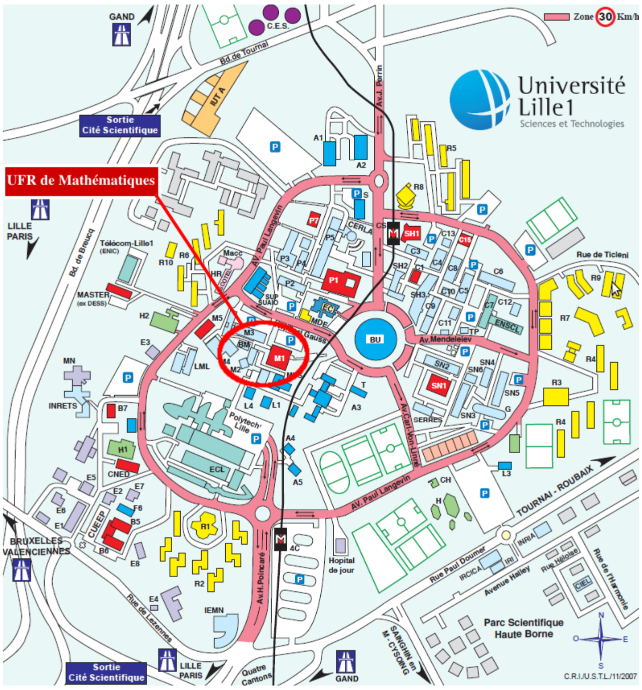
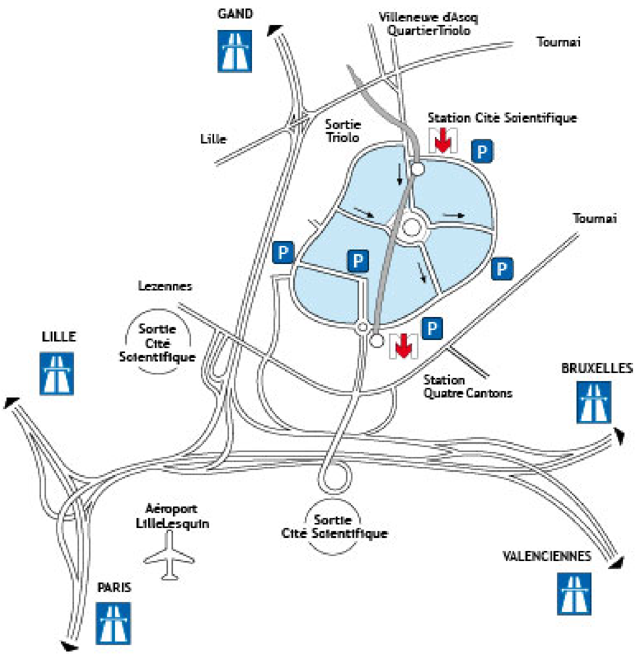
Mathematical Models for Epidemiology and Phylogenetics
Lille, May 30-31 2016.

Organizers: Viet Chi Tran (Lille 1), Amaury Lambert (UPMC - Collège de France).

The conference takes place at Université de Lille, Cité Scientifique, **Amphitheater Bernoulli** in Building “M1-Mathématiques”.

Venue:
From the train station Lille Flandres (train station Lille Europe is next to Lille Flandres, you have a 5 min walk between the two stations), take the Metro Line 1 in the direction 4 Cantons. Leave the metro at the stop Cité Scientifique - Pr. Gabillard.
Once out of the metro station, follow the direction of the library Lilliad - Learning Center, which is the white round building that you can see from the metro station. Following this road leads you to pass the ”Café culturel” and you arrive to the building ”M1 - Mathématiques”.
In the building M1, go up the stairs to the main corridor. The Amphitheater Bernoulli is on the left.

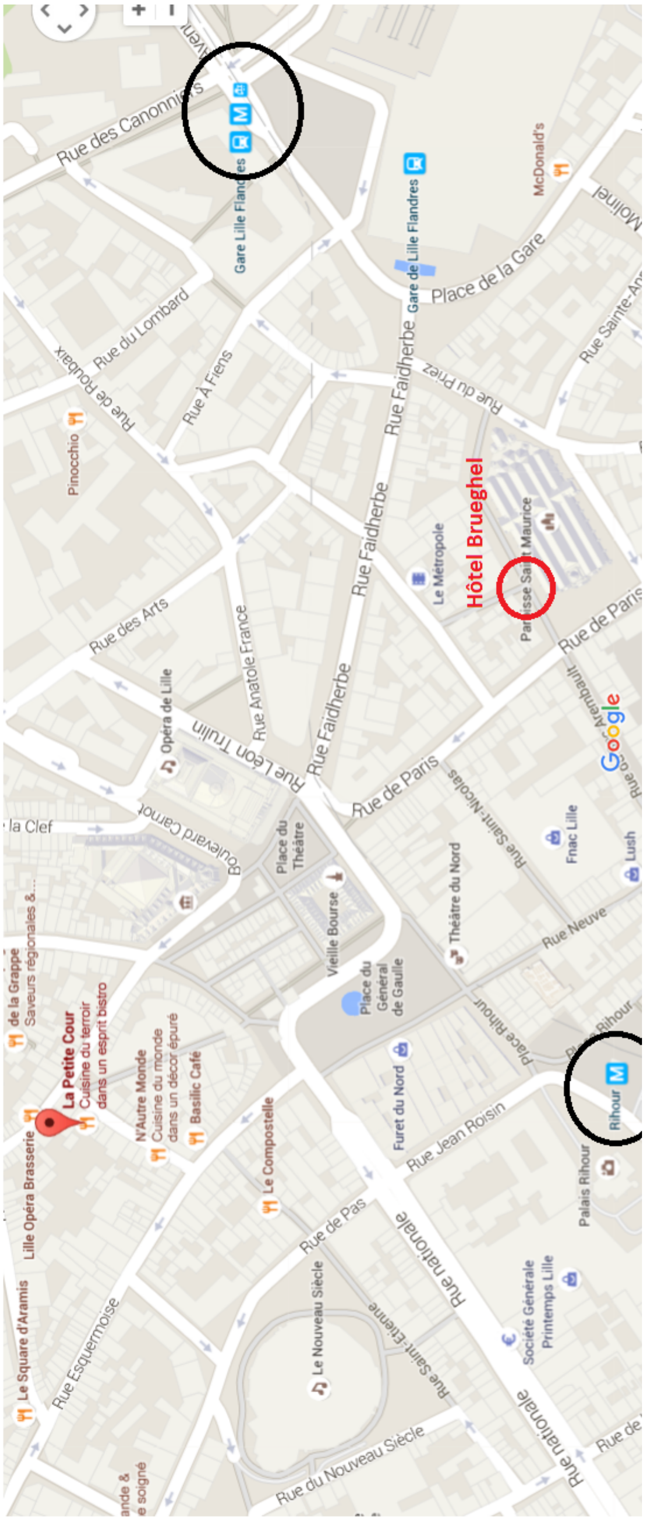




Hotels: see website of the workshop. Invited speakers are at Hôtel Brueghel 5 Parvis Saint-Maurice (see red circle on the map).

Lunches: take place at the Restaurant Sully (R1 in yellow on the campus map).

Restaurant of the workshop (Monday night 19:30) For those who registered, the conference diner will take place at **La Petite Cour**, 17 rue du Curé St-Etienne, in the center of Lille. Tel: 03 20 51 52 81.



Monday, May 30th	
10:30-13:00	**Céline Poux** (Tutorial 1): Introduction to concepts and methods in molecular phylogenetics
Lunch	
14:30-15:30	**Olivier Gascuel** (Tutorial 2): Fast Dating using Least Squares.
16:00-17:00	**Vivian Kouri**: HIV-1 viral variants circulating in Cuba. Implications for disease progression.
17:15-18:15	**Miraine Dávila Felipe**: Joint likelihood of the reconstructed transmission tree and the epidemic size process via time reversal dualities.

Tuesday, May 31st	
9:00-10:30	**Erik Volz** (Tutorial 3): Mathematical models for pathogen gene genealogies
11:00-12:00	**Samantha Lycett**: Inferring transmission patterns in animal disease systems using phylogenetics
12:00-13:00	**Patrick Hoscheit**: The Lambda-skyline process
Lunch	
14:30-15:30	**Olivier Robineau**: Phylogenetic cluster analysis as a tool to understand HIV transmission process : example from a Paris neighborhood.
15:45-16:45	**Philippe Lemey**: Integrating covariates in phylodynamic processes of pathogen sequence and trait evolution

Contents

1 Céline Poux: Introduction to concepts and methods in molecular phylogenetics

In this introduction we will go through all the necessary steps to reconstruct a phylogenetic tree. The first step is to prepare the data for the analyses; this required aligning the data and removing the uninformative sites. Then we must choose the reconstruction methods. Two methodological families exist: the phenetic methods based on the overall resemblance between sequences and the cladistic methods based on the characters (i.e. nucleotidic positions). In this second family we can find the most common methods used now a day: the probabilistic methods (maximum likelihood and Bayesian methods). They require evolutionary models, therefore we will need to assess the different DNA substitution models and choose the best one according to our data. The following step, ones a phylogeny is reconstructed, is to estimate the strength of the reconstruction. This can be done according to several methods that will be presented (bootstrap, posterior probabilities). Finally it is important to critically assess the result of an analysis because several types of errors and artifacts can lead to erroneous topologies. We will review the most common encountered problems. To finish the tutorial I will display some examples of studies using phylogenetic reconstruction in ecology and evolution.

2 Olivier Gascuel: Fast dating using least squares

Phylogenies provide a useful way to understand the evolutionary history of genetic samples, and data sets with more than a thousand taxa are becoming increasingly common, notably with viruses (e.g., human immunodeficiency virus (HIV)). Dating ancestral events is one of the first, essential goals with such data. However, current sophisticated probabilistic approaches struggle to handle data sets of this size. Here,we present very fast dating algorithms, based on a Gaussian model closely related to the Langley-Fitch molecular-clock model. We show that this model is robust to uncorrelated violations of the molecular clock. Our algorithms apply to serial data, where the tips of the tree have been sampled through times. They estimate the substitution rate and the dates of all ancestral nodes. When the input tree is unrooted, they can provide an estimate for the root position, thus representing a new, practical alternative to the standard rooting methods (e.g., midpoint). Our algorithms exploit the tree (recursive) structure of the problem at hand, and the close relationships between least squares and linear algebra. We distinguish between an unconstrained setting and the case where the temporal precedence constraint (i.e., an ancestral node must be older that its daughter nodes) is accounted for. With rooted trees, the former is solved using linear algebra in linear computing time (i.e., proportional to the number of taxa), while the resolution of the latter, constrained setting, is based on an active-set method that runs in nearly linear time. With unrooted trees the computing time becomes (nearly) quadratic (i.e., proportional to the square of the number of taxa). In all cases, very large input trees (¿10,000 taxa) can easily be processed and transformed into time-scaled trees. We compare these algorithms to standard methods (root-to-tip, r8s version of Langley-Fitchmethod, and BEAST). Using simulated data, we show that their estimation accuracy is similar to that of the most sophisticated methods, while their computing time is much faster. We apply these algorithms on a large data set comprising 1194 strains of Influenza virus from the pdm09 H1N1 Human pandemic. Again the results show that these algorithms provide a very fast alternative with results similar to those of other computer programs. These algorithms are implemented in the LSD software (least-squares dating), which can be downloaded from `http://www.atgc-montpellier.fr/LSD/`, along with all our data sets and detailed results.

3 Vivian Kouri: HIV-1 viral variants circulating in Cuba. Implications for disease progression.

Authors: Vivian Kouri[1], Ricardo Khouri[2,4], Lissette Pérez[1], Yoan Alemán[1], Jorge Pérez[1], Carlos Fonseca[1], Lilia M Ortega[1], Jorge Campos[1], Yoeri Schrooten[2], Lore Vinken[2], Jurgen Vercauteren[2], Andrea-Clemencia Pineda-Peña[2], Kristof Theys[2], Sarah Megens[2], Michel Moutschen[5], Nico Pfeifer[6], Johan Van Weyenbergh[2], Celia M Limia[1], Yudira Soto[1], Anne-Mieke Vandamme[2,3], Kristel Van Laethem[2].

1. Laboratory of Sexually Transmitted Diseases, Virology Department, Institute of Tropical Medicine "Pedro Kourí", Havana, Cuba.

2. KU Leuven - University of Leuven, Department of Microbiology and Immunology, Rega Institute for Medical Research, Clinical and Epidemiological Virology, B-3000 Leuven, Belgium.

3. Centro de Malária e Outras Doenças Tropicais and Unidade de Microbiologia, Instituto de Higiene e Medicina Tropical, Universidade Nova de Lisboa, Lisboa, Portugal.

4. LIMI, Centro de Pesquisa Gonçalo Moniz, FIOCRUZ, Salvador-Bahia, Brasil.

5. AIDS Reference Center, Centre Hospitalier Universitaire de Liège, Liège, Belgium.

6. Department of Computational Biology and Applied Algorithmics, Max Planck Institute for Informatics, Campus E1 4, 66123 Saarbrücken, Germany.

Background: The HIV-1 epidemic in Cuba exhibits an extraordinarily high genetic diversity, in contrast to the rest of the Caribbean region. Clinicians reported an increasing trend of rapid progression (RP) (AIDS within 3 years of infection) in Cuba. The objectives of this study were to determine the HIV-1 subtype distribution and evolution and to evaluate the association of rapid progression with epidemiological, clinical, viral and immunological parameters.

Material and Methods: Samples were isolated from 926 HIV-1 patients (342 therapy-naive and 584 therapy-experienced) attending the "Pedro Kouri" Institute in Cuba between January 2008 and December 2014. Additionally, 95 of these patients were further studied for disease progression; 52 Rapid progressors at AIDS diagnosis (AIDS-RP) and 21 without AIDS in the same time frame (non-AIDS). Twenty two patients were sampled at AIDS diagnosis (chronic-AIDS) retrospectively assessed as > 5 years infected.
HIV-1 subtype was determined using Rega Subtyping Tool version 3, and confirmed by manual phylogenetic analysis, using CLUSTAL X and the neighbor-joining method in MEGA version 5. Time trends were investigated using 5 year intervals. For the 95 patients, pol and env sequences were used, transmission cluster analysis, and prediction of resistance, co receptor use and evolutionary fitness. Measurement of cytokines and chemokines in plasma were determined by FlowCytomix using fluorescent beads. Host, immunological and viral predictors of RP were explored through data mining.

Results: The most prevalent HIV-1 genetic forms in this dataset were subtype B (30.9%), BG recombinants (22%) and CRF19_cpx (17.5%). The distribution of subtypes and recombinants was not significantly different between therapy-experienced and therapy-naïve patients. Subtype B infection was associated with male ($p = 0.022$ OR: 1.6; CI: 1.1-2.5) and MSM ($p < 0.001$ OR: 1.2; 95%CI: 1.4-2.9), while subtypes A, F, G and H were associated with heterosexuals ($p < 0.005$). Subtype H was more frequently detected among patients living in the east part of the country ($p = 0.003$ OR: 1.7; CI: 1.2-2.3). The prevalence of subtypes A, C, F, G and H among individuals diagnosed with HIV-1 dropped significantly after 1990 ($p < 0.05$), while CRF BGs (20, 23, 24) significantly increased since 2001 ($p < 0.0001$ OR:2.9; IC:1.9-4.5) and viral variant CRF19_cpx, in samples taken since 2011 (13.5% to 20.2%, $p = 0.0001$, OR:4,33; IC:2,9-6,4). Conversely, subtype B showed a significant parabolic trend, increasing up to 2000, and decreasing again in subsequent years ($p < 0.05$). Among the 95 samples analyzed for determine factors associated to disease progression, CRF19_cpx, oral candidiasis and RANTES levels were identified as strongest predictors of AIDS-RP. CRF19_cpx was more frequently associated with CXCR4 co-receptor use, higher fitness scores in the protease region (PR), and higher viral load at diagnosis.

Conclusions: This study indicates that the genetic diversity of the Cuban HIV-1 epidemic is very high. In recent years, the frequency of local recombinants is increasing while subtype B is decreasing. We propose that CRF19_cpx is evolutionary very fit and causing rapid progression to AIDS in many newly infected patients in Cuba.

4 Miraine Dávila Felipe: Joint likelihood of the reconstructed transmission tree and the epidemic size process via time reversal dualities

In recent years, the availability of pathogen sequences has been constantly increasing, and with it, the interest in using reconstructed phylogenies to infer the parameters controlling the epidemiological mechanisms. Linking these phylogenies to more traditional sources of information can improve our understanding about the dynamics of infectious diseases. We consider in particular the situation where the data consists in incidence time series (number of new cases registered through time) and the reconstructed transmission tree (estimated from pathogen sequences from present-time hosts). We then characterize their joint distribution, assuming these observed statistics are generated from a unique forward in time process and then, are not independent in general. In our approach, the evolution of the population of infectious individuals is described using general branching processes, where the underlying splitting trees encode the transmission history of epidemics. An important aspect of our model, is that no restrictive assumption is made on the distribution of the duration of infection. The results are achieved via contour techniques for random trees and time reversal dualities for Lévy processes.

5 Erik Volz: Mathematical models for pathogen gene genealogies

The genetic diversity of many pathogens is shaped by epidemiological history. But, the dynamics of infectious disease epidemics differ in important ways from demographic processes that have traditionally been studied by population geneticists. In many epidemics, the population size and birth rate changes rapidly in a nonlinear fashion through time. Mathematical models for describing infectious disease dynamics have a long history that has run parallel to the development of modern population genetics, but until recently, there has been little communication between these fields.Interest has grown in developing a new set of mathematical models for genealogies generated by epidemic processes. These methods reveal how the effective population size of a pathogen depends on transmission rates, the number of infected hosts, and the size of the bottleneck at the time of transmission. These mathematical models have also enabled new applications of pathogen genetic data to public health. Pathogen genetic data can be informative about epidemic processes in ways that standard surveillance data are not, especially regarding the source of infections and risk factors for transmission. I will review several approaches to mathematical modeling of pathogen genealogies.

6 Samantha Lycett: Inferring transmission patterns in animal disease systems using phylogenetics

Viral and bacterial pathogens undergo error prone replication, and their genomes accumulate mutations over short time spans, hence can be used to create time scaled trees in order to track the spread of infection. Assuming that one sequence can represent a sampled infected individual, and that individuals may be grouped by location (or host) into "demes", these deme labels can be included as discrete traits upon phylogenetic trees to allow the transition rates between demes to be estimated; which can be interpreted as a transmission network. Alternatively, locations may be included as continuous spatial coordinate traits, diffusion rates estimated and spatial transmission routes inferred.

Here I will describe how time resolved phylogenies augmented with either discrete or continuous geographic information can be used to infer transmission networks from sequence data of fast evolving pathogens, particularly given limited sequence diversity and potentially biased sampling. With these approaches last year's highly pathogenic avian influenza outbreaks are examined, and the route and bird type responsible for the European and North American incursions are inferred. Additionally, using the spread of bovine viral diarrhoea into and around Scotland as an example, I will show how predictors of transmission patterns can be identified, and attempt to estimate the relative importance of local spatial spread vs transmission due to animal movements in a trade network.

7 Patrick Hoscheit: The Lambda-skyline process

The classical skyline process (Pybus et al. 2000), and its many extensions such as the Bayesian Skyline, the Skyride or the Skygrid, are the most commonly used tools for the inference of demographic history from phylogenetic samples. Mathematically, the theory relies on the properties of the Kingman coalescent, which is a random tree describing the genealogy of randomly sampled individuals in a large population. It belongs to a larger class of random tree processes, the so-called Lambda-coalescents (Pitman 1999), which include non-binary trees. In this talk, I will present joint work with Oliver Pybus (University of Oxford) on the application of Lambda-coalescents, especially Beta(2-alpha,alpha)-coalescents to the analysis of multifurcating viral phylogenies.

8 Olivier Robineau: Phylogenetic cluster analysis as a tool to understand HIV transmission process : example from a Paris neighborhood

Clustering analysis is actually extensively used to study HIV epidemiological process. Description of phylogenetic clustering in HIV infected population seems to be of great interrest to describe factor associated with HIV transmission and to define subpopulation of highly risky behaviour individuals. But comparison of individuals belonging to cluster or not or trying to define individuals belonging to the same clusters and their roles in terms of transmission could be challenging in its interpretation. I will describe this issue by using a study based from a population of Men having sex with Men in a Paris neighborhood.

9 Philippe Lemey: Integrating covariates in phylodynamic processes of pathogen sequence and trait evolution

The field of phylodynamics has witnessed a rich development of statistical inference tools with increasing levels of sophistication, but these tools traditionally focus on sequences as their sole data source. Integrating various sources of information, however, promises to deliver more precise insights in infectious diseases and to increase opportunities for statistical hypothesis testing. Here, I will discuss emerging concepts of data integration that are stimulating new advances in Bayesian evolutionary inference methodology. These approaches include connecting sequence to trait evolution, such as for host, phenotypic and geographic sampling information, but also the incorporation of covariates of evolutionary and epidemic processes in the reconstruction procedures. I will highlight how a full Bayesian approach to covariate modelling and testing can generate further insights into sequence and trait evolution in pathogen populations.