MCEB 2023 --- Cargese, Corsica

June 12-16 2023

Preliminary program

KEYNOTES

Probabilistic programming and new inference algorithms: expanding what is possible in Bayesian analysis of evolution and biodiversity

Fredrik Ronquist

https://ronquistlab.github.io/people.html

~

Diversification Rate Estimation from Phylogenies

Sebastian Höhna

https://hoehnalab.github.io/

~

An overview of deep learning approaches in population genetics

Flora Jay

http://flora-jay.blogspot.com/p/research.html

~

Modeling large-scale biotic turnover in the African flora: novel analytical approaches and standing challenges

Isabel Sanmartin

https://rjb.csic.es/personal-cientifico/isabel-sanmartin-bastida/

~

TBA

Céline Scornavacca

https://isem-evolution.fr/en/membre/scornavacca/

~

TBA

Nicola De Maio

https://www.ebi.ac.uk/people/person/nicola-de-maio/

ORAL PRESENTATIONS

Simulation-based inference of demographic parameters under spatial models of isolation by distance

Raphael Leblois* 1 , François Rousset 2 , and Thimothée Virgoulay 1

1 Centre de Biologie pour la Gestion des Populations – Centre de Coopération Internationale en Recherche Agronomique pour le Développement, Institut de Recherche pour le Développement, Institut National de Recherche pour l'Agriculture, l'Alimentation et l'Environnement, Institut Agro Montpellier,
Université de Montpellier – France
2 Institut des Sciences de l'Evolution de Montpellier – Centre National de la Recherche Scientifique –France

Model-based analysis of neutral genetic data allows to indirectly estimate demographic and historical parameters such as population sizes, migration rates or divergence times because those parameters shape the repartition of genetic variability within and between populations over time. In numerous species, dispersal is spatially-limited (individuals preferentially find geographically close mates) and individuals may be spread over a continuous habitat rather than aggregated into discrete panmictic populations. However, inference methods accounting for localized dispersal still bear a number of limitations in terms of biological complexity of the underlying spatial models and of type of information brought by the analyses (e.g. which parameters can be estimated, as well as their biological interpretation).

In this study, we used a recent simulation-based inference method coupled with a spatial genetic data simulator to infer local demographic parameters of population in a continuous habitat. Our results show that we can estimate with good precision more parameters than with previously available methods, notably by independently inferring population density, dispersal rate and the shape of the dispersal distribution. In contrast to competing studies, we reach these results without assuming that the total population size or the habitat size is known. Instead, these results are possible because the simulations do not require coalescent approximations (such as assuming large population size and small migration rate), and because simulation-based methods can exploit summary statistics for which no simple analytical expectation is known. These results highlight the power of simulation-based inference in population genetics and pave the way for new demo-genetic inferences under more realistic spatial population genetic models

Multiple pulses of speciation and introgression characterize a rapid marine radiation

Martin Helmkampf* 1 and Oscar Puebla 1

1 Leibniz Centre for Tropical Marine Research (ZMT) – Germany

Adaptive radiations are an important source of biodiversity, yet few have been studied in diverse and complex environments like coral reefs, where most of the Earth's species richness is found. Hamlets are a species complex of Wider Caribbean reef fishes that have diversified in one of the fastest speciation bursts known in vertebrates. Representing an excellent opportunity to study the evolutionary processes governing adaptive radiations, we sequenced more than 320 genomes from all 18 currently described species, which were collected at 15 sites across their entire range. Genome-wide phylogenetic analyses identified a deep evolutionary split between two lineages – a smaller one consisting of several species endemic to the Gulf of Mexico, and the other containing all Caribbean species nested within the remaining Gulf species. Along with patterns of nucleotide diversity, this topology suggest that the radiation's center of origin lies within the Gulf of Mexico. While some species and populations form highly supported clades (including color variants that have diverged sufficiently to deserve the status of new species), most of the widely distributed species and even populations were found to be para- or polyphyletic. According to D-statistics and phylogenetic network analyses, gene flow between species has been widespread among both the Gulf species and the Caribbean radiation, which may have overwritten much of the underlying tree-like history. In sum, a complex picture emerges of a rapid adaptive radiation characterized by multiple pulses of speciation and fueled by high ancestral variation and widespread introgression. As more population genomic comparisons become available, these features may become apparent as hallmarks of adaptive radiations in general.

Traditional phylogenetic models fail to account for variations in the effective population size

Rui Borges* 1 , Ioanna Kotari 1 , Juraj Bergman 2 , Madeline Chase 3 , Carina Farah Mugal 4 and Carolin Kosiol 5

1 Vetmeduni Vienna – Austria
2 Aarhus University – Denmark
3 Uppsala University – Sweden
4 University of Lyon 1 – UMR5558 LBBE – France
5 University of Saint Andrews – United Kingdom

A substitution represents the emergence and fixation of an allele in a population or species and is the fundamental event from which phylogenetic models of sequence evolution are devised. Because of the increasing availability of genomic sequences, we are now able to take advantage of intraspecific variability when reconstructing the tree of life. As a result, substitutions can be more realistically modeled as the product of mutation, selection, and genetic drift. However, it is still unclear whether this increased complexity affects our measures of evolutionary times and rates. This study seeks to answer this question by contrasting the traditional substitution model with a population genetic equivalent using data from 4385 individuals distributed across 179 populations and representing 17 species of animals, plants, and fungi. We found that when the population genetics dynamic is modeled via the substitution rates, the evolutionary times and rates of the two models are well correlated, suggesting that the phylogenetic model is able to capture the time and pace of its population counterpart. However, a closer inspection of this result showed that the traditional models largely ignore the effect of the effective population size, even when it is explicitly accounted for in the substitution rates. Our findings suggest that superimposing population-genetics results on the substitution rates is an effective strategy to study mutation and selection biases, while other data sources (e.g., life history traits or polymorphisms) may need to be additionally integrated to make the traditional substitution models sensitive to the impact of genetic drift. When combined with the known effect of ancestral population size on generating phylogenomic incongruence due to incomplete lineage sorting, our findings provide further evidence that unaccounted-for variations in the effective population size may be one of the primary causes of errors in phylogenetic analyses at shorter time scales.

The spatial -Fleming-Viot Process and its relation to Birth-Death models

Johannes Wirtz* 1 and Stephane Guindon 1

1 Laboratoire d'Informatique de Robotique et de Microélectronique de Montpellier – CNRS UMR5506 – France

The spatial Lambda-Fleming-Viot Process (SLFV) is a mathematical tool to model the evolution of a population in space and time that has found applications in phylogeography and epideminology. In this talk we will take a look at some key features of the model in two spatial dimensions. Interestingly, when considering different parameter limits, it is possible to view the SLFV as an intermediate between other well-known and well-understood population-genetical models. Here we will focus on its resemblance to the standard birth-death process that emerges in the limit of high intensity and low spatial dispersal variance, and discuss the results of a simulation study we performed. Of particular focus will be the question whether and to which degree the similarity between the two models carries over to their respective tree-generating processes.

Selection on the fly - Application of a Bayesian method to detect targets of selection in Evolve-and-Resequence experiments

Carolin Kosiol* 1

1 Centre for Biological Diversity, University of St Andrews – United Kingdom

Experimental evolution studies are powerful approaches to unveil the evolutionary history of lab populations. Such studies have shed light on how selection changes phenotypes and genotypes by being combined with high-throughput sequencing techniques in so-called Evolve-and-Resequence (E&R) experiments. We present Bait-ER – a fully Bayesian approach based on the Moran model of allele evolution to estimate selection coefficients from E&R experiments. The model has overlapping generations, a feature that describes several experimental designs found in the literature. We tested our method under several different demographic and experimental conditions to assess its accuracy and precision, and it performs well in most scenarios. Furthermore, we analyse allele frequency trajectories in Drosophila pseudoobscura where we altered their sexual selection regime for 200 generations and sequenced pooled populations at 5 time points. The intensity of sexual selection was either relaxed in monogamous populations (M) or elevated in polyandrous lines (E). We found genomic signatures of adaptation for both selection regimes on D. pseudoobscura. There are more significant variants on E lines as expected from stronger sexual selection. However, we found that the response on the X chromosome was substantial in both treatments, only more marked in E and restricted to chromosome arm XR in M. The effective population size N e is lower on the X at the start of the experiment, which might indicate a swift adaptive response at the onset of selection. Additionally, we show that the third chromosome was inparticular affected by elevated polyandry. Its distal end harbours a region showing a strong signal of adaptive divergence in E lines.

Toward single-cell phylodynamics: Quantifying developmental processes based on genetic lineage tracing data

Sophie Seidel* 1, 2, Antoine Zwaans 1,2 and Tanja Stadler 1,2

1 Department of Biosystems Science and Engineering, ETH Zurich, Switzerland
2 Swiss Institute of Bioinformatics (SIB), Lausanne, Switzerland

The cell phylogeny which traces all present-day cells of an organism back to their ancestral zygote contains valuable information on the relationships between the cells, the origin of cell types and their rates of cell division, death and differentiation (1). Until recently, it was impossible to construct such a cell phylogeny for a more complex organism, as evidenced by the fact that only the complete cell phylogeny of *C. elegans* is known to date. In recent years, single-cell lineage tracing methods were developed to record a signal to reconstruct the cell phylogeny. They generally rely on an enzyme to edit a pre-defined genomic region and generate data similar to those classically analysed by phylogenetic methods. However, several traditional assumptions are invalid for this new data type. Therefore, targeted approaches such as TiDeTree (2) have been developed. Nevertheless, these early lineage recorders acquired edits only over a short time, resulting in limited phylogenetic signal in the data. On the experimental side, a newer generation of lineage recorders, for example, TypeWriter (3), is based on repeated insertions allowing for longer editing periods and therefore likely to provide more phylogenetic signal. In this talk, I will present our work on a substitution model and likelihood calculation for this new lineage recording system which we implemented as a software package in BEAST 2. We validated the model on simulated data. Further, we performed phylodynamic analyses to estimate a time-scaled cell phylogeny and the developmental process of the rapid expansion of cells in a 2D culture. Finally, I will highlight some of the successes and challenges we face when quantifying developmental processes based on lineage tracing data.

(1) Phylodynamics for cell biologists; T. Stadler, O.G. Pybus, M.P.H. Stumpf

(2) TiDeTree: a Bayesian phylogenetic framework to estimate single-cell trees and population dynamic parameters from genetic lineage tracing data; S. Seidel, T. Stadler

(3) A time-resolved, multi-symbol molecular recorder via sequential genome editing, J. Choi,..., J. Shendure

Mechanistic phylodynamic models do not provide conclusive evidence that non-avian dinosaurs were in decline before their final extinction

Bethany Allen* 1 , Maria Volkova Oliveira 2 , Tanja Stadler 1 , Timothy Vaughan 1 , and Rachel Warnock 3

1 Department of Biosystems Science and Engineering, ETH Zurich – Switzerland
2 Independent – Switzerland
3 Geozentrum Nordbayern, Friedrich-Alexander-Universität – Germany

Phylodynamic models can be used to estimate diversification trajectories from time-calibrated phylogenies. We applied two such models to phylogenies of non-avian dinosaurs, a clade whose evolutionary history has been widely debated. While some previous authors have suggested that the clade experienced a decline in diversity, potentially starting millions of years before the end-Cretaceous mass extinction, others have suggested that the group remained highly diverse right up until the Cretaceous-Paleogene (K-Pg) boundary. We show that model assumptions, particularly with respect to incomplete sampling, have a large impact on whether dinosaurs appear to have experienced a long-term decline or not. Our results are also highly sensitive to the phylogeny used. Developing comprehensive models of sampling bias, and building larger and more accurate phylogenies, are likely to be necessary steps for us to determine whether dinosaur diversity was or was not in decline prior to the end-Cretaceous mass extinction.

A Bayesian approach to characterize the distribution of adaptive effects from multiple population data.

Lucy Peters∗ 1 and Bertrand Servin 1

1 Génétique Physiologie et Systèmes dÉlevage – Ecole Nationale Vétérinaire de Toulouse, École nationale supérieure agronomique de Toulouse [ENSAT], Institut National de Recherche pour l'Agriculture, l'Alimentation et l'Environnement : UMR1388, Institut National de Recherche pour l'Agriculture, l'Alimentation et l'Environnement – France

Detecting signatures of selection across the genome is a central objective in population genetics and of great importance in the field of evolutionary biology, but also in commercial settings, such as in animal breeding. To distinguish patterns of selection from neutral processes an appropriate null model that can account for complex demographic histories is needed. The FLK statistic (Bonhomme et al. 2010) uses a phylogenetic estimation of the kinship matrix of populations, which integrates historical branching and heterogeneity of genetic drift. Here we extend on the FLK statistic by specifying an alternative model that describes a locus under selection by adding a covariate b (drawn from a Gaussian distribution) that models adaptive effects. We then calculate a locus specific Bayes Factor from FLK's neutral model and our alternative model and use it to detect selection signals across the genome. Our method not only gives an estimate of the relative effect sizes of the bs, it also enables us to identify where within the population tree these effects occur. We then assess our method using simulations and found that it performs well when the bs come from a distribution with sufficiently large variance. We also validate our method using an empirical dataset of 27 French sheep breed from the SMARTER project. We were able to recover several previously identified signals of selection across these populations in addition to identifying new regions under selection.

An analytical derivation of the distribution of distances between heterozygous sites in diploid species to efficiently infer demographic history

Peter Arndt* 1 , Florian Massip 2 , and Michael Sheinman 3

1 Max Planck Institute for Molecular Genetics – Germany
2 CBIO-Centre for Computational Biology – MINES ParisTech, PSL-Research University – France
3 Institute of Advanced Studies, Sevastopol State University – Russia

Heterozygous sites along a diploid genome are not uniformly distributed. On the contrary, their density varies as a consequence of recombination events, and their local density reflects the time to the last common ancestor of the maternal and paternal copy of a genomic region. The distribution of the density of the heterozygous sites therefore carries information about the history of the population size. Despite previous efforts, an exact derivation of the heterozygous sites distribution is still missing. As a consequence, estimating population size variations is difficult and requires several simplifying assumptions. Using a novel theoretical framework we are able to deduce an analytical formula for distribution of distances between heterozygous sites. Our theory can account for arbitrary demographic histories including bottlenecks and more general scenarios where population size is temporally constant during several epochs. In case the population size is constant throughout, the distribution follows a simple function and exhibits a power-law tail $1/r\hat{}a$ with $a = 3$ where r is the distance between heterozygous sites. This prediction is nicely validated when considering heterozygous sites in individuals from African populations. Other populations migrated out of Africa and underwent at least one bottleneck which left a distinctive mark in their interval distribution between heterozygous sites, i.e. an over-representation of intervals of length from 10 to 100 kbp. Our analytical theory for non-constant population sizes reproduces this behavior and can be used to study historic changes in population sizes with high accuracy. The simplicity of our approach makes it easier to analyse demographic histories for diploid species including human, great apes, rodents and flies, requiring only a single unphased genome.

Positive selection in highly recombining genes is not an evidence for a benefic effect of recombination

Julien Joseph* 1 and Thibault Latrille 2

1 Laboratoire de Biométrie et Biologie Evolutive - UMR 5558 – Université Claude Bernard Lyon 1, Institut National de Recherche en Informatique et en Automatique, VetAgro Sup - Institut national d'enseignement supérieur et de recherche en alimentation, santé animale, sciences agronomiques et de l'environnement, Centre National de la Recherche Scientifique – France
2 Department of Computational Biology, Université de Lausanne, Lausanne, – Switzerland

Meiotic recombination is one of the main forces driving the evolution of genomes. It breaks genetic linkage, allowing natural selection to act independently on alleles of different selective values. It is thus theoretically expected that genes with higher recombination rate are under more efficient selection. In this sense, several studies showed that highly recombining genes experienced higher rates of positive selection. The authors concluded that the dissipation of Hill-Robertson interferences decreased the number of deleterious variants and increased the number of adaptive ones. However, it has been shown in many species across the tree of life that recombination can also induce a segregation bias towards GC alleles called GC-Biased gene conversion (gBGC), which can interfere with selection. Using a mutation-selection model, we show that a slight decrease in the genome-wide intensity of gBGC can lead to positive selection in highly recombining genes without invoking any adaptation or Hill-Robertson interferences. We then show that this mechanism is likely to explain the positive selection in highly recombining genes in humans.

Comparative analysis of phylogenetic diversity indices

Kerry Manson* 1

1 University of Canterbury – Christchurch, New Zealand

Phylogenetic diversity indices aim to measure species' individual contributions to overall biodiversity. As such, they can be used to create prioritisation rankings of species for conservation management. Recent work has expanded the number of diversity indices beyond the popular Fair Proportion and Equal Splits indices. This talk covers subsequent approaches to understand and analyse the entire space of diversity indices. We evaluate various competing approaches against a series of properties, such as robustness to changes in phylogeny through extinction and/or small changes in tree topology. This is joint work with Mike Steel.

A combined model for heterogeneous fossilized birth-death inferences

Joëlle Barido-Sottani* 1 and Hélène Morlon 1

1 Institut de biologie de l'ENS Paris – Département de Biologie - ENS Paris, Institut National de la Santé et de la Recherche Médicale, Centre National de la Recherche Scientifique – France

The fossilized birth-death (FBD) process integrates information from the fossil record and has been widely used to improve phylogeny and divergence times estimates. In parallel, rate-heterogeneous models which allow for lineage-specific variations in diversification rates are the subject of increasing interest, as variations in evolutionary rates are generally widespread throughout the Tree of Life. However, so far these two types of model have only been used separately. Here, I present a combined implementation of a rate-heterogeneous FBD model which allow for clade-specific variations both in the evolutionary process but also in the preservation and fossilization process. I explore its behaviour and demonstrate its accuracy on simulated datasets. In particular, I show whether the model can distinguish between scenarios where variations are only present in the preservation or in the diversification process, or where both processes vary independently or jointly.

Phylogenetic Context Using Phylogenetic Outlines

Daniel Huson* 1

1 University of Tuebingen – Germany

Phylogenetic placement can be used to place a draft genome into a reference phylogenetic to determine its taxonomic identity. Here we propose to use phylogenetic outlines, rather than a fixed tree, to place a given draft genome into its phylogenetic context within the GTDB taxonomy. Calculations use mash sketches and Bloom filters to quickly determine related reference genomes, and then use neighbor-net and the outline algorithm to produce a visualization. A phylogenetic outline is a new type of phylogenetic network that is more general than a phylogenetic tree, but significantly less complex than other types of phylogenetic networks. We propose to use such networks, rather than trees, to represent phylogenetic context because they can express uncertainty in the placement of taxa, whereas a tree must always commit to a specific branching pattern. We illustrate the new method using a number of draft genomes of different assembly quality.

Summary tests of introgression are highly sensitive to rate variation across lineages

Cécile Ané* 1 and Lauren Frankel 1

1 University of Wisconsin-Madison – United States

Methods based on summary statistics from subsets of 3 or 4 taxa are popular to detect hybridization from genome-wide sequence data. These methods often carry the assumption of a constant substitution rate across lineages and genes. We quantified the effects of rate variation on the D-statistic (also known as the ABBA-BABA test), the D3 statistic, and HyDe. All three tests are used widely across a range of taxonomic groups, in part because they are very fast to compute. We considered rate variation across species lineages, across genes, their lineage-by-gene interaction, and rate variation across gene-tree edges. For all three methods, we found a marked increase in the false discovery of reticulation when there was rate variation across species lineages. The D3 statistic was the most sensitive: it appeared to more sensitive to a departure from the clock than to the presence of reticulation. For all three tests, the power to detect hybridization events decreased as the number of hybridization events increased, indicating that multiple hybridization events can hide one another if they occur within a small subset of taxa.

Integrative taxonomy using traits and genomic data for species delimitation with deep learning

Manolo Perez* 1,2 , Ricarda Riina 3 , Brant Faircloth 4 , Marcelo Cioffi 1 , and Isabel Sanmartin 3

1 Federal University of São Carlos – Brazil
2 Muséum national d'Histoire naturelle – Institut de Systématique, Évolution, Biodiversité, UMR 7205 ISYEB MNHN – France
3 Real Jardín Botánico de Madrid – Spain
4 Louisiana State University – United States

Recognizing species boundaries in complex speciation scenarios, including gene flow and demographic fluctuations, combined with the plethora of existing species concepts, is a challenge that has recently been brought to attention. Promising recent approaches consider an integrative taxonomy with multiple sources of evidence (e.g., genetic, morphology, geographic distributions), which can be related to diverse properties associated with the dynamics of the speciation continuum. Also, the use of statistical inferential methods for model comparison, such as approximate Bayesian computation, approximate likelihoods, and machine learning, has allowed a better assessment of species boundaries in such cases. However, most approaches involve analyzing genetic and phenotypic/geographical information separately, followed by visual/qualitative comparison. Methods that integrate genetic information with other sources of evidence have been limited to simple evolutionary models and are not able to analyze more than a few hundred loci across a maximum of a few tens of samples. Here, we present a deep learning approach that combines convolutional neural networks and multilayer perceptrons to integrate both genomic data (thousands of loci or single nucleotide polymorphisms, SNPs) and trait information into a unified framework. By using simulated and empirical datasets, we evaluate the power and accuracy of our approach for discriminating among competing allopatric speciation scenarios when varying the number of SNPs and traits, and the impact of missing data. We found that the accuracy of our method was lower for datasets that included only trait data compared to datasets that combined both genomic and trait data types and that considered genomic data alone. However, when we violated the simple allopatric speciation model by including migration, the combined approach had better performance than analyzing only the genomic information. Moreover, using both sources of data can incorporate complementary information associated with different stages of the speciation process. Our approach was able to recover the expected delimitation scenarios in empirical datasets of one plant (Euphorbia balsamifera) and one fish (Lepomis megalotis) species complex. We argue that our method is a flexible and promising approach, allowing for complex scenario comparison and the use of multiple types of data.

Effects of discordance between species and gene trees on phylogenetic diversity conservation

Kristina Wicke* 1 , Mareike Fischer 2 , and Laura Kubatko 3

1 New Jersey Institute of Technology – United States
2 University of Greifswald – Germany
3 The Ohio State University – United States

Phylogenetic diversity indices such as the Fair Proportion (FP) index are frequently discussed as prioritization criteria in biodiversity conservation. They rank species according to their contribution to overall diversity by considering the unique and shared evolutionary history of each species as indicated by its placement in an underlying phylogenetic tree. Traditionally, phylogenetic trees were inferred from single genes and the resulting gene trees were assumed to be a valid estimate for the species tree, i.e., the "true" evolutionary history of the species under consideration. However, nowadays it is common to sequence whole genomes of hundreds or thousands of genes, and it is often the case that conflicting genealogical histories exist in different genes throughout the genome, resulting in discordance between individual gene trees and the species tree. In this talk, we analyze the effects of gene and species tree discordance on prioritization decisions based on the FP index.

POSTERS

A fast and efficient index for rogue taxa detection in large datasets

Paul Zaharias∗ 1 , Frédéric Lemoine 2, and Olivier Gascuel 1

1 Institut de Systématique, Evolution, Biodiversité – Museum National d'Histoire Naturelle, Ecole Pratique des Hautes Etudes, Sorbonne Universite, Centre National de la Recherche Scientifique, Université des Antilles, UMR7205 – France
2 Institut Pasteur, Paris - France

Taxa with unstable phylogenetic positions are usually qualified as ”rogue”. In the context of bootstrap support (e.g., in Maximum Likelihood phylogeny reconstruction), rogue taxa are usually responsible of a drop of phylogenetic support because of their unstable position in the bootstrap trees. The standard approaches to deal with rogues are ”leaf-dropping methods”, where the goal is to find a set of leaves (i.e., potential rogues) to remove in order to maximize the number of (resolved) internal branches in the (bootstrap) consensus tree. Thus, some of these methods allow to quantify the degree of ”rogueness” (i.e., instability) of each taxon, but most of them have been evaluated in the context of improving the resolution of consensus trees. In a previous study we have shown that the Transfer Bootstrap Expectation (TBE) is robust to the presence of rogue taxa such that rogues do not need to be removed prior to computing supports. In this study, we propose a new, fast and efficient measure of rogueness based on the transfer index as well. We compare our index with those existing in the literature on simulated and empirical datasets.

Multiple independent introductions in Bayesian phylodynamics

Ariane Weber* 1,2 , Sanni Översti 1,2 , and Denise Kühnert 1,2,3

1 Max Planck Institute of Geoanthropology, Jena – Germany
2 Max Planck Institute for Evolutionary Anthropology, Leipzig – Germany
3 Robert Koch Institute, Center for Artificial Intelligence in Public Health Research – Germany

Bayesian phylodynamics aims to estimate population dynamics using phylogenetic trees that trace ancestral relationships. While usually a single tree is inferred with its defining parameters, multiple trees can be estimated jointly to increase statistical power, if genetically distinct entities are characterised by the same or overlapping set of parameters. One example in which this is especially relevant is in infectious disease research, when a pathogen is introduced multiple times independently into the studied population. However, it is poorly understood how the phylodynamic methods perform with increasing number and decreasing size of entities. We aim to address this problem in a simulation-based evaluation of two different approaches to quantify the dynamic process jointly from multiple independent introductions in a phylodynamic birth-death sampling setup. The first approach considers the introductions as independent panmictic processes and calculates the joint probability as a product of the entity-specific probabilities. The second additionally models the introduction process into the population. Both methods robustly infer the characterising parameters, with the former being more sensitive to the size distribution of the introductions. Using these results in the analysis of genomic SARS-CoV-2 datasets, we study the transmission dynamics of COVID-19 in Germany, focusing on the inference of variant specific transmission advantages. We thus conclude that the amount and size distribution of jointly analysed entities is of considerable importance.

Algerian honey bees: a case study of the impact of breeding practices on genetic structure

Pierre Faux* 1 , Giovanna Salvatore 2 , Amira Chibani Bahi Amar 3 , Riad Fridi 3 , Nacera Tabet Aoul 3 , Mariasilvia D'andrea 4 , Bertrand Servin 1 , Kamila Canale-Tabet 1 , and Alain Vignal 1

1 Génétique Physiologie et Systèmes dÉlevage – Ecole Nationale Vétérinaire de Toulouse, École nationale supérieure agronomique de Toulouse [ENSAT], Institut National de Recherche pour l'Agriculture, l'Alimentation et l'Environnement : UMR1388, Institut National de Recherche pour l'Agriculture, l'Alimentation et l'Environnement – France
2 Department of Agricultural, Environmental and Food Sciences, University of Molise – Italy
3 Université dÓran 1 Ahmed Ben Bella [Oran] – Algeria
4 Department of Agricultural, Environmental and Food Sciences, University of Molise – Italy

Honey bees (A. mellifera sp.) stand apart from other domesticated species on various biological singularities. Among these, their genetic diversity results from the combined effects of the trade of queens between beekeepers, sometimes over long distances, and their mating with local males, possibly from feral or wild pools. The relative strength of these two effects yields variable patterns of biogeographical distribution. Here, we aimed at assessing how breeding practices would be reflected in the genetic structure of a honey bee population. To that end, we used a sampling of 102 (haploid) males from Algeria, of two honey bee subspecies (A. mellifera intermissa and A. mellifera sahariensis). We first showed that genetic distances correlate well to geographical distances between pairs of individuals. Then, using standard approaches to infer genetic structure, we showed that the whole population was stratified in two main clusters, defined along a geographic cline (Eastern and Western). Each cluster includes samples from both subspecies in comparable proportions. The Western cluster was found to drive the correlation between geographical and genetic distances, while the Eastern cluster included most of the high identical-by-descent relationships between samples. Using ancestral recombination graphs, we further substantiate how different breeding practices shaped the genetic structure and diversity within each cluster. Our study illustrates the potential of advanced statistical modelling for deciphering mating patterns from population genomics data in free ranging species. The outcomes of our study should be considered in local honey bee biodiversity improvement and conservation initiatives.

A primal-dual algorithm for maximum parsimony on networks

Martin Frohn* 1 and Steven Kelk 2

1 Eindhoven University of Technology – Netherlands
2 Maastricht University – Netherlands

Finding the most parsimonious tree inside a phylogenetic network is a NP-hard combinatorial optimization problem. If the network is a rooted tree, then Fitch's well known algorithm calculates an optimal parsimony score in polynomial time. There are combinatorial algorithms in the literature which extend Fitch's algorithm to networks without bounds on the solution quality. We introduce a new extension of Fitch's algorithm which does ensure approximation guarantees on some classes of phylogenetic networks. Specifically, we show (i) that Fitch's algorithm can be recast as a primal-dual algorithm for an half-integral linear program, (ii) how it can be extended to reticulation networks, (iii) how to improve approximation guarantees on binary characters for tree-child networks. These results for classic problems strengthen the traditional link between polyhedral methods and phylogenetics and can aid in the study of other optimization problems on phylogenetic networks.

The patterns of codon usage between chordates and arthropods are different but coevolving with mutational biases.

Ioanna Kotari* 1,2 , Carolin Kosiol 3 , and Rui Borges 1

1 Institut für Populationsgenetik [Vienna] – Austria
2 Vienna Graduate School of Population Genetics, Vienna – Austria
3 Centre for Biological Diversity, University of St Andrews [St Andrews, Fife] – United Kingdom

Different frequencies amongst codons that encode the same amino acid (i.e. synonymous codons) have been observed in multiple species, and studies have focused on uncovering the dynamics that drive such codon usage bias. Two main narratives have been proposed: on one hand, mutations, and on the other, translational selection, are working to produce different frequencies of synonymous codons, with a combined effect being the most likely explanation. However, not many studies have been able to measure and disentangle these effects that leave similar traces on the genome in multiple species. Here, we have developed a codon model that allows the disentangling of mutation, selection on amino acids and synonymous codons, and GC-biased gene conversion (gBGC). It employs a Bayesian estimator to assess the inter-species variability of codon composition found in multiple species in light of those forces. In this study, we analysed a dataset of 415 species belonging to chordates and 191 species of arthropods. We observed that chordates need more synonymous codon categories to explain the empirical codon frequencies compared to arthropods (52 versus 37). This suggests that chordates have more pronounced patterns of codon usage. Additionally, in both phyla, selection forces act more negatively as the GC content of a codon is increasing, indicating that selection at the genome-wide level acts to balance when the mutational processes promote G/C alleles. Methylation at CpG sites seems to also explain patterns of codon usage bias in chordates but not in arthropods. This study shows that while both chordates' and arthropods' codon patterns seem to be dominated by mutational biases, they differ in their extent of codon usage.

A new ancestral range reconstruction model to investigate the role of barriers in shaping global patterns of biodiversity

Thomas Merrien* 1,2

1 University of Helsinki – Finland
2 Finnish Museum of Natural History – Finland

How did Biodiversity arise? What are the mechanisms behind the patterns of diversity we can see? Using comparative phylogenetics, my aim is to understand what the role of barriers is in shaping current patterns of biodiversity. To do so, I will develop a new model of explicit ancestral range reconstruction inspired by the previous Landscape Based Geographical (LBG) model (Bouckaert et al. 2012). I will assume that range evolution is a diffusion process evolving along a phylogeny and across the geographical landscape. The diffusion will be modeled as a continuous time Markov chain (CTMC). The simplest possible model assumes that the dispersal rate is constant over time and geographical landscape. Other flavors of the model can relax the constant-rate assumption by allowing rates to vary across geographical landscape and time. This flexibility enables testing different hypotheses for how spatial barriers and niche dependence affect range evolution over spatio-temporal scales. I will then use simulation and data about dung beetles (Scarabaeinae) and lemurs (Lemuroidea) in Madagascar to test these hypotheses. The model will be available as an R package and can later incorporate new features such as the inclusion of joint reconstruction of biogeography and diversification rates, or the inclusion of competition effects.

Using supervised learning to relate DNA sequence composition from eDNA to ecosystem properties

Letizia Lamperti* 2,3 , Sanchez Theophile 1,2, Stéphanie Manel 3, and Loïc Pellissier 1,2

1 Swiss Federal Institute for Forest, Snow and Landscape Research WSL – Switzerland
2 Department of Environmental Systems Science [ETH Zürich] – Switzerland
3 Centre d'Ecologie Fonctionnelle et Evolutive – École Pratique des Hautes Etudes – France

The current biodiversity crisis calls for new approaches to monitoring the impact of human activities on the biosphere. The rapid development of 'omics' tools, in particular the metabarcoding of environmental DNA (eDNA), has opened a new field for the generation of comprehensive biodiversity data in many regions of the world. However, the development of efficient data processing pipelines has not kept pace with the exponential growth of omics data, limiting the application of eDNA for ecological monitoring. Currently, measures of ecological integrity of ecosystems are manually calculated by human experts based on a wide range of cues, from overall habitat integrity index to taxonomic identifications. Previous studies, including Cordier et al. (2018) and Brantschen et al. (2021), have shown that eDNA combined with machine learning can provide reliable classifications for such ecological assessments of ecosystems. However, these previous attempts either required a reference database or used a MOTUs approach to train the model to score ecosystem properties. By having access to all the information in the raw DNA sequence data, a supervised neural network can directly and better predict complex and integrative ecosystem properties representing ecological integrity or other biotic responses. Indeed, in this project we propose to use a combination of machine learning approaches to transform eDNA metabarcoding data into informative ecological indicators to improve ecosystem monitoring and decision making. We will test whether supervised learning can predict ecosystem properties of interest directly from the DNA sequence composition in eDNA samples. We will train the neural network to relate the DNA sequence composition of eDNA to the quantitative characteristics of the ecosystem. We expect that the neural network will be able to learn to rank ecosystem properties directly from the sequence of eDNA samples. We will evaluate the model by successively calibrating the model with omitted data and comparing the predicted observations in cross-validation. While this is a first set of evaluations of the methodology, many ecosystem quality assessments are done in this way and there are many potential applications of predicting ecosystem properties from DNA sequence data from the training of such an algorithm.

DNAi: processing of environmental DNA using artificial intelligence for ecosystem monitoring

Théophile Sanchez* 1,2 , Letizia Lamperti 2,3 , Steven Stalder 4 , Benjamin Flück 1,2 , Michele Volpi 4 , Stéphanie Manel 3 , Camille Albouy 1,2 , and Loïc Pellissier 1,2

1 Swiss Federal Institute for Forest, Snow and Landscape Research WSL – Switzerland
2 Department of Environmental Systems Science [ETH Zürich] – Switzerland
3 Centre d'Ecologie Fonctionnelle et Evolutive – Ecole Pratique des Hautes Etudes – France
4 Swiss Data Science Center – Switzerland

The use of environmental DNA (eDNA) metabarcoding has recently emerged as a powerful tool for monitoring biodiversity on a large scale. However, the current methods for analyzing eDNA data often face limitations due to incomplete reference databases, and require extensive denoising or/and clustering of raw sequencing data. To address this issue, the DNAi project proposes a new family of deep learning tools that harness the capability of Artificial Neural Networks (ANNs) to process multidimensional and heterogeneous data, enabling the transformation of raw sequences from aquatic ecosystems into meaningful information for monitoring. Firstly, we designed a deep learning scheme that combines species co-occurrences and phylogeny to enhance the taxonomic classification of eDNA sequences. Secondly, we trained an artificial neural network in an unsupervised fashion to create a latent space that ordinates samples, and compared this ordination to ecosystem dynamics. Lastly, we trained a supervised ANN to directly link eDNA samples to ecosystem properties. Overall, the proposed deep learning tools have the potential to extract more information from raw sequences, which could greatly improve our understanding of biodiversity and support more effective ecosystem management efforts.

Surprising Species: information theory, phylogenetics, and food webs

Giulio Valentino Dalla Riva* 1

1 School of Mathematics and Statistics, University of Canterbury – New Zealand

Detecting and analysing the role of macro-evolutionary processes in the structure of ecological networks is challenging, particularly for unipartite food webs. Indeed, current methods are either model-based (and we lack a good model for unipartite food web evolution), or rely on the testing for linear correlation between patristic distances and ecological dissimilarity. Moreover, current methods provide limited information on how different clades, or even single species, fit those models. Here, we propose an alternative approach, building on Information Theory and providing a meaningful interpretation in terms of Minimum Evolution, to test for phylogenetic signal in unipartite food webs. Moreover, we show how that can be used to identify surprising species: species that have an unexpected peculiar ecological role, given their evolutionary history. We apply our methodology to two large food webs, and discuss the results. This is joint work with Daniele Catanzaro.

Ecological class prediction: A new method of discriminant analysis, phylogeny-aware and applicable in large dimension.

Anaïs Duhamel* 1 , Julien Clavel , and Gilles Escarguel

1 Laboratoire de Géologie de Lyon - Terre, Planètes, Environnement [Lyon] – École Normale Supérieure - Lyon, Université Claude Bernard Lyon 1, Centre National de la Recherche Scientifique – France

Understanding the link between ecological and morphological features in extant species is a key issue, notably in paleontology since it allows the inference of extinct species ecologies from their morphological features. Such predictions are classically done using linear discriminant analysis (LDA) : this method fits a model on a training set containing individuals for which both categorical traits (e.g. ecological classes) and continuous traits (e.g. morphology) are known, to be then able to predict the class of an individual for which only continuous traits are known. Morphology is not the only signal which can help to infer past ecologies: the phylogenetic position of an individual can also be highly informative since closely related species often share the same ecology. Moreover, with the rise of 2D and 3D geometric morphometrics, datasets with more traits than species (high-dimensional datasets) are now commonplace, but classical discriminant analysis methods significantly lose statistical power when the number of morphological traits (p) approaches the number of species (n), and are not even computable when p is higher than n. Here we develop a new discriminant analysis which is both phylogeny-informed and applicable to high-dimensional datasets through penalized likelihood techniques. The performances of this newly implemented method were assessed on simulated and empirical datasets. It appears that this new method outperforms, in many situations, conventional discriminant approaches when applied to comparative datasets (e.g., phylogenetically related species).

PoMoSelect and PoMoBalance via RevBayes: Disentangling Modes of Natural Selection

Svitlana Braichenko* 1 , Valeria Montano 1 , Rui Borges 2 , and Carolin Kosiol 1

1 Centre for Biological Diversity, University of St Andrews – United Kingdom
2 Institut für Populationsgenetik, Vetmeduni Vienna – Austria

The interplay between mutation, genetic drift, directional, and balancing selection in shaping populations' diversity is highly convoluted and difficult to disentangle. This requires sophisticated phylogenetic models that have a high degree of flexibility and can handle multi-individual data. For these purposes, our group has developed a polymorphism-aware phylogenetic set of models called PoMos. These models are based on the Moran model and have recently been proven effective in inferring species trees as well as mutational effects, fixation biases and GC-bias rates in great apes and grasshoppers. To make these models more accessible, we implemented them in the open-source Bayesian inference framework RevBayes. The advantage of the framework is its implementation in a graphical model environment and the possibility to compute coverage frequencies for the validation analysis. In this study, we further developed PoMos to study neutral, directional and, for the first time, balancing selection. The key advantage of our novel approach for studying the balancing selection is that PoMos allow for ancestral polymorphisms that can be maintained, and parameters that can measure frequency-dependent selection. We have tested our new method on a set of simulated data with a popular evolutionary framework Slim and a custom Moran model simulator implemented in RevBayes. Furthermore, we investigated real sequences of African human populations to understand the evolutionary history of genomic regions that are known to be under balancing selection driven by malarial parasites.

Treating Gaps as Insertion and Deletions under Maximum Parsimony

Clara Iglhaut* 1,2 , Jūlija Pečerska 3 , Manuel Gil 3 , and Maria Anisimova 3,4

1 Zuricher University of Applied Sciences (ZHAW) – Switzerland
2 University of Zurich – Switzerland
3 Zurich University of Applied Sciences (ZHAW) – Switzerland
4 Swiss Institute of Bioinformatics [Genève] – Switzerland

With the vast amount of genomic sequence data available today, there is a growing need for phylogenetic inference methods that can accurately reconstruct the evolutionary history of datasets with several thousand sequences. While phylogenetic inference tools are continuously improved to be able to process larger and larger datasets, most commonly used methods base their inferences on substitution events only. Insertion and deletion events (indels) are typically not treated as distinct evolutionary events, or the resulting gaps are treated as missing data. We propose ancestral sequence reconstruction and multiple sequence alignment methods based on maximum parsimony that explicitly include indels as distinct events in their inference. Our approach is computationally efficient and has been compared to state-of-the-art inference tools on simulated data with indels. Our results show that indel-aware parsimony is particularly suitable for densely sampled datasets with closely to moderately related sequences, where it achieves comparable MSA quality to probabilistic methods and accurately infers the evolutionary history of the sequences, including indel patterns. To demonstrate the applicability of our methods, we used them to systematically study the indel variation in the HIV-1 env gene that encodes the two glycoproteins, gp120 and gp41. In particular, gp120 is the primary target for neutralizing antibodies generated as a response to HIV infection or vaccination, and its rapid mutation rate contributes to the virus's ability to evade host immune response. Therefore a better understanding of the genetic variability of the region could aid effective vaccine development.

A Phylogenetic Framework for Interspecies Gene Differential Expression Analysis

Paul Bastide* 1 and Mélina Gallopin 2

1 IMAG – Université de Montpellier, CNRS – France
2 Institut de Biologie Intégrative de la Cellule – Commissariat à l'énergie atomique et aux énergies alternatives, Université Paris-Saclay, Centre National de la Recherche Scientifique – France

Comparing the expression levels of othologous genes across closely related species can help us to better understand the mechanisms underlying the evolution of gene expression. RNA-Seq techniques produce highly dispersed count data, and triggered the development of specific statistical methods for differential expression analysis in a single species context (1). For interspercies datasets, the phylogenetic relationships between organisms introduce some correlations between samples, and is usually taken into account through Phylogenetic Comparative Methods (2). Techniques from both literatures have been used for interspecies differential expression analyses, and each have their own strengths and weaknesses, that we illustrated in a benchmark study (3), thanks to a new and realistic simulation tool implemented in compcodeR (4). Combining some of the strengths of both approaches in a new statistical method is the focus of our current work.

(1) Dillies, Rau, Aubert, et al. 2013. Briefings in Bioinformatics. 14(6):671–683.
(2) Harmon 2019. Phylogenetic Comparative Methods: Learning From Trees. Center for Open Science.
(3) Bastide, Soneson, Stern, Lespinet, Gallopin. 2023. A Phylogenetic Framework to Simulate Synthetic Interspecies RNA-Seq Data. Molecular Biology and Evolution. 40(1):msac269.
(4) Soneson. 2014. Bioinformatics. 30(17):2517–2518

Evolution of the plant organelles is characterized by similar substitution rate but different allele dynamics

Marina Khachaturyan* 1,2 , Mario Santer 2 , Thorsten B.h. Reusch 1 , and Tal Dagan 2

1 Marine Evolutionary Ecology, GEOMAR Helmholtz Centre for Ocean Research Kiel, Kiel – Germany
2 Institute of General Microbiology, University of Kiel, Kiel – Germany

Plant cells harbor membrane-bound organelles containing their own genetic material – plastids and mitochondria. As organellar chromosomes are typically occurring in high copy number the mutational supply to organellar genomes is dramatically increased. Yet, the plant organelles were shown to evolve slower than the nucleus. The elevated probability for de novo mutations may be buffered by several factors including segregational drift, preferential uniparental inheritance, or recombination; all of these are expected to shorten significantly the time to loss or fixation of alternative neutral alleles. However, the effect of individual factors on the fate of novel alleles and therefore also the relative and absolute evolutionary rate of the two plant organellar genomes is unresolved. Here we show that mitochondria and plastids have the same substitution rate but different allele dynamics, with plastids remaining in a heteroplasmic state ca. 20 times longer than mitochondria. We analyzed fixed and non-fixed organellar mutations across the marine flowering plant eelgrass (*Zostera marina*) populations distributed worldwide. Using a stochastic agent-based simulation, we demonstrate that copy number is insufficient to explain the elongated allele segregation time in plastids in comparison to mitochondria. We hence invoke an as yet unknown active partitioning of plastids after cell division. Our results expand our understanding of modes of plant organellar DNA inheritance and segregation with potential application to other extrachromosomal genetic elements.

A recombination-aware multitype birth-death process for reconstructing the phylogeographic history of recombining pathogens

Ruth Boersma*1,2 , Arthur Kocher 1,2 , Johannes Krause 1 , and Denise Kühnert 1,2,3

1 Department of Archaeogenetics, Max Planck Institute for Evolutionary Anthropology [Leipzig] – Germany
2 Transmission, Infection, Diversification Evolution group (tide), Max Planck Institute of Geoanthropology [Jena] – Germany
3 Robert Koch Institute [Wildau] – Germany

Genetic recombination is known to play a crucial role in the evolution of many pathogens, making it necessary to account for recombination in phylogenetic reconstruction. However, reconstructing a complete ancestral recombination graph can be difficult due to the computational complexity involved. Since genetic recombination occurs between closely located lineages during co-infection, explicitly considering geographic proximity may aid in reconstructing ancestral recombination graphs by limiting the possible recombination events to those between co-located lineages. Additionally, recombination events can provide valuable geographic information. This makes the joint analysis of both recombination and geographic spread beneficial in uncovering the phylogeographic history of viruses that undergo recombination, such as the Hepatitis B Virus. Furthermore, not accounting for recombination might bias phylodynamic analysis by systematically over- or underestimating the underlying parameters. To address these challenges, I developed a multitype birth-death process with recombination. The model considers subpopulations associated with distinct geographical localities, within which genetic recombination may occur and allows individuals to migrate between these sub-populations. I implement the model in BEAST2, based on the BDMM package (Kühnert et al. 2016). In a simulation study I examine the performance of this model in estimating the correct recombination graph and associated parameters.

Population genetics processes impact Species Trees estimation

Peter Beerli* 1

1 Florida State University – United States

We investigate the effects of DNA mutation models, population growth and gene flow on the power to correctly infer a species from genomic data. These inferences are complicated by the commonly used representation of the sequence data as single nucleotide polymorphisms instead of the full DNA sequences. This data representation choice reduces the potential variability in the data and biases parameters such as the effective population size of a species downwards. Methods that may work for species with small population sizes, such as primates, will most likely fail with species that have large populations sizes such as mosquitoes.

Comparison of Stacks and a custom pipeline for RADseq analysis

Enora Geslain* 1 , Alvaro Cortes Calabuig , Sarah Maes , Gregory Maes , and Filip A M Volckaert

1 Laboratory of Biodiversity and Evolutionary Genomics, KU Leuven – Belgium

RADSeq is a sequencing technique to scan complete genomes of organisms without sequencing them entirely. One of the uses is the genotyping with Single Nucleotide Polymorphisms (SNPs) for applications in population genetics. Different tools exist to search for SNPs from RADseq data. Here, we compare two pipelines: a custom pipeline and the Stacks pipeline developed by Rochette and Catchen. Our conclusion was that it is important to choose the right tools for RADseq analysis because of the large impact they can have on the downstream analysis. In our case, the best tool seems to be Bowtie2 for the mapping. BWA was developed for human genome analysis, which may explain its reduced efficiency for other organisms. However, further filtering might be required to remove the lane effect detected with Bowtie2.

Comparative genomics of lactation

Christophe Lefèvre* 1

1 University of Newcastle – Australia

A complex lactation system with elaborated milk secretion was already in place 200 Million years ago in the ancestors of mammals. During their subsequent radiation, mammals have diversified lactation strategies to accommodate reproductive success and adapt to the environment. There is much to learn about the genetics of lactation from the rich natural resource of animal diversity. Evolutionary analysis can uncover how differential constraints have shaped the lactation system of specific lineages. Comparative genomics makes it possible to study how these evolutionary constraints vary between lineages depending on lactation strategies or environmental and behavioural adaptations, while, at the same time, providing insights on the role of lactation in the programming of mammalian development. We illustrate this approach using transcriptomics and genome analysis on some examples and discuss the deployment of bioinformatics tools for data integration and analysis. These studies provide a broad picture of the evolutionary landscape of lactation, revealing the equivalent importance of conserved metabolic and secretory pathways and either the modular reorganization of existing milk components or the appearance of specific milk proteins. This mix of robustness and flexibility has allowed the adoption of a diversity of lactation strategies under various physiologic and environmental conditions. Thus, both the ancient and highly conserved and the more variable and specific molecular components of lactation are, in part, responsible for the success of the mammals in surviving, adapting, and evolving

Balanced Minimum Evolution as a Minimum Cross-Entropy Estimator

Daniele Catanzaro* 1

1 Center of Operations Research and Econometrics (CORE) Université Catholique de Louvain – Belgium

We show here that Balanced Minimum Evolution (BME) is a minimum cross-entropy estimator and we discuss the biological interpretation of this result. The new perspective extends the previous interpretations of the BME length function described in the literature and may suggest new connections with the maximum log-likelihood estimator.

SPREAD 4: online visualization of pathogen phylogeographic reconstructions

Kanika Nahata* 1 , Filip Bielejec 2 , Juan Monetta 3 , Simon Dellicour 4 , Andrew Rambaut 5 , Marc Suchard 6 , Guy Baele 1 , and Philippe Lemey 1

1 Rega Institute – Belgium
2 Nonce Filip Bielejec, 90-245 Lodz, Poland – Poland
3 Guayabos 1924, Montevideo – Uruguay
4 Spatial Epidemiology Lab (SpELL), Université Libre de Bruxelles – Belgium
5 Institute of Ecology and Evolution, University of Edinburgh – United Kingdom
6 Department of Human Genetics, David Geffen School of Medicine, Department of Biostatistics, Jonathan and Karin Fielding School of Public Health and Department of Biomathematics, David Geffen School of Medicine, University of California – United States

Phylogeographic analyses aim to extract information about pathogen spread from genomic data, and visualizing spatio- temporal reconstructions is a key aspect of this process. We here present SPREAD 4, a feature-rich web-based application that visualizes estimates of pathogen dispersal resulting from Bayesian phylogeographic inference using BEAST on a geographic map, offering zoom-and-filter functionality and smooth animation over time. SPREAD 4 takes as input phylogenies with both discrete and continuous location annotation and offers customized visualization as well as generation of publication-ready figures. SPREAD 4 now features account-based storage and easy sharing of visualisations by means of unique web addresses. SPREAD 4 is intuitive to use and is available online at https://spreadviz.org.