a phenotypic output, such as immune system response to different challenges or pathogen/vector adaptation, is evaluated.

Results since the conception of the expert group, two illustrative examples

Bioinformatics support for the Leishields project
Pt: Gerald Spaeth; RU: Molecular parasitology and signalling; D: Parasites and insect vectors; 2.5 years
LeiShield (www.leishield.org) is an international consortium that involves 14 international institutions (5 Institut Pasteur among them). It includes researchers from disease-endemic countries already involved in the surveillance of leishmaniasis, a neglected tropical disease, and researchers with access to cutting-edge “omics” technologies equipped with the corresponding bioinformatic expertise. GB has been responsible for the coordination of the bioinformatic support of the consortium since 2016. Initially, he developed a Leishmania-specific workflow named L-GERT (Leishmania-GEnome Reporting Tool) for the identification of chromosome and gene copy number variants as well as data visualization. He has analysed a total of 354 genomic sequences either produced in the framework of the consortium or already available to determine the evolvability of Leishmania parasites in culture and during its infectious cycle. His analyses have been instrumental to provide new insights on how genome instability and haplotype selection can drive evolutionary adaptation in Leishmania, resulting in different clinical manifestations of the disease in terms of tissue tropism and drug resistance. This work has produced two different publications in MBio and in Nature Ecology and Evolution (Dumetz et al. 2017; Prieto Barja et al. 2017). Within the same consortium, GB has tackled other questions regarding the geographical sequence variation of different Leishmania species and the re-annotation of recently sequenced strains, which should lead to additional publications in the coming year. Finally, with the collaboration of other Hub staff members, GB has co-organized a symposium and a workshop on sequence data analysis and visualization for the LeiShield partners in 2016. The goal of this course was to enable IPIN collaborators to explore and query their own data to address their specific questions. The successful outcome of this first LeiShield consortium has led to the organization and financing of a second international project in the framework of the H2020 actions. LeiShield-MATI is aimed at investigating the evolutionary adaptation of 3 different cutaneous Leishmania species in insect, rodent, and human hosts across endemic areas of the MATI region (Morocco, Algeria, Tunisia, Iran).

Listeriomics Development of a web platform for visualization and analysis of Listeria omics data
Pt: Pascale Cossart; RU: Bacteria-Cell Interactions; D: Cell biology & infection; 2.5 years
Over the past three decades Listeria has become a model organism for host-pathogen interactions. The number of Listeria “omics” data sets has increased in terms of number and heterogeneity. There are now more than 40 published Listeria genomes, around 400 different transcriptomics data and 10 proteomics studies available. CB has developed a web-based platform named Listeriomics to provide biologists with the capacity to analyse these data through a systems biology approach. It contains two main tools: 1) A genome viewer for displaying gene expression array, tiling array, and RNA-seq data along with proteomics and genomics data 2) An expression atlas, which is a query-based tool which connects every genomic element (genes, smallRNAs, antisense RNAs) to the most relevant “omics” data. These platforms integrate all the genomic, transcriptomic and proteomic data sets ever published on Listeria and allow biologists to analyse them dynamically, and bioinformaticians to have a central database for their analyses. Listeriomics has been published in mSystems (Becavin et al. 2017) and is being used on a daily basis as a hypothesis generation tool. Additionally, it has been used in the discovery of a mini protein Prli42 that is the link
between stress and the stressosome in Listeria, which was also published in Nature Microbiology (Impens et al. 2017) in collaboration with the same research Unit.

For a complete list of the projects of the TEG expert group see Appendix 3.6.

**Platform Services and Activities:**

Since their recruitment, RL and HV have been detached in the BIOMICS platform where they’ve been responsible of the bioinformatic and statistical analysis of ~40 sequencing projects per year and have contributed in 14 publications, in addition to the 3 publications produced in the framework of Hub projects (Figure 1). In order to streamline the execution of her BIOMICS’ projects, RL participated in the development of Sequana, a set of workflows dedicated to high throughput sequence data analysis. The open source code is available under Github and the documentation on Readthedoc. Additionally, RL was involved in the implementation of Sequanix, a graphical interface that facilitates the use of Sequana workflows as well as Snakemake pipelines. With the same objective, HV developed SARtools, an R package for the analysis of RNA-Seq data that provides tools to generate diagnostic graphs, to run the differential analysis and to export the results.

**Inter-group Activities:**

We have collaborated on three different projects with members of the ALPS group in the implementation of visualization tools and interfaces. Similarly, three other projects, which deal mostly with functional annotation, have required the co-ordinated work of members of the FUNGEN and TEG group. We also participate in collaborative projects such as Bioconvert, which involves members of multiple Hub groups, but also research engineers attached to C3BI research Units.

**Teaching and Training Activities:**

All the members have participated as organizers and/or trainers in 15 courses and workshops. Out of these 8 were restricted to the IP and IPIN community (such as the PhD Introduction to Data Analysis and the Hands-on NGS courses), while the remaining (such as the IFB/AVIESAN/Elixir training courses) were extended to French and European students/researchers. These activities, which sum 200h per year on average, were addressed at a mixed audience from M.Sc. students to young researchers. Additionally, we have engaged in the supervision of bachelor, M.Sc. students. Together with them, we have addressed specific methodological and technical questions encountered during our collaborations with the IP research Units. Finally, within the framework of some of our projects, we have also trained our closest experimental collaborators on the basic concepts and tools required for high throughput sequence data analysis.

**Software and tools:**

The following bioinformatic tools/web resources were developed and/or are actively maintained by group members:

- **Listeriomics**: Interactive web platform to browse genomic, transcriptomic and proteomic datasets in Listeria ([https://listeriomics.pasteur.fr](https://listeriomics.pasteur.fr)), Becavin et al., 2018, CB)
- **checkMyIndex**: Application search for a set of compatible indexes for any sequencing experiment ([https://checkmyindex.pasteur.fr/](https://checkmyindex.pasteur.fr/), HV)
- **Sequana/Sequani**: Set of Snakemake HTS pipelines and graphical interface ([https://github.com/sequana/sequana/](https://github.com/sequana/sequana/), Desvillechabrol et al., 2018, RL)
- **CRC mapper**: Core transcriptional regulatory circuits ([https://bitbucket.org/young_computation/crcmapper.git](https://bitbucket.org/young_computation/crcmapper.git), VSA)
- **qaf_demux**: Quality-aware fastq demultiplexer ([https://gitlab.pasteur.fr/bli/qaf_demux](https://gitlab.pasteur.fr/bli/qaf_demux), BL)

**Technology transfer:**

**Patent DI2013-32**: Method to characterise high risk Human Papilloma Viruses in cervix lesions, by quantitative dissection of transcriptomic data (AB).

**Perspectives and strategy:**

In the short term, we aim at completing the publication of 4 manuscripts currently under review, which include, among others, the results of our long-term collaborations with the Nuclear Organization and Oncogenesis, and the Human Genetics and Cognitive Functions research Units. Additionally, we want to identify additional niches of collaboration within IP that may lead to the submission of joint grant proposals with an important bioinformatics and biostatistics component. As a mid-term goal, we will point at enabling research groups to use emerging technologies like epiTranscriptomics, single cell “omics” and long reads sequencing. This will initially require the training of our group members and subsequently the dissemination of the acquired knowledge to the Institut Pasteur research community.

**Publications of TEG expert group**


3.4 Hub Expert Group WINTER: Web Integration

Hervé Ménager, herve.menager@pasteur.fr, https://research.pasteur.fr/en/team/expertise-group-winter-web-integration/

General Overview

Team members:
7 Research engineers including

Hub core (4):
• Hervé Ménager,
• Bryan Brancotte,
• Fabien Mareuil,
• Rachel Torchet

Short term contracts (3):
• Kenzo-Hugo Hillion,
• Jérôme Raciazek (IFB/ELIXIR contract),
• Laurent Bouri (IFB)

Alumni (1):
• Olivia Doppelt-Azeroual

Keywords: Software Development, Web Development, UX Design, Databases, Workflow and Tool Integration

Outlook and highlights:
The WINTER group is a software development team focusing mainly on Web technologies for publishing and sharing scientific tools, analysis, data and workflows. We provide our expertise to the scientists of the campus, covering a broad range of services to design, develop and publish software tools and databases on the Web. As part of the Hub mission, our projects cover a wide variety of scientific topics (Structural Bioinformatics, Transcriptomics, Statistical Genetics, etc.). Moreover, we are in charge of the Galaxy server of the Institut Pasteur, an integration platform to publish and use bioinformatics tools and workflows in a web interface. Finally, our group is heavily involved in collaborations with, most notably, the “Institut Français de Bioinformatique” (e.g. participation to the distributed national environment of services in bioinformatics) and the ELIXIR European infrastructure (e.g. participation to the EXCELERATE program).

Main Projects:

• Development of a web application and new functionalities for the maintenance and curation of iPPI-DB. PI: Olivier Sperandio, Chemoinformatics and proteochemometrics group, Structural Bioinformatics laboratory. http://ippidb.pasteur.fr in development, manuscript in preparation

• Development of a new web interface for ARIA. PI: Benjamin Bardiaux, Structural Bioinformatics laboratory https://gitlab.pasteur.fr/bis-aria/ariaweb


• ELIXIR Tools Platform - improve the discovery, quality and sustainability of software resources. External infrastructure project, PI: Hervé Ménager http://bio.tools, 5 publications in various journal

Figure 1: the 8 achieved projects of WINTER. Projects submitted by a Pasteur department are represented on the left side of the panel. Specific projects, depending on C3BI, Institut Pasteur International Network (IPIN) or other centres (like Pasteur Platforms), are on the right. "Achieved projects" include closed projects (with publication, manuscript submitted, manuscript in preparation or no publication expected) and running project with publications. Number of publications is indicated above bars. It should be noted that the majority of the projects in WINTER are long term software development projects, and therefore still either not finished or not published. In addition to these projects, we have published 9 other papers in the context of external collaborations (ELIXIR, IFB or MADICS)
Introduction:
The Institut Pasteur has been providing a Web-based access to bioinformatics tools and data for now almost 20 years. It facilitates the access to IP computing resources and software tools for scientists, inside and outside the Institut Pasteur, bypassing the complex technicalities of the computing cluster usage and providing a more user-friendly interface. It also enables advanced data query and visualizations through universally portable interfaces, and promotes the bioinformatics resources developed at the Institut Pasteur through their publication on highly visible platforms. It is a popular and convenient way to easily access tools and data in reference research institutions such as the SIB, the EBI, or the NCBI. Our expert group provides the technical skills that helps creating and maintaining these applications, in close collaboration with other groups from the Hub and research teams.

Illustrative examples of results since the conception of the expert group:

JASS: an online tool for the joint analysis of GWAS summary statistics
PI: Hugues Aschard, ‘Statistical Genetics’ GS group, C3BI, in collaboration with SABER and ALPS groups - 2 PMs
The Statistical Genetics group developed an effective and robust multivariate method to get GWAS summary statistics. This new approach addresses the major barriers of existing approaches, i.e. the presence of correlation between studies that would exists when multiple GWAS analyses share samples. This method consists in a robust test of GWAS summary statistics. The goal of the JASS project is to develop an innovative web interface allowing for the computing of joint statistics, and the dynamic visualization of (i) the associations between the chosen SNPs and the selected phenotypes, and (ii) the computed joint association. The JASS software is already publicly distributed, and available as a web server on http://jass.pasteur.fr. Preliminary results have been presented at the IGES-2017 conference. The software has been presented at the RECOMB, BOSC-2018, and JOBIM conferences, through posters and an oral presentation. A manuscript describing the software is in preparation.

Development of a web application and new functionalities for the maintenance and curation of iPPI-DB
PI: Olivier Sperandio, Unit ‘Structural Bioinformatics’, Department ‘Structural Biology and Chemistry’ - 12 PMs
In order to boost the identification of low-molecular-weight drugs on protein–protein interactions (PPI), it is essential to properly collect and annotate experimental data about successful examples. This allows the scientific community to derive trends about privileged physicochemical properties and chemotypes that maximize the likelihood of promoting a given chemical probe to the most advanced stages of development. iPPI-DB is a public and web-accessible database that contains the structure, physicochemical characteristics, pharmacological data and profile of the PPI targets of several hundred modulators of PPIs. The purpose of our current project is first to develop a new version of the database, and its attached web application, in order to facilitate queries of the users and updates of the database. Our second contribution in this project is to expand the functionalities of iPPI-DB, by including the addition of a new target-centric mode to query the database. The software has been already presented at JOBIM 2018. Once ready for production, the database will be available on http://ippidb.pasteur.fr. A series of papers, covering the new version of the application itself, as well as the content updates for the database, is in preparation.

Development of a new web interface for ARIA
PI Benjamin Bardiaux, Unit ‘Structural Bioinformatics’, Department ‘Structural Biology and Chemistry’ - 5 PMs
The software ARIA (Ambiguous Restraints for Iterative Assignment) automates the treatment of NMR data and calculation of protein structures by molecular dynamics simulation. ARIA is well established in the structural biology community, with a large user base. However, its usage is hindered by the complexity of the deployment and execution of the software. In fact, as a typical analysis protocol, ARIA involves the conversion of NMR input raw data into XML format with a specific configuration file and command line, as well as the configuration of the structure calculation itself, using a GUI or by writing an XML configuration file with hundreds of parameters. To enhance the visibility and usability of the software, we designed a new web interface where users are able to easily manage data, perform calculations, and analyse the results of their ARIA calculations. This application combines a Django server that provides a simple yet highly flexible interface to configure ARIA analyses, with a Galaxy server that handles the execution of the ARIA jobs. The code for the web interface is already available on https://gitlab.pasteur.fr/bis-aria/ariaweb, and the application itself will be publicly available on http://aria.pasteur.fr. For a complete list of the projects of the WINTER expert group see Appendix 3.7.
Platform Services and Activities:

**galaxy.pasteur.fr** - Galaxy is an integrated platform that enables the execution of bioinformatics tools, or the construction of complex automated pipelines (workflows), through a web interface. The Galaxy server of Pasteur is public since 2016, and includes 540 tools, covering several topics including NGS, Metagenomics, RNA-Seq, ChIP-SEQ, and Phylogeny. Since the creation of the Hub, 205 tools have been added, including 13 developed at Institut Pasteur. It also provides through its API an access to the computing infrastructure to more specialized applications such as SHAMAN, ng-Phylogeny and Booster. Although the impact of the platform in terms of publications cannot be easily evaluated, in 2017 only, Galaxy Pasteur counted more than 2000 users, who launched more than 44000 jobs.

**C3BI website** - In collaboration with Damien Mornico (FunGen) and Christophe Malabat (Hub head), Rachel Torchet has been in charge of the design and the maintenance of the C3BI website, which includes a presentation of the C3BI platform and research activities, and the Hub’s project management system.

**ELIXIR and IFB participation** - Hervé Ménager acts as the representative for the Institut Pasteur, and as the “Workflows and Workbenches” pillar of the tools platform, for the ELIXIR infrastructure. The C3BI/Hub has actively participated in various initiatives within the IFB, and has been funded by the ELIXIR/EXCELERATE program to participate to the development of the bio.tools registry, and to the “Workflow Interoperability Implementation Study” of the interoperability platform.

**Inter-group Activities:**

Since our main expertise deals mainly with software development, a large part of our projects is realized in collaboration with other groups, such as JASS with SABER and ALPS. In addition, many analysis projects are conducted in collaboration with the TEG, FunGen, and SABER groups. Furthermore, we are also involved in the activities of the Python Working group (PyWog), and Rachel Torchet, one of our research engineers, is leading the Design Kit group that aims at coordinating the communication material within the Hub.

**Teaching and Training Activities:**

The members of the WINTER group participate on a regular basis to various courses and trainings. We have so far accumulated a number of 138 h of teaching and training. Some courses cover very general topics on biology or bioinformatics (e.g. Pasteur Biology for Non-Biologists and Genomes Analysis Courses), while others are more specifically leveraging on the specific skills of our expert group or the projects we participate in (e.g. Galaxy and CWL trainings).

**Software and tools:**

**Salmonella CRISPR Typing** - Tool to subtype Salmonella strains: https://galaxy.pasteur.fr/root?tool_id=toolshed.pasteur.fr/repos/khillion/salmonella_crispr_typing/salmonella_crispr_typing/1.0.1


**ReGaTE** - software component enabling the automated publication of Galaxy tools and workflows into the ELIXIR Tools and Data Services Registry: https://github.com/C3BI-pasteur-fr/ReGaTE, https://doi.org/10.1093/pjigascience/pnx022


**EDAM ontology (collaboration)** - the ontology of Bioinformatics operations, types of data, data identifiers, data formats, and topics - http://edamontology.org, https://github.com/edamontology/

**CWL standard (collaboration)** - specification for describing analysis workflows and tools in a way that makes them portable and scalable - https://www.commonwl.org/, https://github.com/common-workflow-language/

**Galaxy (collaboration)** - open source, web-based platform for data intensive biomedical research https://usegalaxy.org/, https://github.com/galaxyproject/ (2000 users, who launched more than 440000 jobs)

**Perspectives and strategy:**

Since the creation of the WINTER group, we launched multiple ambitious projects in collaboration with other expert groups from the Hub and research Units of the Institut Pasteur. Meanwhile, our involvement in community projects like Galaxy, and infrastructures like the IFB and ELIXIR, provides us with opportunities of collaborations with key partners, and increases our visibility within the mission of the C3BI. We will continue offering to the scientists of the Institut Pasteur our expertise in the development and deployment of applications, through multiple complementary actions. First, we will offer trainings on the technologies and practices required to use and build robust and sustainable web resources. Second, we will systematically include in our projects activities such as user experience design, continuous testing and integration, to deliver usable and...
robust applications. Finally, we will increase our expertise in database design and management, technical skills which are often needed in our projects and highly demanded within Institut Pasteur scientific community.

Grants and Awards

- ELIXIR Workflow Implementation Study: Institut Français de Bioinformatique, 10-month salary: J. Raciazek, PIs: R. Finn (EBI) J. Chilton (PSU), M. Crusoe (CWL project), H. Ménager
- IFB Resources Catalog: Institut Français de Bioinformatique, 1.5 years salary: L. Bouri, PIs: J. Van Helden (IFB), H. Ménager

Publications of WINTER expert group:


3.5 Hub Expert Group ALPS: Algorithmics and programming in science


General Overview

Team members (As of July 14, 2018):
4 Research engineers including
- Hub core (2):
  - Pierre Lechat
  - Yoann Dufresne
- Embedded in research Units (2):
  - Bertrand Néron (Microbial Evolutionary Genomics 40%)
  - Nicolas Maillet (Structural Virology - 80%)
- Former members:
  - Paul STRETENOWICH (M.Sc. students)
  - Baptiste CAPBLANCQ (M.Sc. students)
  - Alexandre UYTTENHOVE (M.Sc. students)

Keywords: Algorithmics, Development, Visualisation, Data Structure, Scaling

Outlook and highlights

ALPS expert group is composed of 4 bioinformaticians and uses advanced computer skills in programming, algorithmics, visualization and software development to help addressing issues of other Pasteur’s teams by creating new software and methodologies. Scaling complex computation problems and visualizing results are key points in current research, especially in human health where the time factor can be of vital importance. Within the last 2 years, the group has developed 9 computational tools and 1 database, and has participated in 12 publications, in various biological fields. Our implication in software development is sustained and 3 software will be included in publications in the near future, most of those have been developed in collaboration with other Hub groups. ALPS members are actively involved in teaching activities, the overall teaching load for our expert group totals up to 200 hours of lecturing within the past two years within Institut Pasteur teaching program.

Main Projects

- Antibody sequencing by MS/MS: Automation of analysis. PI: Bertrand SAUNIER, Unit ‘Structural Virology’, dpt. ‘Virology’. In development,
- Prediction of proteases-induced cleavage sites on amino acid sequences, internal HUB project. https://gitlab.pasteur.fr/nmaillet/rpg, manuscript in preparation.

Figure 1: the 24 achieved projects of ALPS. Projects submitted by a Pasteur department are represented on the left side of the panel. Specific projects, depending on C3BI, Institut Pasteur International Network (RIIP) or other centres (like Pasteur Platforms), are on the right. "Achieved projects" corresponds to closed projects with publication, manuscript submitted, manuscript in preparation or no publication expected. Number of publications is indicated above bars.
Introduction

ALPS expert group uses advanced computational skills in different areas of computer science: programming, algorithmics, data structure, grammars, visualization, development, etc. The main goal of ALPS is to use this wide range of skills to develop new tools and methodologies that will help various research Units at the Institut Pasteur to address important biological problems. In fact, the main target users of our software are not only experts in bioinformatics but also and mainly researchers in biological and medical science within and beyond the Pasteur campus. Our skills and expertise lie within the conception and development of open source software dedicated to specific questions, the design of visualization tools providing easy access to relevant information, the creation of new algorithms to address challenging problems and scaling (memory / time) to cope execution time of complex problems with classical computers. We are committed to follow high standards for software development with, for example, continuous integration, automatic on-line documentation and good availability of source (GitHub, pip, etc).

So far, ALPS was involved in 34 projects leading to 9 computational tools and 1 database, and has participated in 12 publications.

Results since the conception of the expert group, three illustrative examples

Counter RNAseq Window: Identification of promoter-overlapping antisense transcription

PI: Gwenaël Badris-Breard, Unit ‘Genetics of Macromolecular Interactions’, department of Genomes and Genetics.

Gwenaël Badris-Breard, from the ‘Genetics of Macromolecular Interactions’ team, is designing experiments to highlight how regulation of pervasive transcription by antisense transcription plays a major role in gene regulation of budding yeast. This effect has already been shown for some genes. To do so, they are creating mutants that stabilize asRNA. In these mutants, hidden pervasive transcripts are escaping the nuclear NNS surveillance and then accumulating in the cytoplasm: they can thus be quantified by RNAseq. In order to analyse these experiments, the team needed to quantify the asRNA that hybrid around a specific region (here the Transcription Start Site or TSS), and then distinguish the coverage on each strand. As no software, to the best of our knowledge, fulfills these specifications, we decided to create a dedicated one. The result is a package named CRAW (Counter RNA seq Window) composed of two scripts. The first one (craw_coverage) computes the coverage on each strand around a reference position, here the TSS. All coverages are then aligned using this reference position. The second script (craw_htmp) graphically represents these coverages. Data from the RNAseq experiments were analysed with the CRAW package and show that TSS-overlapping antisense-mediated transcriptional interference is a frequent mechanism used in full gene repression.

SynTView: A tool to integrate and visualize large scale data for the comparative genomics analysis of bacterial populations

PI: Carmen Buchrieser, Unit ‘Biology Of Intracellular Bacteria’, dpt of Genomes and Genetics

SynTView is a multi-view genome browser for next-generation comparative microorganism genomics. The software is characterised by the presentation of syntenic organisations of microbial genomes and the visualisation of polymorphism data obtained from next generation sequencing. This web tool is very helpful in the analysis of a large number of strains, bringing together phylogeny, polymorphisms, larger variants such as indels, coverage, as well as functional annotations and strains meta-data. It allows dynamic interaction and flexibility between many specialised views. In the study of the clones of Legionella pneumophila (David S et al., Genome Research, 2016) a circular representation of the five major disease-associated STs with each genome shown as a concentric circle allows to identify the main recombined regions (that belong to higher density of SNPs) between the clones and the five major STs. So far, SynTView has been involved in 4 projects/publications (cited by 9 publications).

Rapid Peptides Generator: In silico protein digestion

HUB, Center for Bioinformatics, Biostatistics and Integrative Biology

Proteases, also known as proteolytic enzymes, have been studied for more than 80 years. Those enzymes are widely used in industry, medicine and as a biological research tool, for example in protein characterization or more generally in proteomics and proteogenomics. Recently, interest in proteases has gained importance due to advancements in mass spectrometry techniques used in proteomics and proteogenomics. In “bottom-up” analysis, using tandem mass spectrometry (MS/MS), optimal peptide size range is 600–5,000Da (3) when proteins size is usually more than 10000 Da. Therefore, for bottom-up approaches, protein digestions are required. Performing multiple digestions can increase overall confidence in protein identification if cleaving sites are different. Yet, it is not always easy to determine which combination of enzymes will lead to a
set of peptides suitable for MS/MS analysis. RPG is dedicated to predict proteases-induced cleavage sites on sequences. Currently, 42 enzymes and chemicals are included in RPG. The main innovation of this software is the possibility for users to easily design new enzymes, using a simple yet powerful grammar. This grammar allows the user to design complex enzymes like trypsin or thrombin, including many exceptions and different cleavage sites. User-defined enzymes are then interpreted by RPG and included in the local installation of the software. For a complete list of the projects of the ALPS expert group see Appendix 3.8.

**Platform Services and Activities**

The ongoing project “Antibody sequencing by MS/MS: Automation of analysis is in collaboration with the Proteomics platform, led by Mariette Matondo. This project is in the main topic of Emerging Diseases of Institut Pasteur.

**Inter-group Activities**

Many ALPS projects are collaborative with other Hub groups, for example JASS, VarView and Beautiful Pattern Matching projects mentioned earlier but also SHAMAN, Listeriomics, etc. ALPS is actively collaborating with all other Hub groups. Apart from those projects, all ALPS members are involved in the Python development working group (Pywog) led by Thomas Cokelaer and Bertrand Néron (ALPS). Some ALPS members also participate to the Design system group, led by Rachel Torchet. Recently, YD started a journal club on Algorithmics, where we can present in a flash format (whiteboard) new methodologies.

**Teaching and Training Activities**

ALPS group is actively and continuously involved in teaching activities. During the past two academic-years, we have contributed in the curriculum and teaching of various courses, for a total of 215 hours. The focus of our teaching activities is computer-science (Unix, Python, Data Structures, etc): a number of the courses taught have been designed by ALPS members. Moreover, we have given courses in metagenomics, assembly and NGS data analysis. ALPS was also involved in the running of the mandatory course “Introduction to Data Analysis”, for all new Pasteur PhD students. Finally, we supervised 4 students from engineering schools, three in the second year and one in the third year of studies.

**Software and tools**

CRAW: Counter RNA seq Window is a package to compute and visualize the coverage of RNA seq experiments. ([https://gitlab.pasteur.fr/bneron/craw](https://gitlab.pasteur.fr/bneron/craw))


RPG: tool dedicated to predict proteases-induced cleavage sites on amino acid sequences. ([https://gitlab.pasteur.fr/nmaillet/rpg](https://gitlab.pasteur.fr/nmaillet/rpg))

IntegronFinder: bioinformatics tool to finds integrons in DNA sequences. ([https://github.com/gem-pasteur/Integron_Finder](https://github.com/gem-pasteur/Integron_Finder))

JASS: online tool for the joint analysis of GWAS summary statistics. ([http://jass.pasteur.fr](http://jass.pasteur.fr))

MacSyFinder / MacSyView: tools for the detection of macromolecular systems in protein datasets using systems modelling and similarity search. ([https://github.com/gem-pasteur/macsyfinder](https://github.com/gem-pasteur/macsyfinder))

CRISPRCasFinder: tool and a database allowing the identification of both CRISPR arrays and Cas proteins. ([https://crisprcas.i2bc.paris-saclay.fr](https://crisprcas.i2bc.paris-saclay.fr))


BPM: tool for rapid bacterial strain identification. ([https://gitlab.pasteur.fr/ydufresn/bpm](https://gitlab.pasteur.fr/ydufresn/bpm))

Listeriomics: interactive web platform to browse genomic, transcriptomic and proteomic datasets in Listeria ([https://listeriomics.pasteur.fr](https://listeriomics.pasteur.fr))

**Technology transfer**

SynTView: Dynamic Browser for Microbial Synteny and Polymorphism Information (sell to GENOSTAR)

GenoList: integrated environment for comparative analysis of microbial genomes (installed in DANONE with their own organisms)

**Perspectives and strategy:**

ALPS perspective for next years is dual: increasing the overall capacity in computer-science of researchers and identifying and addressing the global needs of the Pasteur community. The last C3BI retreat allowed to identify a need for strengthening programming skills within the researcher community (see Appendix 1.8). In order to address this, ALPS group plans to design an innovative course on algorithmic, based on peer-programming and self-training: each course would consist in a 2-hour session where a problem would be stated and each person, organized in teams of two in peer-programmers, will try to solve it. At the end of the session, all solutions would be analysed, and key points in algorithmic would be highlighted. Those courses would also be a good opportunity to promote good programming practices. We also plan to organize some hackathons on
specific topics, such as new languages, new technologies, etc. Finally, ALPS team will continue its effort in fulfilling the needs of research teams in Pasteur, not only by addressing submitted issues, but also by identifying new ones. Through its tight links with the other expert groups of the Hub and multiple teams on the campus, ALPS is in a unique position to identify main issues and needs in terms of software development and to address them by ad-hoc development, technologies and methodologies.

**Publications of ALPS EXPERT GROUP**


3.6  Hub Expert Group FUNGEN: FUNctional GENomics

Natalia Pietrosemoli, natalia.pietrosemoli@pasteur.fr
https://research.pasteur.fr/en/team/expertise-group-fungen-functional-genomics/

General Overview

Team members:
6 research engineers including

Hub core (3):
- Stéphane Descorps-Declère
- Damien Mornico
- Natalia Pietrosemoli

Embedded in research Units (2):
- Etienne Kornobis (Epigenetic Regulation - 40%; Functional Genetics of Infectious Diseases - 40%)
- Corinne Maufrais (Fungal Biology and Pathogenicity - 40%; RNA Biology and Fungal Pathogens Unit - 40%)

Detached in the following Unit (1):
- Varun Khanna (Genetics of Macromolecular Interactions - 50%)

Intern M.Sc (1):
- Maëlle Daunesse (2-year program).

Alumni (2):
- Louis Jones (RE).
- Maxime Borry (M.Sc. student).

Keywords: functional analysis, structural and functional annotation, comparative genomics, multi-omic data integration, structure and genome dynamics, teaching, project management system.

Outlook and highlights

Functional Genomics (FunGen) is a group of research engineers whose objective is to provide support to the C3BI and the Institut Pasteur (IP). Prior to the creation of the C3BI, some FunGen members previously worked in other IP research Units, thus, their integration facilitated the transfer not only of technical expertise but also of valuable organizational knowledge specific to research activities at IP. FunGen provides IP’s scientific community a high level of expertise in several bioinformatics domains. We mainly analyse -omics data derived from high throughput sequencing (HTS) projects: from raw sequences to the gene (and protein) function. Our goal is to help scientists disentangle the complex relationship between genotype and phenotype at different scales: from single mutations to complete genomes. Our main areas of expertise include structural and functional annotation, comparative genomics, genome fluidity and evolution. In addition to these services and research activities, we are highly involved in several collective tasks, including the development and maintenance of the Hub’s projects management system (BISTRO), and a seminar management tool to disseminate C3BI’s seminar’s information. Finally, we coordinate many of C3BI and IP courses. To date, we have collaborated on 61 projects with 38 IP research Units resulting in 23 publications. We have also co-organized and taught in 15 training programs (about 190 hours per year).

Main Projects

- **Candida albicans** genome diversity: strengthening existing resources for re-sequencing, developing new resources for haplotype mapping. Christophe D’Enfert. Fungal Biology and Pathogenicity; Department of Mycology [CM; 3 years; 2 publications, 1 book chapter]
- **Molecular analysis of muscle stem cells.** Shahragim Tajbakhsh. Stem Cells and Development; Department of Developmental and Stem Cell Biology. [NP; 1 year; 1 first author publication; one interactive web tool]
- **High throughput CRISPR-Cas9 screen in bacteria.** David Bikard. Synthetic Biology; Department of Microbiology. [VK; 1 year; 1 publication]

Figure 1: The 61 achieved projects by FunGen. The 61 projects by FunGen. Number of projects submitted by the 11 departments of the Institut Pasteur, C3BI, Institut Pasteur International Network (IPIN) or other centres (such as Institut Pasteur Platforms) are represented on the y axis. Number of publications is indicated above bars. In addition to these 19 publications, four papers (LJ, EK), one tool (EK) and one book chapter (NP) resulted from external collaborations.
Introduction:
FunGen is one of the 6 expert groups of the Hub. The main mission of FunGen is to offer the wide range of expertise of its members (biologists, bioinformaticians, engineers, biophysicists) related to functional genomics in order to provide high quality service to IP’s community. FunGen’s projects encompass a wide variety of methodological and technical approaches, including handling and analysis of NGS data (whole genome, RNA-seq) and the corresponding down-stream analysis of the enriched biological pathways. Additionally, we perform syntactic, functional, and relational annotation of genomes, as well as variant calling (e.g. identification of single nucleotide polymorphisms (SNPs) and structural variants (SVs) using de novo genome assembly and read mapping strategies). Our projects also include orthologous inferences and comparative genomics studies. We are currently interested in combining multi-omics data to help uncover biological signatures, which emerge sometimes only after combining different layers of biological information and following careful examination from the functional perspective. Our analyses include not only state-of-the-art languages and tools, but also our own ad-hoc developed software solutions. As part of the Hub and C3BI missions, we contribute in developing and maintaining the online project management system and the information management system that helps automatically disseminate information regarding the bi-monthly seminars of the C3BI. Finally, we coordinate and teach many of IP and C3BI trainings.

Results since the conception of the expert group, three illustrative examples

*Candida albicans* genome diversity: strengthening existing resources for re-sequencing, developing new resources for haplotype mapping. PI: Christophe D’Enfert. RU: Fungal Biology and Pathogenicity. D: Mycology.

*C. albicans* is responsible for the majority of life-threatening fungal infections occurring in hospitalized patients. *C. albicans*’s population includes at least 18 genetic clusters whose origin remains unclear. This project comprised the analysis of the genomic sequences of 182 *C. albicans* isolates, constituting the largest dataset to date for this major fungal pathogen. C. Mauvais developed a pipeline to thoroughly analyse Illumina resequencing data, which was used for 1) analysing additional isolates in the frame of an international project aimed at establishing genotype-phenotype correlations; and 2) for characterizing isolates that have undergone in vitro or in vivo evolution (Ropars et al. 2018) (d’Enfert et al. 2017) (Feri et al. 2016). A third manuscript, “Generating genomic platforms to study *C. albicans* pathogenesis” is currently under revision for publication in Nucleic Acids Research. CM is also responsible for the analysis of long-read sequencing data (PacBio) for strains representative of the major genetic clusters and define the haplotypes for these strains.


This project analysed gene expression profiles of activated and quiescent states of mouse satellite cells to define a consensus molecular signature of the quiescent state. A large compendium of expression data, derived from nine microarray datasets (three novel microarray datasets and six publicly available datasets) offered the first comparison and integration of independent studies of the quiescent state of mouse satellite cells. Comparison of available data offered challenges related to the inherent diversity of datasets and biological conditions. Thus, N. Pietrosemoli developed a standardized workflow to homogenize the normalization, filtering, and quality control steps for the analysis of gene expression profiles allowing the identification of differentially expressed genes and the subsequent gene set enrichment analysis (Pietrosemoli et al. 2017). The publication includes interactive web tool called Sherpa (SHiny ExploRation tool for transcriptomic Analysis) to provide comprehensive access to the individual datasets analysed in a homogeneous manner. This server allows users to (i) identify differentially expressed genes of the individual datasets, (ii) identify the enriched gene sets of the individual datasets, and (iii) effectively compare the chosen datasets. Sherpa is adaptable and serves as a repository for the integration and analysis of future transcriptomic data.

High throughput CRISPR-Cas9 screen in bacteria. PI: David Bikard. RU: Synthetic Biology. D: Microbiology.

This project analysed data from a very recent technology which has led to many breakthroughs in genome editing and the control of gene expression: Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR) derived from the bacterial immune system, and the associated cas genes. In particular, the Cas9 protein has proven to be a very versatile RNA-guided
nuclease. Here, a genome-wide library of guide RNAs was used to direct the catalytically dead Cas9 (dCas9) to block gene transcription in *Escherichia coli*. The results showed that the last 5 nucleotides of guide RNAs could induce strong fitness defects independently of the rest of the sequence or the target position. These results also highlight the fact that off-targets with as little as nine nucleotides of homology to the guide RNA can strongly block gene expression. Altogether this study provides important design rules to safely use dCas9 in *E. coli* (Cui et al. 2018a).

For a complete list of the projects of the FunGen expert group see Appendix 3.9.

**Platform Services and Activities:**

FunGen actively participates in the submission of genomic data to databanks, sequencing files (fastQ) and annotated assemblies (EMBL files); a mandatory step in the publication process of studies undertaken by IP research teams of the Institut Pasteur. Thus, over 300 bacterial strains or viral genomes have been published in the European databases of the EBI (ENA), NCBI and DDBJ with the help of our group.

D. Mornico, in collaboration with R. Torchet (WINTER expert group) and C. Malabat (Hub), develops and maintains the Hub projects’ management system (BISTRO). This is a very complex and versatile system designed to accurately track the real-time progress of all the submitted projects. Since its development, BISTRO has handled all information related to over 280 projects. D. Mornico also developed a seminar management tool designed to optimize the dissemination related to C3BI’s bi-monthly seminar information and to manage the flow of external guests into the IP’s campus.

N. Pietrosemoli has closely collaborated with the Proteomics Platform (M. Matondo) at IP and Q. Giai Gianetto (statistician detached to the proteomics platform) in several projects that involved pipeline development and benchmarking for the differential and functional analysis of proteomic data.

**Teaching and Training Activities:**

FunGen holds monthly trainings open to the international scientific community. So far, we have co-supervised several B.Sc., M1 and M2 students, and upon individual requests, we have also prepared ad-hoc trainings for PhD students and researchers from research Units of IP interested in learning complex bioinformatics analyses (C. Picard; analysis of NGS data, J. Schaeffer, E. Prina, H. Lecoeur; functional analysis of transcriptomic data; L. Audebert; analysis of NGS data, J. Schaeffer, E. Prina, H. Lecoeur; functional analysis of proteomic data).

Inter-group Activities:

FunGen actively collaborates with TEG (teaching, functional analysis), SABER (omic data integration), WINTER (Hub and C3Bis project and seminar management), but exchange with all groups on a regular basis. Some FunGen members also participate in collaborative projects, which involve members of several Hub groups as well as research engineers attached to C3Bi’s research Units, such as Bioconda (https://bioconda.github.io/) and Bioconvert (https://github.com/biokit/bioconvert).

**Software and tools:**

- **coGSEA:** comparative Gene Set Enrichment Analysis. R package Gene Set Enrichment Analysis and comparison among methods [coGSEA](https://github.com/maxbor/coGSEA).
- **BIOCONDA:** channel for the [conda](https://github.com/bioconda/bioconda) package manager specializing in bioinformatics software. [https://bioconda.github.io/](https://bioconda.github.io/). Published in Nat. Methods 2018. doi: 10.1038/s41592-018-0046-7
- **BISTRO:** Beautiful InterFace and System To Report and Organize [https://c3bi.pasteur.fr/hub/projects/?projects=table](https://c3bi.pasteur.fr/hub/projects/?projects=table).
- **Seminar management tool:** [https://c3bi.pasteur.fr/seminar-registrations/](https://c3bi.pasteur.fr/seminar-registrations/)

**Grants:**

In the framework of a collaboration between the Microbiology and Cell Science Dept. of University of Florida, the LABGeM Génoscope of the National Sequencing Centre (CEA) and the Hub, FunGen was granted a *Workshop Fellowship* from the INCEPTION programme to partially finance a workshop that we organized: “Linking Gene and Function by Integrative Approaches”.

**Perspectives and strategy:**

As an immediate goal, FunGen members will devote many efforts towards publishing 9 manuscripts currently under review, including results from a functional analysis of proteomic data on the Rift Valley Virus, the functional study of transcriptomic data from Ebola infected patients, *de novo* sequencing and analysis of mycobacteria, and a comparative analysis of the gene expression in the presence of infection.
expression profiles of human cell clones to understand the contribution of SUMO protein as a multifaceted modifier of chromatin.

We are fully committed to responding to IP’s needs on bioinformatics and biostatistics. Indeed, as a result of an on-going collaboration, we have identified the potential need within several IP research Units of expertise in the analysis of metabolomics data. To this end, we plan to enrol several FunGen members in trainings related to this topic. We will also renew our engagement in disseminating bioinformatics and biostatistical knowledge in IP’s community by proposing novel courses, for example, on good programming practices applied to bioinformatics and reproducible research.

**Publication list of FunGen expert group:**


### Hub Expert Group GIPhy: Genome Informatics and Phylogenetics

**Alexis Criscuolo, alexis.criscuolo@pasteur.fr, [https://tinyurl.com/0-GIPhy-0](https://tinyurl.com/0-GIPhy-0)**

### General Overview

**Team members:** 3

3 Research engineers including:

- **Embedded in research Units (3):**
  - Alexis Criscuolo (PIBnet – 80%)
  - Julien Guglielmini (PIBnet - 40%; Molecular Biology of Gene in Extremophiles - 40%)
  - Thomas Bigot (Biology of Infection – 80%)

**Keywords:** Phylogenetics, Evolution, Genomics, Classification, Genotyping

### Outlook and highlights:

The expertise group ‘Genome Informatics and Phylogenetics’ (GIPhy) is composed of 3 research engineers. It is highly involved in scientific research topics focusing on biological classifications. Therefore, projects regarding important themes such as systematics, taxonomy, homology and related fields are specifically addressed by the members of this dedicated group. Due to the importance of classification for pathogenic bacteria, every member of GIPhy is contributing significantly to the public health mission of the Institut Pasteur (e.g. PIBnet initiative) with the main purpose of modernizing the activities of the associated Units and valorising their biological resources. Since its creation, the group GIPhy has collaborated on 20 scientific projects submitted by 14 Units, from 8 of the 11 departments of the Institut Pasteur, leading to 17 publications in peer-reviewed international journals (see Figure 1 and publications list below).

### Representative Projects

- **Pasteur International Bioresources Network (PIBnet) bioinformatics:** whole-genome sequencing of microbial agents for disease surveillance, outbreak investigation, epidemiology and population biology. Principal investigator (PI): Sylvain Brisse, Unit ‘Biodiversity and Epidemiology of Bacterial Pathogens’, dpt. ‘Infection and Epidemiology’ [AC 3.5 years; JG 2.5 years; 11 publications in peer-reviewed international journals; 2 websites]
- **Identification of new or unexpected pathogens, including viruses, bacteria, fungi and parasites associated with acute or progressive diseases.** PI: Marc Eloit, Unit ‘Biology of Infection’, dpt. ‘Cell Biology and Infection’ [TB 2 years; 1 database]
- **Evolutionary relationships between giant viruses and eukaryotes.** PI: Patrick Forterre, Unit ‘Molecular Biology of Gene in Extremophiles’, dpt. ‘Microbiology’ [JG 2.5 years; 2 international conferences]

![Figure 1: The 20 achieved projects of GIPhy. Number of projects submitted by the 11 departments of the Institut Pasteur, C3BI, Institut Pasteur International Network (IPIN) or other centres (such as Institut Pasteur Platforms) are represented on the y-axis. Number of publications are indicated above bars. In addition to these 15 publications, AC published 2 other papers (Antunes et al. 2016, Yazouli et al. 2017) from collaborations that are not based on Hub projects.]
Detailed description

![Timeline of people hired by the Institut Pasteur to join the group GIPhy](Image)

**Figure 2: Timeline of people hired by the Institut Pasteur to join the group GIPhy**

**Introduction:**

GIPhy expert group (3 research engineers recruited from 2014 to 2016; see Figure 2) has developed its expertise to help and assist the IP laboratories for performing different types of bioinformatics analyses, e.g. assembling genomes in order to better explore the diversity of life from a systematic point of view (Nicholson et al. 2017), building core-gene sets leading to robust genotyping schemes (Moura et al. 2016, Bouchez et al. 2018), determining the presence/absence pattern of specific genes (Drini et al. 2016), implementing bioinformatics strategies for characterising new species (Hurtado-Ortiz et al. 2017, Criscuolo et al. 2018), or performing phylogenetic analyses to infer accurate gene or species trees (Dahmane et al. 2016, Perrin et al. 2017). The important involvement of the group in projects related to public health demand has played a significant role in improving the missions of several public health centres, e.g. epidemiological surveillance (Moura et al. 2017), reference characterisation of pathogenic strains (Breurec et al. 2016, Garcia-Hermoso et al. 2018), evolutionary studies (Maury et al. 2016), or pathogenesis research (Mazuet et al. 2017, Yazouli et al. 2017).

**Results since the conception of the expert group: three illustrative examples**

**Pasteur International Bioresources Network (PIBnet) bioinformatics: whole-genome sequencing of microbial agents for disease surveillance, outbreak investigation, epidemiology and population biology.**

**PI:** Sylvain Brisse, Unit ‘Biodiversity and Epidemiology of Bacterial Pathogens’, dpt. ‘Infection and Epidemiology’. [AC, JG]

Since 2014, the PIBnet initiative is a joint effort to upgrade the activities of different Units — National Reference Centres (NRC), World Health Organization Collaborating Centres (WHOCC), Collection of the Institut Pasteur (CIP), Laboratory for Urgent Response to Biological Threats (CIBU)— including biological resource value-creation and microbial characterisation approaches. Half of the GIPhy workforce over the past years has been involved in this initiative with the objective to facilitate the transition of the involved Units toward the regular use of whole genome sequencing data. The implication of the GIPhy group has led to the improvement of the daily epidemiological surveillance activity of several NRCs (development of new methods for inferring and exploiting genotyping schemes). To this aim, core-gene Multi-Locus Sequence Typing (cgMLST) schemes have been built for different pathogenic bacteria (e.g. *Listeria monocytogenes*, *Elizabethkingia anophelis*, *Bordetella pertussis*) and have been made publicly available via a website for typing new strains (Institut Pasteur MLST; see below).

Moreover, the showcase of a large number of bacterial strains kept by the CIP has been achieved through the collection of the Institut Pasteur (CIP), Laboratory for Urgent Response to Biological Threats (CIBU) — including biological resource value-creation and microbial characterisation approaches. Half of the GIPhy workforce over the past years has been involved in this initiative with the objective to facilitate the transition of the involved Units toward the regular use of whole genome sequencing data. The implication of the GIPhy group has led to the improvement of the daily epidemiological surveillance activity of several NRCs (development of new methods for inferring and exploiting genotyping schemes). To this aim, core-gene Multi-Locus Sequence Typing (cgMLST) schemes have been built for different pathogenic bacteria (e.g. *Listeria monocytogenes*, *Elizabethkingia anophelis*, *Bordetella pertussis*) and have been made publicly available via a website for typing new strains (Institut Pasteur MLST; see below).

**Identification of new or unexpected pathogens, including viruses, bacteria, fungi and parasites associated with acute or progressive diseases.**

**PI:** Marc Eloit, Unit ‘Biology of Infection’, dpt. ‘Cell Biology and Infection’. [TB]

In the context of the recent public health trends aiming at exploiting high-throughput sequencing (HTS) data for medical purposes, this project started in 2016 with the ambition of developing efficient bioinformatics procedures for determining the pathogens that are associated with diseases of unknown aetiology. First focusing on viral pathogens, TB has been involved in this project for developing a pipeline that is able to efficiently determine the taxonomic composition of HTS reads of viral origin. High accuracy has been reached through the development of a novel database (RVDB-prot; see below). This database has been implemented in collaboration with members of the Center for Biologics Evaluation and Research (CBER, U.S. FDA) to improve virus detection from HTS data. In collaboration with several Units (IPIN, Parisian hospital Units - APHP), the inferred pipeline is currently in its test phase to identify known viruses within unexpected species, particularly those in contact with human (e.g. blood-sucking arthropods, bats).

**Evolutionary relationships between giant viruses and eukaryotes.**

**PI:** Patrick Forterre, Unit ‘Molecular Biology of Gene in Extremophiles’, dpt. ‘Microbiology’. [JG]

The phylogenetic position and status of NucleoCytoplasmic Large DNA viruses (NCLDV), to which belong the famous “giant viruses”, is a trendy topic leading to controversial debates among taxonomists, evolutionists and virologists. As the recently
increasing number of available NCLDV genomes opens new opportunities to study their evolutionary history and relatedness within the overall living world, JG has been involved since 2016 in this project in order to achieve the systematic revision of these viruses. Comparative genomic studies of these fast-evolving and poorly annotated genomes have allowed the inference of a robust core-gene set that was used for reconstructing and updating species trees. These results, presented in 2 international scientific conferences, now open the way to promising studies that will contribute to the resolution of current debates, such as the monophyletic status of the NCLDV, their contribution to early eukaryotic evolution, or their usefulness for rooting the phylogenetic tree of eukaryotes.

Other projects. [AC, JG, TB]

The GIPhy group has also collaborated (or is currently collaborating) in 17 other projects, mainly related to sequence classification and phylogenetic inference, involving 11 Units distributed among 8 of the 11 departments of the Institut Pasteur and leading to 4 publications in international peer-reviewed journals (see Figure 1 and publication list below). For a complete list of the projects of the GIPHY expert group see Appendix 3.10.

Inter-group Activities:

Six scientific achievements were conducted in collaboration with other expertise groups of the Hub (ALPS: 3 [NM, PL, YD]; FUNGEN: 2 [CM, LI]; WINTER: 1 [ODA, FM]) they have led to 1 co-authored poster to a scientific conference, as well as 3 publications in international peer-reviewed journals (Drini et al. 2016, Moura et al. 2016, Lago et al. 2017).

Platform Services and Activities:

In the context of the PIBnet program (see above), the dedicated Mutualized Platform of Microbiology (P2M) was created for allowing the different involved Units to have rapid and adequate HTS capacity in their daily mission of epidemiological surveillance. As the GIPhy group is the main provider of bioinformatics solutions for the PIBnet members, AC takes in charge the analysis of the HTS data generated by the P2M (sequencing read pre-processing, de novo genome assembly, sequencing accuracy assessment). Currently, the P2M is sequencing ~900 samples per month, leading to the analysis and distribution of ~11,000 genomes/year since its creation (2015, Feb.).

Teaching and Training Activities:

The GIPhy group (AC, JG) has participated in teaching phylogenetic inference methods and techniques in the Institut Pasteur (Parisian campus and RIIP).

Software, tools, databases, websites:

- **Institut Pasteur MLST**: a website that hosts databases of the (core-genome) MultiLocus Sequence Typing (MLST) schemes implemented for strain nomenclatures in the context of the PIBnet program (bigpdb.pasteur.fr)
- **RVDB-prot**: reference viral coding sequence and associated HMM database developed for enhancing virus detection from HTS data (rvdb-prot.pasteur.fr)
- **PhyloM: bacteria**: reference phylogenetic markers (along with multiple sequence alignments and position specific scoring matrices) that are well-suited for bacterial tree reconstruction (giphy.pasteur.fr/PhyloM/bacteria)
- **GIPhy tools**: a set of bioinformatics tools for sequence and phylogenetic analyses, either dedicated to the Institut Pasteur computing facilities or publicly available (gitlab.pasteur.fr/GIPhy)

Technology transfer:

In collaboration with the group headed by Marc Eloit (see above), an innovative development was performed by one member of the GIPhy group [TB] to infer specific oligonucleotide sets determined from known but worthless genomes (e.g. reagent, host). These sets were provided under confidentiality contract to NuGEN Technologies, Inc., for developing HTS approaches based on prior oligonucleotides representative of genome sequences to be depleted.

Perspectives and strategy:

The expert GIPhy group will continue its main mission of collaboration in order to allow the Units of the Institut Pasteur to benefit from the GIPhy high level of expertise in biological classification. Moreover, its bioinformatics achievements in the field of public health will be decisive in the Institut Pasteur strategic context of emerging disease control and antibiotic resistance studies. The GIPhy group will also continue to participate to the Hub mission of scientific assistance in order to implement practical solutions for the current bioinformatics challenges faced by the research Units of the Institut Pasteur in the domains of genome analysis, phylogenetic inference and bacterial strain typing. Initially developed for the needs of the Institut Pasteur, a significant effort will be made to distribute these programs to the whole scientific community.

Publications of GIPhy expert group:


