2020 annual workshop on Statistical Methods for Post Genomic Data (SMPGD)

Venue: Auditorium François Jacob, Institut Pasteur, Paris

Program

Thursday, January 23

09:00 Registration

09:30 Introduction

09:45 Keynote: **Julia Gog**, *University of Cambridge, Cambridge, UK*
"The dynamics of influenza evolution and spatial spread"

10:45 Coffee break

Session 1 – Mathematical models for evolution & epidemiology

organized by Amandine Veber

11:15 Miraine Davila Felipe: Phylogenetic bootstrap based on the transfer distance: applications to (large) pathogen datasets

11:45 Oliver Ratmann: Quantifying HIV transmission flow from cross-sectional viral phylogenetic deep sequence data: a population-based study in Rakai, Uganda

12:15 Simon Cauchemez: Epidemic forecasting to support policy making

12:45 Lunch Break

14:00 Keynote: **Laurent Jacob**, *LBBE, Lyon, FR*
"Learning with pangenomes"

Session 2 – Deep learning algorithms for computational biology

organized by Chloé Azencott

15:00 Flora Jay: Bits to Bases: Creating Artificial Genomes using Generative Adversarial Networks and Restricted Boltzmann Machines

15:30 Nataliya Sokolovska: Disease classification using omics data

16:00 Coffee break

INCEPTION GdR BiM GDR Statistique & Santé INRA SCIENCE & IMPACT

16:30 Slim Lotfi: kernelPSI: a powerful post-selection inference framework for nonlinear association testing in genome-wide association studies

16:50 Ghislain Durif: Genome-wide local ancestry inference in admixed individuals with scalable penalized nearest neighbor algorithm

17:10 Wessel Van Wieringen: Are reconstructed molecular networks reproducible?

17:30 Laura Cantini: Benchmarking of multi-omics joint dimensionality reduction (DR) approaches for cancer study

17:50 Cocktail and Poster session

19:00 End

INCEPTION GdR BiM GDR Statistique & Santé INRA SCIENCE & IMPACT

Friday, January 24

9:00 Keynote: **Louis Lambrechts**, *Institut Pasteur, Paris, FR*
"Dissecting the genetic basis of natural variation in mosquito susceptibility to arbovirus infection"

10:00 Cosimo Lupo: V-gene insertions and deletions during the affinity maturation process in BCR repertoires

10:20 Hélène Ruffieux: A global-local variational approach for detecting hotspots in molecular quantitative trait locus studies

10:40 Coffee break

Session 3 – Evolution of inter-species interactions
organized by Bastien Boussau et Damien de Vienne

11:10 Hélène Morlon: Accounting for interspecific interaction in phylogenetic comparative methods

11:40 Alessandra Carbone: Protein coevolution and the viral world

12:10 Emmanuelle Jousselin: The coming and going of obligate nutritional endosymbionts in aphids: insights from phylogenomic approaches

12:40 Lunch break

14:00 Keynote: **Simona Cocco**, *Ecole des Neurosciences, Paris, FR*
"Extracting features from protein sequence data with Restricted Boltzmann Machines"

15:00 Hugues Richard: Evolutionary Inference in the Context of Alternative Splicing

15:20 Ulysse Herbach: Modeling the dynamics of circulating tumor DNA for detecting resistance to targeted therapies: a phylogenetic approach

15:40 Boris Hejblum: A variance component score test for flexible RNA-seq data differential analysis

16:00 Marie Verbanck: HOPS: a quantitative score reveals pervasive horizontal pleiotropy in human genetic variation is driven by extreme polygenicity of human traits and diseases

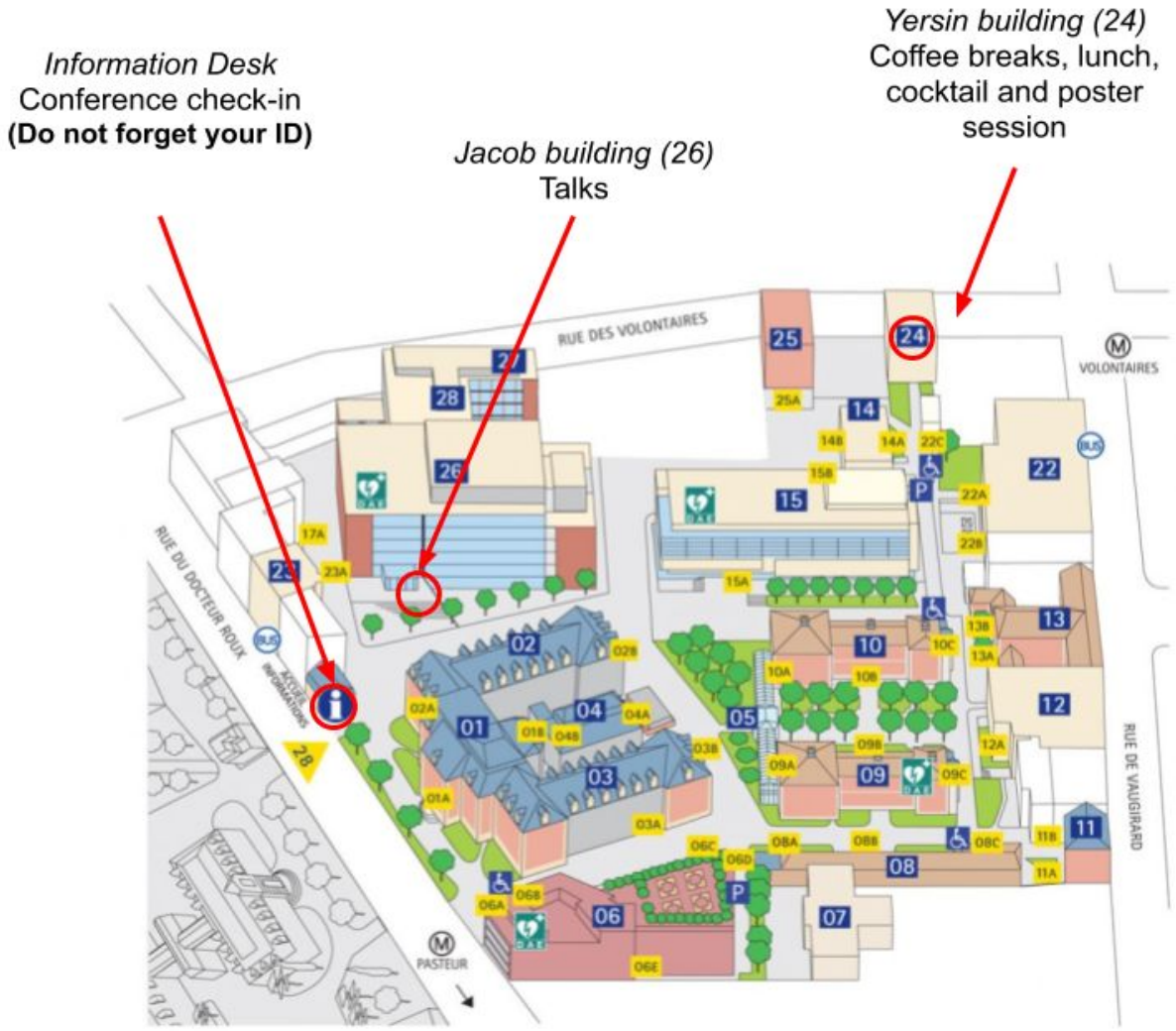
16:20 End

INCEPTION GdR BiM GDR Statistique & Santé INRA SCIENCE & IMPACT

2020 annual workshop on Statistical Methods for Post Genomic Data (SMPGD)

Venue: Auditorium François Jacob, Institut Pasteur, Paris

Map



NB: the conference room is located on the ground floor of the François Jacob building

2020 annual workshop on Statistical Methods for Post Genomic Data (SMPGD)

Venue: Auditorium François Jacob, Institut Pasteur, Paris

Abstracts - Talks

Benchmarking of multi-omics joint dimensionality reduction (DR) approaches for cancer study

Laura Cantini , Pooya Zaker, Aurelien Nald , Denis Thieffry , Elisabeth Remy, Anaïs Baudot

Dimensionality Reduction (DR), decomposing data into low-dimensional spaces while preserving most of their information content, is among the most prevalent machine learning techniques in data mining. With the advent of high-throughput technologies, high-dimensional data have become a standard in biology, emphasizing the use of DR. This phenomenon is particularly pronounced in cancer biology, where consortia have profiled thousands of patients for multiple molecular assays ("multi-omics"), including at the emerging single-cell scale. DR approaches have been mainly applied to single omics data leading to cancer subtyping, tumor sub-clones quantification and immune infiltration quantification. Recently, DR approaches designed to jointly analyze multiple omics have been proposed. Integrative DR methods are based on various mathematical assumptions, ranging from extensions of CCA, tensors, or more general data fusion approaches, which makes difficult to choose which method to apply.
In this context, we here in-depth benchmark multi-omics DR approaches using: i) artificial multi-omics cancer data ii) multi-omics bulk data from 10 different cancer types downloaded from TCGA iii) multi-omics single-cell data from cancer cell lines In (i), the capability of the various methods to predict the clustering ground truth was found strongly sensible to the size of the clusters, with intNMF, RGCCA, MCIA and JIVE being the more robust methods. For (ii), MCIA, RGCCA, MOFA and JIVE more consistently identified factors associated to survival, clinical annotations and biological annotations. Finally in (iii), despite never being applied to single-cell data, tICA and MSFA outperformed other methods for their ability to cluster single cells based on their cell line of origin. Overall, our results show that RGCCA, MCIA and JIVE perform consistently better across the three scenarios. This suggests that a mathematical formulation, based on the search of omic-specific factors whose inter-dependence is maximized, better approximates the nature of multi-omics data.

Protein coevolution and the viral world

Alessandra Carbone

A fundamental question in computational biology is the extraction of evolutionary information from protein sequences. This information relates to protein-protein binding sites and the mechanical and allosteric properties of proteins. We will present a computational approach for co-evolution analysis and apply it to the reconstruction of viral genome protein networks. We will show how the interaction network of the hepatitis C virus genome proteins can be reconstructed at a residue / domain resolution from genotype sequences and will briefly present recent experimental work demonstrating the fusion of HCV as a unique mechanism. This work provides a proof of concept for a broader exploration of viral protein-mediated processes and highlights coevolution as a valuable tool for guiding the design of viral inhibitors. In a second example, we will show how, a generalization of the method applied to the hepatitis B virus, provides important information on primary and secondary mutations in response to antiviral drugs.

Epidemic forecasting to support policy making

Simon Cauchemez

I will present modelling techniques that can be used to forecast the trajectory of infectious disease epidemics with a view to support policy making and planning. The talk will be illustrated with applications to dengue and influenza viruses and will highlight different types of challenges (e.g. absence of historical data on which models can be trained, uncertainty about key model parameters and the observation process). I will discuss how these forecasts are used by policy makers and planners, current limitations and future developments.

Extracting features from protein sequence data with Restricted Boltzmann Machines

Simona Cocco

In this talk, I will present inference/machine-learning methods to extract constraints acting on protein sequences from evolutionary sequence-data collected in Multi-sequence Alignments of protein families. I will present results on extracting information on protein structure, fitness cost due to sequence mutations, and protein design from the models inferred from data.

References: [1] Inverse Statistical Physics of Protein Sequences: A Key Issues Review. S. Cocco, C. Feinauer, M. Figliuzzi, R. Monasson, M. Weigt. Reports on Progress in Physics 81, 032601 (2018). [2] Learning protein constitutive motifs from sequence data J. Tubiana, S. Cocco, R. Monasson, eLife 2019;8:e39397 (2019).

V-gene insertions and deletions during the affinity maturation process in BCR repertoires

Lupo Cosimo, Mora Thierry, Walczak Aleksandra

The ability of the immune system to recognize and kill a huge range of external pathogens is ensured by a high diversity in the binding sites of membrane Receptors in B-Cell lymphocytes (BCR). The resulting repertoire of BCRs is updated and increased via a 2-step stochastic process for the creation (recombination) and the evolution (affinity maturation) of each nucleotide sequence encoding the receptors. The common picture of the recombination process involves a random choice of the genes from the germline DNA, plus some nucleotide deletions and insertions (briefly, indels) at the junctions of such genes. Instead, the affinity maturation of the sequence involves some context- and position-dependent point mutations, namely the exchange of some nucleotide bases. Our analysis focuses on the possibility of experiencing indels not just at the junctions between the germline genes in the recombination process, but directly in the bulk of the most variable (V) gene in the chain during the affinity maturation stage, further enhancing the variability of the repertoire. These indels appear prominently in the BCR of both healthy people and HIV-responding broadly neutralising antibodies. Within a fully probabilistic framework, we aim at introducing an effective model for the occurrence of such indels, inferring the parameters of such model from real data through maximum-likelihood approaches. This model can then be used to evaluate the likelihood of indels in real data from other patients, so to spot eg diseases that could be more likely to favour indels. Finally, this model can also be used to generate more reliable synthetic sequences, in turn useful for data analysis and comparison with further real repertoires.

Phylogenetic bootstrap based on the transfer distance: applications to (large) pathogen datasets

Miraine Davila Felipe

The phylogenetic bootstrap is a method to assess the robustness of phylogenies inferred from sequence data. Its usefulness, simplicity and interpretability have made it extremely popular in evolutionary studies since it was introduced by Joseph Felsenstein in 1985. However, it is commonly acknowledged that Felsenstein's bootstrap is not appropriate for large datasets, which are now frequent thanks to high-throughput sequencing technologies. The main reason for such shortcoming 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. Here we propose, as an alternative to the classical bootstrap, to compare a given branch in the original tree with the bootstrap branches using a smother dissimilarity index. This new index is based on the transfer distance (TD) between bipartitions of a common set. We provide empirical and theoretical results showing that our method can be a powerful tool to assess the branch support of phylogenies inferred from a large number of sequences. In particular, we study the distribution and asymptotic behavior of this new support measure for some relevant null models of phylogenetic trees (Yule, PDA, Caterpillar and totally balanced trees).

Genome-wide local ancestry inference in admixed individuals with scalable penalized nearest neighbor algorithm

Ghislain Durif, Mairal Julien, Blum Michael

In most Eukaryotes species, the transmission of genetic materials between generations is achieved through sexual reproduction. During this process, each individual inherits half of their genome from both their parents. Thanks to genetic recombination, the genome of an individual is a non-uniform combination of the genetic material of their ancestors. This process directly impacts individual phenotype transmission and species or population evolution.

During inter-population (or inter-species) breeding events, descendants inherit an admixture of genetic materials from both source populations (or species). The study of genome-wide locus ancestry (i.e. determining the population of origin of each locus) can be done with local ancestry inference (LAI). It can be very useful to characterize admixture events (time, proportion) during a species history. Local ancetry inference can also be used to study biological adaptation and penotypic variation, or to explore population-specific disease predisposition.

We present Loter2, a machine-learning-based library for genome-wide local ancestry inference, derived from Loter (https://github.com/bcm-uga/Loter). Our method uses a locus-based penalised nearest-neighbor-like approach to determine the local ancestry of each locus in haplotypes from admixed individuals, by using reference haplotypes of individuals from different potential source populations. Loter2 aims at finding for each admixed haplotypes the closest reference haplotype regarding SNP similarity. The resolution is achieved with an efficient and scalable dynamic programming algorithm (with linear complexity). Loter2 implements a specific penalized optimization scheme to account for (i) reference population intra-variability (with a penalty on switches between reference populations), (ii) phasing error in haplotypes (authorizing switches between homologous haplotypes), (iii) a priori locus similarity between admixed and reference populations (based on locus-specific supervised learning of local ancestry). In addition, we use a bagging technique to get more robust results and to avoid hyper-parameter tuning (simplified usage).

Loter2 is able to process haplotype data where haplotype estimation (or phasing) is done in silico by processing SNP genotype data (or directly obtained with haplotype sequencing). For instance, in diploid species, each locus can be homozygous (both ancestral allele or both derived allele) or heterozygous (ancestral and derived allele). We used the phasing software Beagle (Browning & Browning, 2016) in the experiments.

Performance and comparison to state-of-the-art approaches for local ancestry inference with Loter2 are proposed based on the analysis of simulated genotype data, generated with the software msprime (Kelleher et al. 2016), using human chromosome recombination maps and realistic scenarios of admixture events during human species history.

INCEPTION GdR BiM GDR Statistique & Santé INRA SCIENCE & IMPACT

The dynamics of influenza evolution and spatial spread

Julia Gog

This talk will be focussed on influenza. For seasonal influenza (the usual type of 'flu that circulates each winter), the key challenges are in understanding viral evolution in highly dynamic population models. For pandemics, past and recent data allows detailed exploration of the spatial dynamics, but raises many questions. In this talk, I will introduce some types models used for influenza evolution and spread. Then I will introduce some of the data from the "BBC pandemic" study, and introduce some potential problems of interest to those working on probabilistic models.

A variance component score test for flexible RNA-Seq data differential analysis

Boris Hejblum, Marine Gauthier, Rodolphe Thiebaut, Denis Agniel

Gene expression measurement through RNA-sequencing keeps producing ever richer high-throughput data for transcriptomics studies. As such studies grow in size, frequency, and importance, there is an urgent need for statistical methods that better control the type-I error and the subsequent False Discovery Rate. We model transformed RNA-seq counts as continuous variables using nonparametric regression to account for their inherent heteroscedasticity, in a principled, model-free, and efficient manner. We rely on a powerful variance component score test that can deal with both covariates adjustment and data heteroscedasticity to identify the genes whose expression is significantly associated with one, or even several factors of interest, in complex experimental designs (including longitudinal data), and that can also be used to perform Gene Set Analysis. Our test statistic has a simple form and limiting distribution, which can be computed quickly, and an exact permutation procedure is also derived for small sample sizes. We show that our test has good statistical properties in simulations, with an increase in stability and power when compared to state-of-the-art methods voom-limma, edgeR, and DESeq2. In particular, we show that those three methods can all fail to control the type I error when the sample size becomes larger, while our method behaves as expected. We also apply our proposed method to a public dataset studying tuberculosis.

Modeling dynamics of circulating tumor DNA for detecting resistance to targeted therapies: a phylogenetic approach

Ulysse Herbach, Alexandre Harle, Coralie Fritsch, Aurelie Muller-Gueudin, Aline Kurtzmann, Pierre Vallois, Anne Gegout-Petit, Nicolas Champagnat

Targeted therapies represent a real advance in the treatment of patients with cancer. Most of these therapies are kinase inhibitors and require precise analysis of tumor DNA mutations to ensure the absence of primary resistance. Although tumours are often genetically heterogeneous with the presence of many subclones, they release "circulating" cell-free DNA (cfDNA) that can be directly extracted from basic blood samples: as sensitivity of measurements improves, such liquid biopsies increasingly appear as a mirror of tumour heterogeneity. In this context, we describe a promising statistical approach to analyze longitudinal cfDNA data, with the purpose of gaining a deeper understanding of the mechanism by which resistance develops in specific patients. While addressing the now classic problem of reconstructing the associated phylogenetic tree (Roth et al., 2014 ; Malikic et al., 2015), this approach also describes production of cfDNA from the temporal dynamics of cells, in order to best exploit the longitudinal structure of the data (Khan et al., 2018)

Learning with pangenomes

Laurent Jacob

As the number and variety of sequenced genomes grows, representing them by comparison to a single reference leads to an increasing level of approximation, discarding accessory genes, rearrangements and repeated regions. This problem is particularly acute when studying microbial genomes or metagenomes, and hinders essential statistical tasks such as GWAS or prediction in this context. I will discuss genome representations which are well suited to statistical analysis when genomes are ill-suited to alignment or even assembly.

Bits to Bases: Creating Artificial Genomes using Generative Adversarial Networks and Restricted Boltzmann Machines

Flora Jay

Generative models have shown breakthroughs in a wide spectrum of domains due to recent advancements in machine learning algorithms and increased computational power. Despite these impressive achievements, the ability of generative models to create realistic synthetic data is still under-exploited in genetics and absent from population genetics. Yet a known limitation of this field is the reduced access to many genetic databases due to concerns about violations of individual privacy, although they would provide a rich resource for data mining and integration towards advancing genetic studies. We demonstrated that deep generative adversarial networks (GANs) and restricted Boltzmann machines (RBMs) can be trained to learn the high dimensional distributions of real genomic datasets and generate novel high-

quality artificial genomes (AGs) with little privacy loss.
We show that our generated AGs replicate characteristics of the source dataset such as allele frequencies, linkage disequilibrium, pairwise haplotype distances and population structure. Moreover, they can also inherit complex features such as signals of selection and genotype-phenotype associations. To illustrate the promising outcomes of our method, we showed that imputation quality for low frequency alleles can be improved by augmenting reference panels with AGs and that the RBM latent space provides a relevant encoding of the data, hence allowing further exploration of the reference dataset and providing features that could help solving supervised tasks.

The coming and going of obligate nutritional endosymbionts in aphids: insights from phylogenomic approaches

Emmanuelle Jousselin, Coeur d'acier, A., Clamens, A-L., Orvain, C., Cruaud ,C., Barbe, V., and A. Manzano-Marin

Many insects depend on maternally-inherited bacterial endosymbionts to provide essential nutrients and these interactions often result in long-term associations over evolutionary time scales. For instance Aphids (Hemiptera: Aphididae) have been associated for 150 Ma with the bacterium *Buchnera aphidicola* and phylogenetic analyses show that this bacterium has cospeciated with its hosts since then. Until recently this bacterium was thought to be the sole provider of essential amino acids and vitamins of aphids. However recent studies have shown that, in some aphid species, *Buchnera aphidicola* co-exists with a younger bacterial partner which is also essential for their host and whose identity varies according to aphid lineages. These newcomers offer a unique opportunity to understand how obligate association with bacterial symbionts arise and the processes driving their long-term evolution. Here we investigate the origin and evolutionary stability of two new obligate bacterial endosymbionts (*Serratia* and Erwinia) of aphids of the Cinara genus. Using bacteria whole genome data, aphid mitochondrial genomes and complex evolutionary models that are known to reduce long-branch attraction artefacts we reconstruct the phylogenetic histories of aphids and their new symbionts. We show that the association with *Erwinia* arose from a single event of symbiont lifestyle shift, from free-living to intracellular symbiont. This event resulted in drastic genome reduction, long-term genome stasis in *Erwinia*, and co-divergence with the aphid hosts. On the other hand the evolutionary history of the obligate association with *Serratia* has been more unstable: reconciliation analyses suggest that *Serratia* has been lost and regained several times during the diversification of *Cinara* aphids. Every acquisition resulted from a shift from a facultative symbiont life-style to obligate association and was associated with rapid genome erosion in *Serratia*. We discuss how the process of genomic erosion affecting endosymbionts once they become obligate partners can hamper phylogenetic reconstructions but also its putative role in the succession of symbionts.

Dissecting the genetic basis of natural variation in mosquito susceptibility to arbovirus infection

Louis Lambrechts

The mosquito Aedes aegypti is the primary vector of several arthropod-borne viruses (arboviruses) of public health significance such as dengue and Zika viruses. In nature, there is substantial variation in Ae. aegypti susceptibility to arbovirus infection at the population and individual levels. Despite their importance to understand arbovirus transmission patterns and to inform the development of novel mosquito-centered arbovirus control strategies, the genes that cause this variation are unknown. In this talk, I will illustrate the value of unbiased genetic screens to unravel the molecular determinants of Ae. aegypti susceptibility to dengue and Zika virus infections. Such genetic surveys in natural populations pave the way for identifying the mechanisms underlying the ability of mosquitoes to carry arboviruses.

kernelPSI: a powerful post-selection inference framework for nonlinear association testing in genome-wide association studies

Slim Lotfi, Chatelain Clément, Azencott Chloé-Agathe, Vert Jean-Philippe

We present the results of an extensive study, in which we demonstrate the use of kernelPSI for genome-wide association studies (GWAS). kernelPSI is a statistical tool to perform post-selection inference (PSI) for nonlinear variable selection. The nonlinearity is modeled through quadratic kernel association scores, which are a quadratic form of the response vector. The latter scores allow the incorporation of nonlinear effects and interactions among covariates. In the context of GWAS, kernelPSI assesses the effect of a predetermined genomic region e.g. gene, or regulatory region, while simultaneously identifying the causal loci within. This can facilitate the downstream biological interpretation. Notably, the identification of causal loci is a major limitation of the sequence kernel association test (SKAT), a state-of-the-art method for association testing of genomic regions. Moreover, in kernelPSI, we generalize the SKAT statistic to a broader family of association scores, hence providing a general and flexible framework to measure the association between a given locus and a phenotype of interest. Another added benefit of our framework in comparison to SKAT is its greater statistical power, as shown in several experimental settings.
kernelPSI is a two-step approach: a number of putative loci are selected in a supervised manner in the first step, and their aggregated effect on the phenotype is tested in the second step. The selection step introduces a statistical bias in the subsequent hypothesis testing. For instance, if the most associated loci are selected in the first step, the significance of their overall effect is likely to be overestimated. To answer this problem, we develop a PSI methodology in order to derive valid empirical p-values. This is achieved thanks to a constrained sampling of replicates of the response vector. We then compare the statistics of the response to those of the replicates to obtain the desired p-values. In our case study, we apply kernelPSI to a set of obesity-related phenotypes such as body mass index (BMI), weight and fat distributions. Our selection of such phenotypes was motivated by the breadth of information available in large biobanks. Yet,

kernelPSI can be applied to any continuous response.
The theoretical foundations of kernelPSI have been published in a previous work (Slim, L. et al., 2019), in addition to an eponymous R package (https://cran.r-project.org/package=kernelPSI) which implements our PSI framework with different association scores.

Accounting for interspecific interaction in phylogenetic comparative methods

Hélène Morlon

Phylogenetic Comparative Methods (PCMs) are key to our understanding of species and traits diversification. An underlying assumption of almost all PCMs, however, is that evolution on each branch of a phylogenetic tree is independent from evolution on all the other branches. While this assumption is convenient mathematically, it precludes accounting for, and quantifying the effect of interspecific interactions on macroevolutionary dynamics. I will present developments that aim to relax the assumption of interspecific independency in PCMs. Next, I will focus on the effect of interspecific competition. I will show that slowdowns in macroevolutionary rates, that are often used to detect the signal of competition during evolutionary radiations, are in fact a poor indicator of the macroevolutionary consequences of competition. Phylogenetic signal in trait data and support for models accounting for competition-driven character displacement provide more robust tests of the effect of interspecific competition. I will illustrate the utility of such models with an empirical application on ecological and social trait evolution in a large bird radiation.

Quantifying HIV transmission flow from cross-sectional viral phylogenetic deep sequence data: a population-based study in Rakai, Uganda

Oliver Ratmann, Joseph Kagaayi, Matthew Hall, Tanya Golubchick, Godfrey Kigozi, Xiaoyue Xi, Chris Wymant, Gertrude Nakigozi, Lucie Abeler-Dörner, David Bonsall, Astrid Gall, Anne Hoppe, Paul Kellam, Jeremiah Bazaale, Sarah Kalibbala, Oliver Laeyendecker, Justin Lessler, Fred Nalugoda, Larry W. Chang, Tulio de Oliveira, Deenan Pillay, Thomas C. Quinn, Steven J. Reynolds, Simon E.F. Spencer, Robert Ssekubugu, David Serwadda, Maria J. Wawer, Ronald H. Gray, Christophe Fraser, M. Kate Grabowski, the Rakai Health Sciences Program and the Pangea HIV Consortium.

To prevent new infections with human immunodeficiency virus type 1 (HIV-1) in sub-Saharan Africa, UNAIDS recommends targeting interventions to populations that are at high risk of acquiring and passing on the virus. However, it is often unclear who and where these “source” populations are. Viral deep-sequence data have recently been established as a reliable source of information for inferring the direction of transmission at an accuracy sufficient for population-level analyses. We present a novel semi-parametric Bayesian modelling framework of viral sequence data to quantify HIV transmission flows and detect source and sink population groups that directly incorporates this information. Our model uses a high-dimensional set of Poisson regression equations, adjusting for sampling heterogeneity known from cross-sectional

surveillance to avoid selection bias in the flow estimates. We illustrate the methodology on population-based sequence data obtained from infected individuals participating in the Rakai Community Cohort Study in south-eastern Uganda, indicating that the method provides new opportunities for identifying the drivers of HIV spread.

Evolutionary Inference in the Context of Alternative Splicing

Hugues Richard, Diego Javier Zea, Adel Ait-Hamlat, Alexander Korzec, Sofya Laskina, Antoine Labeeuw, Elodie Laine

Alternative splicing (AS) has the potential to greatly expand the proteome in eukaryotes by producing several transcript isoforms from the same gene. Although these mechanisms are well described at the genomic level, little is known about their contribution to protein evolution and their impact at the protein structure level. Here, we address both issues by proposing a set of methods to reconstruct the evolutionary history of transcripts and to jointly model the tertiary structures of the corresponding protein isoforms. We reconstruct AS evolution across multiple species and combine it with structural information in order to functionally annotate the different transcripts.¬† We present our contributions to the problem of inferring evolutionary scenarios and phylogenies in the context of alternative splicing. We first proposed a method for the inference of phylogenies on a set of transcript observed in a set of species, which is implemented in the PhyloSofS tool (Ait-Amlat et al, in revision). We then extended this method into a principled framework designed to cope with a more general data structure: the splice graph. A splice graph is a directed acyclic graph where nodes are exons and transcripts are paths in the graph. The alternative paths in the splice graph summarize the set of AS events occurring over a gene. The problem of reconstructing ancestral splice graphs has to our knowledge never been done before, and comes with its own combinatorics and modeling challenges. We proposed an efficient method to compute maximum likelihood estimates of possible ancestral splice graphs. Doing inference over splice graphs rather than transcripts have various advantages: a gene is optimally described as a collection of AS events rather than a list of transcripts and the evidence from sequencing data -still sparse, can be easily integrated. We present preliminary results on a set of 14 genes over 9 species where functionally distinct alternative splicing events have been experimentally determined and where we reconstructed phylogenies while adding RNA-Seq evidence from thousands of samples.

A global-local variational approach for detecting hotspots in molecular quantitative trait locus studies

Hélène Ruffieux, Richardson Sylvia, Bottolo Leonardo

We tackle modelling and inference for variable selection in regression problems with many predictors and many responses. We focus on detecting hotspots, i.e., predictors associated with several responses. Such a task is critical in statistical genetics, as hotspot genetic variants shape the architecture of the genome by remotely controlling the expression of many gene products and may initiate decisive functional mechanisms underlying disease endpoints.Existing hierarchical regression approaches designed to model hotspots suffer from two limitations: their discrimination of hotspots is sensitive to the choice of top-level scale parameters for the propensity of predictors to be hotspots, and they do not scale to large predictor and response spaces, e.g., with thousands to hundreds of thousands of variables in genetic applications.We address these shortcomings by introducing a flexible hierarchical regression framework that is tailored to the detection of hotspots and scalable to the above dimensions. Our novel framework allows information-sharing across outcomes and variants, thereby enhancing the detection of weak effects, and directly controls the hotspot propensity via a dedicated top-level representation. In particular, it implements a fully Bayesian model for hotspots based on the horseshoe shrinkage prior: its global-local formulation shrinks noise globally and hence accommodates the highly sparse nature of genetic analyses, while being robust to individual signals, thus leaving the effects of hotspots unshrunk. Inference is carried out using a fast variational algorithm coupled with a novel simulated annealing procedure that allows efficient exploration of multimodal distributions. We illustrate the merits of our approach in an expression quantitative trait locus (eQTL) study of monocytes after immune stimulation. This is joint work with Sylvia Richardson and Leonardo Bottolo. Software is available at https://github.com/hruffieux/atlasqtl.

Disease classification using omics data

Nataliya Sokolovska

Deep learning (DL) techniques have shown unprecedented success when applied to images. In cases where the sample size is much bigger than the number of features, the DL often outperforms other machine learning techniques, often through the use of Convolutional Neural Networks (CNNs). However, in many bioinformatics fields (including metagenomics), we encounter the opposite situation where the number of features is significantly greater than the number of observations. In these situations, applying DL techniques would lead to severe over-fitting. We aim to improve the classification of various diseases with metagenomic data through the use of CNNs applied to omics data represented as images, and we discuss possible approaches.

HOPS: a quantitative score reveals pervasive horizontal pleiotropy in human genetic variation is driven by extreme polygenicity of human traits and diseases

Marie Verbanck, Daniel Jordan, Do Ron

Horizontal pleiotropy, where one variant has independent effects on multiple traits, is important for our understanding of the genetic architecture of human phenotypes. A challenge is distinguishing horizontal pleiotropy from its counterpart, vertical pleiotropy, which is defined as a genetic variant having an effect on one phenotype which in turn has a downstream effect on another phenotype. Due to this challenge, it is unknown the extent of horizontal pleiotropy that is present in human genetic variation. We developed the HOrizontal Pleiotropy Score (HOPS), a novel method to quantify horizontal pleiotropy through a variant-level pleiotropy score using summary statistics from published genome-wide association studies (GWASs). HOPS relies on applying a statistical whitening procedure to a set of input variant-trait associations, which removes correlations between traits caused by vertical pleiotropy and normalizes effect sizes across all traits. Using the whitened association Z-scores, we measure two related but distinct components of horizontal pleiotropy: the total magnitude of effect on whitened traits and the total number of whitened traits affected by a variant. After validation on simulations, this two-component quantitative pleiotropy score can be used to measure horizontal pleiotropy for all variants in the human genome. When applied to 372 heritable medical phenotypes measured in 337,119 humans from the UK Biobank, HOPS detected a significant excess of horizontal pleiotropy. This signal of horizontal pleiotropy was pervasive throughout the human genome and across a wide range of phenotypes, but was especially prominent among highly polygenic phenotypes. We further showed that horizontal pleiotropy was significantly enriched in actively transcribed regions and active regulatory regions and was correlated with the number of genes and tissues for which the variant is an eQTL. In addition, we found that variants that are eQTLs for genes whose orthologs are associated with multiple measurable phenotypes in mice or yeast have higher pleiotropy scores, demonstrating that our pleiotropy score is also related to pleiotropy in model organisms. Finally, we identified thousands of loci with extreme horizontal pleiotropy, a majority of which had never been reported in any published GWAS. Our results highlight the central role horizontal pleiotropy plays in the genetic architecture of human phenotypes, and the importance of modeling horizontal pleiotropy in genomic medicine. The HOrizontal Pleiotropy Score (HOPS) is available on Github at https://github.com/rondolab/HOPS.

Are reconstructed molecular networks reproducible?

Van Wieringen Wessel, Chen Yao

The ability of the immune system to recognize and kill a huge range of external pathogens is ensured by a high diversity in the binding sites of membrane Receptors in B-Cell lymphocytes (BCR). The resulting repertoire of BCRs is updated and increased via a 2-step stochastic process for the creation (recombination) and the evolution (affinity maturation) of each nucleotide sequence encoding the receptors. The common picture of the recombination process involves a random choice of the genes from the germline DNA, plus some nucleotide deletions and insertions (briefly, indels) at the junctions of such genes. Instead, the affinity maturation of the sequence involves some context- and position-dependent point mutations, namely the exchange of some nucleotide bases. Our analysis focuses on the possibility of experiencing indels not just at the junctions between the germline genes in the recombination process, but directly in the bulk of the most variable (V) gene in the chain during the affinity maturation stage, further enhancing the variability of the repertoire. These indels appear prominently in the BCR of both healthy people and HIV-responding broadly neutralising antibodies. Within a fully probabilistic framework, we aim at introducing an effective model for the occurrence of such indels, inferring the parameters of such model from real data through maximum-likelihood approaches. This model can then be used to evaluate the likelihood of indels in real data from other patients, so to spot eg diseases that could be more likely to favour indels. Finally, this model can also be used to generate more reliable synthetic sequences, in turn useful for data analysis and comparison with further real repertoires. This is joint work with Yao Chen.

2020 annual workshop on Statistical Methods for Post Genomic Data (SMPGD)

Venue: Auditorium François Jacob, Institut Pasteur, Paris

Abstracts - Posters

CloneSig: Joint Inference of intra-tumor heterogeneity and signature deconvolution in tumor bulk sequencing data

Judith Abecassis, Fabien Reyal, Jean-Philippe Vert

The possibility to sequence DNA in cancer samples has triggered much effort recently to identify the forces at the genomic level that shape tumorigenesis and cancer progression. It has resulted in novel understanding or clarification of two important aspects of cancer genomics: (i) intra-tumor heterogeneity (ITH), as captured by the variability in observed prevalences of somatic mutations within a tumor, and (ii) mutational processes, as revealed by the distribution of the types of somatic mutation and their immediate nucleotide context. These two aspects are not independent from each other, as different mutational processes can be involved in different subclones, but current computational approaches to study them largely ignore this dependency. In particular, sequential methods that first estimate subclones and then analyze the mutational processes active in each clone can easily miss changes in mutational processes if the clonal decomposition step fails, and conversely information regarding mutational signatures is overlooked during the subclonal reconstruction. To address current limitations, we present CloneSig, a new computational method to jointly infer ITH and mutational processes in a tumor from bulk-sequencing data, including whole-exome sequencing (WES) data, by leveraging their dependency. We show through an extensive benchmark on simulated samples that CloneSig is always as good as or better than state-of-the-art methods for ITH inference and detection of mutational processes. We then apply CloneSig to a large cohort of 8,954 tumors with WES data from the cancer genome atlas (TCGA), where we obtain results coherent with previous studies on whole-genome sequencing (WGS) data, as well as new promising findings. This validates the applicability of CloneSig to WES data, paving the way to its use in a clinical setting where WES is increasingly deployed nowadays.

Gene prioritization for Mendelian disorders integrating disease-associated gene lists, patient omic data and gene functional networks.

Felix Brechtmann, Patricia Goldberg Figueira, Ziga Avsec, Julien Gagneur

Rare diseases affect 6 to 8% of the European population, the equivalent of about 30 million individuals. Often, these diseases have a genetic basis. Pinpointing the genetic cause of a rare disease is essential to stratify patients, enables preimplantation and preconception genetic screens and provide a rationale for therapy developments. However, the majority of patients undergoing exome or genome sequencing do not receive a genetic diagnosis. One reason for this is that there are still many disease-associated genes to be discovered. Historically, two flavours of gene prioritization techniques have been established. Some algorithms prioritize genes based on their relatedness to known disease-causing genes. Other algorithms, including burden tests, make use of genome or exome sequencing databases and gene sets (pathways) to identify genes or pathways enriched for deleterious variants in cases versus controls. We will show current progress on GeneProf, an algorithm that improves the gene prioritization by integrating both strategies. Application to gene prioritization for mitochondrial diseases will be shown. Also, extensions of the model to integrate RNA-seq data will be discussed.

Combining network-guided GWAS to discover susceptibility mechanisms for breast cancer

Hector Climente-Gonzalez, Christine Lonjou, Fabienne Lesueur, Investigators Genesis, Dominique Stoppa-Lyonnet, Nadine Andrieu, Chloé-Agathe Azencott

Systems biology provides a comprehensive approach to biomarker discovery and biological hypothesis building. It does so by jointly considering the statistical association between a gene and a phenotype, obtained experimentally, and the biological context of each gene, represented as a network. In this work we study the utility of six network methods to discover new biomarkers in GWAS data by searching subnetworks highly associated to a phenotype. We interrogate a familial breast cancer GWAS focused on BRCA1/2 negative French women. By trading statistical astringency for biological meaningfulness, most network methods get more compelling results than standard SNP- and gene-level analyses, recovering causal subnetworks tightly related to cancer susceptibility. We perform an in-depth benchmarking of the methods with regards to the size of the solution subnetwork, their utility as biomarkers, and the stability and the runtime of the methods. Interestingly, a combination solution subnetworks provided a concise subnetwork of 93 genes, enriched in known familial breast cancer susceptibility genes (BABAM1, BLM, CASP8, FGFR2, and TOX3, Fisher's exact test p-value = 7.8×10^{-5}) and more central than average. Additionally, it includes subnetworks of mechanisms related to cancer, like protein folding (HSPA1A, HSPA1B, and HSPA1L) or mitochondrial ribosomes (MRPS30, MRPS31, MRPS18B). We also observed a general dysregulation in the neighborhood of COPS5, a gene related to multiple hallmarks of cancer.

Leveraging persistent genetic effects using the Conditional False Discovery Rate boosts the power to detect genotype-environment interactions

Rachel Moore, **Loukia Georgatou-Politou**, James Liley, Oliver Stegle, Inès Barroso

To address this and to enable assessing GxE effects on a genome wide scale without the aforementioned limitations, we here propose a strategy that adapts the conditional False Discovery Rate (cFDR) (Andreassen, O. A. et al., 2013 ; Liley, J. & Wallace, C., 2019 ; Liley, J. & Wallace, C., 2015) to test for GxE using our recently proposed StructLMM model (Moore et al., 2019), a method that tests for interactions between variants and complex multivariate environments. For a given hypothesis, cFDR is defined as the posterior probability of being non-null given the p-value of the tested hypothesis and the corresponding covariate value, and hence this strategy avoids the need to define arbitrary association thresholds. To validate our strategy, we assessed and compared cFDR's power with the conventional FDR multiple testing correction as well as two-step filtering procedures for different selection thresholds using simulated data. The results showed that overall cFDR performs better than any of its comparison partners. Next, we applied cFDR to UK Biobank (UKBB) data to test for GxE effects on BMI. We considered data from 252,153 unrelated UKBB individuals of European ancestry for which BMI and 64 lifestyle environmental factors were available. We randomly split the data into a discovery set of 126,077 and a validation set of 126,076 individuals and considered 7,515,856 variants. Using the discovery subset, we identified known and novel GxE signals, many of which replicated in the validation dataset. When applying cFDR to the full UK Biobank dataset, we identified 140 loci with significant GxE effects. In contrast, the genome-wide FDR and the conventional two-step filtering approaches identified 6 and 23 loci respectively. Finally, it is important to note that while we applied cFDR in combination with the StructLMM interaction test, it is equally valid to use it with any other method that tests for GxE effects and therefore has wide applicability.

Interpreting black box models through variable importance

Jonathan Ish-Horowicz, Sarah Filippi, Seth Flaxmn

While the predictive performance of non-parametric models such as Gaussian processes and neural networks is well-established across a variety of domains, our ability to explain and interpret these methods is limited - they operate as black boxes. This lack of interpretability means that variable importance, a key question in biomedical applications, cannot be determined. Here, we present RATE [Crawford et al 2019, Annals of Applied Statistics; Ish-Horowicz et al 2019, arXiv:1901.09839], a method for establishing variable importance for probabilistic black box models. In the context of such models a feature is rarely important on its own, so our strategy is specifically designed to leverage partial covariance structures and incorporate variable interactions into our proposed variable ranking. RATE applies an information theoretic criterion to the posterior distribution of effect sizes to assess variable significance and its performance is demonstrated on simulation studies and examples from statistical genetics.

COSMONET: An R package for survival analysis using screening-network methods

Antonella Luliano

Identifying relevant genomic features that can act as prognostic markers for building predictive models for survival analysis is a central theme in personalised medicine. However, the high dimension of genome-wide omic data, the strong correlation among the features and the low sample size greatly increase the difficulty of cancer survival analysis demanding for the development of specific statistical methods and software. Here, we present a novel R package, 'COSMONET', that implements a multistage computational-statistical procedure combining screening techniques and network-regression methods able to identify prognostic gene signatures and predict patient survival outcome. In particular, COSMONET implements (i) three different screening approaches to reduce the initial dimension from an high-dimensional space p to a moderate scale d and (ii) two network-penalized Cox-regression regression algorithms to fit the observed survival times and genome-wide expression profiles, and (iii) a prediction step based on the evaluation of a prognostic index. Moreover, COSMONET provides a number of graphical functions for visualize survival curves, gene signatures, sub-networks and many others. We demonstrate the use of our package by applying it to a breast cancer dataset downloaded from the TCGA portal.

A statistical and bioinformatic framework for functional interpretation of single-cell RNA-Seq

Loredana Martignetti, Akira Cortal, Antonio Rausell

Rapid advances in single-cell approaches have opened up new ways of studying cell-to-cell variability in different contexts, including personalized medicine to characterize disease pathogenesis and outcome. Dedicated bioinformatic and statistical methods need to be developed to deal with the peculiarities of single-cell measurements in order to extract robust and relevant information (Stegle et al., 2015). Here we propose a statistical and bioinformatic framework, called Sample-ID, based on Singular Value Decomposition, developed to extract a sample identity card in the form of unbiased gene signature to characterize the observed transcriptional heterogeneity within a sample. Per-sample signatures, or sample fingerprints, allow, in clinical setting, to “blast” query patient samples against reference libraries from genetically-characterized patients, thus favoring molecular diagnosis. Sample-ID has been systematically applied to single-cell RNA-seq datasets from (i) the Human Cell Atlas project, profiling healthy human organs and tissues, and (ii) in-house collections of rare disease patients profiling the affected organs and tissues.

How to use local score for scan statistics with undefined window size to highlight atypical concentration of events in continuous sequences

Sabine Mercier

Scan statistics, sliding windows and local score are three usual methods used to highlight atypical regions in sequences. In Glaz et al. (2009), traditional scan statistics of a sequence of length n, which can be discrete or continuous, is defined as the maximum over 0 is defined as the maximal count of observed event in a window with a prescribed length w. Counts and are often modeled using a Poisson distribution. Results on the distribution of scan statistics have already been established (see Guenin (2013) for a review). For *A* a discrete sequence of length *n* taking its values in a component alphabet ({A,C,G,T} for DNA for example). Sliding windows consist in intervals of a given size *m* that literally “slides” across the studied. They are mapped to files containing signal or annotations of interest. They can so be considered as scan statistics but they differ from them because the considered values can be different from counts of observations. For a given scoring function f that attributes a real or integer value to each component, chosen depending on the studied context (nucleotide frequency variability for profile sequences, hydrophobic scoring scheme for transmembrane sequences, for examples) the sliding windows statistic corresponds to the maximum over i=1,...,n-m+1 of f(A(i)+...+f(A-i+m-1). Thresholds of the statistics are established in practice using Monte Carlo approach. Local score approach consists in sliding windows without any restriction on the window size and is defined as follows. For a discrete sequence A of length n and a given scoring function f, the local score of the sequence is defined by the maximum over every possible beginings i and every possible ends j of f(A(i))+...+f(A(j)). Different results on the distribution of the local score exists (see Karlin el al. (1990, 1992) and Mercier (2001, 2007, 2019) for example). Local scores do not need a previous knowledge of the length of the region of interest we want to highlight but is only defined for discrete sequence. They are interesting when non positive scores are possible. Scan statistics and sliding windows are based on a fixed length window. Scan statistics can be used for continuous sequences and focus on event counts. I will present an approach to highlight region of unknown length with atypical concentration of events in a continuous sequence. This method allows, conditionally to the total number of events in the sequence, to use the local score statistic for which a *p*-value can be computed in diverse contexts.

TADreg: A versatile regression framework for TAD identification, differential analysis and predictions

Raphael Mourad

In higher eukaryotes, the three-dimensional (3D) organization of the genome is intimately related to numerous key biological functions including gene expression, DNA repair and DNA replication regulations. Alteration of this 3D organization is detrimental to the organism and can give rise to a broad range of diseases such as cancers. In particular, topologically associating domains (TADs) represent a pervasive structural feature of the genome organization.Here, we propose a versatile framework which not only identify TADs in a fast and efficient manner, but also detect differential TAD borders across conditions for which few methods exist. Moreover, the regression can predict Hi-C data in the case of structural variants, thereby allowing studying the deleterious impact of de novo enhancer-promoter interactions. The framework is biologically relevant since it models long-range interactions depending on the presence of insulating elements with varying strength, and model parameters have an intuitive interpretation and are easy to visualize.

Efficient Simulation of spreading processes on adaptive networks

Malysheva Nadezhda

Epidemiological process modeling and simulation are powerful tools allowing to predict disease prevalence and optimize public health strategies. The class of adaptive network models become a significant part of modern epidemiological modelling study. Infectious diseases are transmitted through contacts, which dynamically change in time. On the other hand, contact behavior of the individual can dramatically change after his infection. Traditional approaches (e.g. SIS, SIR modeling) as well as static network approaches shown to lacks important temporal information about causal paths that underlie the spreading process, consequently providing false conclusions for the control of the spreading process. In the class of adaptive network models, the network structure itself changes dynamically in response to the dynamics of the spreading. Epidemic spreading has been extensively studied on these types of networks to understand the influence of social contact structure on disease prevalence.
The dynamics of contagion spreading on adaptive networks are usually complex, and analytical results can only be obtained in a very few particular cases. This makes numerical studies based on stochastic simulations indispensable. Several computational approaches are available for today: The stochastic simulation algorithm (SSA) allows the exact numerical simulation of contagion spreading on adaptive networks. However, it may generate computational overhead, when the contact dynamics are considerably faster than the contagion dynamics. The computational overhead becomes even more observable, when the outcome of intervention strategies, like vaccination, prophylaxis or “treatment as prevention” are studied which further lower the per-contact contagion transmission probability.

Existing Inexact methods discretize the time and then perform parallel updates of the contact structure and the contagion spreading (akin to a tau-leaping procedure). Due to the parallel updates of the contacts and spreading dynamics some essential information regarding causal pathways could be missed, what can dramatically affect the outcome prediction.
The method proposed by Vestergaard assumes deterministic contact dynamics, e.g. a temporal set of static contact configurations is used, which may be derived from a set of data. However, in original implementation the underlying contact network is not adaptive any longer; i.e., the network configurations affect the spreading dynamics, but not vice versa.
We propose an efficient stochastic simulation method that allows to sufficiently predict events related to the spreading process while network dynamics is captured approximately with desired accuracy. This method benefiting in performance in cases when both the contact dynamics, as well as the spreading dynamics are governed by inter-dependent stochastic processes. Particularly, in the case when contact dynamics are considerably faster than the contagion dynamics (for example, in sexually transmitted diseases like HIV) when existing computational approaches either fail dramatically or cause computational overhead.

Multi-task group Lasso for Genome Wide Association Studies

Asma Nouira, Chloé-Agathe Azencott

Population stratification refers to the presence of differences in allele frequencies between subpopulations within samples due to different ancestry. The presence of population stratification is one of the major challenges in Genome Wide Association Studies (GWAS) as it increases type I error and leads to ambiguous results. This occurs when allele frequencies differences in cases and controls are due to differences in ancestry rather than association between genotype and phenotype. Thus, it is possible that different subpopulations do not share the same causal Single-Nucleotide Polymorphisms (SNPs) associated with the disease. Several methods have been developed to adjust for population stratification, such as genomic control, structured association and PCA-based methods. However, the adjustment could be uniform for all SNPs, which can result in over correction of some population-specific causal SNPs. From a biological perspective, the SNPs are correlated due to Linkage Disequilibrium (LD), which is related to the distance between markers. Association studies on single SNP level can miss some causal regions due to the small size of population samples compared to the number of markers. Considering SNPs in the same LD block jointly can alleviate this issue. (Dehman et al., 2015) introduce an approach for the partition of LD blocks and incorporate it in a single task group Lasso regression. Our main motivation is to propose an efficient framework for selecting SNPs at the block level associated with the disease in admixed populations. In this contribution, we introduce a multi-task feature selection approach where each task corresponds to a subpopulation. To this end, we combine the multi-task Lasso approach (at a single SNP level) of (Puniyani et al., 2010) with the (Dehman et al., 2015) group Lasso.
Our algorithm provides the selection of a shared features set across all tasks, and also the selection of a set of population specific SNPs. We illustrate the efficiency of the proposed

method on a realistic simulated dataset following the LD patterns of the HapMap data. We also compare the proposed multi-task group Lasso approach to PCA-based adjustment methods which include the top principal components as covariates in regression models, such as EIGENSTRAT (Price et al., 2006) and statistical tests based on logistic regression.

UMI-VarCal: a new UMI-based variant caller that efficiently improves low-frequency variant detection in paired-end sequencing NGS libraries

Vincent Sater, Pierre-Julien Viailly, Thierry Lecroq, Philippe Ruminy, Elise Prieur-Gaston, Mathieu Viennot, Helene Dauchel, Fabrice Jardin

Motivation: Next Generation Sequencing (NGS) has become the go-to standard method for the detection of Single Nucleotide Variants (SNV) in tumor cells. The use of such technologies requires a PCR amplification step and a sequencing step, steps in which technical artifacts are introduced at very low frequencies. These artifacts are often confused with true low-frequency variants that can be found in tumor cells and cell-free DNA. The recent use of Unique Molecular Identifiers (UMI) in targeted sequencing protocols has offered a trustworthy approach to filter out artifactual variants and accurately call low frequency variants. However, the integration of UMI analysis in the variant calling process lead to developing tools that are significantly slower and more memory consuming than raw-based variant callers.
Results: Here we present UMI-VarCal, a UMI-based variant caller for targeted sequencing data with remarkably higher sensitivity compared to raw-reads-based variant callers. Being developed with performance in mind, UMI-VarCal stands out from the crowd by not relying on Samtools to do its pileup. Instead, at its core runs an innovative homemade pileup algorithm specifically designed to treat the UMI tags present in the reads. After the pileup, a Poisson statistical test is applied at every position to determine whether the allele frequency of the variant is significantly higher than the background error noise. Finally, an analysis of UMI tags is performed and a strand bias filter is applied to achieve better accuracy. We illustrate the results obtained using UMI-VarCal through the sequencing of both biopsy and plasma samples and we show how UMI-VarCal is both faster and more sensitive than other publicly available solutions.