Computational Biology Department Days 2021

Detailed Program

Organizing team:

Mélanie Bernardini-Ridel

Rayan Chikhi

Vincent Guillemot

Christophe Zimmer

Table of Contents

Featured Talks

Hybrid Ai For Biomedical Image Understanding - Towards Explainability

Contact: isabelle.bloch@telecom-paris.fr

Isabelle Bloch, Sorbonne Universite, CNRS, LIP6, Paris

This presentation will focus on hybrid AI, as a step towards explainability, more specifically in the domain of spatial reasoning and biomedical image understanding. In this domain, we have to deal with both knowledge and information extracted from images, with a potential semantic gap between symbolic knowledge representation and concrete image information. Besides these knowledge representation aspects, reasoning will be addressed. In particular model-based methods will be presented, relying on structural representations such as graphs, ontologies, formal concepts, logical knowledge bases, endowed with fuzzy models of spatial relations, and on mathematical morphology. Reasoning methods are then based on graph matching, abductive reasoning, constraint satisfaction problems, etc. Applications in biomedical image understanding, expressed as spatial reasoning problems, combined with deep learning, will illustrate the presented topics and the usefulness of combining several approaches.

Deep Learning For Biological Sequences

Contact: jpvert@google.com

Jean-Philippe VERT, Google

In recent years, deep learning has revolutionized natural language processing (NLP), and is increasingly used to analyze biological sequences including DNA, RNA and proteins. While many deep learning architectures and techniques successful in NLP can be directly applied to biological sequences, there are also specificities in biological sequences that should be taken into account to adapt NLP techniques to that context. In this talk I will discuss two such specificities, namely 1) the fact that a double-stranded DNA sequence can be represented by two reverse-complement sequences, and 2) the fact that a natural way to compare homologous biological sequences is to align them. In each case, I will show how the biological constraints can lead to specific models, and illustrate empirically the benefits of incorporating such prior knowledge on several tasks such as protein-DNA binding prediction, protein homology detection or correction of long read sequencing.

Talks

Another Brick To The Bastion Of Bacterial Gwas : Adapt Measures To Infer Relatedness Between Strains

Contact: arthur.frouin@pasteur.fr

Arthur FROUIN, Department of Computational Biology Institut Pasteur ; Fabien LAPORTE, Department of Computational Biology Institut Pasteur ; Mer Molecules Sante Universite Catholique de l'Ouest Angers ; Mylene MAURY, Biology of Infection Unit, National Reference Center and WHO Collaborating Center Listeria Institut Pasteur Inserm U1117 ; Etienne PANTIN, Human Evolutionary Genetics Institut Pasteur UMR 2000 CNRS ; Lukas HAFNER, Biology of Infection Unit, National Reference Center and WHO Collaborating Center Listeria Institut Pasteur Inserm U1117 ; Alexandre LECLERCQ, Biology of Infection Unit, National Reference Center and WHO Collaborating Center Listeria Institut Pasteur Inserm U1117 ; Lluis QUINTANA MURCI, Human Evolutionary Genetics Institut Pasteur UMR 2000 CNRS ; Rayan CHIKHI, Department of Computational Biology Institut Pasteur ; Marc LECUIT, Biology of Infection Unit, National Reference Center and WHO Collaborating Center Listeria Institut Pasteur Inserm U1117 ; Hugues ASCHARD, Department of Computational Biology Institut Pasteur

Genome-wide Association Studies (GWAS) have been central to identify genetic variations associated with complex human phenotypes. There is now tremendous interest for implementing GWAS-like approach to genomes of pathogenic bacteria in order to advance our understanding of infectious diseases. However, bacterial genomes harbour complex structure and long-range linkage disequilibrium (LD), making such analyses extremely challenging.

DBGWAS, a novel method for baterial GWAS, was designed using unitigs derived from compacted De Bruijn Graph nodes, rather than SNP from core genome or kmers, to estimate pairwise genetic relatedness between strain. DBGWAS then use linear mixed model, a widely used tool in human GWAS, for association test between unitigs and phenotype. Because of the specitifies of the complex structure underlying bacteria's genetics, we argue that some parameterizations of DBGWAS, strongly inspired by those used in human GWAS, may be adapted and refined.

In this work, we compared alternative ad hoc algorithms for deriving pairwise genetic relatedness between strain using simulation based on whole genome sequencing of 3718 strains from the MONALISA cohort, a unique prospective cohort that systematically collects listeria strains in France. The simulations showed encouraging resultats for our genetic relatedness measurement methods adapted to bacteria, and highlight the need to adapt the tools developed in human genetics before their application to bacterial GWAS.

Spatial Organisation Of Genomes

Contact: axel.cournac@pasteur.fr

Axel COURNAC, Institut Pasteur, Universite de Paris, CNRS UMR3525, Regulation Spatiale des Genomes, F-75015 Paris, France.

The spatial organization of chromosomes and entire genomes is rich in structures that can be connected to certain functions of the cell such as gene expression and replication. In recent years, in parallel with the use of microscopy techniques, chromosome contact technologies (Hi-C, 3C) have been developed. These approaches, thanks to high-throughput sequencing, allow to quantify the frequencies of physical contact between 2 loci within a chromosome or between different chromosomes. They reveal 3D structures that were previously invisible such as chromosomal loops or domains. I will present several recent or ongoing computational projects of the team such as the use of computer vision to detect and quantify any types of patterns in chromosome contact maps as well as the meta-analysis of various contact data to understand the contact profile of a molecular parasite within chromosomes.

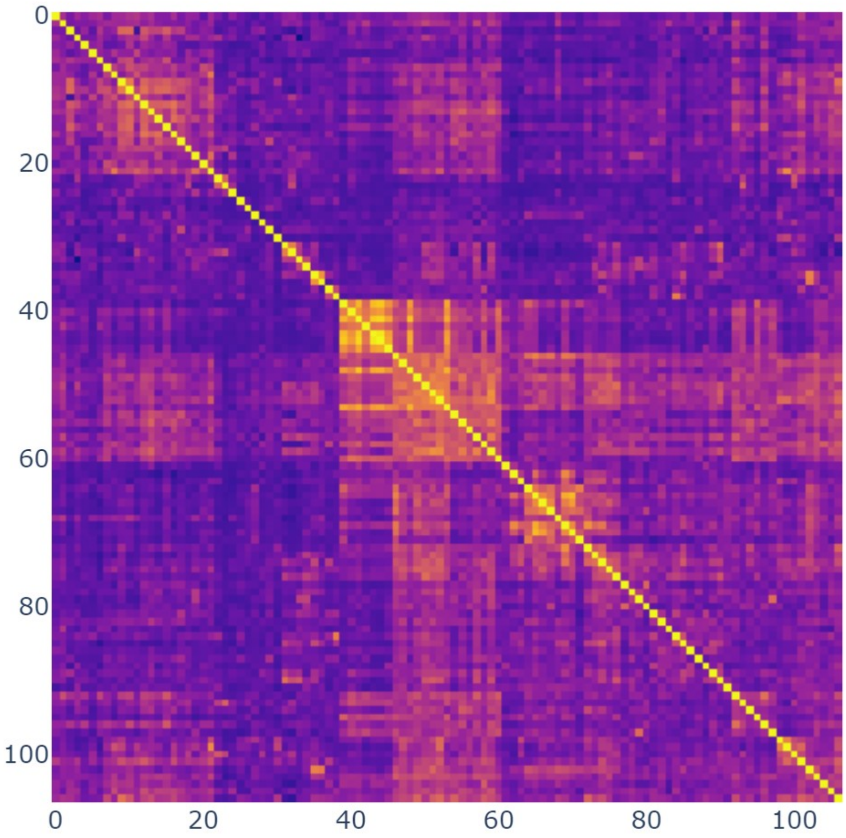
Dissecting The Predictors Of Microbiome Variability

Contact: christophe.boetto@pasteur.fr

Christophe BOETTO, PhD student G5 Statistical Genetics; Hugues Aschard, Head of G5 Statistical Genetics

The human microbiota has become a very active area of research, with hundreds of studies published in the past few years. Major initiative such as the Human Microbiome Project and large-scale genome-wide association studies (GWAS) of the gut microbiota revealed numerous important features associated with changes in the microbiomes. However, these studies also highlighted important challenges in deciphering the complex host-microbiome relationships, and highlighted the need to complement standard univariate analyses with innovative multivariate approaches able to capture the overall microbiome variability.

Here, we propose to study changes in the microbiome composition by focusing on changes in the microbiome correlation matrices conditional on a predictor. We first conducted series of simulation demonstrating that existing matrix comparison methods are invalid when applied to microbiome-like data, displaying severe type 1 error when presented with highly correlated and non-normal data. To address these issues we developed a new statistical test that uses the standard multiple regression framework applied to products of bacteria quantification. We further adapted the approach to address the dimensionality limitation commonly encountered in microbiome data. Altogether the approach shows strong performances in simulated data. It is currently being applied to 1,000 healthy participants from the Milieu Interieur cohort to study the effect of a range of potential factors.



Gut microbiome correlation matrix of people above 50 years old

Temporal Predictions Using Biological Neural Networks

Contact: david.digregorio@pasteur.fr

Alessandro Barri, Giovanni Diana, and David DiGregorio (Institut Pasteur)

The brain is comprised of billions of neurons and trillions of synapses, wired with exquisite precision and beauty. But how can we decipher the brain's messages that make us move, feel, think, and even anticipate the future? It is believed that the brain implements algorithms similar to a computer. Our exquisite knowledge of the brain's wiring diagram inspired artificial networks from which we evolved robust architectures and powerful algorithms (e.g. backpropagation), which no longer resemble biological mechanisms. Can a dialogue between the field of neuroscience and computer science inspire new computer algorithms or give insight into how the brain works?

I will introduce the basic biological building blocks of a biological neural circuit and show an example of a biologically inspired machine learning algorithm. I will show how a simple two-layer perceptron model, when "upgraded" with dynamic synaptic weights, can be used to anticipate upcoming information about the internal and external world. In the discussion period, I hope to discuss whether the utility of our findings, which we think are essential for fine-tuning our movements, could also predict time-series information either in biological systems or beyond (yes, the stock market).

Search For Gene Signatures In Ovarian Cancer

Contact: emil.zakiev@pasteur.fr

Emile ZAKIEV, Johann DREO, Anna Vaharautio, and Benno Schwikowski; HERCULES and PARIS consortiums

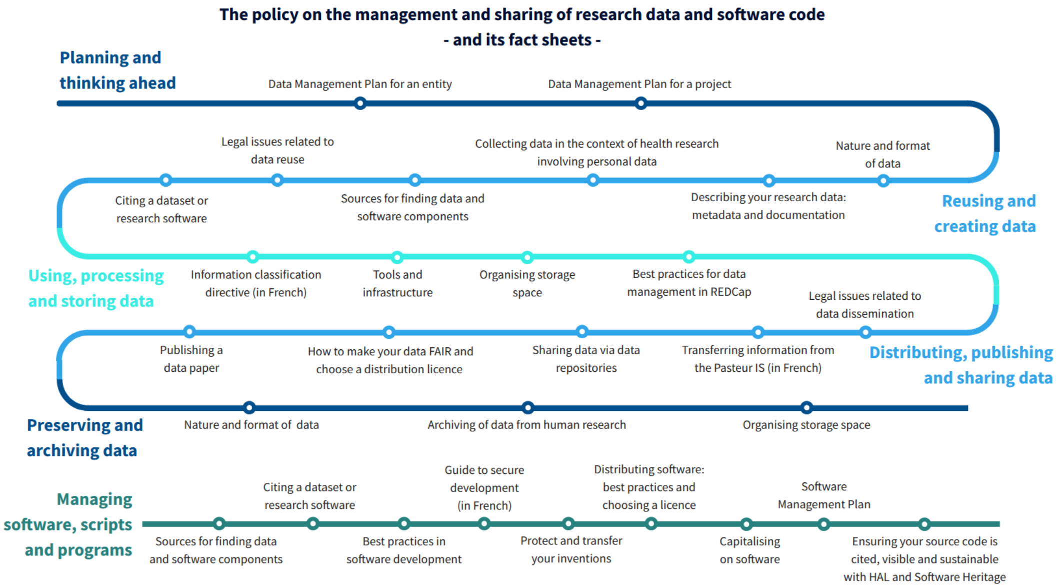
Ovarian Cancer is a complex disease that first manifests with nonspecific pelvic or abdominal symptoms. Common diagnostic tests like transvaginal ultrasonography and serum cancer antigen 125 level are not specific for ovarian cancer. These factors combined make 70 percent of ovarian cancer cases diagnosed at stage III or IV. Despite the recent advancements in the platinum-based chemotherapy treatments, most women diagnosed with ovarian cancer develop recurrent disease and chemother- apy resistance, despite initially responding to treatment. Poly (ADP-ribose) poly- merase inhibitors, while initially promising, fell short of the expectations as resistance to them is ubiquitous in the current clinical practice. Focusing on targetable molec- ular alterations on the level of single-cell transcriptome in the context of personalized medicine is the current go-to approach for deciphering the resistance of ovarian cancer to chemotherapy. We shall call such transcriptomic alterations as signatures. In simple terms, a transcriptomic signature of ovarian cancer is a set of genes that are together implicated in the development of chemotherapy drug resistance. Significant challenge is the batch effect in large-scale single-cell RNA sequencing (scRNAseq) data sets. Data produced at different times contain batch effects that may compromise the integration and interpretation of the data. Building on the clinical expertise of members of HERCULES and PARIS consortiums and boasting an unprecedented set of longitudinal (i.e. before and after chemotherapy) single-cell RNAseq tumor samples, we devised an optimizational computational approach to circumvent the batch effect issues entirely by focusing on each sample individually, while keeping in mind the big picture of recurrent patterns across multiple samples. Join us in our journey to decipher the ovarian cancer chemotherapy resistance signatures, as we wade through the turbulent waters of stochasticity of scRNAseq dropouts. Watch us trying to strike the balance between exploration and exploitation and separate the wheat from the chaff in the world of 40,000 genes and 25,000 cells.

Management And Sharing Of Research Data And Software Code At The Institut Pasteur

Contact: fanny.sebire@pasteur.fr

Anne-Caroline DELETOILLE, Data Management Core Facility ; Fanny SEBIRE, CeRIS

The Data Management Core Facility was created in February 2020. The aim is to pool the needs of the research entities for data managers, to increase their competence in this field and to participate in the structuring of data management on the campus. To facilitate this structuring, a project for the implementation of a policy on the management and sharing of research data and software code was launched in 2019, sponsored by the SGS (scientific secretariat). The policy is official since May 2021. This policy sets out the Institut Pasteur's guidelines on the management and sharing of research data and software code. It summarises the best practices that the Institut Pasteur requires or recommends researchers to implement throughout the research process. In particular, it includes a specific part for code and software management. The aim is to facilitate the sharing and reuse of data and software code, according to the FAIR (Findable, Accessible, Interoperable, Reusable) principles. The policy refers to fact sheets to give scientists the operational resources they need to implement the best practices.


The policy on the management and sharing of research data and software code
- and its fact sheets -
Planning and thinking ahead
Data Management Plan for an entity
Data Management Plan for a project
Legal issues related to data reuse
Collecting data in the context of health research involving personal data
Nature and format of data
Citing a dataset or research software
Sources for finding data and software components
Describing your research data: metadata and documentation
Reusing and creating data
Using, processing and storing data
Information classification directive (in French)
Tools and infrastructure
Organising storage space
Best practices for data management in REDCap
Legal issues related to data dissemination
Publishing a data paper
How to make your data FAIR and choose a distribution licence
Sharing data via data repositories
Transferring information from the Pasteur IS (in French)
Distributing, publishing and sharing data
Preserving and archiving data
Nature and format of data
Archiving of data from human research
Organising storage space
Managing software, scripts and programs
Citing a dataset or research software
Guide to secure development (in French)
Distributing software: best practices and choosing a licence
Software Management Plan
Sources for finding data and software components
Best practices in software development
Protect and transfer your inventions
Capitalising on software
Ensuring your source code is cited, visible and sustainable with HAL and Software Heritage

Overview of the topics covered by the Institut Pasteur's policy on the management and sharing of research data and software code

Hub Presentation + A New Route For Integron Cassette Dissemination Among Bacterial Genomes

Contact: gael.millot@pasteur.fr

Celine Loot, Gael A Millot, Egill Richard, Claire Vit, Jean Cury, Baptiste Darracq, Vincent Parissi, Eduardo PC Rocha and Didier Mazel

The presentation will be divided in two parts: a general presentation of the Hub, followed by an example of Hub support for the team of Didier Mazel. The Hub was created in early 2015 to contribute to the research effort in computational biology at Institut Pasteur and to provide support to the campus in bioinformatics and (bio)statistics. The kind of support depends on the needs of scientists at the IP. Our activities stretch from specific advice on data analysis or experimental planning, up to long-term collaborative projects. Another important role of the Hub is to prepare and deliver training sessions to students, postdocs and permanent staff on the IP campus. Hub members also contribute to the development, adaptation, improvement and benchmarking of methods, tools, and databases either in the context of collaborative projects, or on their own initiatives. As a platform, the Hub also supports external bioinformatics networks, playing an active role in many national and international network instances. The project submitted by Céline Loot in the Didier Mazel team is an example of how the Hub can support Pasteur teams. Integrons are genetic elements found in bacteria and made of genes embedded in several cassettes under the control of a unique promoter. A remarkable feature of the integron system is its capacity to shuffle its cassettes and therefore modulate their expression level, which facilitates adaptation of the bacteria to new/stressful environments. It was also suspected that integron cassette shuffling could promote cassette dissemination into the bacterial genome. To test this assumption, an experimental system was developed such that the Escherichia coli bacteria survive to selective medium only if a designed integron cassette present on a plasmid is excised and is integrated into the bacterial genome. After culture on selective medium, the genome of bacteria was recovered, was PCR enriched for the cassette-genome junctions and was Illumina sequenced. A specific bioinformatic pipeline was developed to adapt to the PCR enrichment step. Results showed that integron cassette is inserted at high frequency in the genome (1e-3 resistant colony), in more than 20,000 different sites, avoiding the promoter and 5' regions of essential genes. Insertion seems dependent on a 5'-GWT-3' consensus site. Thus, integrons can randomly invade the bacterial genomes and behave as transposons. This represents a new route for the bacterial genomic diversification.

Ancient Pathway With Surprising Roles For Metabolic Homeostasis

Contact: ganna.panasyuk@inserm.fr

Ganna PANASYUK, Laboratory of Nutrient Sensing Mechanisms, Inserm U1151/CNRS UMR 8253, Institute Necker Enfants Malades (INEM), Paris 75015, France

Metabolic demands rhythmically fluctuate relying on coordination between nutrient sensing and the circadian clock. We show that class 3 Phosphatidylinositol-3-kinase, best known for its essential role in endocytosis and lysosomal degradation by autophagy, has novel role in assuring metabolic rhythmicity in liver. We demonstrate that class 3 PI3K inactivation in liver manifests in defect of rhythmic nucleotide synthesis. Mechanistically, we discover novel nuclear roles of class 3 PI3K in controlling transcriptional activity of the circadian clock and promoting timed expression of the clock repressor Rev-Erbα as well as enzymes of de novo nucleotide synthesis. Thus, we show that class 3 PI3K nutrient sensing pathway has unexplored roles in fine-tuning transcriptional activity of the circadian clock allowing to maintain energy homeostasis.

Gender-Based Disparities And Biases In Science: Observational Study Of A Virtual Conference

Contact: hanna.julienne@pasteur.fr

Junhanlu Zhang, Institut Pasteur, Universite de Paris, Department of Computational Biology,Bioinformatics and Biostatistics HUB, F-75015 Paris, France ; Rachel Torchet, Institut Pasteur, Universite de Paris, Department of Computational Biology, Bioinformatics and Biostatistics HUB, F-75015 Paris, France ; Hanna Julienne, Institut Pasteur, Universite de Paris, Department of Computational Biology, StatisticalGenetics Group, F-75015 Paris, France

Most scientists would agree that success in science should solely be determined by the merit. However, success in STEM fields (Science, Technology, Engineering and Math) is still profoundly influenced by other factors such as race, gender and socioeconomic status. For instance, numerous studies documented the gender bias throughout the publication process: women publish less than men[1], are less likely to be in the first position among authors who contributed equally[2], and are less cited than men[3]. Gender disparities are also noticeable on less externally constrained behaviours such as the number of questions asked in scientific conferences[4][5]. In this context, quantifying and providing means to alleviate these inequities is of the upmost importance.

As an interdisciplinary team composed of diverse skillsets (anthropology, statistics and UX design), we observed gender asking behaviours during the 2021 JOBIM virtual conference (Journées Ouvertes en Biologie et Mathématiques). We gathered quantitative and qualitative data including: a registration survey with detailed demographic information, post-conference survey on question asking motivations, live observations and in depth interviews of participants. Quantitative analysis highlighted several new findings such as an important fraction of the audience identifying as LGBTQIA+ and an increased attendance of women in virtual JOBIM conferences compared with in person conferences. Notably, the observations revealed a persisting under-representation of questions asked by women (p-value : 3.1e-05). In depth interviews of participants highlighted several barriers to oral expression encountered by gender minorities in STEM. Examples of recurring themes in interviews were negative reactions to woman speech, discouragement of gender minorities to pursue a career in research and sexual harassment.

Informed by the study, a set of guideline for conference organizers has been written along with a scientific manuscript.

G5 Statistical Genetics

Contact: hugues.Aschard@pasteur.fr

Hugues ASCHARD, Hanna JULIENNE, Arthur FROUIN, Christophe BOETTO, Antoine AUVERGNE, Leo HENCHES

The Statistical Genetics G5 work focuses on the development and application of integrative approaches to improve our understanding of the genetic component of multifactorial diseases in human. Here, I will briefly introduce the general context of our ongoing research, and the challenges linked to the high polygenicity of human traits and diseases.

From Single Cells To Populations And Back: Optogenetic Control Of Microbial Communities

Contact: jakob.ruess@inria.fr

Jakob RUESS, Inria Paris and Institut Pasteur

At the single-cell level, biochemical processes are inherently stochastic. For many natural systems, the resulting cell-to-cell variability is exploited by microbial communities, for instance to create bet-hedging strategies or division of labor in isogenic populations. In synthetic biology, on the other hand, cell-to-cell variability is typically seen as a nuisance since it often leads to a lack of robustness and to unexpected outcomes when rationally designed circuits are implemented in live cells. Here, I will present an optogenetic differentiation system in yeast that achieves differentiation of a single strain into genetically distinct subpopulations via recombination-based genetic rewiring. The system operates on population heterogeneity to ensure that the population fraction that recombines is tunable by varying the duration and/or intensity of light stimuli. This enables the creation and dynamic control of synthetic microbial consortia but leads to complex couplings of stochastic single-cell processes and population dynamics. To be able to predict emerging population dynamics from a specification of the single-cell circuit for such coupled systems, I will present a mathematical modeling framework that couples a Kolmogorov forward equation for single-cell processes (chemical master equation or Fokker-Planck equation) to auxiliary population processes such as heterogeneous growth or selection. I will then demonstrate that such multi-scale models can indeed be used to explain and characterize our differentiation circuit and to in silico predict how emerging population dynamics change when cell-to-cell variability is altered by expressing circuit components from plasmids.

Enhancing Single Molecule Localization Microscopy With Deep Learning

Contact: jibai@pasteur.fr

Jiachuan BAI, 1 ; Wei OUYANG, 2 ; Manish SINGH, 1 ; Benoit LELANDAIS, 1 ; Christophe ZIMMER, 1 ; ; 1. Imaging and Modeling Unit, Department of Computational Biology, Institut Pasteur, Paris, France ; 2. Science for Life Laboratory, School of Engineering Sciences in Chemistry, Biotechnology and Health, KTH - Royal Institute of Technology, Stockholm, Sweden

Single-molecule localization microscopy (SMLM) allows to computationally reconstruct super-resolution images with resolutions of ~20 nm, enabling broad applications in cell biology. However, with standard reconstruction methods, SMLM typically requires $\sim 10^{4-10}5$ low-resolution frames to generate a single super-resolution image, hence the temporal resolution of SMLM is very poor. Therefore applications of SMLM to live cells are typically restricted to slow dynamics. To improve the temporal resolution of SMLM, we previously developed a deep learning approach (ANNA-PALM) that reduces the number of required frames to image microtubules or other structures by up to ~100-fold. However, so far, ANNA-PALM was only demonstrated for fixed-cell imaging. Furthermore, ANNA-PALM reconstructions can exhibit artifacts when applied to images taken under different experimental conditions than the training images.

To improve the robustness of ANNA-PALM reconstructions, we first set out to collect a much larger amount and diversity of SMLM training data. For this purpose, we created an online platform (shareloc.xyz) to help share SMLM data acquired by the community. We retrained ANNA-PALM models on 50-92 SMLM images of microtubules obtained by four different teams and obtained reconstructions of significantly higher quality than the original model (trained on only 7 images from one team), as evaluated on data from independent teams.

Extending ANNA-PALM to live-cell imaging is challenging, in part because unlike for fixed cells, ground truth is not available. To address this, we implemented a strategy that creates chimeric data based on fixed cell experimental images and simulated motions. We also explore changes in deep learning architectures to account for dynamic information. We will report ongoing analyses to investigate how reconstruction quality is affected by sampling rate and velocity.

Hub Presentation + A New Route For Integron Cassette Dissemination Among Bacterial Genomes

Contact: marie-agnes.dillies@pasteur.fr

Marie-Agnes DILLIES, Gael MILLOT, Hub de Bioinformatique et Biostatistique, Institut Pasteur

The presentation will be divided in two parts: a general presentation of the Hub, followed by an example of Hub support for the team of Didier Mazel. The Hub was created in early 2015 to contribute to the research effort in computational biology at Institut Pasteur and to provide support to the campus in bioinformatics and (bio)statistics. The kind of support depends on the needs of scientists at the IP. Our activities stretch from specific advice on data analysis or experimental planning, up to long-term collaborative projects. Another important role of the Hub is to prepare and deliver training sessions to students, postdocs and permanent staff on the IP campus. Hub members also contribute to the development, adaptation, improvement and benchmarking of methods, tools, and databases either in the context of collaborative projects, or on their own initiatives. As a platform, the Hub also supports external bioinformatics networks, playing an active role in many national and international network instances. The project submitted by Céline Loot in the Didier Mazel team is an example of how the Hub can support Pasteur teams. Integrons are genetic elements found in bacteria and made of genes embedded in several cassettes under the control of a unique promoter. A remarkable feature of the integron system is its capacity to shuffle its cassettes and therefore modulate their expression level, which facilitates adaptation of the bacteria to new/stressful environments. It was also suspected that integron cassette shuffling could promote cassette dissemination into the bacterial genome. To test this assumption, an experimental system was developed such that the Escherichia coli bacteria survive to selective medium only if a designed integron cassette present on a plasmid is excised and is integrated into the bacterial genome. After culture on selective medium, the genome of bacteria was recovered, was PCR enriched for the cassette-genome junctions and was Illumina sequenced. A specific bioinformatic pipeline was developed to adapt to the PCR enrichment step. Results showed that integron cassette is inserted at high frequency in the genome (1e-3 resistant colony), in more than 20,000 different sites, avoiding the promoter and 5' regions of essential genes. Insertion seems dependent on a 5'-GWT-3' consensus site. Thus, integrons can randomly invade the bacterial genomes and behave as transposons. This represents a new route for the bacterial genomic diversification.

The Computational Challenges Of Microbial Paleogenomics

Contact: nicolas.rascovan@pasteur.fr

Nicolas Rascovan, Institut Pasteur, Universite de Paris, CNRS UMR 2000, Microbial Paleogenomics Unit, F-75015 Paris, France

The Microbial Paleogenomics Unit was created in September 2020 and our main goal is to use ancient DNA data recovered from archaeological samples to investigate the history and evolution of human populations and their associated microbes (including pathogens). As a recently emerging field, the analysis of ancient genomic data, either human or microbial, has multiple computational challenges to face in the coming years. At the level of ancient microbial genomes, some of the challenges include the de novo assembly of paleogenomic data, reducing the biases of reference-based genomic reconstruction, assessing the gene content of ancient strains, and dealing with the analysis of highly recombinant species. At the paleometagenomic level, (e.g., the analysis of ancient oral microbiomes), it remains highly challenging cleaning such data from environmental and modern contaminant sequences, as well as performing de novo assembly and binning of metagenome-assembled-genomes (MAGs). Finally, despite the possibility of recovering oral metagenomes, pathogen/commensal genomes and human genomes from ancient human samples, these layers of information remain mostly studied separately. Therefore, there is a great challenge ahead for developing new analytical frameworks that can integrate all this information to get a much broader and complete understanding of ancient human populations and how events from the past impacted on our present.

Spatial Transcriptomics Of A Hematopoiesis Niche Of The Zebrafish Embryo

Contact: olivier.mirabeau@pasteur.fr

Olivier MIRABEAU

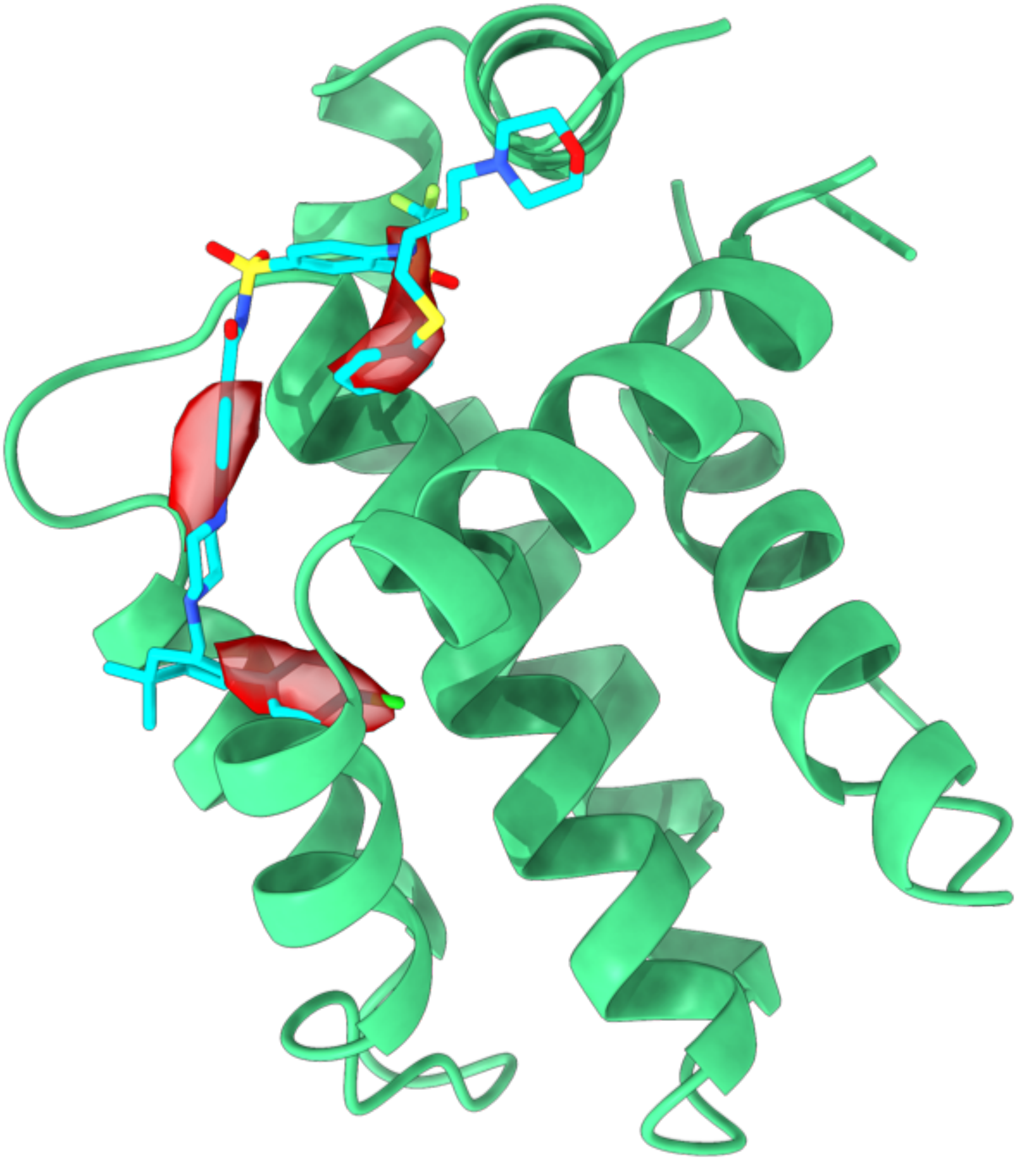
This pilot project is developed with the department of Developmental and Stem cell Biology, in collaboration with the CBUTechS. With it, we wish to characterise spatially the hematopoeitic lineage of the Zebrafish caudal hematopoietic tissue (CHT), a tissue compartment harbouring a hematopoiesis niche that is analogous to the mammalian fetal liver. For this we used the 10X Genomics Visium on different slices of CHT. The Visium technology allows to map back the gene expression data to its location within the tissue and obtain spatially resolved transcriptomics data on up to about 5000 distinct spots distributed as 78X64 grid at a resolution of 1-10 cells per spot. The transcriptomics signal measured at spots is a superposition of signals shared by other spots, some of which are common to multiple cell types and can be used as general quality control measures and some of which are specific to cell types and can be used to define the identity and proportion of cell types in each of the spots. To algebraically retrieve these different components we used a deconvolution technique called Independent Component Analysis (ICA). I will present the results of this analysis and highlight future challenges in image and omics data analysis that naturally emerge from these first experiments.

Olivier Sperandio

Contact: olivier.sperandio@pasteur.fr

Olivier Sperandio

Drug discovery still suffers from the paucity of therapeutic targets and from a shrinking pipeline for the development of new chemical entities. In this presentation, I will describe our use of data driven approaches to facilitate the identification of new therapeutic compounds on opportune but intricate targets such as protein-protein interactions (PPIs). These approaches are driven by the quality data we collect within iPPI-DB (https://ippidb.pasteur.fr/), our database of protein-protein interactions modulators and complex 3D structures. Such data are then used to train either machine learning models to design dedicated chemical libraries, or deep learning models to identify the most promising locations (binding sites) at the surface of PPIs.



Ligandability prediction (red surface) performed on Bcl-2 (pdb 4lvt) surface. The red surface patches of ligandability are localized around known hot spots of the Bcl-2/Bax interaction.

The Sequence Bioinformatics Group + Human Structural Variants Detection Using Accurate Long Reads

Contact: rayan.chikhi@pasteur.fr

Rayan Chikhi, Luca Denti

We will present the activities of the Sequence Bioinformatics G5. A highlight will be given on Luca Denti’s project with a new detection method for human structural variations using accurate PacBio HiFi long reads.

Metagenomic Tools And Analysis Of The Perinatal Vaginal Microbiota

Contact: skennedy@pasteur.fr

Sean KENNEDY, Metagenomic Signatures - Dept. of Com. Biol. ; Agnes BAUD, Metagenomic Signatures - Dept. of Com. Biol. ; Kenzo-Hugo HILLION, Metagenomic Signatures - Dept. of Com. Biol. ;

Metagenomic analysis of complex environments requires accurate identification and quantification of community members. This is especially true in clinical metagenomics where treatments and risk assessments frequently involve subtle shifts in microbial community structure and function. Our study of the vaginal microbiota of pregnant women has further highlighted the fact that accurate detection of species, normalization of abundance data and calculations of diversity is critical to achieving a better understanding of its role in human health. Here we present data from the analysis of 750 mother-infant pairs as part of the InSPIRe project. To accurately analyze this data we developed a method of generating a project-specific Kraken2 database which showed increased precision while retaining high sensitivity. Our results show that stratification by dominant species reveals additional community structure associated with infection risk. Taxonomic diversity was found to be particularly important in studying the perinatal vaginal microbiota. We implemented Faith's Phylogenetic Diversity and Weighted UniFrac analyses for shotgun metagenomics data to show significant links to clinical variables. Finally, we used random forest machine learning to build and test a model for Streptococcus agalactiae infection.

Stochastic Dynamics Of Two Dna Loci On A Compacted Chromosome

Contact: thomas.gregor@pasteur.fr

David B. Bruckner, Lev Barinov, Hongtao Chen, Thomas Gregor

Chromosomes are highly compacted and organized to fit into the eukaryotic nucleus. For many functional processes, including the initiation of transcription, pair-wise interactions of distantal chromosomal elements, such as enhancers and promoters, are essential. Previous theory based on simple polymer models successfully captures the dynamics of single loci in terms of sub-diffusive motion in the viscoelastic nucleoplasm. However, how these approaches extend to the joint motion of pairs of chromosomal loci remains unclear. Using a live imaging assay to simultaneously measure positions of pairs of enhancers and promoters in thousands of nuclei of the developing fly embryo, we analyze the two-point correlations of these pairs of DNA loci. Our analysis reveals long-ranged correlations with striking deviation from simple polymer models. Based on a scaling approach, we show how these findings can be reconciled based on the compaction of the chromosome, highlighting the key role of polymer packing in determining the coupled dynamics of chromosomal loci.

Hpc @Pasteur, Humble Past, Challenging Present And Bright Future

Contact: youssef.ghorbal@pasteur.fr

Youssef Ghorbal, HPC core facility

High Performance Computing (HPC) was always part of the toolbox Pasteur scientists leveraged to tackle challenging problems. Over the years, HPC activity came in a lot of shapes and colors. It ranged from small systems nested in specific labs, to bigger and shared resources in departements, to become a dedicated core facility offering HPC services to all scientists @pasteur. This talk will give a brief overview of HPC offering over the past. It will also expose the current challenges of HPC activity and will conclude with some insights of what to expect for the future.

Posters

Parameter Inference For Stochastic Models

Contact: andela.davidovic@pasteur.fr

Andjela DAVIDOVIC, Institut Pasteur, ; Remy CHAIT, University of Exeter, ; Gregory BATT, Inria Paris, ; Jakob RUESS, Inria Paris

Understanding and characterising biochemical processes inside single cells requires experimental platforms that allow one to perturb and observe the dynamics of such processes as well as computational methods to build and parameterise models from the collected data. Here we investigate the following questions: How can we calculate likelihoods and infer parameters of stochastic kinetic models from data sets in which each cell receives a different input perturbation? How does the computational efficiency of parameter inference methods scale with the number of inputs and the number of measurement times? Is it more informative to diversify input perturbations but to observe only few cells for each input or should we rather ensure that many cells are observed for only few inputs? In order to answer these questions we use the CcaS/CcaR optogenetic system driving the expression of a fluorescent reporter protein as primary case study.

Using Rich Metadata To Study Pathogen Spreads

Contact: anna.zhukova@pasteur.fr

Anna ZHUKOVA, Evolutionary Bioinformatics Team & Bioinformatics and Biostatistics Hub ; Department of Computational Biology, Institut Pasteur & Universite de Paris

From where was Sars-CoV-2 introduced to Europe? What is the expected number of cases directly infected by an Ebola-positive individual? Is a drug-resistant virus less transmissible? These and other questions impacting health policies, historical knowledge and our understanding of epidemics can be addressed with phylodynamic analyses of pathogen sequences. This involves reconstructing phylogenetic trees and using them to estimate epidemiological parameters (e.g. R0), infer the timeline and locations of epidemic spreads, or detect clusters of public health interest (e.g. transmission of HIV-1 drug-resistant strains).

Many of these phylodynamic tasks are challenging from the computer science prospective. For instance, already reconstructing the maximum likelihood phylogenetic tree topology is computationally intractable (hence requires heuristics); many of the phylodynamic models are asymptotically unidentifiable; and it is generally impossible to accurately estimate both the root state and the rates of state changes along the tree branches from the data observed at the tips.

The good news however is that despite being extremely challenging as general computer science problems, phylodynamic tasks also have their biological component. Using biological knowledge about a problem, one can add constraints to significantly simplify many of the computational issues. In this poster I will focus on how rich metadata can be used to inform phylodynamic analyses. I will present examples of adapting the methods for phylogeographic and phylodynamic reconstructions to use the constraints coming from biological/epidemiological/historical knowledge, and their applications to studies of spatio-temporal spreads of HIV-1 and Sars-CoV-2 and of drug-resistance patterns in HIV-1.

The Genetic Network Of Neuranatomical Phenotypes Underlying Psychiatric Diseases

Contact: antoine.auvergne@pasteur.fr

Antoine AUVERGNE, G5 Statistique genetique

The study of neuroanatomical phenotypes and their associated genetic and environmental factors has become a central component of ongoing research but several of the methods used so far are reaching limitations and new strategies and extension of existing methods are now needed. The main objective of this thesis is to build upon previous works to develop and apply a new powerful and computationally efficient method for inferring genetic network underlying intricated neuroanatomical phenotypes, and to assess the role of these networks in psychiatric diseases.

Sle Map: A Consensus Map Of Sle Patients Into A Low Dimensional Space

Contact: behnam.yousefi@pasteur.fr

Behnam Yousefi, Benno Schwikowski

Human transcriptome profiles typically contain gene expression values for many thousands of genes, thus representing points in a high-dimensional feature space. A principal technical challenge in understanding the pathways involved in the disease and characterizing inter-individual variation is to reduce this high dimensionality while preserving biologically relevant information. This results in an interpretable lower-dimensional space, which can reveal patient sub-populations corresponding to different forms of disease that can be treated in a more targeted fashion than the generic form of the disease.

In this work, we develop a novel computational pipeline to [i] map the blood gene expression profile of systemic Lupus erythematosus (SLE) patients into a robust, semantically rich and low dimensional space, and to [ii] find a SLE patient stratification that is congruent across different publicly available datasets. Our approach is to extract patterns within patient data, through which we aim to identify the underlying biological pathways associated with disease progression, and thus to map the complex disease dynamics, reflected in a high-dimensional space of trascriptome abundances, into a reduced and thus more understandable space.

We identify and download transcriptomic data from ten different publicly available human SLE data sets in Gene Expression Omnibus. In all datasets, the gene expression data is acquired from the whole blood cells. It is argued that blood cells provide an informative and easily accessible means to study the immune systems

The subspace we chose to project ten publicly available SLE blood transcriptomic datasets into has four dimensions, each representing key axes of variation. The axes were labeled "Interferon/lymphocyte/erythrocyte/inflammation" activity after determining enriched biological functions of the four gene sets ("modules") that were computationally determined as most suitable for capturing the variability across the cohorts.

Decoding The Central Dogma

Contact: benjamin.zoller@pasteur.fr

Benjamin Zoller, Po-Ta Chen, Michal Levo, Thomas Gregor

How precise expression patterns emerge from transcriptional bursts in individual nuclei is a decade old question that remains puzzling today. To bridge the central dogma, we measure transcriptional activity of individual endogenous loci in living fly embryos, achieving single RNA detection sensitivity. We demonstrate that dynamical pattern establishment is governed by the transcription rate, whose maximum is shared among patterning genes. By deconvolving single-nucleus time traces, we reconstruct the Pol II loading events enabling quantification of bursts underlying the transcription rate regulation. We show that the bursting phenotypes of the patterning genes are tightly constrained across space, time and nuclear cycles. We further infer the spatiotemporal regulation of the bursting kinetics and identify the key regulatory parameter as the fraction of time a gene is transcriptionally active, regardless of gene identity, boundary position, or enhancer-promoter architecture. These results point to a shared simplicity underlying the apparently complex processes of early embryonic patterning and indicate a path to general rules in transcriptional regulation.

Zola-3d: A Software For Optimal 3d Smlm With Arbitrary Psf

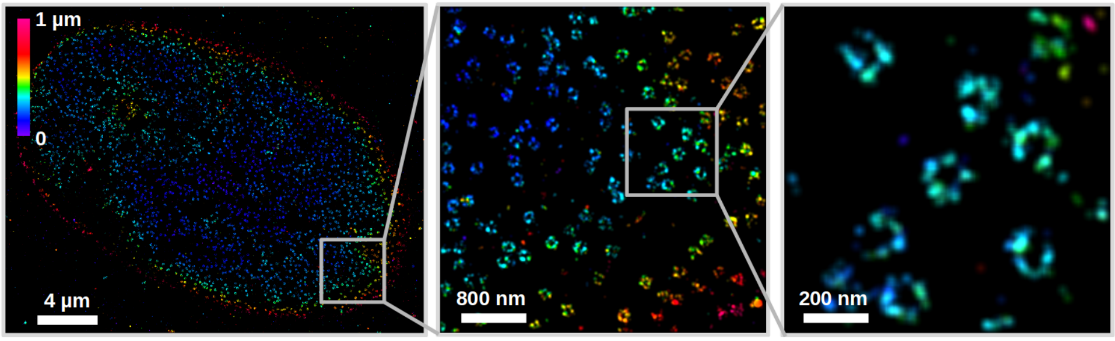
Contact: benoit.lelandais@pasteur.fr

Benoit LELANDAIS, Mickael LELEK, Christophe ZIMMER, Imaging and Modeling Unit

Super-resolution microscopy based on single molecule localization (SMLM) typically achieves lateral resolutions of 20 nm in 2D images. Point spread function (PSF) engineering is widely used as a means to enable 3D SMLM reconstructions from 2D image sequences obtained without scanning. The most common engineered PSF is astigmatism [1], which enables axial resolutions of about 50 nm over an axial range of ~1 µm. Larger axial ranges can be achieved with more sophisticated PSFs, such as double helix [2], saddle-point or tetrapod PSFs [3]. PSF engineering that improves the axial range necessarily entails a reduction of lateral and axial resolutions. This degradation of resolution can be compensated by combining two opposing objectives that collect photons from both sides of the sample [4,5]. However, imaging deep inside the sample can lead to spherical aberrations, particularly in dual-objective systems. To reconstruct 3D super-resolution images with optimal resolutions with such complex and aberration-prone microscopes, a generic localization method based on a flexible and accurate modeling of the PSF is needed.

We recently developed ZOLA-3D, an easy-to-use Fiji plugin that enables 3D SMLM image reconstruction using a realistic modelling of the PSF, and which accounts for depth-dependent spherical aberrations [6]. ZOLA was initially designed for PSFs obtained with continuous phase masks such as astigmatism, saddle-point or tetrapod PSFs. Here, we present an extension of ZOLA that makes it compatible with discontinuous phase masks, used e.g. for the double-helix PSF, and with dual objective systems. We will show and compare ZOLA reconstructions of 3D images obtained with astigmatism, saddle-point, tetrapod or double-helix PSFs, as well as a dual-objective system.

[1] B. Huang et al. Science, vol 319, 2008. [2] S.R.P. Pavani et al. Proceedings of the National Academy of Science, vol 106, 2009. [3] P.N. Petrov et al. Opt. Express, vol 25, 2017. [4] S. Ram et al. Opt. Express, vol 8, 2009. [5] Xu et al. Nature Methods, vol 9, 2012. [6] A. Aristov et al. Nature Communications, vol. 9, 2018.



Super-resolution 3D image of nuclear pores reconstructed with ZOLA-3D. Imaging was performed with a dual-objective setup combined with astigmatism.

Macsyfinder V2 Story Of Sucessful Colaboration

Contact: bneron@pasteur.fr

Bertrand Neron, Institut Pasteur, Universite de Paris, Bioinformatic and Biostatistic Hub ; Marie Touchon, Institut Pasteur, Universite de Paris, CNRS UMR3525, Microbial Evolutionary Genomics ; Remi Denise, Institut Pasteur, Universite de Paris, CNRS UMR3525, Microbial Evolutionary Genomics ; Sophie Abby, Universite Grenoble Alpes, CNRS, Grenoble INP, TIMC-IMAG, Grenoble, France ; Eduardo Rocha, Institut Pasteur, Universite de Paris, CNRS UMR3525, Microbial Evolutionary Genomics

MacSyFinder is a program to model and detect macromolecular systems, genetic pathways… in protein datasets. In prokaryotes, these systems have often evolutionarily conserved properties: they are made of conserved components, and are encoded in compact loci (conserved genetic architecture). The user models these systems with MacSyFinder to reflect these conserved features, and to allow their efficient detection.

MacSyFinder V2 is developed in python following the development good practices: • the code source is versioned and freely accessible • released under GPLv3 license • fully tested (unit and functional test) • documented (user and developer documentation)

MacSyFinder has been developed inspired by agile methodology: short iterations. Each iteration is constitute of design, development, test and review phases. This methodology is very powerful and adapted in research environment but some key aspect must be respected otherwise the project will fail.

Genome Organization Regulation & Expression

Contact: cchica@pasteur.fr

Victoire BAILLET, Hub de Bioinformatique et Biostatistique ; Bernd JAGLA, Hub de Bioinformatique et Biostatistique, Cytometry and Biomarker UTechS, Center for Technological Resources and Research (C2RT) ; Rachel LEGENDRE, Hub de Bioinformatique et Biostatistique ; Yann LOE-MIE, Hub de Bioinformatique et Biostatistique, Nuclear Organization and Oncogenesis Unit ; Meije MATHE, Hub de Bioinformatique et Biostatistique ; Olivier MIRABEAU, Hub de Bioinformatique et Biostatistique ; Adrien PAIN, Hub de Bioinformatique et Biostatistique ; Claudia CHICA, Hub de Bioinformatique et Biostatistique ; ;

We seek to apply and develop methods to quantify the effect of transcriptomic and epigenomic variation in establishing a phenotypic output, e.g. the maintenance/loss of a specific cellular state, or an immune system response. In parallel, and considering the growing interest of the IP research community, we want to deepen our expertise in: (i) OMIC data integration as a tool to generate testable hypotheses on regulatory mechanisms and propose targeted experiments to infer causal function. (ii) Single cell OMICs, to extend our current knowledge on scRNAseq and gain insight on the analysis of combined single-cell experiments, where the inference of the correlation among OMIC layers is greatly facilitated.

We also ensure knowledge transfer by engaging in the advanced training sessions/workshops organised by the Hub and by developing tools that that facilitate the interaction of the experimentalists with the OMIC dataset.

A non exhaustive list of our current projects include: Multidimensional OMIC analysis of bacterial infection memory. Identification and characterisation of specific cell population within complex tissues. Dissection of tissue interactions that shape organ function through multi-OMIC single cell approaches. Impact of enhancer polymorphism in the evolution of malaria resistant of close species of mosquitoes.

Identification Of Consensus Whole Blood Transcriptomic Gene Modules In Sjogren's Syndrome Patients

Contact: cheima.boudjeniba@gmail.com

Cheima BOUDJENIBA, Servier, Institut Pasteur and Universite de Paris. ; Etienne BIRMELE, Universite de Strasbourg. ; Benno SCHWIKOWSKI, Institut Pasteur. ; Etienne BECHT, Servier. ;

Sjögren's syndrome is the second most common autoimmune disease, characterized by lymphoid infiltration and production of different autoantibodies responsible for dry mouth, eyes and in the most severe cases organ failures. SjS is a multi-organ and systemic disease with considerable heterogeneous clinical manifestations among individuals. Besides, its pathophysiology remains unknown, with no commercialised treatment avalaible.

Currently, following the paradigm of Precision Medicine, most of the studies focused on setting up a symptom-based or molecular-based stratification of patients to identify biomarkers and develop new treatments, adjusted to the biological characteristics of the patients.

To complement these studies, we retrieved gene modules from four transcriptomic datasets proceeding from blood samples via clustering method coupled to a dimension reduction technic. More precisely, a graph was used to model the genes' pairwise affinity for each cohort's transcriptomic matrix. Then, Similarity Network Fusion (SNF) was used to merge the four resulting graphs. Finally, the resulting consensus graph was projected into a sublinear space modelled by a Gaussian mixture model (GMM). The 12 resulting consensus modules (CM) were then biologically characterized using GSEA and correlation to clinical features.

Genomic Determinants Of Natural Antisense Transcripts In A Compact Eukaryotic Genome

Contact: damien.mornico@pasteur.fr

Damien Mornico(4) and Nancy Guillen (1,2,3); ; 1 Institut Pasteur, Unite Biologie Cellulaire du Parasitisme,F-75015 Paris, France; ; 2 INSERM U786, F-75015 Paris, France; ; 3 Centre National de la Recherche Scientifique, CNRS ERL9195, F-75015 Paris, France ; 4 Hub de Bioinformatique et Biostatistique - Departement Biologie Computationnelle, Institut Pasteur, Paris, France ;

Cis-natural antisense transcripts (NATs) are RNAs that contain sequences which are complementary to other endogenous RNAs and transcribed from opposing DNA strands at the same genomic locus. NATs have been widely studied in prokaryotes and in eukaryotes. While the role of most of these RNAs remains unknown, numerous functions have been highlighted, including the capacity to regulate gene transcription. Entamoeba belongs to the Amoebozoa kingdom, which represents one of the earliest branches from the last common ancestor of all eukaryotes and is phylogenetically distinct from ‘model organisms’ of animals, fungi and plants. E. histolytica has a dense protein-coding genome: intergenic regions are short and thus, regulatory mechanisms are limited. In this study, we show that numerous genes of E. histolytica have at least one cis-NAT localized in the 3’ part of the gene on the opposite strand in relation with the sense transcripts. Some features such as transcription start sites (TSS) or polyadenylation sites (Poly(A)) are directly reverse encoded in gene coding sequences. Antisense TSS are mainly located on the third base of the stop codon and the same mRNA Poly(A) motifs are retrieved in the coding sequence. The cis-NAT organisation provides a potential compact regulatory system for gene transcription. Thus, we studied gene expression during different stress and environmental conditions and demonstrated that NATs were globally up-regulated. More precisely, few gene transcriptions seem correlated or anti-correlated with the corresponding NAT expression. This data indicates that the most abundant NATs does not influence the rate of mRNA sense transcription, suggesting a complex role for these specific antisense transcripts in gene expression

Primary Sjögren’s Syndrome: What A Second Look On Assess Data Can Tell Us

Contact: diana.trutschel@pasteur.fr

Diana Trutschel, Institut Pasteur, Universite de Paris, Department of Computational Biology, Systems Biology Group; J. Eric Gottenberg, Les Hopitaux Universitaires de Strasbourg, Rhumatologie; Benno Schwikowski, Institut Pasteur, Universite de Paris, Department of Computational Biology, Systems Biology Group

Background: Primary Sjögren’s Syndrome (pSS) is a systemic chronic autoimmune disorder characterized by fatigue, dryness and pain. Among patients with pSS is a higher mortality rate compared to healthy people. Although research has improved the understanding of the disease, no significant progress has been made in the treatment. The type I interferon (IFN) system has a pivotal role in the disease process and is associated with the clinical and immunological phenotype. Objective: To more precisely characterize the role of IFN, we studied circulating blood IFNα concentrations and their correlation to autoantibody presentation on gene expression. IFNα levels are associated with autoantibody levels, like anti-SSA and anti-SSB, presentation, as well as disease activity. The objective was to dissect and quantify their relation to gene expression. Method: Data from patients with pSS from the French multi-center prospective clinical cohort ASSESS were analyzed, whereby transcriptomic data as well as blood measurement were available. A model search workflow was used to identify the best fitting model for each gene and estimate coefficients for all included factors. Linear mixed regression models were adjusted for the effect of age as a fixed factor and for hospital center effects as a random factor. Results: Most of the IFNα-associated genes are correlated to autoantibody presentation. Only 4 genes could be identified, which are solely associated by the autoantibody presentation, 31 genes which are not additionally associated to the autoantibody status. Furthermore, in general the gene expression increases with increasing IFNα blood concentration, but the degree of increase correlates with autoantibodies levels. Discussion: Since the influence of autoantibody presence besides the predominance of IFNα on the gene expression could be shown, further studies should investigate the impact of other biomarkers like the rheumatoid factor on the association of IFNα to gene expression. A better understanding of the interplay of all blood proteins on gene expression may lead to more targeted drugs.

The Four Horsemen Of Neglecting Experimental Design

Contact: elise.jacquemet@pasteur.fr

Elise JACQUEMET, Pascal CAMPAGNE, Emeline PERTHAME, Stevenn VOLANT, Thomas OBADIA, Hugo VARET, Francois LAURENT ; Affiliation for all : Institut Pasteur, Universite de Paris, Hub de Bioinformatique et Biostatistique, F-75015 Paris, France

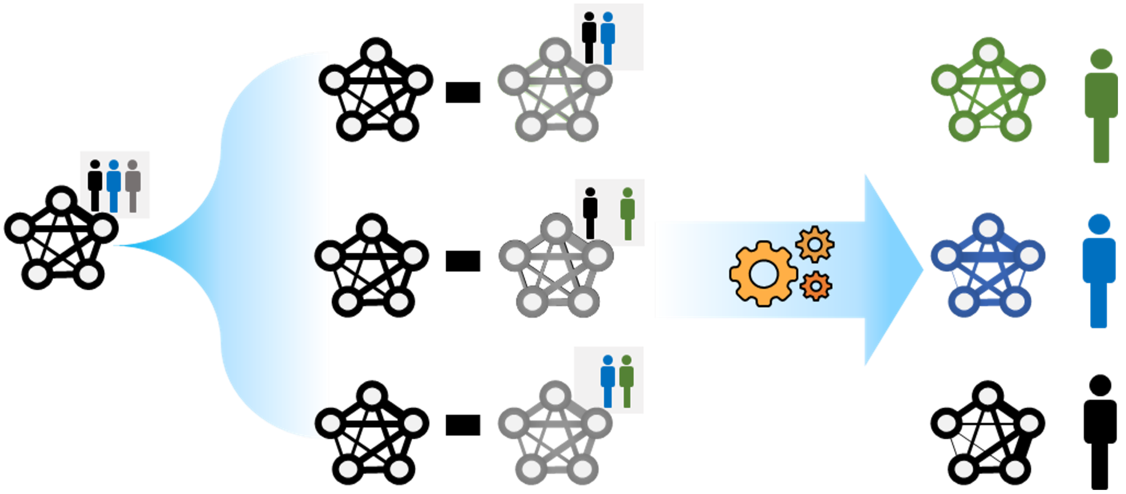
Reproducibility in experimental results has become a central concern of biological sciences. Increasing attention must be brought on how to carefully design experiments and choose appropriate statistical tools. Controlled experiments may be hard to conduct as they are subject to many technical constraints that need to be dealt with, especially in animal experimentation. Difficulties are such that (i) experimental outcomes may not be properly designed, statistically speaking; (ii) basic statistical tools may not be suited to properly analyse data that are partly shaped by technical constraints. Both issues are not rare in studies made at Pasteur, and their impact on results are seldom considered. A downside of neglecting experimental design is an uncontrolled inflation of false-positives that may occur well above the typical 5% control threshold. This inflation is further aggravated by low statistical power, that depends on both sample size and experimental effects. Lack of power insidiously leads to overestimating the magnitude of experimental effects, all the more when they are coupled with confounding factors due to ill-defined designs. Here we discuss four problems we often encounter in our daily life of data analysts. We illustrated why & how they may drastically alter the validity of some scientific findings.

Significance Assessment In Individual Specific Networks

Contact: federico.melograna@kuleuven.be

Federico MELOGRANA, BIO3 - Laboratory for Systems Medicine, KU Leuven ; Fabio STELLA, Department of Informatics, Systems and Communication, University of Milan-Bicocca ; Kristel VAN STEEN, 1BIO3 - Laboratory for Systems Medicine, KU Leuven and GIGA-R Medical Genomics, University of Liege

Individual-specific networks (ISNs) are graphical structures of measurement with individual-specific node or edge values, or both. ISNs are promising in the context of Precision Medicine and can be used to infer subnetworks or sets of genes linked to individual-relevant pathways, or to highlight associations between an individual's network properties and external data. Focusing on ISNs with individual-specific edge weights (computed as in Kuijjer et al., 2019), evaluating their statistical significance remains an under investigation problem. Given that an edge is not a stand-alone feature but belongs to a – possibly intricate – network structure, we propose and evaluate several strategies on modules, with single edge as a particular case. These include leave-one-out techniques with a linear (LOO-ISN), and non-linear (MultiLOO-ISN) aggregation across ISN edges, based on resampling procedures from a multivariate normal and assessing the impact on the corresponding correlation matrix. We also employ a customized Cook's distance approach by iterative linear modelling of the edge weights in the targeted module. In view of accommodating generic ISNs with flexible edge weight definitions, we empirically evaluate the aforementioned methods against outlier detection techniques, including DBSCAN, kNN and Spoutlier (Sugiyama et al., 2013). We grid-explore 168 different settings, repeating each combination 200 times to reduce errors and noise.; varying sample and module sizes, number of outliers and outlier distributions. Heterogeneity in results increases with module size and proportion of outlying individuals; There is a sizeable advantage towards ensemble techniques than their single-shot counterparts, with Cook's distance showing excellent to good performance in all scenarios. Overall, our study shows the value of using network structures in ISNs to establish the significance of an individual. Significance assessment is vital to determine the added value of ISNs over similar networks across samples, for risk assessment, disease diagnosis or management. We have to acknowledge that statistical significance is different from medical impact; However, moving from edge-by-edge significance assessment to subnetwork may be a correct step toward biological relevance



Depicted pipeline of individual-specific networks (ISNs). From the entire samples a graphical representation is firstly extracted, then via a leave-one-out (LOO) procedure, the impact of an individual is estimated and thus constitutes the core value used to build ISNs.

Machine Learning Techniques For Drug Response Prediction

Contact: firoozbakht.bme@gmail.com

Farzaneh Firoozbakht, Behnam Yousefi, Federico Melograno, Benno Schwikowski

Machine learning (ML) is a specific subset of artificial intelligence that allows automatic learning from data, and has contributed to a wide range of genomics research. One particular application of ML is drug response prediction (DRP), in which phenotypic responses of biological samples are predicted on the basis of their molecular profiles, with predictors often providing mechanistic insight. Efforts to understand, and predict, drug responses in a data-driven manner have led to a proliferation of ML methods. Here we first provide a systematic classification of machine learning-based drug response prediction methods using a large number of research papers. These methods can generally be classified as single-drug learning (SDL) and multi-drug learning (MDL). Briefly, to predict the response to a given drug, MDL leverages data from other drugs; SDL exclusively uses data about the given drug. SDL and MDL differ in their capabilities and validation process.

Then, to investigate the association between microbial profile and the drug response, we further focus on the gut microbiome, and propose a ML method to predict the drug response given the microbiome profile of each individual. The data we consider In this study is the microbial profile of fecal samples for UC and CD patients along with their response to Vedolizumab, Ustekinumab and anti-TNFs drugs. We impute individual specific networks (ISNs) of microbial co-occurrence, in which the nodes represent microbes and the edges are weighted by their co-occurrence. We next use these ISNs to train a graph-based neural network for the prediction of drug response. To this end, we use graph convolutional neural networks to extract deep features from the structure of graphs.

Using Single Molecule Fish To Exploring Transcriptional Regulation At Multiple Scales

Contact: fmueller@pasteur.fr

Christian WEBER, Institut Pastur; Arthur IMBER Arthur, MINES PariTech; Wei OUYANG, KTH Stockolm; Edouard BERTRAND, IGH Montpellier; Thomas WALTER, MINES PariTech, Christophe ZIMMER, Institut Pasteur; Florian MUELLER, Institut PASTEUR

Regulation of RNA abundance and localization is a key step in gene expression control. Single-molecule RNA fluorescence in-situ hybridization (smFISH) is a widely used single-cell-single-molecule imaging technique enabling quantitative studies of gene expression and its regulatory mechanisms. Here, we present the currently available approaches in our lab both for experiments and analysis, as well as ongoing developments.

Reproducibility Of Bioinformatic Workflows

Contact: frederic.lemoine@pasteur.fr

Frederic LEMOINE, Institut Pasteur, Gael MILLOT, Institut Pasteur, Bertrand NERON, Institut Pasteur

Usual bioinformatic analyses are more and more complex and computer intensive. First of all, they are made of many steps involving a great diversity of tools that are in constantl evolution. Then, their structure is more and more complex, involving many parallel steps and branching structures. Finally, the size and diversity of analyzed data make their optimization and execution more difficult. As a result, bioinformatic workflows are often difficult to implement, describe, share, and reproduce. In this poster, we describe some of these challenges, and different ways to facilitate the implementation and the execution of bioinformatic workflows, in particular by using workflow management systems, container technologies, and versioning systems.

Building A Metagenome-Assembled Genomic Database Of Oral Microbes For High-Resolution Metagenomic

Contact: gabriel.ponce-soto@pasteur.fr

Gabriel Yaxal PONCE SOTO, Institut Pasteur, Universite de Paris, CNRS UMR 2000, Microbial Paleogenomics Unit, F-75015 Paris, France ; Nicolas RASCOVAN, Institut Pasteur, Universite de Paris, CNRS UMR 2000, Microbial Paleogenomics Unit, F-75015 Paris, France

The oral microbiome is situated at the gateway between the environment and the organism and plays multiple roles in the local and overall health of the host. While different metagenomic studies have permitted to generate a quite comprehensive knowledge of the microbial taxa that make up the oral microbiome, there is still a lot to learn about the oral microbiome variation at multiple levels (pangenomic, taxonomic and functional) and scales (individuals, populations, time and geographies). Our approach to investigating oral microbiomes at these scales and levels will be to generate a comprehensive and non-redundant database of oral microbial genomes, to give a view of the global structure of the human oral microbiome, but also, to understand the evolution of its components over time. With this genomic database, we aim at increasing the mappability of oral metagenomic data to be able to study the oral microbiome at the whole-genome scale. We will do so by compiling and reanalyzing public databases and metagenomic datasets comprising a wide range of studies across the globe, and individuals with different health statuses, age-range, and ethnicity. We will reconstruct Metagenome-Assembled Genomes (MAGs) from all samples and cluster them into operational taxonomic units (OTU) by their average nucleotide identity (ANI). Then the pangenomic diversity of each OTU will be assessed to define their core and accessory genomes, which can be used to trace their phylogeographic distributions and horizontal gene transfer events, respectively. Finally, to bring a temporal context to the analysis of the human oral microbiome diversity, we will also generate whole-genome data from ancient oral microbiomes by mapping public ancient metagenomes against the generated database to investigate if there were shifts in diversity and phylotypes over time and to reconstruct the spread history of human oral microbes.

Activities Of The Web Integration Group In The Hub

Contact: herve.menager@pasteur.fr

Bryan BRANCOTTE, Hub de Bioinformatique et Biostatistique ; ; Hippolyte KENGNI, Institut Francais de Bioinformatique ; ; Lucie LAMOTHE, Institut Francais de Bioinformatique ; ; Fabien MAREUIL, Hub de Bioinformatique et Biostatistique ; ; Remi PLANEL, Hub de Bioinformatique et Biostatistique ; ; Rachel TORCHET, Hub de Bioinformatique et Biostatistique ; ; Herve MENAGER, Hub de Bioinformatique et Biostatistique

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, maintain, 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.).

Over the past few years, our group has created more than 15 web applications and databases, in collaboration with research units and other groups of the Hub, and with the support of the IT department, including for instance:

- iPPI-DB: a database of modulators of protein-protein interactions, created with the team of Olivier Sperandio, from the Structural Bioinformatics Unit. The data are retrieved from the literature either peer reviewed scientific articles or world patents. A large variety of data is stored within iPPI-DB: structural, pharmacological, binding and activity profile, pharmacokinetic and cytotoxicity when available, as well as some data about the PPI targets themselves.

- CRISPR-browser: a genome browser to visualize the results of CRISPR-dCas9 screens in bacteria, created with David Bikard from the Synthetic Biology Group. You can upload your own data or navigate through published datasets.

- Modelisation-COVID: an information website that publishes the work carried out by the Mathematical Modelling of Infectious Diseases Unit on COVID-19.

Moreover, we oversee the Galaxy server of the Institut Pasteur. 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 611 tools (including 31 developed at the Institut Pasteur), covering multiple types of analyses, including NGS, Metagenomics, RNA-Seq, ChIP-SEQ, and Phylogeny. It also provides through its API an access to the computing infrastructure to more specialized applications such as SHAMAN, ng-Phylogeny, ARIAWeb, and Booster. In 2021, Galaxy Pasteur counted more than 3387 registered users and an average of 75000 jobs per month are launched.

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).

Ct-Based Diagnosis Of Covid-19 With Deep Learning: A Study From Multi-Source And Large Datasets

Contact: hoa.nguyen-thi-thanh@pasteur.fr

Hoa NGUYEN, Imaging and Modeling Unit, Institut Pasteur ; Cedric THEPENIER, Institut de recherche biomedicale des armees, Bretigny-sur-orge ; Ilan OBADI, Centre Hospitalier Intercommunal, Poissy Saint Germain-en-Laye ; Benoit LELANDAIS, Imaging and Modeling Unit, Institut Pasteur ; Mohamed Amine CHAABOUDII, Centre Hospitalier Intercommunal, Poissy Saint Germain-en-Laye ; Julia GROSS, Centre Hospitalier Intercommunal, Poissy Saint Germain-en-Laye ; Wei OUYANG, Science for Life Laboratory, School of Engineering Sciences in Chemistry, Biotechnology and Health, KTH - Royal Institute of Technology, Stockholm, Sweden ; Robert CARLIER ; Christophe ZIMMER, Imaging and Modeling Unit, Institut Pasteur

As of November 2021, COVID-19 has affected over 254 million people worldwide, causing at least 5 million deaths. Hundreds of artificial intelligence-based studies have been published to accelerate and refine the diagnosis of pneumonia caused by COVID-19. However, the robustness of these models has usually not been thoroughly demonstrated. In addition, most methods do not provide tools to help explain the model's predictions. Moreover, tools allowing to easily retrain models on a specific set of data are mostly lacking.

In an attempt to overcome these limitations, we first collected data from 7 different sources across multiple countries to train and validate the model's generalizability with a large and diverse data set. Our data set totals 75,208 CT scans from 18,272 individuals, including 2,599 COVID-19 scans, and trained 3D convolutional neural networks to diagnose COVID-19 on a subset of 3,858 CT scans. When tested on nearly 71 thousand scans, our model achieves a sensitivity of 90%, a specificity of 94%, and an area under the curve (AUC) of 97%, close to the state-of-the-art. Second, we implemented attention maps to highlight the discriminating features used for prediction and subjected them to evaluation by one radiologist. In 64 out of 70 positive cases, the attention map highlighted features validated by the radiologist as COVID-19 lesions. However, in 38/70 cases the map also highlighted other features that require further investigation. Subsequently, we will build a user-friendly Imjoy plugin to allow users without programming skills to evaluate the trained model and re-train it using their own data and resources.

Setting Up Of The Metabolomics Core Facility At The Institut Pasteur

Contact: hugo.varet@pasteur.fr

Hugo VARET, Hub Bioinformatics & Biostatistics, Metabolomics Core Facility ; Lise BOULARD, Metabolomics Core Facility ; Kathleen ROUSSEAU, Metabolomics Core Facility ; Sandrine AROS, Metabolomics Core Facility

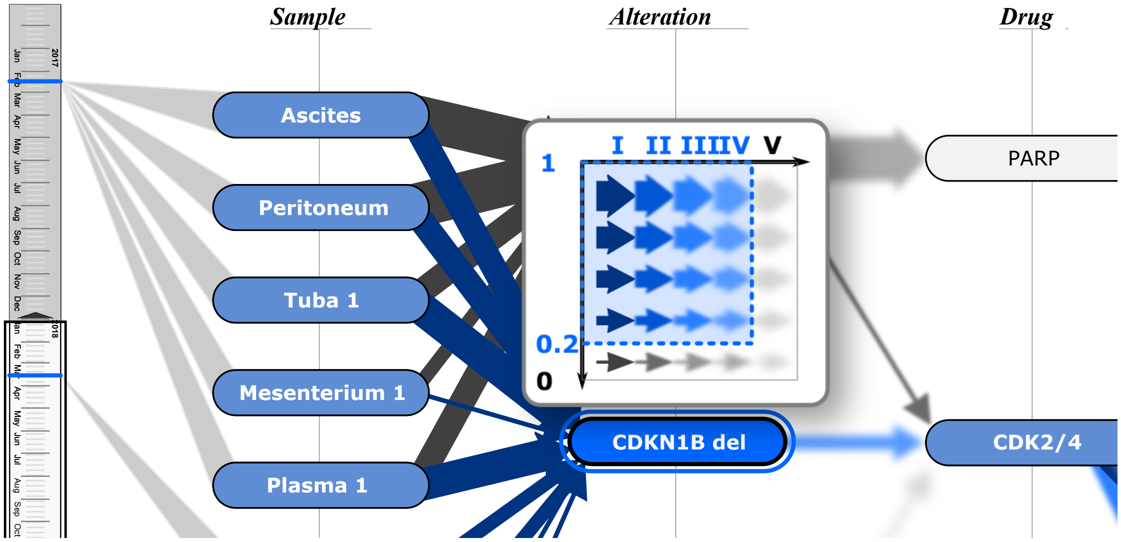
This poster introduces the Metabolomics Core Facility that have been created late 2020. It highlights our current challenges: building the platform and bioinformatic/statistical tools to analyze the data we will generate.

Oncodash: Decision Support Platform For Tumour Boards, Using (Interactive) Bayesian Network

Contact: johann.dreo@pasteur.fr

Johann DREO, Institut Pasteur, Universite de Paris, Computational Biology, Systems Biology ; Oceane FOURQUET, Institut Pasteur, Universite de Paris, Computational Biology, Systems Biology ; Benno SCHWIKOWSKI, Institut Pasteur, Universite de Paris, Computational Biology, Systems Biology

We design a software platform that allows tumour boards to leverage cutting-edge information about their patients, using explainable AI tools in a modern HMI. This platform notably aims at presenting a bayesian (credal) network to oncologists, allowing them to interact with causal inference models. This open-source web-based software is developed within the Decider project, which targets high-grade serous ovarian cancer and provides a large set of omics and clinical data.



Interactive manipulation of a bayesian network in Oncodash.

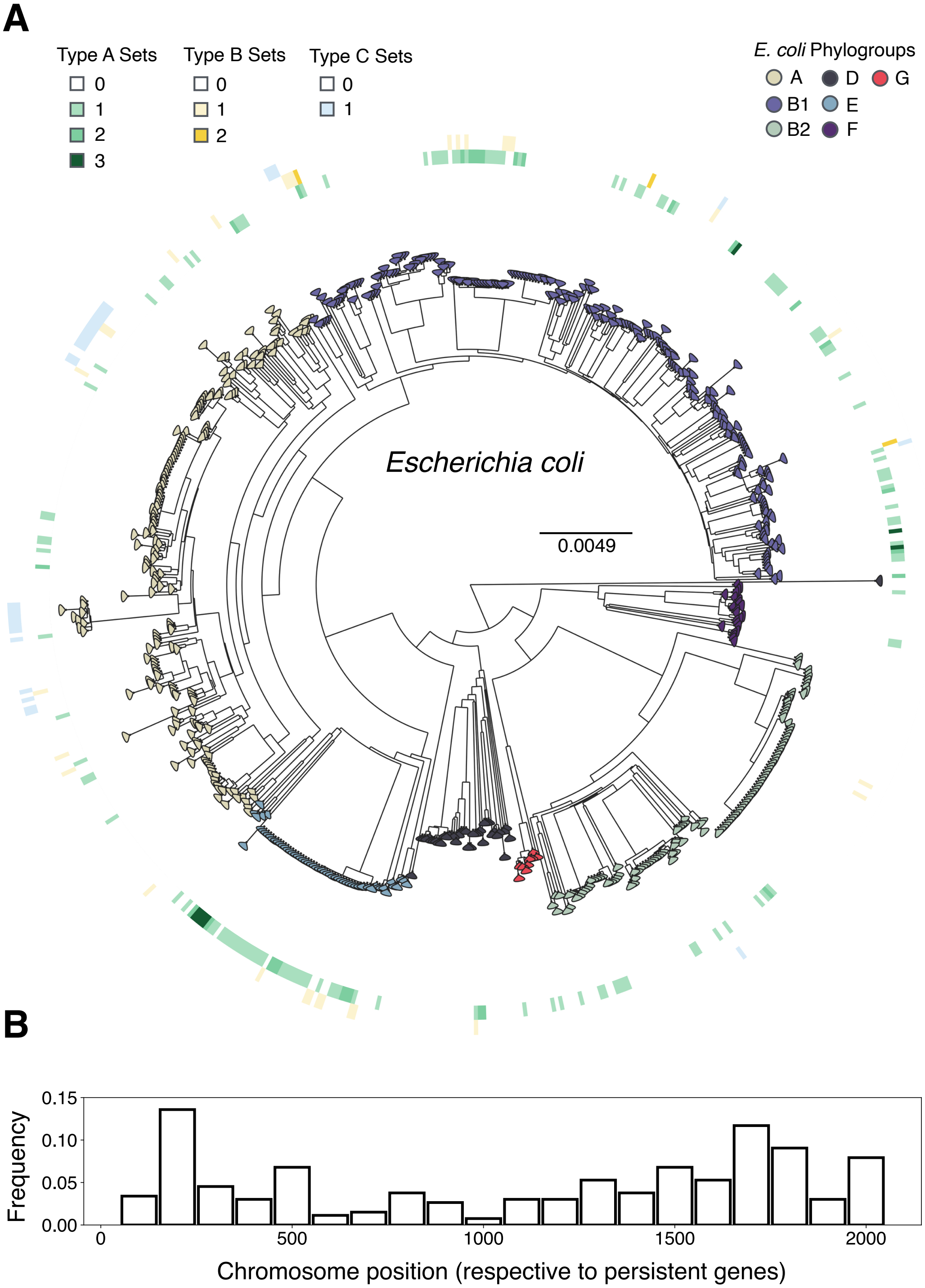
To Catch A Hijacker: Abundance, Evolution And Genetic Diversity Of P4-Like Bacteriophage Satellites

Contact: jorge-andre.sousa@pasteur.fr

Jorge MOURA DE SOUSA, Microbial Evolutionary Genomics, Institut Pasteur, CNRS, UMR3525 ; Eduardo P.C. ROCHA, Microbial Evolutionary Genomics, Institut Pasteur, CNRS, UMR3525

Bacteriophages (phages) use their bacterial hosts to replicate, which usually ends in bacterial death and the release of virions containing the phage genome. However, phages have their own parasites, satellite mobile elements that cannot produce virions and instead hijack those of functional (so-called helper) phages. The best studied phage satellites are P4 in Escherichia coli, Staphylococcus aureus pathogenicity islands and phage-inducible chromosomal islands in Bacillales and enterobacteria, and phage-inducible chromosomal islands-like elements in Vibrio spp. From these, the satellite-helper system that is best understood in molecular terms is the pair P4-P2. P2 is a fully functional enterobacteriophage, of the Myoviridae family, that can be exploited by the satellite P4. The co-infection of an E. coli bacterium with both P2 and P4 provides the context for the parasitism of the latter. P4 was initially regarded as a phage-plasmid, i.e. a phage that can reside in the lysogenic cells as a plasmid, and to-date there was still no clear evidence on whether P4 is phage, a plasmid or a completely distinct mobile element. Nevertheless, decades of studies have revealed the key functions of genes encoded by P4.

Despite the exquisite knowledge obtained on the molecular interactions between P4 and P2, we know little about the diversity and distribution of the P4 family of satellites, which also hinders the study of their evolution. Is P4 part of a family of mobile genetic elements with conserved gene repertoires and genetic organizations? What is the core genome and the genetic organization of such a family? Are members of the family abundant in bacterial genomes? Have P4-like elements recently derived from other mobile genetic elements, e.g., phages, or are they ancient? To provide answers to these questions, we searched bacterial genomes for regions with clusters of homologs to the key components of P4. Given the lack of available methods to detect P4-like satellites, we studied the composition of these clusters to uncover and characterize a putative family of P4-like satellites. We quantified their abundance, which resulted in the identification of ca. 1000 novel P4-like mobile elements. This allowed us to study their prevalence and association with bacterial hosts. We also characterized the composition and organization of the gene repertoires of P4-like elements, as well as the phylogenetic history of key genes. Our results uncover the hidden diversity of P4-like satellites and highlight their role as a distinct mobile element in the microbial world.



Distribution of P4-like satellite elements in the Escherichia coli species tree (A), and the spatial localisation of these elements in the bacterial genomes (B).

Pseudo-Embryos As A Novel Quantitative Mammalian Model

Contact: leah.friedman@pasteur.fr

Leah FRIEDMAN ; Melody MERLE ; Corinne CHUREAU ; Jerome WONG NG ; Thomas GREGOR ; Affiliation: Physics of Biological Functions, Department of Developmental and Stem cell Biology, Institut Pasteur, CNRS UMR3738, 75015 Paris, France

We are developing a novel quantitative model to study mammalian embryonic development on the basis of mouse embryonic stem cell-derived pseudo-embryos, termed gastruloids. These structures can be grown at varying sizes and display an identical gene expression program to early mouse embryos. We find that size and gene expression patterns in gastruloids are reproducible and that gene expression patterns scale with gastruloid size. To further quantify the parameters involved in reproducibility and scaling with higher precision, we develop a method to determine the volume and number of cells of individual gastruloids. We also explore how machine learning methods can help understand the different stages and structures of gastruloid development.

Perspective Of Polygenic Risk Scores

Contact: leo.henches@pasteur.fr

Leo HENCHES, IP | Hugues Aschard, IP

Polygenic risk score (PRS) based on thousands of genetic variants has become a central tool for genetic prediction in multifactorial diseases. The prospect of using these PRS in clinical care has received increasing attention. Numerous studies have been conducted to investigate the strength and limitations of PRS and explored solutions to improve performances. However, while the accuracy of existing PRS will likely continue to increase with increasing sample size of genome- wide association studies and the development of new, more powerful methods, questions remain on how much prediction can be achieved in the future. Here, we used a few examples to synthesize previous results and theory and provide a perspective on future PRS studies.

Information Flow In A Gene Locus

Contact: leone.debarge@pasteur.fr

Leone DEBARGE, Benjamin ZOLLER, Isma BENNABI, Thomas GREGOR, Physics of Biological Function, DSCB department

Gene expression is a highly regulated process. A key mechanism for the regulation is the interaction between the gene promotor with its cis-regulatory region such as enhancers, but it is still unknown how enhancer and promotor interact. Recent studies using Fluorescent In Situ Hybridization (FISH) in Drosophila reveal a larger distance between enhancer and promotor during transcription (on the order of a hundred nanometers) than what a direct molecular contact would suggest. How then is the information about the state of transcriptional activity transmitted from the enhancer to the promotor ? In this project, we study the interplay between enhancer and promotor in mammals using FISH in gastruloids, and we aim to explain the transmission of information through such large distances by modelling the chromatin as a system poised at criticality in a liquid liquid phase transition.

Resistant Subpopulations In Ovarian Cancer: Tools For Their Identification And Treatment Selection

Contact: mara.santarelli@pasteur.fr

Mara SANTARELLI, Institut Pasteur, Universite de Paris, Department of Computational Biology, Systems Biology Group ; Sorbonne Universite, CNRS, LIP6, F-75005 Paris, France ; Carola DOERR, Sorbonne Universite, CNRS, LIP6, F-75005 Paris, France ; Benno SCHWIKOWSKI, Institut Pasteur, Universite de Paris, Department of Computational Biology, Systems Biology Group

High grade serous ovarian cancer (HGSOC) is the most malignant and most frequently encountered type of ovarian cancer. The 5-year survival rate is only ~ 40-50 % and has remained largely unchanged in the past 30 years, since the approval of platinum-based chemotherapy which is still the standard treatment for HGSOC patients. Although initially responding to chemotherapy, eventually most patients will develop recurrent disease and resistance to this treatment. Thus, novel therapies are urgently needed.

The PARIS (Precision drugs Against Resistance in Subpopulations) project combines preclinical research with advanced bioinformatics and medical ethics research to evolve the state of the art in the personalised treatment of chemoresistant HGSOC. Within this project, our work comprises two objectives: First, we aim at identifying and characterizing cell subpopulations that underlie the resistance to chemotherapy. For this, we will develop algorithms for the identification of resistance-related gene signatures in scRNA-seq data of HGSOC tumors and matched organoids. Second, on the basis of the inferred gene signatures, we will develop predictors for drugs that target these chemoresistance-associated subpopulations.

With our research effort, we hope to improve the prognosis for HGSOC patients.

Systems Bioinformatics, Sysbio

Contact: natalia.pietrosemoli@pasteur.fr

Giovanni BUSSOTTI, Vincent GUILLEMOT , Helene LOPEZ-MAESTRE, Damien MORNICO, Natalia PIETROSEMOLI. Hub de Bioinformatique et Biostatistique - Departement Biologie Computationnelle - Institut Pasteur, Universite de Paris ; 25-28 Rue du Docteur Roux, 75015 Paris, France ;

We incorporate systems biology approaches to characterise complex biological interactions from a holistic perspective. Our team offers IP's scientists high-level bioinformatics expertise in large-scale multi-OMICs data analysis derived from diverse high throughput technologies such as: genomics, transcriptomics, proteomics, and metabolomics. Our methods and pipelines allow to address complex study designs integrating multiple technologies, samples, donors, time points, experimental protocols and treatments. This offers unprecedented opportunities for disentangling the molecular mechanisms of biological systems, helping scientists uncover complex genotype-phenotype relationships at different scales, from single mutations to complete genomes. Our analyses may include network theory-derived approaches, and usually require the development of robust statistical and bioinformatic tools for the integration of diverse OMICs data types, each characterised by intrinsic sources of technical and biological noise.

Integrative Models For Decision Support System In Ovarian Cancer Care

Contact: oceane.fourquet@pasteur.fr

Oceane FOURQUET, Institut Pasteur, Universite de Paris, Department of Computational Biology, Systems Biology Group ; Sorbonne Universite, CNRS, LIP6, F-75005 Paris, France; ; Martin KREJCA, Sorbonne Universite, CNRS, LIP6, F-75005 Paris, France;; Carola DOERR, Sorbonne Universite, CNRS, LIP6, F-75005 Paris, France`; ; Benno SCHWIKOWSKI, Institut Pasteur, Universite de Paris, Department of Computational Biology, Systems Biology Group;

In Europe, over 40 000 women die of ovarian cancer every year. In addition to surgery, most patients are treated with platinum-based chemotherapy. Unfortunately, the effect of the chemotherapy often decreases during the treatment cycles, and currently there are few effective treatments to those patients who develop resistance to platinum-based drugs. The survival of these patients has not improved much in the past decades and new solutions are urgently needed.

As part of the DECIDER project (https://www.deciderproject.eu/), we develop tools to support clinical decision making in high-grade serous ovarian cancer by integrating various data (ex: histopathology, imaging, omics …). Together with 15 other research groups and companies, we build models to predict the efficacy of possible treatments.

Our group focuses on two aspects: 1. Prediction of relevant phenotypes using monotonic ensemble models, a new type of regression model that represents an increased level of complexity, compared to classical models, while remaining interpretable in a complex clinical context. 2. Integration of monotonic ensemble models together with other predictive models into a global inference model (credal network) as a data-driven integrative, but also interpretable model to improve clinical decision making.

Syntviewjs: A Dynamical Viewer For The Microbial Genome Analysis

Contact: plechat@pasteur.fr

Rachel Bellone, Pierre Le-Bury, Christophe Becavin, Catherine Dauga, Olivier Dussurget, Javier Pizarro-Cerda and Pierre Lechat

SynTView is a published interactive multi-view genome browser for next-generation comparative microorganism genomics. SynTViewJS is the rewrite of the software in javascript with the addition of new features. The software is characterised by the presentation of syntenic organisations of microbial genomes and the visualisation of polymorphism data obtained from next generation sequencing. SynTViewJS is built as a generic genome browser including sub-maps that hold information about genomic objects. After selecting genomes of interest, the users can explore them visually by genomic location, or directly go to specific genes by name. Several genomic maps can be stacked ordered by a phylogenetic tree according to biological metadata on top of each other. The creation of a SynTView website 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. SynTViewJS is designed to visualise information about polymorphism across a large number of bacterial strains. The SNP maps allow the user to navigate through polymorphism data sets. The non javascript tool has been used in many projects such as the study of Legionella 3 bacterial strains. I will show in the poster the study of the mutational dynamics of chikungunya virus as a function of temperature with visibility filters (mutation frequency, specificity …) with the possibility of zooming to the sequence. SynTView has been also integrated to the Listeriomics web site, a platform for visualizing and analysing every heterogeneous Listeria “omics” dataset published to date and will be integrated soon in Yersiniomics (same platform dedicated to Yersinia dataset). The tool can be uploaded to a website and the data made accessible on a server or directly added by drag and drop. Source code is available at https://gitlab.pasteur.fr/plechat/syntviewjs.

16s Rrna And Shotgun Sequencing To Characterize The Gut Microbiome

Contact: violeta.basten-romero@pasteur.fr

Violeta BASTEN ROMERO, Christophe BOETTO

Introduction and first comparison of the 16S rRNA and shotgun metagenomic sequencing methods to analyze and obtain taxonomic classification of human gut microbiome.