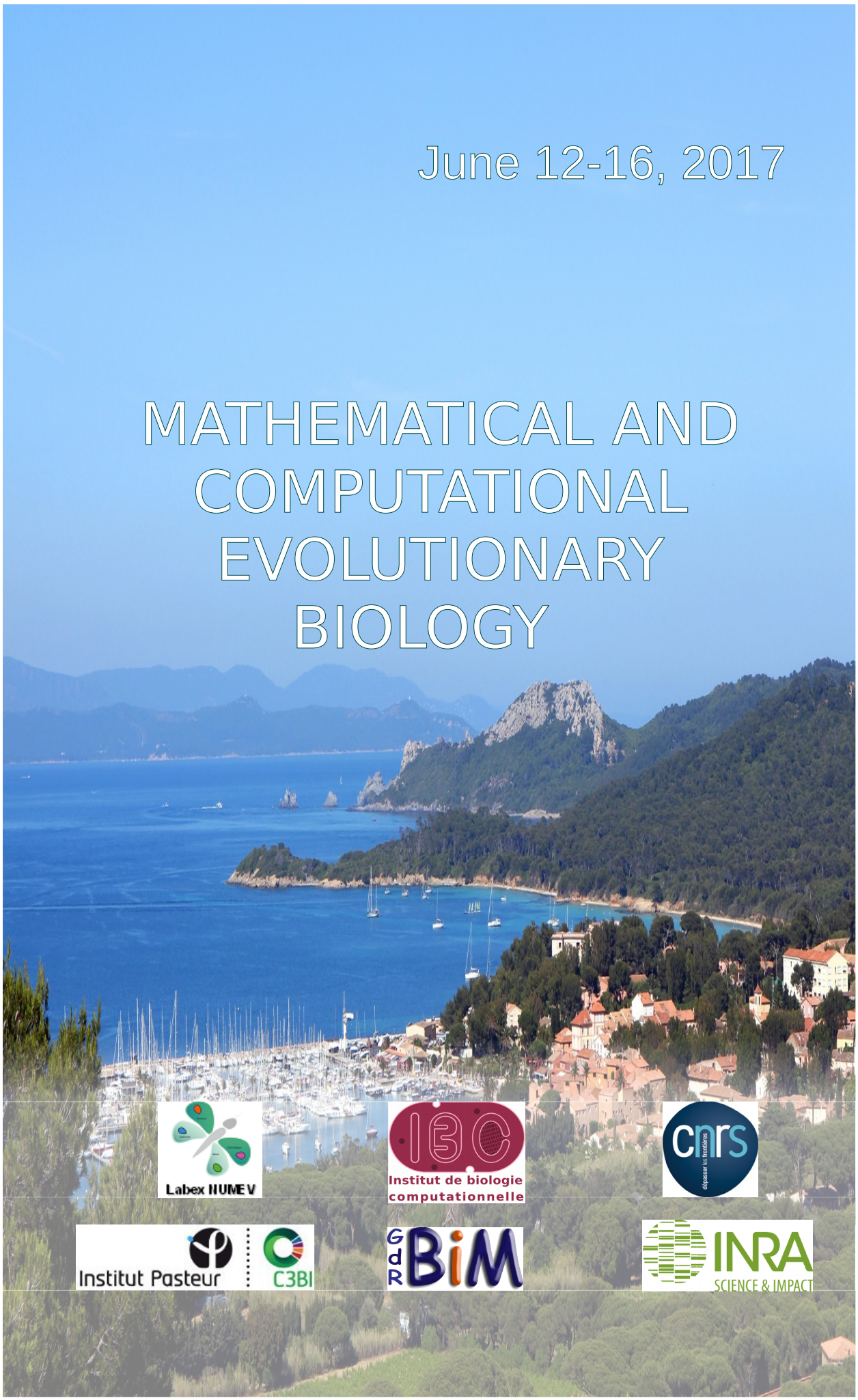
June 12-16, 2017

MATHEMATICAL AND COMPUTATIONAL EVOLUTIONARY BIOLOGY



INFORMATION

Meeting Point

Directly at IGESA Center for the drinks at « le théâtre de verdure » (next to reception) on Monday 7pm.
However, you can arrive at IGESA and get the key from 4pm

Aim to get the 6.30pm ferry (or earlier) at La Tour Fondue.

In case of problems :

Olivier Gascuel : +33 (0)6 48 12 14 82

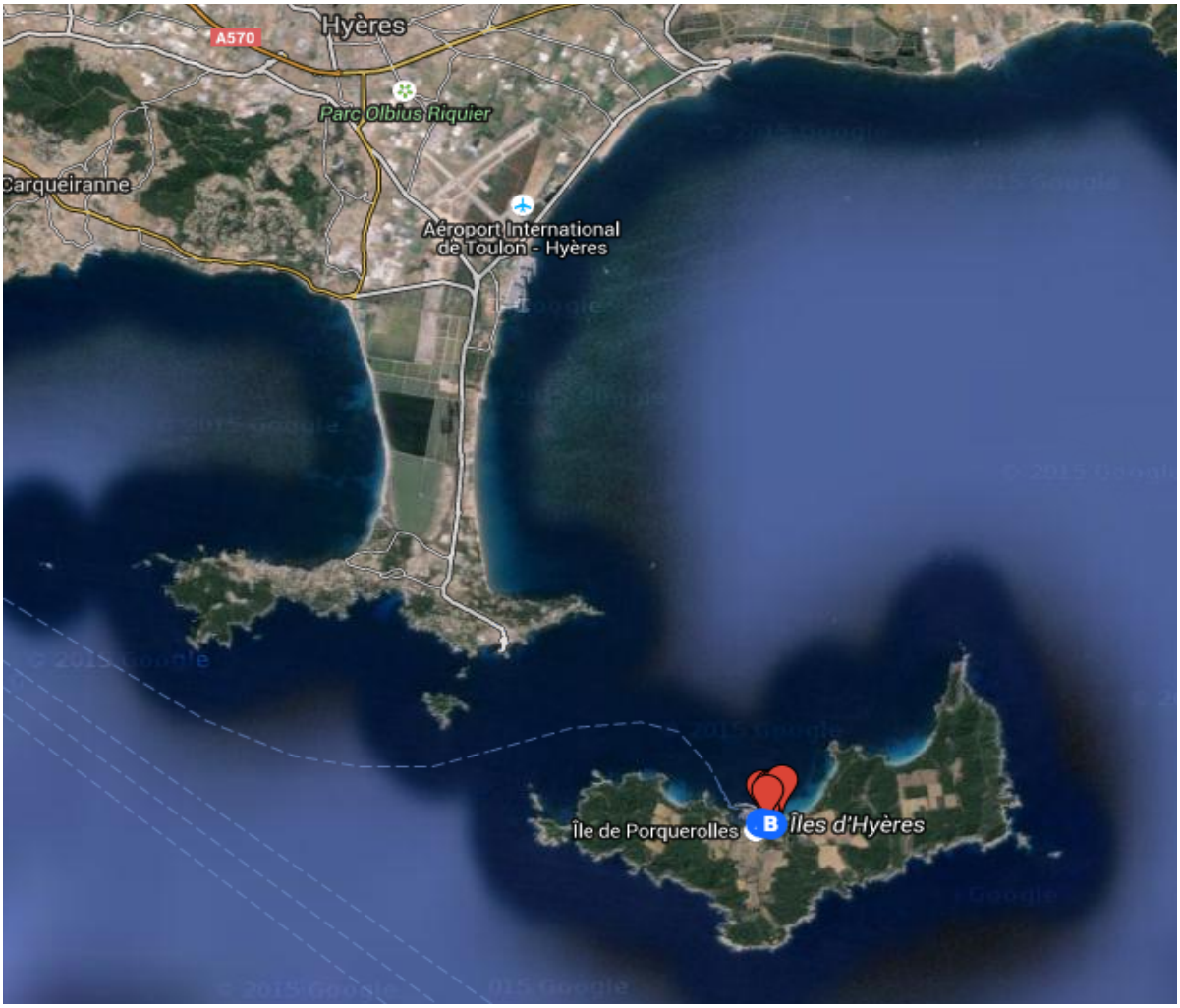
Fabio Pardi : +33 (0)6 83 23 20 14

Krister Swenson : +33 (0)7 81 71 60 90

=> inside France use 0 to start (06 48...), outside france use +33 without the 0 (336 48...)

Location

The conference will be held at the ***IGESA*** centre, in the village of Porquerolles, five minutes from the harbour and the beach. It is located in a nature reserve, the Port-Cros National Park, one of the most intact coastal areas in the Mediterranean.



Practical information

Accommodations :

Hôtel Club IGESA
Rue de la Douane
Ile de Porquerolles
83 400 HYERES
Tel : 04 94 12 31 80

Coordonnées GPS : 43.000402,6.205076

WI-FI : available but very limited speed

Ferry :

TLV-TVM : 04 94 58 21 81
Return ticket: 19,50€
La Tour Fondue -> Porquerolles :
7:30, 9:00, 9:30, 10:00, 10:30, 11:00, 11:30, 12:00, 12:30, 13:30, 14:30, 15:30, 16:30, 17:30, 18:30
Porquerolles -> La Tour Fondue :
7:00, 8:30, 9:30, 10:00, 10:30, 11:00, 11:30, 14:00, 15:00, 16:00, 17:00, 18:00, 19:00

After hours shuttle :
One way ticket : 18,50€ (buy your ticket directly on board)
La Tour Fondue -> Porquerolles : 19 :45, 23 :15
Porquerolles -> La Tour Fondue : 20:00, 23:30

Taxi-Boat:

Taxi Boat « Le Pelican » : 06 09 52 31 19
The taxi-boat costs 16.50€ per person, when at least 6 people are booked on it.

Car Park on Tour Fondue :

Cars are not allowed on the island, so if you come by car you'll have to leave it in a car park at La Tour Fondue. You can book a place.
Car park Coulomb (watchman 24/24) : 04 94 58 14 67 (24h : 14€)
Car park des Iles (videosurveillance) : 04 94 58 90 78 (24h : 15€)
Car park Vinci (watchman 24/24) : 04 94 01 99 28 (24h : 16€)

Bus Hyères-La Tour Fondue :

If you come by train or plane to Hyeres, you can take the bus line 67 on « Reseau Mistral »
More information is available with your mobile : http://m.reseaumistral.com/
Or call a taxi : Taxi Hyères 04 94 00 60 00

PROGRAM

Monday, June 12

> 19:30 : Welcoming Drink

> 20h00 : Dinner

Tuesday, June 13

> 09h00 : Welcome

> 09h15 – 10h30 : *KEYNOTE* *p.8*

Marc SUCHARD (David Geffen School of Medicine at UCLA, Departments of Biomathematics, Biostatistics and Human Geneticq)
« Phylogenetic factor analysis»

> 10h30 : Coffee break

> 11h00 – 12h20 : 4x 20 min *TALKS* (including questions)

Peter BEERLI (Florida State University, USA) *p.9*
« Population divergence estimation using individual lineage label switching »

Damien DEVIENNE (Laboratoire de Biométrie et Biologie Evolutive, UMR5558, FR) *p.10*
« Lifemap, Exploring the Entire Tree of Life. And after? »

Liangliang WANG (Simon Fraser University, Burnaby, CA) *p.15*
« An adaptive sequential Monte Carlo sampler for Bayesian phylogenetic non-clock tree inference »

Alex GAVRYUSHKIN (Department of Biosystems Science and Engineering, ETH Zurich, CH) *p.11*
« Inferring genetic interactions from competition experiments »

> 12h20 : Lunch

> 14h00 – 15h00 : 3x20 min *TALKS* (including questions)

Anna ZHUKOVA (C3BI, USR 3756 Institut Pasteur et CNRS, FR) *p.15*
« A phylodynamic model of pathogen drug resistance emergence and transmission »

Florian CLEMENTE (Institut de Biologie Computationnelle, FR) *p.9*
« Estimation of the sex-ratio in population trees from allele frequency data »

Nicola DE MIAO (University of Oxford, UK) *p.10*
« The bacterial sequential Markov coalescent »

> 15h00 : Break

> 15h15 – 16h30 : *KEYNOTE* *p.8*

Anna-Sapfo MALASPINAS (Institute of Ecology and Evolution, University of Bern, CH)
« A genomic history of Aboriginal Australia »

> 16h30 - 18h30 : Freetime, beach...

> 18h30-20h00 : *POSTERS* *p.16-23*
Wine and discussion (posters 1-18)

> 20h00 : Dinner

Wednesday, June 14

> 09h15 – 10h30 : *KEYNOTE* *p.7*

Elizabeth ALLMAN (Department of Mathematics and Statistics, University of Alaska Fairbanks, USA)
« Split scores for phylogenetic trees and applications »

> 10h30 : Coffee break

> 11h00 – 12h20 : 4X20 min *TALKS* *(including questions)* *p.9*

Erik VOLZ (Imperial College London, UK) *p.15*
« Scalable relaxed clock dating »

Carina MUGAL (Uppsala University, Uppsala, SW) *p.13*
« Time-dependent estimates of dN/dS and its implications for molecular ecology and genetic studies »

Robin THOMPSON (University of Oxford, UK) *p.14*
« Accounting for donor viral diversity gives high estimates of the number of HIV founder virions among recipients »

Jakub TRUSZKOWSKI (EMBL-EBI and Cancer Research UK, University of Cambridge, UK) *p.14*
« Reconstructing phylogenies from single-cell sequencing data »

> 12h20 : Lunch

> 14h00 - 20h00: Free afternoon (beach, hiking, theorems, etc.)

> 20h00 : Dinner

Thursday, June 15

> 09h15 – 10h30 : *KEYNOTE* *p.8*

Yun SONG (Computer Science Division and Department of Statistics, UC Berkeley; Departments of Mathematics and Biology, University of Pennsylvania, USA)
« The major determinants of translation speed and their evolutionary signatures »

> 10h30 : Coffee break

> 11h00 – 12h00 : 3x 20 min *TALKS* (including questions)

Caroline COLIJN (Imperial College London, UK) *p.9*
« Connection genetic and clinical information to understand pathogen transmission »

Fabian FREUND (University of Hohenheim, Stuttgart, DE) *p.11*
« A new codon substitution model to better estimate evolutionary processes »

Itay MAYROSE (Tel Aviv University, IL) *p.12*
« A Likelihood Method for Detecting Trait-Dependent Shifts in the Rate of Molecular Evolution »

> 12h00 : Lunch

> 14h00 – 15h00 : 3x20 min *TALKS* (including questions)

Rohan METHA (Stanford University, CA, USA) *p.12*
« The probability of monophyly of a sample of gene lineages on a species tree »

Willy RODRIGUEZ (INRA, Université Paris-Saclay, FR) *p.13*
« The Inverse Instantaneous Coalescence Rate (IICR) as a new summary statistic in population genetics »

Alexandra GAVRYUSHKINA (ETH Zürich, CH) *p.12*
« Fossilized birth-death process under different modes of speciation and fossil stratigraphic ranges »

> 15h00 - 15h15 : Break

> 15h15 – 16h30 : *KEYNOTE* *p.7*

Stéphane DRAY (Laboratoire de Biométrie et Biologie Evolutive, Lyon, FR)
« Environment, traits, space and phylogeny: integrating constraints in the analysis of ecological data tables »

> 16h30 - 18h30 : Freetime, beach...

> 18h30 - 20h00 : *POSTERS* *p.23-31*
Wine and discussion (posters 19-37)

> 20h00 : Dinner

Friday, June 16

> 09h15 – 10h30 : *KEYNOTE* *p.7*

Guy BAELE (Rega Institute / KU Leuven - Evolutionary and Computational Virology Section, BE)
« Computational approaches for analysing partitioned data in pathogen phylodynamics »

> 10h30 : Coffee break

> 11h00 – 12h15 : *KEYNOTE* *p.7*

Barbara HOLLAND (Theoretical Phylogenetics Group, School of Mathematics and Physics, University of Tasmania, AU)
« Desirable properties of models for phylogenetic analysis »

> 12h15 : Lunch and then farewell session

> 14h00 : First ferry to « La Tour Fondue »

KEYNOTE SPEAKERS

> ***Elizabeth ALLMAN***
Department of Mathematics and Statistics, University of Alaska Fairbanks, USA

Split scores for phylogenetic trees and applications

From DNA sequence data collected from n taxa, we construct a `splitscore' under the assumption that the aligned sequences have evolvedunder the general Markov model (GM) on an evolutionary tree. Thissplit score, based on theoretical properties for the GM model ontrees, can be computed efficiently from genomic scale data using thesingular value decomposition. In this talk, we describe this splitscore and illustrate how it might be used to detect true splits in theevolutionary tree relating taxa, and shifts in the evolutionaryprocess along a chromosome.

> **Guy BAELE**
Rega Institute / KU Leuven - Evolutionary and Computational Virology Section, BE

Computational approaches for analysing partitioned data in pathogen phylodynamics

Integrating various sources of information offers a promising avenue for research into pathogen phylodynamics, deliver more precise insights and increasing opportunities for statistical hypothesis testing. The incorporation of covariates of evolutionary and epidemic processes in the reconstruction process is often computationally demanding, but can be facilitated by using specific hardware and high-performance computational libraries. Further, advances in sequencing technology continue to deliver increasingly large molecular sequence data sets that are often heavily partitioned in order to accurately model the underlying evolutionary processes. In phylogenetic analyses, partitioning strategies involve estimating conditionally independent models of molecular evolution for different genes and different positions within those genes, requiring a large number of evolutionary parameters that have to be estimated, leading to an increased computational burden for such analyses. Along with the increase in dimensions of sequence data, the past two decades have also seen the rise of multi-core processors, enabling massively parallel computations that are not yet fully exploited by many software packages for multipartite analyses. I will discuss a Markov chain Monte Carlo (MCMC) approach using an adaptive multivariate transition kernel to estimate in parallel a large number of parameters, split across partitioned data, by exploiting multi-core processing. These implementations are all part of the BEAST code base and the BEAGLE library, widely used open source software packages that allow performing Bayesian phylogenetic inference.

> **Stéphane DRAY**
Laboratoire de Biométrie et Biologie Evolutive (LBBE), Lyon, FR

Environment, traits, space and phylogeny: integrating constraints in the analysis of ecological data tables

In ecology, a community is defined as an assemblage of individuals belonging to different species and co-occuring at a given place and at a given time. Community ecology aims to describe the composition of species assemblages and to disentangle the effects of various processes (abiotic or biotic factors) that shapes species distributions at multiple spatial scales. As experiments are rarely feasible, studies in community ecology are usually based on field works where assemblages are described by the measurements of species abundances in sampling sites. These data tables are then usually summarized by multivariate methods to identify similarities (i) between sites according to their species compositions and (ii) between species according to their distributions. On the other hand, sampling sites can also be described by environmental descriptors (e.g., climatic variables) and this information can be considered to evaluate the effect of environmental gradients (niche filtering) on the structuring of ecological communities. This could be achieved by constrained (or canonical) multivariate methods that allows to focus on the effect of a set of predictors (here, environment) on the structure of species communities. In my talk, I will describe this principle of constrained analysis and show how it can be generalized to integrate various sources of external information gathered on sampling sites (environment, spatial information) and/or species (traits, phylogeny).

> **Barbara HOLLAND**
Theoretical Phylogenetics Group, School of Mathematics and Physics, University of Tasmania, AU

Desirable properties of models for phylogenetic analysis

These days the most widely used methods of phylogenetic inference are based on explicit models of how sequences (DNA nucleotides, amino acids, or codons) evolve. Recently our research group in Hobart has been interested in a property of sequence evolution models that seems fundamental, but

surprisingly it is not attained by many popular models. This property is that of closure under multiplication. The closure property ensures that if I take a transition matrix from a particular class of models (e.g. F81) and multiply it by another matrix from the same class, the resulting process is still in the same class. Such a property becomes important if we believe that evolution is likely to be a heterogeneous process that might act differently in different branches of the tree. In this talk I will discuss which models have this property, and which don't. I will give some results from simulation studies that explore the consequences of lack of closure for phylogenetic inference in terms of both branch lengths and topological accuracy. I will also discuss the consequences of lack of closure for codon models when they are used to estimate the ratio of synonymous to non-synonymous changes.

> **Anna-Sapfo MALASPINAS**
Institute of Ecology and Evolution, University of Bern, CH

A genomic history of Aboriginal Australia

The population history of Aboriginal Australians remains largely uncharacterized. Here we generate high-coverage genomes for 83 Aboriginal Australians (speakers of Pama–Nyungan languages) and 25 Papuans from the New Guinea Highlands. We find that Papuan and Aboriginal Australian ancestors diversified 25–40 thousand years ago (kya), suggestingpre-Holocene population structure in the ancient continent of Sahul (Australia, New Guinea and Tasmania). However, all of the studied Aboriginal Australians descend from a single founding population that differentiated ~10–32 kya. We infer a population expansion in northeast Australia during the Holocene epoch (past 10,000 years) associated withlimited gene flow from this region to the rest of Australia, consistent with the spread of the Pama–Nyungan languages. We estimate that Aboriginal Australians and Papuans diverged from Eurasians 51–72 kya, following a single out-of-Africa dispersal, and subsequently admixed with archaic populations. Finally, we report evidence of selection in AboriginalAustralians potentially associated with living in the desert.

> **Yun SONG**
Computer Science Division and Department of Statistics, UC Berkeley; Departments of Mathematics and Biology, University of Pennsylvania, USA

The major determinants of translation speed and their evolutionary signatures

Translation elongation speed is quite heterogeneous along the transcript. Previous studies have shown that multiple factors are involved in regulating the speed, but the observed heterogeneity remains only partially explained. In this talk, I will describe a method based on probabilistic modeling to extract useful quantitative information from ribosome profiling data and offer some new insights into the major determinants of the translation dynamics. I will then describe evolutionary signatures that are consistent with our finding.

> **Marc SUCHARD**
David Geffen School of Medicine at UCLA, Departments of Biomathematics, Biostatistics and Human Genetics, USA

Phylogenetic factor analysis

Phylogenetic comparative methods explore the relationships between quantitative traits adjusting for shared evolutionary history. This adjustment often occurs through a Brownian diffusion process along the branches of the phylogeny that generates model residuals or the traits themselves. For high-dimensional traits, inferring all pair-wise correlations within the multivariate diffusion is limiting. To circumvent this problem, we propose phylogenetic factor analysis (PFA) that assumes a small unknown number of independent evolutionary factors arise along the phylogeny and these factors generate clusters of dependent traits. Set in a Bayesian framework, PFA provides measures of uncertainty on the factor number and groupings, combines both continuous and discrete traits, integrates over missing measurements and incorporates phylogenetic uncertainty with the help of molecular sequences. We develop Gibbs samplers based on dynamic programming to estimate the PFA posterior distribution, over three-fold faster than for multivariate diffusion and a further order-of-magnitude more efficientlyin the presence of latent traits. We further propose a novel marginal likelihood estimator for previously impractical models with discrete data and find that PFA also provides a better fit than multivariate diffusionin evolutionary questions in columbine flower development, placental reproduction transitions and triggerfish fin morphometry.

TALKS

> **Peter Beerli**[1]; Haleh Ashki[1,2]; Michal Palczewski[1,3]
[1] Department of Scientific Computing, Florida State University, Tallahassee FL, USA;
[2] Institute for Health Policy Studies, University of Californian, San Francisco CA, USA;
[3] currently at Google, Inc., Mountain View CA, USA

Population divergence estimation using individual lineage label switching

Divergence time estimation from multilocus genetic data have become common in population genetics and phylogenetics. We present a new Bayes inference method that treats the divergence time as a random variable. The divergence time is calculated from an assembly of splitting events on individual lineages in a genealogy. The waiting time for such a splitting event is drawn from a hazard function of the truncated normal distribution. This allows easy integration into the standard coalescence framework used in programs such as MIGRATE. We explore the accuracy of the new inference method with simulated population splittings over a wide range of divergence time values and with two datasets of the Zika and Chikungunya pathogens; the geographic analyses of the expansion of both pathogens follows an Africa to Asia to Americas trajectory, corroborating analyses based only on the dates of incidences. Evaluations of simple divergence models show high accuracy, whereas the accuracy of the results of isolation with migration (IM) models depend on the magnitude of the immigration rate and potentially on the number of samples. Looking backwards in time, high immigration rates lead to a time of the most recent common ancestor of the sample that predates the divergence time, thus looses any potential signal of the divergence event in the sample data. This reduced accuracy with high immigration rates is problematic.

> **Florian Clemente**[1]; Mathieu Gautier[1][2]; Renaud Vitalis[1][2]
[1] Institut de Biologie Computationnelle (IBC), Montpellier, France;
[2] Centre de Biologie pour la Gestion des Populations, INRA, France

Estimation of the sex-ratio in population trees from allele frequency data.

The relative female and male contributions to demography are of great importance to better understand the history and dynamics of populations. While earlier studies relied on uniparental markers to investigate sex-specific questions, the increasing amount of sequence data now enables us to take advantage of thousands of independent loci from autosomes and the X chromosome. Here, we develop a novel method to estimate the effective sex-ratio, which profits from jointly estimating the branch lengths from autosomes and X chromosomes for a given tree topology. Our method is based on the Kimura diffusion approximation for genetic drift to explicitly model the change of allele frequencies along the branches. In its Bayesian framework, it can be applied to genome-wide SNP data for model and non-model organisms. We show via simulations that parameters are inferred robustly even under scenarios that violate the model assumptions. However, usage of ascertained SNPs may be highly misleading in such kind of parameter inference. In a human sample of non-African populations, we find a male bias in Oceanians that may reflect complex marriage patterns in Aboriginal Australians.

> **Caroline Colijn** [1]; Jennify Gardy [2]; Christophe Fraser [3]; Xavier Didelot [1]
[1] Imperial College London,
[2] British Columbia Centre for Disease Control (UBC);
[3] Oxford Big Data institute

Connection genetic and clinical information to understand pathogen transmission

There has been increasing interest in using short-term genetic variation in pathogens to trace chains of transmission. Indeed, many promises have been made that whole-genome sequencing of pathogens will allow us to reconstruct precisely who infected whom, with a reduced need for time-consuming outbreak investigations. The idea is that, just as in the childrens' game of 'telephone', a pathogen moving from host to host accrues small amount of genetic variation that can be used to trace the infection through the population. However, the phylogenetic trees that describe the pathogen's genetic variation are not directly informative about who infected whom; a phylogenetic tree is not a transmission tree. Using sequences to infer transmission is challenging: variation does not occur in direct proportion to the time elapsed, individual courses of infection and infectiveness vary, and the result is that even with whole-genome sequencing, considerable uncertainty remains. We have developed a Bayesian approach to inferring transmission using sequence data together with clinical information. It uses a "colouring" approach to explore the space of transmission trees that are consistent with a fixed phylogenetic tree. The key ingredient is the likelihood of a branching (transmission) tree given a finite time and incomplete sampling, in a general epidemiological model. Together with the colouring approach to project transmission trees

onto phylogenetic trees, this allows inference of who infected whom, with uncertainty, in an ongoing outbreak in which not all the cases are known. The approach takes in-host variation in the pathogen into account. We have implemented this approach using reversible-jump MCMC to account for the variable numbers of unsampled cases, and made the method available in the R package TransPhylo. We demonstrate the method using tuberculosis data, and illustrate the uncertainty that remains in who infected whom even when whole genome sequences are available.

> **Nicola De Maio**[1]; Daniel Wilson[1]
[1] University of Oxford

The bacterial sequential Markov coalescent

Bacteria can exchange and acquire new genetic material from other organisms directly and via the environment. This process, known as bacterial recombination, has a strong impact on the evolution of bacteria, for example leading to the spread of antibiotic resistance across clades and species, and to the avoidance of clonal interference. Recombination hinders phylogenetic and transmission inference because it creates patterns of substitutions that are not consistent with the hypothesis of a single evolutionary tree (homoplasies). Bacterial recombination is typically modelled as statistically akin to the gene conversion process of eukaryotes, i.e., using the coalescent with gene conversion (CGC). However, this model can be very computationally demanding as it requires to account for the correlations of evolutionary histories of even distant loci. So, with the increasing popularity of whole genome sequencing, the need has emerged for a new and faster approach to model and simulate bacterial evolution at genomic scales. We present a new model that approximates the coalescent with gene conversion: the bacterial sequential Markov coalescent (BSMC). Our approach is based on a similar idea to the the sequential Markov coalescent (SMC), an approximation of the coalescent with recombination. However, bacterial recombination poses hurdles to a sequential Markov approximation, as it leads to strong correlations and linkage disequilibrium across very distant sites in the genome. Our BSMC overcomes these difficulties and shows both a considerable reduction in computational demand compared the exact CGC, and very similar patterns in the simulated data. We use the BSMC within an Approximate Bayesian Computation (ABC) inference scheme and show that we can correctly recover parameters simulated under the exact CGC, which further showcases the accuracy of our approximation. We also use this ABC approach to infer recombination rate, mutation rate, and recombination tract length from a whole genome alignment of Bacillus cereus. Lastly, we implemented our BSMC model within a new simulation software FastSimBac. In addition to the decreased computational demand compared to previous bacterial genome evolution simulators, FastSimBac also provides a much more general set of options for evolutionary scenarios, allowing population structure with migration, speciations, population size changes, and recombination hotspots. FastSimBac is available from https://bitbucket.org/nicofmay/fastsimbac and is distributed as open source under the terms of the GNU General Public Licence.

> **Damien M. de Vienne**[1]
[1]Univ Lyon, Université Claude Bernard Lyon 1, CNRS, Laboratoire de Biométrie et Biologie Evolutive, UMR5558, Villeurbanne, Lyon, France

Lifemap, Exploring the Entire Tree of Life. And after?

The visualization of trees with thousands to millions of tips is a challenge that gets more important every day. Trees are growing larger and larger, due to intensive sequencing programs, improved tree reconstruction methods, and to the desire of researchers and public to visualize the complete evolutionary history of life on earth, the so-called Tree of Life (ToL). Some solutions to the problem of visualizing very large trees have been proposed in the last 5 years, with tools such as OneZoom [1] or DeepTree [2] that propose deep-zooming approaches. However, these tools have some limitations that prevent their use for visualizing, for instance, the ToL: incapacity to display multifuractions (one node connected to more than two descendants), cpu overload, or impossibility to display more than a few hundreds of thousands of leaves. To solve these issues, I developed Lifemap, a tool that allows a smooth and interactive exploration of the entire Tree of Life or any tree with millions of leaves. Lifemap is inspired by advances made in cartography, in the context of the OpenStreetMap project (an open-source equivalent to Google Maps). It also solves the problem of multifurcation in a deep-zooming context by proposing a new representation of trees based on half-circles fractal-like structures. Since its publication last December in PLOS Biology[3], Lifemap is becoming a reference for teaching evolution at high-school and universities. However, future developments of Lifemap should also be towards researchers, especially in comparative genomics. In a near future, Lifemap could for instance (i) become a useful solution for accessing diverse resources that are of interest in our everyday work (trees, sequences, protein structures, genomes, ...) by linking to external databases (NCBI, Uniprot, PDB, ...), and (ii) allow the visualization of genomic features that are never visualized at this scale on a phylogeny (GC%, Genome sizes, etc.). After presenting Lifemap, I will discuss recent advances and possible improvements

along these lines.

[1] Rosindell J, Harmon LJ (2012) OneZoom: A Fractal Explorer for the Tree of Life. PLOS Biology 10(10): e1001406. doi: 10.1371/journal.pbio.1001406
[2] Block F, Horn M, Phillips BC, Diamond J, Evans M, Shen C (2012) The DeepTree Exhibit: Visualizing the Tree of Life to Facilitate Informal Learning. IEEE Transactions on Visualization and Computer Graphics (TVCG), 18(12), Page(s): 2789 - 2798, 2012, ISSN: 1077-2626.
[3] de Vienne DM (2016) Lifemap: Exploring the Entire Tree of Life. PLOS Biology 14(12): e2001624. doi: 10.1371/journal.pbio.2001624

> **Fabian Freund** [1], Arno Siri-Jégousse [2], Jere Koskela [3]
[1] Institute of Plant Breeding, Seed Science and Population Genetics, University of Hohenheim, Stuttgart, Germany;
[2] Departamento de Probabilidad y Estadística, IIMAS, UNAM, Mexico City, Mexico;
[3] Institute of Mathematics, TU Berlin, Germany

Beyond the site-frequency spectrum: Low-dimensional statistics to distinguish between coalescent models

Consider a genetic locus without recombination which is selectively neutral. The standard model for the genealogy of a sample (of alleles) taken from a fixed-size, large population is Kingman's n-coalescent, a bifurcating random tree with n leaves and a Markovian structure of merging branches. To model genealogies of samples from structured populations or populations moderately fluctuating in size, similar bifurcating random trees can be used as genealogy models. However, for samples e.g. from populations undergoing sweepstake reproduction, extreme population size changes or rapid selection at other loci, theory predicts that multiple-merger n-coalescents should be better fitting as genealogy models (reviewed in [TL]). Multiple-merger coalescents are random trees with n leaves and a Markovian structure of merging branches similar to Kingman's n-coalescent, but allow for multifurcations. Several approaches have been introduced to distinguish between these different n-coalescents based on genetic data, see a (non-exhaustive) list in [GAE]. While likelihood methods based on the full genetic data (see e.g. [SBB]) are computationally expensive, lower-dimensional approaches are usually based on (summaries of) the site-frequency spectrum (SFS), the frequency spectrum of the mutations observed in the sample. Mutations are placed on the genealogy via a Poisson point process with homogeneous rate. The SFS does not contain information about how different mutations are shared by sampled alleles. We analyse the probabilistic properties of several statistics that measure such information in different n-coalescents, including asymptotics for n tending to infinity. We consider, for a chosen allele, the number of other alleles that share all mutations of the chosen allele that are not private. Private mutations are only present in the chosen allele itself. Additionally, we consider several one-dimensional statistics summarising the genetic information within different subsamples of 3 alleles. Complementary, we use simulations to assess the statistical properties for non-asymptotical tests based on these statistics.

References:

[TL] Tellier, Aurélien, and Christophe Lemaire. "Coalescence 2.0: a multiple branching of recent theoretical developments and their applications." Molecular ecology 23.11 (2014): 2637-2652.
[GAE] Grant, W. Stewart, Einar Árnason, and Bjarki Eldon. "New DNA coalescent models and old population genetics software." ICES Journal of Marine Science: Journal du Conseil (2016): fsw076.
[SBB] Steinrücken, Matthias, Matthias Birkner, and Jochen Blath. "Analysis of DNA sequence variation within marine species using Beta-coalescents." Theoretical population biology 87 (2013): 15-24.

> **Alex Gavryushkin**[1], Kristina Crona[2], , Devin Greene[2], Niko Beerenwinkel[1]
[1] Department of Biosystems Science and Engineering, ETH Zurich, Switzerland;
[2] Department of Mathematics, American University, Washington DC, USA

Inferring genetic interactions from competition experiments

Fitness is a central concept in evolutionary biology. The fate of evolving populations depends on the underlying fitness landscape, i.e., the collection of fitness values for all genotypes in the population, and the fitness landscape encodes all higher order epistatic gene interactions. In practice, however, it is hardly possible to measure fitness for all genotypes in a natural population. We present a mathematical framework and computational tools to make inference about epistatic gene interactions, when the fitness landscape is only incompletely determined, for example, due to imprecise measurements or partial observations. The method allows for robust inference of epistasis for a broad range of empirical data that can be summarized by pairwise fitness comparisons of genotypes, rather than the actual fitness values. We demonstrate that genetic interactions can often be inferred for fitness rank orders, where all genotypes are ordered according to fitness, and even for partial fitness orders, where only a subset of fitness comparisons is available. We provide a complete characterization of rank and partial orders that imply higher order epistasis and an efficient algorithm for detecting such interactions. Application of our method to empirical data revealed higher order interactions for a diverse set of genetic systems, including HIV, malaria, and antibiotic resistance.

> Tanja Stadler[1,2], **Alexandra Gavryushkina** [1,2], Rachel C. M. Warnock[1,2], Alexei J. Drummond[3], Tracy A. Heath[4]
[1] Department of Biosystems Science and Engineering (BSSE), ETH Zürich, Switzerland
[2] Swiss Institute of Bioinformatics (SIB), Switzerland,
[3] Department of Computer Science, University of Auckland, New Zealand;
[4] Department of Ecology, Evolution, & Organismal Biology, Iowa State University, USA

Fossilized birth-death process under different modes of speciation and fossil stratigraphic ranges.

In search of objective ways to transform fossil records to calibration densities for dating molecular phylogenies several new methods have been developed. These methods are based on employing the fossilized birth-death process as a model for simultaneous processes of speciation and fossil sampling. A sequence of events: fossilization, preservation, discovery, and inclusion in an analysis is modeled as a single “fossil sampling” event which occurs according to a Poisson process along lineages in a phylogeny. The existing applications of this model do not allow multiple fossils to be assigned to a single species. However, paleontological data bases are often comprised of stratigraphic ranges which represent multiple fossil specimens of the same species. The direct modeling of the stratigraphic range data would facilitate a more accurate inference. We have extended the fossilized birth-death process to model different modes of speciation and calculated the probability density of the sampled phylogenies with fossil samples assigned to distinct species. This probability density can further be used in a phylogenetic inference of dated species phylogenies or/and macroevolutionary parameters. Also, such a model can be used to infer transmission trees where several sequences per infected individual are sampled.

> Karin EL [1,2], Wicke S [3], Pupko T [2], **Mayrose I** [1]
[1] Department of Molecular Biology & Ecology of Plants, George S. Wise Faculty of Life Sciences, Tel Aviv University, Israel

A Likelihood Method for Detecting Trait-Dependent Shifts in the Rate of Molecular Evolution

Rate heterogeneity within groups of organisms is known to exist even when closely related taxa are examined. A wide variety of phylogenetic and dating methods have been developed that aim either to test for the existence of rate variation or to correct for its bias. However, none of the existing methods track the evolution of features that account for observed rate heterogeneity. Here, I will describe TraitRate [1], a likelihood model that combines two well-established sets of models: those that describe sequence evolution and those that describe the evolution of traits into one likelihood framework. The combined model can be used to determine whether shifts in the rate of molecular sequence evolution are associated with species’ intrinsic characteristics, such as a particular life-history trait, morphological feature, or habitat association. I will then describe an extension of the method that further allows the detection of specific sequence sites whose evolutionary rate is most noticeably affected following a transition in the analyzed character trait, suggesting a shift in functional/structural constraints [2]. The use of this method will be exemplified by applying it to study the evolutionary changes of plastid plant genomes that transition to a heterotrophic lifestyle.

[1] Mayrose, I, SP Otto. 2011. A likelihood method for detecting trait-dependent shifts in the rate of molecular evolution. Molecular Biology and Evolution. 28:759-770.
[2] Karin EL, Wicke S, Pupko T, Mayrose I. 2017. An integrated model of phenotypic trait changes and site-specific sequence evolution. Systematic Biology. In Press.

> **Rohan Mehta**[1]; David Bryant[2]; Noah Rosenberg[1]
[1] Stanford University, Stanford, CA, United States of America;
[2] University of Otago, Dunedin, New Zealand

The probability of monophyly of a sample of gene lineages on a species tree.

Monophyletic groups---groups that consist of all the descendants of a common ancestor---are important objects of study in fields that concern genealogy or population history, including phylogeography, species delimitation, and phylogenetics. Recent work has investigated mathematical aspects of monophyletic groups under coalescent models, generating predictions about the properties of monophyly in relation to the genealogy of a set of populations or species. We derive the probability that a group of individuals within a larger genealogy is monophyletic conditional on a species tree of any size and shape. We also extend two-species computations to compute the probability of reciprocal monophyly for samples from three or four species. We analyze the effects of species tree height, branch lengths, and sample size on monophyly probabilities. We also use an example dataset from the study of maize domestication to demonstrate that the theoretical probabilities we obtain are comparable to those found in computations from data. Finally, we present user-friendly software for computing monophyly probabilities under the assumptions of coalescent models.

> **Carina F. Mugal**[1]; Ingemar Kaj[2]; Verena E. Kutschera[1]; Jochen B.W. Wolf[1]; Hans Ellegren[1]
[1] Department of Ecology and Genetics, Uppsala University, Uppsala, Sweden;
[2] Department of Mathematics, Uppsala University, Uppsala, Sweden

Time-dependent estimates of dN/dS and its implications for molecular ecology and genetic studies

The ratio of divergence at non-synonymous and synonymous sites, dN/dS, is a widely used measure in molecular ecology and genetics to investigate the extent to which natural selection modulates gene sequence evolution. Its estimation is based on a phylogenetic approach and the computation of sequence divergence between codon sequences of related species. This approach relies on the indirect assumption that sequence divergence and species divergence are identical, an assumption, which is generally violated and reasonable only for distantly related species. The violation of the underlying assumption leads to a time-dependence of rate estimates, and biases estimates in particular for closely related species, where the impact of ancestral and lineage-specific polymorphisms is substantial (Mugal C.F. et al. Mol Biol Evol. 2014, 31, p212-231). Such time-dependence of estimates of molecular evolutionary rates has been brought up earlier in the context of the molecular clock (Garcia-Moreno J., J. Avian Biology 2004, 35, p465–468; Ho S.Y. et al., Mol. Biol Evol. 2005, 22, p1561-1568), and was initially put forward as a biological phenomenon rather than a methodological artifact. However, after several debates and discussions on this topic, also methodological artifacts were put forward as potential explanations, such as the effect of ancestral polymorphisms (Ho S.Y. et al., Mol. Ecol. 2011, 20, p3087-3101). To address this issue analytically, we here use Poisson random field models and formulate a codon model that is firmly anchored in population genetic theory. The derivation of an analytical expression of the time-dependent allele frequency spectrum (Kaj I. and Mugal C.F. Theor. Pop. Biol. 2016, 111, p51-64), allows us to express sequence divergence not only as a function of divergence time but also as a function of sample size. The dependence on sample size now enables us to show that the use of polymorphism data can assist the estimation of dN/dS for closely related species. Finally, we discuss the impact of the time-dependence of dN/dS on molecular ecology and genetic studies, and illustrate the problematic by studying a phylogeny of crows, which involves the estimation of dN/dS at different time-scales.

> Olivier Mazet[1]; **Willy Rodríguez**[2]; Simona Grusea[1]; Simon Boitard[3, 4]; and Lounès Chikhi[5, 6, 7]
[1] Université de Toulouse, Institut National des Sciences Appliquées, Institut de Mathématiques de Toulouse, Toulouse, France
[2] Génétique Quantitative et Evolution-Le Moulon, Institut National de la Recherche Agronomique/Université Paris-Sud/CNRS/AgroParisTech, Université Paris-Saclay, F-91190 Gif-sur-Yvette, France
[3] UMR7205 Institut de Systématique, Évolution et Biodiversité, École Pratique des Hautes Études & Muséum National d'Histoire Naturelle & CNRS and Université Pierre et Marie Curie, Paris, France
[4] UMR1313 Génétique Animale et Biologie Intégrative, Institut National de la Recherche Agronomique and AgroParisTech, Jouy-en-Josas, France
[5] CNRS, Université Paul Sabatier, ENFA, UMR 5174 EDB (Laboratoire Évolution and Diversité Biologique), Toulouse, France
[6] Université de Toulouse, UPS, EDB, Toulouse, France
[7] Instituto Gulbenkian de Ciência, Oeiras, Portugal

The Inverse Instantaneous Coalescence Rate (IICR) as a new summary statistic in population genetics.

The rapid development of DNA sequencing technologies is expanding the horizons of population genetic studies. It is expected that genomic data will increase our ability to reconstruct the history of populations. While this increase in genetic information will likely improve our capacity to reconstruct the demographic history of populations, it also poses big challenges. In some cases, hypotheses made by the models may lead to erroneous conclusions about history of the population under study. Recent works have shown that DNA patterns expected in individuals coming from structured populations correspond to those who come from unstructured populations with changes in size through time[1, 2]. As a consequence, it is often difficult to determine whether demographic events such as expansions or contractions (bottlenecks) inferred from genetic data, are real or due to the fact that populations are structured in nature[3]. Moreover, few inferential methods allowing to reconstruct past changes in population size take into account structure effects. For instance, it is increasingly recognized that popular methods like the PSMC[4] and MSMC[5] are sensitive to the effects of structure. In this talk, I will present some arguments based on the coalescent theory that allow to understand the presence of (sometimes spurious) signals of population size changes when the data deviate from the panmixia hypothesis. I will introduce the IICR (Inverse Instantaneous Coalescence Rate) as a different interpretation for population size changes in more general models (e.g. structured models). I will illustrate how the notion of IICR allows to exactly quantify the changes in population size that will be inferred when the PSMC is applied to a constant-size structured population. Finally, I will give some insights on how the IICR can be used as a summary

statistic computed by any inference method (like PSMC or MSMC). This idea makes possible to extend the application of these methods to infer parameters on a wide range of more general models in population genetics.

[1] Wakeley, J. (1999). Nonequilibrium migration in human history. Genetics, 153(4):1863–1871.
[2] Heller, R., Chikhi, L., and Siegismund, H. R. (2013). The confounding effect of population structure on bayesian skyline plot inferences of demographic history. PLoS ONE, 8(5):e62992.
[3] Mazet, O., Rodriguez, W., Grusea, S., Boitard, S., and Chikhi, L. (2016). On the importance of being structured: instantaneous coalescence rates and human evolution - lessons for ancestral population size inference ?. Heredity
[4] Li H, Durbin R (2011). Inference of human population hitory from individual whole-genome sequences. Nature, 475, 493–496.
[5] Schiffels S, Durbin R (2014). Inferring human population size and separation history from multiple genome sequences. Nature Genetics, 46, 919–925.

> **Robin Thompson**[1]; Chris Wymant[2]; Jayna Raghwani[1]; Christophe Fraser[2]; Katrina Lythgoe[1]
[1] Department of Zoology (University of Oxford), Oxford, UK
[2] Big Data Institute (University of Oxford), Oxford, UK

Accounting for donor viral diversity gives high estimates of the number of HIV founder virions among recipients

After observations that most sexually transmitted HIV infections are initiated by single strains, it was hoped that signatures of transmission would be identified and used as targets for vaccine development. Selection occurring at the transmission bottleneck can solve the paradox of a very small probability of transmission per contact but multiple transmitted/founder (T/F) strains when successful transmission does occur in 20-40 percent of new infections. However, genotypic and phenotypic signatures of transmission have been difficult to find. Using a probabilistic modeling approach, we show that selection need not be invoked to explain this apparent paradox. If transmission is only possible for a minority of contacts and changing viral diversity in the donor throughout their course of infection is accounted for, it is possible to resolve the low probability of transmission and new infections being founded by multiple strains 20-40 percent of the time. We apply our modeling framework to published whole-genome deep sequencing data, enabling us to infer the distributions of the number of virions and the number of distinct strains that establish new infections in a population. Crucially, we find that the number of T/F virions and strains are not identical quantities, and there is not necessarily always a positive correlation depending on the assumptions made about contact rates in the host population. This is important, since different studies have suggested that either the initiating volume of virus or the number of T/F strains are predictors of set point viral load. Our model illustrates that decoupling the number of T/F strains and virions is important for making accurate predictions of quantities characterizing infection. Furthermore, our results show that it is possible for most individuals to be infected by one or a small number of strains despite the more complex quasispecies observed in data measured in individuals later in infection, without requiring selection at transmission. This is due to changing diversity in donors throughout their courses of infection combined with the stochastic nature of exactly which strains transfer at each transmission event.

> **Jakub Truszkowski** [1,2], Nick Goldman[1], Simon Tavare[2]
[1] EMBL - European Bioinformatics Institute, Hinxton, Cambridgeshire
[2] Cancer Research UK - Cambridge Institute, University of Cambridge

Reconstructing phylogenies from single-cell sequencing data

Advances in sequencing technology are making it possible to sequence the genomes of individual cells. Single-cell sequencing provides an opportunity to survey the genomic heterogeneity of cells within an organism, both in healthy development and in cancer. Most cell divisions introduce mutations as a result of DNA replication errors. Consequently, the history of cell divisions in the organism is encoded in the genomes of individual cells, and could be reconstructed by phylogenetic methods. In this talk, we present a method for reconstructing cell evolutionary histories while accounting for the stochastic nature of sequencing errors. We show that the problem can be reduced to finding a series of graph cuts in a certain graph. In simulations, our method outperforms standard phylogenetic methods for this task. On real data sets from human cancers and healthy mice, our method produces plausible trees. A comparison of our reconstructed phylogenies with variant allele frequency data from bulk sequencing reveals high levels of agreement between the trees and the constraints implied by variant allele frequencies, and also allows us to further improve our reconstructed trees by identifying sites with high levels of sequencing error.

> **Erik M. Volz**[1]; Sergei Kosakovsky-Pond[2]; Simon D.W. Frost[3]
[1] Department of Infectious Disease Epidemiology, Imperial College London, UK
[2] Department of Biology, Temple University, Philidelphia, USA
[3] Department of Veterinary Medicine, Cambridge University, UK

Scalable relaxed clock dating

Many large data sets of fast-evolving viruses are not well fitted by molecular clock models that assume a constant substitution rate through time. Estimation of relaxed molecular clocks using state-of-the-art Bayesian methods is computationally expensive and not scalable to large data sets. We build on recent advances in maximum likelihood and least-squares phylogenetic and molecular clock dating methods to develop a fast relaxed-clock method based on a Gamma-Poisson mixture model of substitution rates. Our method estimates a distinct substitution rate for every lineage in the phylogeny while being scalable to large phylogenies. Unknown lineage sample dates can be estimated as well as unknown root position. We estimate confidence intervals for rates, dates, and tip dates using a parametric bootstrap approach. This method is implemented as an open-source R package 'treedater'. We further present bioinformatic pipelines for rapid phylodynamic analysis of viral sequence data and demonstrate this approach using a large whole genome sequence alignment of Ebola virus from the 2014-2015 epidemic in Western Africa.

> **Liangliang Wang**[1]; Alexandre Bouchard-Côté[2]
[1] Simon Fraser University, Burnaby, Canada
[2] University of British Columbia, Vancouver, Canada

An adaptive sequential Monte Carlo sampler for Bayesian phylogenetic non-clock tree inference

Bayesian phylogenetics, which approximates a posterior distribution of phylogenetic trees, has become more and more popular with the development of Monte Carlo methods. Standard Bayesian estimation of phylogenetic trees can handle rich evolutionary models but requires expensive Markov chain Monte Carlo (MCMC) simulations, which may suffer from two difficulties, the curse of dimensionality and the local-trap problem. Our previous work in [1] has shown that sequential Monte Carlo (SMC) methods can serve as a good alternative to MCMC in posterior inference over phylogenetic trees. In this talk, I will present our recent work on an SMC sampler for general non-clock trees that can incorporate the MCMC kernels from the rich literature of Bayesian phylogenetics. We illustrate our method using simulation studies and real data analysis.

[1] Liangliang Wang, Alexandre Bouchard-Côté, and Arnaud Doucet. Bayesian phylogenetic inference using a combinatorial sequential Monte Carlo method. Journal of the American Statistical Association, 110(512):1362–1374, 2015.

> **Anna Zhukova** [1], Stéphane Hué [2], Olivier Gascuel [1]
[1] Unité Bioinformatique Evolutive, C3BI, USR 3756 Institut Pasteur et CNRS, Paris, France
[2] London School of Hygiene & Tropical Medicine, London, UK.

A phylodynamic model of pathogen drug resistance emergence and transmission

Drug resistance mutations (DRMs) emerge in genetic sequences of pathogens through selective pressure during therapy. Drug resistances can be transmitted and reduce the chances of long-lasting successful treatment. The rates at which DRMs are acquired and transmitted, and at which transmitted DRMs persist are multifactorial and vary considerably, depending on the pathogen (e.g. HIV, Mycobacterium tuberculosis), treatment and resistance mutation. To improve our understanding of pathogen transmission and drug resistance mechanisms, we developped a Markov chain-based model that describes both the state evolution along the tree branches, and the transmission process. The states correspond to a patient-pathogen pair: treatment-experienced or -naive patient, infected with drug-sensitive or drug-resistant pathogen strains. The state transitions can be caused by three events: start of treatment, acquirement of DRM due to treatment, DRM loss in the absence of drug selective pressure. Our model also accounts for partial sampling of the population. Overall, it includes 8 parameters: sampling, state transition and transmission rates (for different states of a patient-virus pair). Our model has several important, non-standard properties. Notably, the characteristics of the transmission tree depend on the state evolution along the branches; for example, as under successful treatment the pathogen presence in the patient is generally undetectable, the transmission rate from the “treatment-experienced drug-sensitive” state is low, and consequently there are few inner nodes in this state in the transmission tree. Another property is that our model is not ergodic: The ‘’vertical’’ stationary distribution of states on a long path in the tree (corresponding to a pathogen lineage over time) is different from the ‘’horizontal’’ stationary state distribution of an infected population at a given time point. We describe a mathematical study of this model. We provide the equations linking the described equilibrium frequencies to the model parameters, and an analytical solution to compute the likelihood of a sampled transmission tree with known tip states. We show using simulated trees that maximum-likelihood optimization allows for accurate estimation of the model parameters, including the sampling rate, which is an uncommon feature among phylodynamic models.

POSTERS

Poster 1

> **Ilana Arbisser** [1], Ethan Jewett [2], Noah Rosenberg [1]
[1] Stanford University; Stanford, California, US
[2] University of California, Berkeley; Berkeley, California, US

On the joint distribution of tree height and length under the coalescent

Many statistics that examine genetic variation depend on the underlying shapes of genealogical trees. Under the coalescent model, we investigate the joint distribution of two of the simplest quantities that describe genealogical tree shape: tree height and tree length. We derive a recursive formula for their exact joint distribution under a simple demographic model of a constant-sized population. We obtain approximations for the mean and variance of the ratio of tree height to tree length, using them to show that this ratio converges in probability to 0 as the sample size increases without bound. Next, using simulations, we examine the joint distribution of height and length under demographic models with population growth and population subdivision. We interpret the joint distribution in relation to problems of interest in data analysis, including inference of the time to the most recent common ancestor and tests of neutrality. The results assist in understanding the influences on two fundamental features of tree shape.

Poster 2

> **Diepreye** [1]; Louis GrandJean [2,3] ; Carolijn Colijn [1]
[1] Department of Mathematics, Imperial College London ;
[2] Universidad Peruana Cayetano Heredia, Avenida Honorio Delgado, San Martin de Porras, Lima, Peru,
[3] Academic Health Sciences Center, Imperial College London

Detecting disruptive sites in tuberculosis genome

Evolutionary tools, especially phylogenetic trees, are increasingly being used to study short-term variation in infectious pathogens, with the aim of improv- ing our ability to understand pathogen adaptation and control infections. The Mycobacterium tuberculosis genome consists of about ~ 4000 genes and approximately a tenth of these are in two families of genes that constitute a highly repetitive class of antigenic genes which are poorly characterized. These repeti- tive regions are highly variable, and so are a potential source of useful phyloge- netic information, but they are prone to considerable sequencing and alignment error, and so are often removed before reconstruction of a phylogenetic tree (and onward analysis). Furthermore, sites known to be under selection, such as sites where mutations confer antibiotic resistance, are often removed prior to tree inference, because their evolution is likely not to be consistent with a tree model due to convergent evolution (homoplasy). However, there is no consensus in the literature as to what sites or regions of a multiple sequence alignment should be included and which should be excluded before a phylogenetic analysis, with some researchers excluding repetitive regions and resistance sites and oth- ers not. Here we present a phylogeny-based method to detect phylogenetically disruptive sites. We identify sites which, when removed, result in a substantially altered phylogeny. Our method is a sliding window approach across the multiple sequence alignment, in which sites in the window are dropped from the align- ment, a tree is then inferred and compared (using an appropriate metric) to a reference tree which is reconstructed on the entire alignment. The method thus associates each site with a tree distance. We take a classification approach to the question of which sites should be removed prior to phylogenetic analysis, using the tree distances as an input and testing whether we can automatically identify disruptive sites. We test our method on simulated data in which the disruptive sites are known, and on a tuberculosis dataset that is enriched for resistance. In both cases, our method is able to identify sites that are phylogenetically dis- ruptive; these include sites under antibiotic selection and hypervariable genomic regions. We illustrate the effect excluding these sites has on onward inference of the phylogeny and the homoplasy. We compare our approach with the influence function of [Bar-hen et al, MBE, 2008], which is based on tree likelihoods, and we discuss the differing results in the context of closely related bacterial sequences.

Poster 3

> **Joëlle Barido-Sottani** [1]; Tanja Stadler [1]
[1] ETH Zurich, Department of Biosystems Science and Engineering (BSSE), Basel, Switzerland

Detection of HIV transmission clusters from phylogenetic trees through a multi-states birth-death model

A lot of infectious diseases, including sexually-transmitted pathogens like HIV, are not transmitted at random in the susceptible population, but across a network of contacts. This network is usually composed

of several clusters, parts of the transmission network that are strongly inter-connected but have few links between each other. Identifying these clusters and thus the main groups in the transmission network has multiple applications, for instance allowing public health officials to target the populations most vulnerable to infection. Clustering methods aiming to identify clusters in a phylogeny built from viral sequences already exist, but they are restricted to monophyletic clades and cannot detect nested clusters. They also require the user to specify a cutpoint as input. Villandre et al. (2016) tested those methods on simulated epidemics and found that their performance was strongly dependent both on the cutpoint settings and on the non-violation of the monophyletic assumption. In this work we present a new method to identify clusters of transmission, based on detecting changes in the transmission rate. During the progression of an epidemic, the infection will first spread quickly after its introduction into a cluster. The rate of transmission will then go down as the population of susceptibles in the cluster is progressively exhausted. Eventually an introduction event into a new cluster may occur through one of the inter-cluster connections. The rate of transmission will then go up suddenly as a new, entirely susceptible, population group becomes accessible. Our proposed multi-states birth-death model can detect these sudden increases and thus sort the lineages of the transmission tree in different states, corresponding to the different population clusters.

Poster 4

> Cécile Ané[1,2]; **Paul Bastide**[3,4]; Mahendra Mariadassou[4]; Stéphane Robin[3]
[1] Department of Statistics, University of Wisconsin-Madison, WI, 53706, USA;
[2] Department of Botany, University of Wisconsin-Madison, WI, 53706, USA;
[3] UMR MIA-Paris, AgroParisTech, INRA, Université Paris-Saclay, 75005, Paris, France;
[4] MaIAGE, INRA, Université Paris-Saclay, 78352 Jouy-en-Josas, France

Detection of adaptive shifts for multivariate Ornstein-Uhlenbeck models of trait evolution with correlations and missing data

The goal of Phylogenetic Comparative Methods (PCM) is to study the distribution of several quantitative traits among related species. The classical framework consists of a multivariate random process running along the branches of a phylogenetic tree, that describes the shared evolutionary history of the studied species [3]. One popular choice is the Ornstein-Uhlenbeck (OU) process [4]. Compared to the Brownian Motion (BM), the OU has a tendency to revert to a given state, that can be interpreted as an optimal state for the species in a given environment, each trait having its own optimal value. Assuming a phylogenetic niche conservatism, those optimal states should change only a few times along the phylogeny, in association with dramatic events, such as migration, or climate change. Our goal is to automatically detect the position of these shifts on the phylogenetic tree. The general OU is difficult to study, even when the position of the shifts is known a priori [2]. A first simplifying assumption could be that all the traits are evolving independently [5, 6]. However, this is rarely the case in practice, and we show that phylogenetic Principal Component Analysis [7] fails to decorrelate traits in a shifted dataset, even under a BM hypothesis. To address this issue, we used a "scalar OU" (scOU) process, that relaxes this independence assumption, allowing the traits to evolve in a correlated fashion. For the model to be tractable, this however comes at the cost of assuming that all the traits revert to their own optimal value with the same speed. Building on the univariate case [1], we designed a multivariate method that allows for an efficient maximum likelihood reconstruction of a shifted scOU process on a fixed phylogeny. The inference is based on a re-scaling of the tree and on an incomplete-data formulation of the model, that allows for missing values, and makes data imputation or ancestral state reconstruction possible. As only extant species are observed, some identifiability problems in the localization of the shifts on the tree arise, which we already studied in the univariate setting. The number of shifts is selected through a penalized likelihood procedure, with a tailored penalty that takes those identifiability issues into account. Thanks to an upward-downward algorithm, adapted from Felsenstein's pruning algorithm, and to an efficient implementation, our method is quite fast, and can scale up to datasets with a large number of species (more than 1000). It is available as an R package on the CRAN (PhylogeneticEM).

[1] Bastide et al. (2016). J. R. Stat. Soc. B.
[2] Clavel et al. (2015). Methods Ecol Evol, 6(11), 1311–1319.
[3] Felsenstein (1985). The American Naturalist, 125(1):1-15.
[4] Hansen (1997). Evolution, 51(5):1341-1351
[5] Ingram & Mahler (2013). Methods Ecol Evol, 4(5), 416–425.
[6] Khabbazian et al. (2016). Methods Ecol Evol, 7(7), 811–824. [7] Revell (2009). Evolution, 63(12), 3258–3268.

Poster 5

> **Sarah Bastkowski** [1]
[1] The Earlham Institute, Norwich, UK

SPECTRE: a Suite of PhylogEnetiC Tools for Reticulate Evolution - The Release

The use of phylogenetic networks is becoming increasingly popular among biologists as a means of

inferring reticulate evolutionary patterns that cannot be represented by a tree. Split systems, collections of weighted bipartitions of taxa, have provided a mechanism to do this but research into efficient algorithms for computing planar split networks that best represent the data is ongoing.There is currently a lack of robust, open-source implementations of associated data structures and algorithms for computing split systems and visualising these planar split networks.To address this we present Spectre, a readily available, open-source library of data structures written in Java, that comes complete with new implementations of several pre-published algorithms and an interactive graphical interface for visualing planar split networks.Longer running tools can also be executed at the command line on servers or in High Performance Computing (HPC) environments.

Poster 6

> **Arnaud Becheler**[1]; Camille Coron[2]; Stéphane Dupas[3]
[1] Laboratoire Evolution Genomes Comportement Ecologie (EGCE), Gif-sur-Yvette, France;
[2] Laboratoire de Mathématiques d'Orsay (LMO), Orsay, France;
[3] Laboratoire Evolution Genomes Comportement Ecologie (EGCE), Gif-sur-Yvette, France;

Demogenetics model for invasive processes

Because biological invasions are processes well delimited in both space and time, they offer a unique framework in which complex models can be studied by numerical simulations. We use the Approximated Bayesian Computation method (ABC) to study the Vespa velutina invasion, estimating the parameters of spatially explicit demographic and genetic probabilistic model. The population growth in each landscape unit is described by a function of the local environmental features, whereas the migration flux between populations are sampled in laws which densities are functions of the geographical distance. Some of the parameters of these functions are unknown and should be estimated. Conditionally to the demography, a coalescence process allows for simulating the genetic history of the sample. Once the simulation done, the ABC method allows for accepting/rejecting the parameters values as a function of the data plausibility they generate.

Poster 7

> María Inés Fariello[1,2], **Simon Boitard**[3], Sabine Mercier[4,5], Magali SanCristobal[3]
[1] Facultad de Ingeniería, Universidad de la República, Montevideo, Uruguay.
[2] Unidad de Bioinformática, Institut Pasteur, Montevideo, Uruguay.
[3] INRA, GenPhySE, Castanet-Tolosan, France.
[4] Université de Toulouse II, UFR SES, Département Mathématique-Informatique, Toulouse, France.
[5] Université de Toulouse, UMR5219, Institut de Mathématiques, Toulouse, France.

Integrating results from single marker tests in genome scans for selection : the local score approach

Detecting genomic footprints of selection is an important step in the understanding of evolution. Accounting for the correlation between markers in genome scans increases detection power, but haplotype-based methods require individual genotypes and are not applicable on pool-sequenced samples. We propose to take advantage of the local score approach to account for linkage disequilibrium in genome scans for selection, accumulating (possibly small) signals from single markers over a genomic segment, to clearly pinpoint a selection signal. Using computer simulations, we demonstrate that this approach detects selection with higher power than several state-of-the-art single marker, windowing or haplotype-based approaches. We illustrate this on a benchmark data set including individual genotypes, for which we obtain similar results with the local score approach and one haplotype-based approach. Finally, we apply the local score approach to Pool-Seq data obtained from a divergent selection experiment on behavior in quail, and obtain precise and biologically coherent selection signals : while competing methods fail to highlight any clear selection signature, our method detects several regions involving genes known to act on social responsiveness or autistic traits. Although we focus here on the detection of positive selection from multiple population data, the local score approach is general and can be applied to other genome scans for selection or other genome-wide analyzes such as GWAS.

Poster 8

> **Stephen M Crotty**[1,2], Bui Q Minh[1], Barbara R Holland[3], Lars S Jermiin[4], Nigel G Bean[2], Jonathan Tuke[2], Arndt von Haeseler[1]
[1] Centre for Integrative Bioinformatics Vienna, Vienna, Austria,
[2]University of Adelaide, Adelaide, SA, Australia,
[3] University of Tasmania, Hobart, Tasmania, Australia,
[4] Australian National University, Canberra, ACT, Australia

GHOST: A mixture model for phylogenetic inference of heterogeneously evolved sequence data

Heterogeneous evolutionary processes have cast a shadow over the reliability of phylogenetic inference for as long as it has been attempted. These processes bring with them the inevitable consequence of model

misspecification, which one would obviously like to minimize. Much work has been done in this area and mixture models that account for rate heterogeneity amongst sites have been in widespread use for some time. These models however are too restrictive to truly represent heterotachous evolution. At the cost of complexity, we introduce a more general mixture model capable of recovering tree and model parameters from sequence alignments evolved under heterotachous conditions. We then apply our model to a real dataset, where it recovers the subtle phylogenetic signal associated with the convergent evolution in a sodium channel gene of two geographically distinct lineages of electric fish.

Poster 9

> **Victoria Culshaw**[1], Thiago Rangel[2], Isabel Sanmartín[3]
[1] Real Jardín Botánico, Madrid, CSIC,
[2] LETS, Univeridade Federal de Goiás,
[3] Real Jardín Botánico, Madrid, CSIC

Exploring the relationship between present-day distributions and evolutionary history using a novel spatial simulation approach

Using a non-stochastic host/parasite model associated to a non-informative landscape, Moore dispersal kernel and a reproduction equation, Hassell et al. (1991) demonstrated that for such a model only four spatially-stable distributions could be formed: a single and double headed spirals, a crystal lattice, and chaos. Yet, if a significant number of permanent barriers were distributed across the landscape, only the chaos distribution persisted. Given long enough time, no physical barrier is permanent, so “what kind of spatially stable distributions occur in a landscape with temporal barriers to dispersal and persistence?” Introducing evolutionary events (EE) to this type of models is not new. Past studies (Rangel et al., 2007; Gotelli et al., 2009) have done this by specifying the number of EEs in the simulation, or by including a ‘species’ measurement and observing species presence and how many EEs occurred at a set time. These studies, however, assume independence of evolutionary events (EE), making it difficult to test whether the empirical phylogenetic tree is linked to the actual distribution or the same tree could have arisen through any other distribution. Here, I introduce a new simulation framework that incorporates the effects of temporary barriers, a dispersal kernel, and evolutionary information from phylogenetic trees, so that EEs in the simulation are not independent. To showcase the power of the model, we used the African "Rand Flora" (RF) as a case study. The RF pattern describes plant clades that share a common disjunct distribution on the margins of the continent. These clades belong to different angiosperm families and exhibit different life-history and dispersal traits, but they share the tropical lowlands and the arid desert regions as common climatic barriers against population persistence and dispersal. The RF pattern has been addressed using phylogenetic comparative methods to discriminate between vicariance (geographic fragmentation) vs. long-distance dispersal (Sanmartín et al. 2010). Using our simulation framework, we expect to demonstrate that the RF pattern is significantly linked to the evolutionary events depicted in the phylogenetic trees, and that any other distribution, though possible, is unlikely because the pattern is replicated across multiple unrelated groups. Hassell et al. (1991). “Spatial structure and chaos in insect population dynamics, Letters to Nature Vol 353: pp255-258. Gotelli et al. (2009) “Patterns and causes of species richness: a general simulation model for macroecology”, Eco. Letters Vol 12: pp 873-886 Rangel et al. (2007) “Species Richness and Evolutionary Niche Dynamics: A Spatial Pattern-Oriented Simulation Experiment”, American Society of Naturalists Vol 170: pp 602-616 Sanmartín et al. (2010) “Bayesian Island Biogeography in a Continental Setting: The Rand Flora Case”, Bio. Letters Vol 6: pp 703-707.

Poster 10

> **Rémi Denise** [1,2,3]; Sophie Abby[4, 5]; Eduardo PC Rocha[1,2,3]
[1] Institut Pasteur, Paris, France;
[2] CNRS UMR3525, Paris, France;
[3] C3BI, Paris, France
[4] Université Grenoble Alpes, Grenoble, France;
[5] CNRS TIMC-IMAG F-38000, Grenoble, France

Co-option of complex molecular systems in bacterial membranes

Protein secretion systems are spread in many bacterial and archaeal species and are important for bacterial virulence. These systems are complex machineries made of many different proteins that interact together to allow other proteins to pass through the cell wall and be secreted outside of the cell. The proteins constituting these systems show various evolutionary rates and patterns of conservation within the secretion systems. Many of these systems were co-opted in complex evolutionary processes from other molecular structures. Biochemical, phylogenetic, and structural evidence show that the family of molecular machineries including type II secretion system (T2SS, involved in protein secretion), type IV pilus (T4P, involved in cell motility, adherence and virulence), Tad pilus (idem), the competence apparatus (Com, involved in natural transformation) and the archaeal flagellum (Archaellum, motility) share

homologous genes and have a similar genetic organization. We designed custom comparative genomic tools to detect and distinguish them in genome sequences, based on their particular components and genetic organization. This allows us to investigate the evolutionary origins of these machineries by phylogenetic and comparative genomics approaches, and thus to decipher some mechanisms of co-option involved in the diversification of microbial cellular machineries. Another goal would be the use of these phylogenetic analyses to facilitate the discrimination between related systems, and produce tools to perform the automatic annotation of an unknown system We have identified the key components of the systems of interest on a dataset of more than 5750 complete bacterial genomes. For each key component, I've established a phylogeny, and am now trying to reconcile them and understand the biological reasons of their discordance. Ultimately, I would like to put together the different phylogenies and the patterns of presence and absence of components to produce an integrated evolutionary model of the evolution of the systems. The systems analyzed are probably among the most complex network of molecular co-options analyzed to date and should provide an excellent basis to (i) infer the frequency of horizontal transfer of each type of derived molecular system, (ii) distinguish it from events of parallel co-option, and (iii) study the evolution of the genetic organization of the loci encoding these systems in the light of their evolutionary history.

Poster 11

> **Frantz Depaulis**[1]; Sandrine Adiba[1]; Olivier Chaillon[2]; Ludovic Orlando[3]. Tanawan Samleerat[2], Catherine Hannï[3], Francis Barin[2]
[1] Institut de Biologie de l'Ecole Normale Supérieur (IBENS), Paris, France
[2] Université François Rabelais, Tours, France

Time series in population genetics: from macro to micro evolution.

I will present three instances of time series analyses in population genetics on various time scales. In the first two cases (i) & (ii) mutations accumulate between sampling time points and provide the source of information. The largest of that two time scales (i) concerns ancient DNA data ([-130; -20 kyr]) from cave bear, a European extinct species [1]. I will show how some basic coalescent analyses (probabilistic gene genealogies) are affected by serial sampling over time. Lineages cannot have any common ancestor between their sampling time points, thus extending tip branches on the trees and the number of mutations private to a single individual. This pattern may be confused with demographic expansion or natural selection effects. The additional mutations also increase apparent population size and mutation rates. Difference of time sampling schemes between subsamples may also increase their apparent genetic distance. A much shorter time scale (ii) involves within host HIV evolution with applications on mother to child infection timing (1-2 years data) [2]. I will illustrate how estimates of infection timing are improved by the time series structure of the data. Although in absolute time units the latter time range (ii) is much shorter than the former (i), the virus mutation rates are much greater and their generation times are much shorter. As a consequence the two cases show similar levels of genetic variation. Finally, the last case (iii) involves three months of experimental evolution on Escherichia coli bacteria faced to an amoeba predator. Here the mutation rate of a selectively neutral marker is negligible. However, the frequency of a neutral variant varies as a function of the reproductive rate and ultimately, the population size and we derive such estimates. Finally, the genome scale shows a number of mutations that arise linked to one of the neutral variant. Sporadically, selection of mutations may occur and correlatively accelerates the frequency variations of the linked neutral variant. Such patterns are used to derive selective parameters. The naturally variable experimental conditions are likely to lead to time heterogeneous selective regime. We address how robust are more simple population genetics models to such conditions. The frequency trajectory approach provides most different results than more instantaneous and more direct ones from competition experiments. I shall discuss a number of reasons for such a discrepancy.

[1] Depaulis F, Orlando L, Hanni C 2009. Using classical population genetics tools with heterochroneous data: time matters! PLoS One 4: e5541.
[2] Chaillon A, Samleerat T, ……..,Depaulis F, Barin F 2014. Estimating the Timing of Mother-to-Child Transmission of the Human Immunodeficiency Virus Type 1 Using a Viral Molecular Evolution Model. PLoS ONE 9: e90421.

Poster 12

> **Linda Dib** [1,2], Xavier meyer [2,3], Daniele, Silvestro [2,3], David Gfeller [1,2], Nicolas Salamin [2,3]
[1] Ludwig Center for Cancer Research, Lausanne, Switzerland
[2] Swiss Institute of Bioinformatics, Quartier Sorge, Lausanne, Switzerland
[3] Department of Ecology and Evolution, University of Lausanne, 1015 Lausanne, Switzerland

A new model that considers dependant sites when infering phylogenies

Several evolutionary-based Markov models are now available to understand the evolution of protein coding sequences. Some describe the evolution of individual sites when others investigate codon evolution to predict positively selected sites. A recent study contributed in the development of evolutionary-based models by proposing the first Markov model that describes the evolution of mutually

evolving sites (coevolving sites) [Dib et al., 2014] . The diversity of evolutionary-based models makes it difficult to have a complete picture of the processes affecting the evolution of a protein. In this paper, I am proposing a new comprehensive framework that combines these Markov models to provide a unified picture of the processes that affect protein evolution. The model analyses both individually evolving and coevolving sites with no prior knowledge of coevolving sites number, rate of substitutions, or structural and functional protein constraints. Using this framework, the model corrects for branch lengths estimates that assume that sites evolve independently, and highlight sevolutionary constraint sites that maintain a protein function or structure. This project is novel and could not be realised earlier since it makes use of the recent advances in Bayesian statistics and the Markov model designed previously to study coevolving sites. Moreover, this unique framework that studies all the processes of evolution of a protein is essential to solve the protein structure-function-sequence paradigm and will reconcile both molecular evolution and molecular structure fields understanding of sites constraints. This comprehensive framework will be essential to unravel complex evolutionary dynamics across vertebrate species and characterise constrained sites that once mutated drive protein structure or functional failure. Dib L, Silvestro D, Salamin N. (2014). Evolutionary footprint of coevolving positions in genes. Bioinformatics, 30 (9): 1241-1249.

Poster 13

> **Tomochika Fujisawa**[1]; Amr Aswad[2]; Timothy Barraclough[2]
[1] Kyoto University, Kyoto, Japan
[2] Imperial College London, London, UK

A scalable multilocus species delimitation using Baysian model selection

Multilocus sequence data provide far greater power to resolve species limits than the single locus data typically used for broad surveys of clades. However, current statistical methods based on a multispecies coalescent framework are computationally demanding, because of the number of possible delimitations that must be compared and time-consuming likelihood calculations. New methods are therefore needed to open up the power of multilocus approaches to larger systematic surveys. Here, we present a rapid and scalable method that introduces 2 new innovations. First, the method reduces the complexity of likelihood calculations by decomposing the tree into rooted triplets. The distribution of topologies for a triplet across multiple loci has a uniform trinomial distribution when the 3 individuals belong to the same species, but a skewed distribution if they belong to separate species with a form that is specified by the multispecies coalescent.

A Bayesian model comparison framework was developed and the best delimitation found by comparing the product of posterior probabilities of all triplets. The second innovation is a new dynamic programming algorithm for finding the optimum delimitation from all those compatible with a guide tree by successively analyzing subtrees defined by each node. This algorithm removes the need for heuristic searches used by current methods, and guarantees that the best solution is found and potentially could be used in other systematic applications. We assessed the performance of the method with simulated, published, and newly generated data. Analyses of simulated data demonstrate that the combined method has favorable statistical properties and scalability with increasing sample sizes. Analyses of empirical data from both eukaryotes and prokaryotes demonstrate its potential for delimiting species in real cases. [Bacterial species, Bayesian model comparison, Dynamic programming, Multilocus species delimitation

Poster 14

> **Andreas Futschik** [1], Sonja Zehetmayer [2]
[1] Dept. of Applied Statistics, JKU Linz,
[2] Dept. of Medical Statistics, Medical Univ. Vienna

Statistical Tests for Genomic Time Series Data

Genomic time series data arise for instance in experimental evolution, where several populations of fast reproducing organisms, such as yeast or Drosophila, are kept under different environmental conditions. The goal is to find genomic signatures of adaptive selection. Recent experiments revealed however, that different experimental populations do not necessarily show the same genomic signatures of adaptation. This may be for instance due to the random loss of beneficial variants due to genetic drift, due to initial linkage with deleterious alleles, or to different possibilities for adaptation. We investigate different possibilities for identifying selection under such circumstances, and propose new powerful tests against the null hypothesis of neutrality across all replicates.

Poster 15

> **Elsa Guillot** [1][2], Marc Robison-Réchavi[1][2], Jerôme Goudet[1][2]
[1] Département Évolution et Écologie (DEE), Université de Lausanne, Suisse
[2] Swiss Institute of Bioinformatics (SIB) Université de Lausanne, Suisse

Selection in human, the lost signal

While population genetics permit to detect signal of selection between populations, molecular evolution techniques unravel positive selection between distant species. Bridging the gap between these two fields, this study aims at finding new signal of selection on the human branch of the evolutionary trees. Based on previous approaches [1] we develop a new computational method, to detect selection between close species such as Humans and Chimps.

[1] The distribution of fitness effects of new mutations. Adam Eyre-Walker, Peter D. Keightley Nat Rev Genet. 2007 Aug; 8(8): 610–618

Poster 16

> Lucas Michon[1], Marie Sémon[2], **Laurent Guéguen**[3]
[1] INSA Lyon, Villeurbanne, France;
[2] LBMC, ENS Lyon, Lyon, France;
[3] LBBE, Univ. Lyon 1, Lyon, France

Modeling and analysing the evolution of gene expression time-series
Since gene expression data are broadly available, it is possible perform inter-specific gene expression comparisons. Beyond such comparisons, the evolutive study of gene expression gains in popularity. However, to reach a better understanding of the biological processes, it is necessary to consider gene expression in a more realistic way, which means as a temporal and dynamic feature. This is markedly important in studying development : during organ development, the expression of a gene can vary a lot, and the morphological differences between homologous organs of different species will depend on this temporal dynamics. Models of evolution of gene expression are similar to -- and inspired from -- models of evolution of quantitative traits in ecology, as they have been studied since at least 20 years. However, studying the evolution of gene expression time-series requires other methods. An expression time-series in a species is not only a succession of expression levels, but is also characterized by specific time points that depend on the specific developmental rate of the organ. So, for each gene, it is necessary to recover homologous relations between time points in different species. Using an approach similar to "dynamic time warping", we propose a method to recover these relations while comparing globally time-series, taking into account the phylogenetic background, and we evaluate gene specifically the relevance of this approach. Then, we extend the usual modeling of evolution of quantitative traits to model of the evolution of gene expression time-series.

Poster 17

> **Philipp Hermann**, Andreas Futschik
Department of Applied Statistics, Linz, Austria

Estimating Variable Recombination Rates from Population Genetic Data

Recombination is a natural process in meiosis which increases genetic variation by producing new haplotypes. Populations with higher recombination rates are seen to be more flexible to adapt to new environments. Recombination rates are heterogeneous between species and also across the DNA sequence of a population. Large fractions of recombination events are concentrated on short intervals which are called hotspots [1]. Molecular and evolutionary mechanisms of the process of recombination can be better understood with accurate estimates of the recombination rate in different regions of the genome [2,3]. Precise knowledge of the recombination rate variation along the DNA sequence also improves inference from polymorphism data about e.g. positive selection [5]. Two particularly popular methods for estimating the variable recombination rate as a function of the DNA position use Bayesian approaches. More specifically, a composite likelihood is used within a reversible jump MCMC framework. Two software packages, LDHat [2,4] and LDHelmet [3] are available for this purpose. An improvement of local recombination rate estimators can be obtained via optimizing the trade-off between bias and variance [6]. In our current work we estimate local recombination rates with relevant summary statistics as explanatory variables in a regression model. In order to estimate locally varying recombination rates we apply a frequentist segmentation algorithm with type I error control [7]. A bias correction improves the quality of fit of our estimates and we compare our method in terms of false positive as well as true positive rates and the RMSE with existing software packages. We will also present an application of our method on the human genome. References

[1] Baudat, F., Imai, Y., and de Massy, B. (2013). Meiotic recombination in mammals: localization and regulation. Nature reviews. Genetics, 14(11):794-806.
[2] McVean, G. A. T., Myers, S. R., Hunt, S., Deloukas, P., Bentley, D. R., and Donnelly, P. (2004). The fine-scale structure of recombination rate variation in the human genome. Science, 304(5670):581-584.
[3] Chan, A. H., Jenkins, P. A., and Song, Y. S. (2012). Genome-Wide Fine-Scale Recombination Rate Variation in Drosophila melanogaster. PLoS Genetics, 8(12):e1003090.
[4] Auton, A. and McVean, G. (2007). Recombination rate estimation in the presence of hotspots. Genome Research, 17(8):1219–1227.
[5] Sabeti, P. C., Schaffner, S. F., Fry, B., Lohmueller, J., Varilly, P., Shamovsky, O., Palma, A., Mikkelsen, T. S., Altshuler, D., and Lander, E. S. (2006). Positive natural selection in the human lineage. Science (New York, N.Y.), 312(5780):1614-20.
[6] Gärtner, K. and Futschik, A. (2016). Improved Versions of Common Estimators of the Recombination Rate. Journal of

Computational Biology, 23(9):756-768.
[7] Futschik, A., Hotz, T., Munk, A., and Sieling, H. (2014). Multiscale DNA partitioning: Statistical evidence for segments. Bioinformatics, 30(16):2255–2262.

Poster 18

> **Valentin Hivert**[1] ;Mathieu Gautier[1];Renaud Vitalis[1]
[1]Centre de Biologie pour la Gestion des Populations (CBGP), Montpellier, France

A hierarchical bayesian model for measuring the extent of local adaptation from haplotype data.

The recent advent of high throughput sequencing and genotyping technologies (Next Generation Sequencing, NGS) enables the comparison of patterns of polymorphisms at a very large number of markers, which makes it possible to characterize genomic regions involved in the adaptation of organisms to their environment. Here, we present some recent developments to SelEstim (Vitalis & al., 2014), a hierarchical bayesian model that identifies and measures genomic signatures of selection from gene frequency data. In particular, we extend the model to analyse multi-allelic markers.This allows to use haplotype data, defined by means of unsupervised classification methods (Scheet & Stephens 2006), and considering haplotype blocks as multi-allelic markers. We expect this approach, which accounts for linkage disequilibrium information across individual markers, to be more powerful than those that consider independent SNPs. We will illustrate our results with some analyses conducted on simulated data, comparing the information brought by haplotype data relatively to analyses using SNP data alone. We will also discuss some potential extensions of this method, that would allow to test for correlations between haplotype frequencies and environmental variables, which could ultimately be used to predict the potential of evolution for some particular populations (e.g., in the context of invasion dynamics). References : Vitalis, R., M. Gautier, K.J. Dawson, and M.A. Beaumont, 2014 Detecting and Measuring Selection from Gene Frequency Data. Genetics 196 : 799-817. Scheet, P., and M. Stephens 2006 A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. American journal of human genetics 78 : 629-644

Poster 19

> ***Sota Ishikawa****[1],[2],[3], Tomochika Fujisawa[1],[3], François Chevenet[3], Wataru Iwasaki[2], Olivier Gascuel[1],[3]*
[1] Unité Bioinformatique Evolutive, C3BI - USR 3756 Institut Pasteur et CNRS, Paris, FRANCE;
[2] Graduate School of Science, the University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, 113-0033, Tokyo, Japan;
[3] Méthodes et Algorithmes pour la Bioinformatique (MAB), IBC et LIRMM, UMR 5506 – CNRS et Université de Montpellier, France

A probabilistic model-based prediction of virus character evolution on large phylogenies

In this study we focus on the phylogenetic approach to trace the origin and evolution of virus epidemics, by combining large virus trees with extrinsic characters (e.g. geographic location, risk group, presence of a given resistance mutation). The PhyloType software (Chevenet et al. Bioinformatics 2013) is one of the solutions as it combines ancestral character reconstruction (ACR) using maximum parsimony (MP), with numerical criteria to select sequence clusters with statistical significance. The MP version of PhyloType is remarkably fast and scales well on extremely large virus phylogenies. However, it lacks of robustness, because of the binary nature of MP. To solve the problem, we developed a probabilistic, model-based approach and implemented the two standard maximum-likelihood (ML) method of ACR, based on joint and marginal posterior probabilities of the character states at every tree node. Both methods are accurate, but the ‘joint’ method proposes a unique prediction (even when several have similar probabilities), and the ‘marginal’ method does not make a decision (even when some of the states clearly emerge while the others have low probabilities). Thus, we implemented a new approach to provide predictions in between these two extremes. This method starts with the marginal predictions and simplifies these by removing the most unlikely character states step-by-step, until one reaches the one-character-per-node prediction (= joint prediction). Various criteria are used to select the most appropriate solution encountered along this path. The method is still fast, with O(n^2) time complexity where n is the number of tree tips, thus making it possible to deal with ten thousands tips in reasonable computing time. In this poster presentation, we will show the performance of our new ACR program based on the results from simulated datasets in a variety of conditions, including large numbers of tips and character states. We’ll also present the results and user-friendly graphics obtained with a large HIV dataset (3,036 tips, 14 geographical characters). Acknowledgement: This work is supported by the VIROGENESIS H2020 European Project (N° 634650).

Poster 20

> **Prabhav Kalaghatgi**[1]; Thomas Lengauer[1]
[1] Max-Planck-Institut für Informatik, Saarbrücken, Germany.

Designing efficient phylogenetic inference algorithms by exploiting the correspondence between phylogenies and minimum spanning trees.

In 1999, Atteson[1] showed that NJ is bad at inferring minimally balanced trees (minBal trees). In 2011, Choi et al.[2] introduced a minimum spanning tree (MST) based tree construction algorithm called CLG and showed that it inferred minBal trees with high accuracy. In order for CLG to be consistent, the MST must share the following correspondence (condition 1) with the phylogeny. For each edge of the MST, the split that is induced by the edge must be present in the phylogeny. We observed that condition 1 does not necessarily hold true if the distance graph has multiple MSTs. In order to ensure that condition 1 always holds true we introduced the notion of vertex-ranked MSTs. Vertex-ranked MSTs are constructed by breaking ties between equally weighted edges on the basis of the ranks of the the incident vertices. We note that it is important to carefully select the vertex ranking, especially if the distances are additive in clock-like minBal trees. We argue that the optimal vertex-ranked MST should have the minimum number of leaves. In recent unpublished work[3], we provide a O(n^2log n) algorithm for constructing a vertex-ranked MST with the minimum number of leaves (ML-VR-MST). While CLG may outperform NJ for minBal trees, the performance trend is reversed if the tree is a maximally balanced (maxBal) tree. We designed an algorithm called MJ that performs reasonably well for both type of trees. MJ is a agglmorative algorithm that performs the following steps iteratively. First we construct a ML-VR-MST M. Subsequently, we search for cherries in the neighborhood of the leaves of M. Once a pair of taxa is selected as a cherry, the vertices in M that correspond to the taxa pair are removed and a new vertex corresponding the hidden ancestor of the taxa is added. Distances from the newly introduced vertex to the remaining vertices are computed and a new ML-VR-MST is constructed using the updated distance matrix. The rewiring of the MST that takes place in each iteration, improves the reconstruction accuracy for maxBal trees. We simulated 64-taxa large maxBal and minBal trees and set branch lengths to 0.01 subs/site. Sequences of length 50 nt were evolved under a Jukes-Cantor model of substitution. We performed 1000 trials and for each internal edge in the model trees we computed edge recovery which is the proportion of the estimated trees that contained the edge. For maxBal and minBal trees the median edge recovery is 0.8 and 0.3 for NJ, 0.6 and 0.89 for CLG, and 0.82 and 0.72 for MJ, respectively.

References: [1] Atteson, K. 1999. The Performance of Neighbor-Joining Methods of Phylogenetic Reconstruction. Algorithmica, 25(2-3): 251-278.
[2] Choi, M. J., Tan, V. Y. F., Anandkumar, A., and Willsky, A. S. 2011. Learning Latent Tree Graphical Models. JMLR, 12: 1771–1812.
[3] Kalaghatgi, P., and Lengauer, T. Selecting optimal minimum spanning trees that share a topological correspondence with phylogenetic trees. arXiv:1701.02844v1 [math.CO].

Poster 21

> **Frederic Lemoine** [1], Jean-Baka Domelevo-Entfellner [2], Eduan Wilkinson [3], Olivier Gascuel [1,4]
[1] Unité Bioinformatique Evolutive, C3BI, USR 3756 Institut Pasteur et CNRS, Paris, France,
[2] Univ. Western Cap,
[3] Univ. of Durban,
[4] IBC – LIRMM, UMR5506, Université de Montpellier & CNRS

Boosting Felsenstein's Bootstrap

Felsenstein's article describing the application of the bootstrap principle to evolutionary trees is one of the 100 most cited papers of all time. This statistical method based on resampling and replications, is used in an impressive number of studies to assess the robustness of phylogenetic inferences. However, with the ever growing size of the sequence datasets available today, its usefulness is clearly lowered, as it poorly supports the tree branches, especially the deep ones that generally are the main focus of evolutionary studies. We propose a revised version of Felsenstein's bootstrap, where the presence of a branch in replications is measured using a refined "transfer" distance, as opposed to the original version using a binary presence/absence index. The resulting transfer bootstrap supports are higher than Felsenstein's bootstrap supports, but do not induce falsely supported branches, as shown using simulated data. Our method is applied to large Mammals and HIV datasets. We show that it reveals the main signal contained in these data, especially regarding deep branches. Moreover, our new branch support is easily interpreted, and we provide computational tools to obtain this support from standard data.

Poster 22

> **Ivan Levkivskyi**, Anna-Sapfo Malaspinas
Institute of Ecology and Evolution, Biology, Bern, Switzerland

Multidimensional scaling (MDS) analysis, spectral decomposition and coalescent theory

Population structure plays an important role in determining the evolutionary history of a group. In recent years, the unprecedented increase in sequencing data has opened up a wide range of possibilities to

investigate population histories – provided one can handle such large amounts of data. Methods based on non-parametric multidimensional statistics (more specifically principal components analysis, PCA) were first applied to genetic data more than 30 years ago. PCA has since become a standard tool in population genetics owing in particular to the low computational demand of such analyses. In this work, we investigate a related statistical approach, namely multidimensional scaling (MDS). Following a recent study by McVean [1], we first derive analytical results for an arbitrary number of populations that relate the Euclidean distances between points on the MDS plots with pairwise coalescent times under simple demographic scenarios. We implement a related method that allows to compute demographic parameters under such scenarios. We explore analytically and with simulation the sensitivity of specific pairwise genetic distances to sequencing error and study the rate at which simulated data converges towards theoretical predictions as a function of the number of SNPs. Finally, we apply our method to a recently published whole genome dataset [2] and further characterize the human colonization of Australia.

[1] McVean, G., 2009. A Genealogical Interpretation of Principal Components Analysis. PLoS Genet 5, e1000686. doi:10.1371/journal.pgen.1000686

[2] Malaspinas A-S, Westaway MC, Muller C, Sousa VC, Lao O, Alves I, Bergström A, Athanasiadis G, Cheng JY, Crawford JE, et al. 2016. A genomic history of Aboriginal Australia. Nature 538:207–214.

Poster 23

> **Benjamin Linard**[1], Krister Swenson[1], Fabio Pardi[1]
LIRMM - equipe MAB, Batiment 5 - 860 rue de St Priest

Rapid phylogenetic placement through k-mer positionning estimated through ancestral reconstruction

Metagenomic projects, wheter they are related to environmental communities or medical diagnostics are constantly scaling up. In such approaches, the limited content of the database often limits the read association to specific marker genes and clades which are genomically well-known, leaving a large proportion of the metagenome to the state of "dark matter". With the democratization of metagenomic approaches, the dark matter fraction being more and more seen as a source of novel genomes, often partially prokaryotic but more and more regarded as a complex metavirome [1]. Few tools are adapted to the exploration of this "dark matter", and while k-mer classification can produce interesting results [2], it is generally explored through local alignement to very large references databases (mostly, large Blast searhes to the NCBI) [3-4]. Still, while aligning millions of reads though Blast can be seen as an "easy" approach, it remains a limiting step in many labs, as it will require huge computations and blast alignements between distant sequences remains limited. An alternative remains in using phylogenetic placement, which benefits from [1] Virus Res. 2017 Feb 9. pii: S0168-1702(16)30801-2. doi: 10.1016/j.virusres.2017.02.002. [Epub ahead of print] Origins and challenges of viral dark matter. Krishnamurthy SR, Wang D. [2] Comput Struct Biotechnol J. 2017; 15: 48–55. Published online 2016 Dec 5. doi: 10.1016/j.csbj.2016.11.005 PMCID: PMC5148923 Bioinformatics strategies for taxonomy independent binning and visualization of sequences in shotgun metagenomics Karel Sedlar,∗ Kristyna Kupkova, and Ivo Provaznik [3] MEGAN Community Edition - Interactive Exploration and Analysis of Large-Scale Microbiome Sequencing Data. Huson DH, Beier S, Flade I, Górska A, El-Hadidi M, Mitra S, Ruscheweyh HJ, Tappu R. PLoS Comput Biol. 2016 Jun 21;12(6):e1004957. doi: 10.1371/journal.pcbi.1004957. PMID: 27327495 [4] Metagenome Skimming of Insect Specimen Pools: Potential for Comparative Genomics. Linard B, Crampton-Platt A, Gillett CP, Timmermans MJ, Vogler AP. Genome Biol Evol. 2015 May 14;7(6):1474-89. doi: 10.1093/gbe/evv086. PMID: 25979752

Poster 24

> **Guillaume Louvel**[1]; Eric Lewitus[1]; Hélène Morlon[1]; Hugues Roest Crollius[1]
[1] Institut de Biologie de l'Ecole Normale Supérieure

Genomic markers of species diversification in vertebrates

Evolutionary biology notably aims at two different goals: reconstructing past biological history, as well as explaining the mechanisms of evolution. These two topics are intertwined but may imply different scales, and different modes of study. For example, there are many ecological models to explain patterns of biodiversity or the formation of new species, for example the Metabolic Theory of Ecology (Brown et al. (2004)) or the Neutral Theory of Biodiversity (Hubbell (2001)). There are currently several hypotheses on the mechanisms of species divergence, from genomic incompatibilities to divergent selection pressures (Coyne and Orr (1998)) and several case studies already allowed to bring empirical support to some of these ideas in specific contexts. However, given the current availability of annotated full genomes for many non-model organisms sampling various branches of the vertebrate phylogenetic tree, it becomes possible to combine this genomic data with the patterns of speciation brought by the analysis of complete phylogenies (Lewitus and Morlon (2016)). Our work tries to implement this general overview of how genomic features are implied in species diversification. We first focus on gene duplications, mapping them on the species phylogenetic trees by different functional categories. We also tried to date each

duplications more precisely using dS (synonymous substitution rate) calculations, in order to obtain a finer estimation of the rate of gene duplication through time and lineages. As the estimated rates seem to show, this first method is sensitive to multiple factors (fast evolving branches, quality of gene alignments, etc) making it quite imprecise. We are currently working on improving these dating method, and aim towards more sophisticated models of gene evolution that could estimate duplication ages (either adapting existing models or developping one). We will start working on such a method in the following months and hope to present its results at the meetings in porquerolles, either as a talk or a poster. Our broader aim is to compare the duplication rates with diversification rates of taxons, and to question, quantify the role of gene duplication in evolution. Brown, James H., James F. Gillooly, Andrew P. Allen, Van M. Savage, and Geoffrey B. West. 2004. “TOWARD A METABOLIC THEORY OF ECOLOGY.” _Ecology_ 85 (7). Ecological Society of America: 1771–89. Coyne, J A, and H A Orr. 1998. “The evolutionary genetics of speciation.” _Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences_ 353 (1366). The Royal Society: 287–305. Hubbell, Stephen P. 2001. _The unified neutral theory of biodiversity and biogeography (MPB-32)_. Vol. 32. Princeton University Press. Lewitus, Eric, and Hélène Morlon. 2016. “Natural Constraints to Species Diversification.” Edited by Anthony D. Barnosky. _PLOS Biology_ 14 (8). Public Library of Science: e1002532.

Poster 25

> **Marc Manceau**[1,2], Amaury Lambert[1,3], Hélène Morlon[2]
[1] Center for Interdisciplinary Research in Biology (CIRB), Paris, France;
[2] Institut de Biologie de l'Ecole Normale Supérieure (IBENS), Paris, France;
[3] Laboratoire de Probabilités et Modèles Aléatoires de l'Université Pierre et Marie Curie (LPMA), Paris, France

Modeling molecular evolution with fast adaptive divergence at speciation events

Models of molecular evolution have been described and used since the 60's to build phylogenies using molecular sequence data. Starting from a 'strict clock' hypothesis assuming constant molecular rates, the models have later been expanded to account for variations in rates both among loci and throughout the phylogenetic tree. However, despite the widespread idea that fast molecular divergence occurs at speciation events, for example as a result of disruptive adaptation, such effects have not yet been incorporated in models of molecular evolution. Here, we model episodes of fast molecular divergence happening at speciation events in a birth-death model of cladogenesis. We derive the likelihood of this model and analyze the signature that rapid adaptive events leave on molecular data at the macroevolutionary time-scale. Our model could be used to scan genes likely to play an important role in the speciation process.

Poster 26

> **Sebastian Matuszewski**[1]; Marcel Hildebrandt [1]; Jeffrey D. Jensen [1]
[1] École Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland

Coalescent Processes with Skewed Offspring Distributions and Non-equilibrium Demography

Many population genetic statistics and subsequent inference are based on the Kingman coalescent and the Wright–Fisher (WF) model. These methods are robust to some violations of WF model assumptions such as constant population size and random mating, and have been extended to accommodate selection and population structure. However, it has been suggested that violations of the assumption of a small variance in offspring number in the WF model lead to erroneous inference of population genetic parameters when analyzed under the Kingman coalescent [1]. A more general coalescent class of models, so-called multiple merger coalescent (MMC) models, can account for these violations, particularly for skewed offspring distributions, by allowing more than two lineages to coalesce at a time. These models thus provide a better fit to observed patterns of genetic variation in a range of organisms, particularly in marine species such as sardines [2]. Coalescent trees under the MMC tend to show more pronounced star-like genealogies with longer branches, similar to the effects of recent population growth, and their site frequency spectra (SFS) are skewed toward an excess of low-frequency and high-frequency variants [1]. While the two classes of models can be distinguished by the weight of the right tail of the SFS [3], models that account for both skewed offspring distributions and population growth are largely missing. Here, we develop an extended Moran model with exponential population growth, and show that the underlying ancestral process converges to a time-inhomogeneous psi-coalescent. However, by applying a non-linear change of time scale – analogous to the Kingman coalescent – we show that the ancestral process can be rescaled to its time-homogeneous analog, allowing the process to be simulated quickly and efficiently. Furthermore, based on a recently developed approach [4], we develop a pseudo-likelihood framework for the joint estimation of the coalescent and growth parameters, and show that both can be estimated accurately if the genealogy reflects the expected topology of the ancestral process. However, variance in tree topology, reflective of the random ancestry, complicates precise demography estimation. In particular,

when reproductive skew is large, the variance of the estimator prohibits a reliable estimation of the underlying demography.

[1] Eldon B and Wakeley J. Genetics, 172(4):2621–2633, 2006.
[2] Niwa HS, Nashida K, and Yanagimoto T. ICES Journal of Marine Science: Journal du Conseil, 73(9):2181-2189, 2016.
[3] Eldon B, Birkner M, Blath J, and Freund F. Genetics, 199(3):841–856, 2015.
[4] Spence J, Kamm J, and Song Y. Genetics, 202(4):1549–1561, 2016.

Poster 27

> **Cornelia Metzig**[1], Caroline Colijn[1]
[1]Imperial College London, Department of Mathematics

Non-Neutrality in Phylogenies – A Simulation Study on Dynamic Networks

Understanding whether and how transmission patterns are revealed by branching patterns in phylogenetic trees for pathogens remains a challenging research question. Besides the diversification of the pathogen, branching pat- terns depend strongly on the host contact structure as it shapes opportunities for the pathogen to reproduce. A range of characteristics have been used to summarise phylogenetic branching patterns, including for example the number of small substructures (cherries, pitchforks) in the trees, measures of tree imbal- ance (Sackin index, Colless index), and features derived from network science (diameter, closeness). These capture the shapes of phylogenetic trees. In ad- dition, in a tree branching “neutrally”, every contemporaneous branch has the same probability to branch text. We introduce a statistic to test this neutral property. We investigate the link between different host contact network hypotheses and these features of simulated phylogenetic trees (of a pathogen spreading on the network), with the aim of identifying which measures are most informative about the underlying network. We use a novel dynamic contact network model exhibiting realistic features of social networks (short average pathlength, skewed degree distribution, clus- tering and positive assortiveness) to simulate pathogen transmission and re- sulting pathogen phylogenetic trees. The network’s skewed degree distribution can be derived theoretically. The model follows the general idea of preferential attachment like the Barabasi-Albert model, but whereas the BA model grows indefinitely, our network has stationary size, since it includes both entry and exit of people (and of partnerships). We focus on the role of “population turnover”, i.e. the rate at which people enter and exit, which cannot be addressed using static network models, but is important in human populations. Turnover affects a variety of tree statistics, but it can be distinguished most clearly by the neutrality measure, especially for low turnover rates where other measures like the Colless index cannot discriminate between different rates. There is likely to be no universal most informative measure of phylogenetic branching that that best suits all parameter regions or underlying generative models, so a variety of statistics will be needed.

Poster 28

> **Raphaël Mourad**
Centre de Biologie Intégrative (CBI), Toulouse, France

Phylogenetic analysis of 3D genome

Introduction Chromosomal DNA is tightly packed in three-dimension (3D) such that a 2-meter long human genome can fit into a microscopic nucleus. Recent studies have revealed that such packing of DNA is not random but instead structured into topologically associating domains (TADs). Those TADs are essential to numerous key processes in the cell, and their disruption can lead to genetic diseases and cancers. TADs are stable across different cell types and highly conserved across species. A growing body of evidence supports the role of CTCF, a highly conserved protein in metazoans, in mediating TADs and long-range contacts. Comparative analysis further showed that CTCF motif position and orientation are conserved across species and that divergence of CTCF binding is correlated with divergence of internal 3D domain structure. Results Here we proposed to use phylogenetic analysis by ancestral character inference to determine the importance of CTCF motif turnover in shaping 3D genome and in regulating gene expression by long-range contacts. For this purpose, we used CTCF motif position frequency matrices from the JASPAR database and called CTCF binding sites over mammal genome assemblies including: hg19, gorGor3, rheMac3, panTro4, ponAbe2, bosTau7, equCab2, oryCun2, susScr2, mm10, canFam3, rn6. We then converted all coordinates to hg19 using liftOver. We focused on HoxD gene cluster that controls the body plan of an embryo along the cranio-caudal, and which is divided into two contiguous TADs by a border. Long-range contacts between gene promoters and regulatory elements are constrained within these TADs. Enhancers beyond the 5’ preferentially contact ’posterior’ genes, whereas those at the 3’ mostly interact with ’anterior’ genes. Analysis of the TAD border revealed that apes shared almost the same CTCF motif positions and orientations, except Pongo abelli. We also showed that several motifs were specific to the ape clade and were likely to be acquired during the emergence of the clade. Conclusion Using a TAD border example, we demonstrated that 3D genome can be analyzed by phylogenetic analysis in order to reconstruct ancestral 3D genome, helping to better understand the emergence of new clades driven by gain/loss of long-range contacts that regulate gene expression.

Further analyses should systematically study 3D genome phylogenetics over all conserved TAD border.

Poster 29

> **Jarosław Paszek**, Paweł Górecki
Faculty of Mathematics, Informatics and Mechanics, University of Warsaw, Poland

Inferring genomic duplication events

Discovering the location of gene duplications and multiple gene duplication episodes is a fundamental issue in evolutionary molecular biology. The phenomenon of whole-genome duplication (WGD) has crucial impact on the evolution of crops and is thoroughly studied. The main approaches of detecting WGDs are: synteny based (comparing locations of genes on chromosomes and their collinearity), Ks method (inferring from paralog distribution) and phylogenetic (probabilistic like [6]). We aim towards phylogenetic combinatorial optimization approach. The idea was introduced by Guigó et al. in 1996 [1] which is to map gene duplication events from a collection of rooted, binary gene family trees onto theirs corresponding rooted binary species tree in such a way that the total number of multiple gene duplication episodes is minimized. Existing approaches vary in the two fundamental aspects: the choice of evolutionary scenarios that model allowed locations of duplications in the species tree, and the rules of clustering gene duplications from gene trees into a single multiple duplication event [1-5]. Here we study the method of clustering called minimum episodes (see [4]) for several models of allowed evolutionary scenarios (see [1-3]) with a focus on interval models in which every gene duplication has an interval consisting of allowed locations in the species tree. We present mathematical foundations for general genomic duplication problems. Next, we propose the first linear time and space algorithm for minimum episodes clustering jointly for any interval model and the algorithm for the most general model (from [3]) in which every evolutionary scenario is allowed. We also present a comparative study of different models of genomic duplication based on simulated and empirical datasets. We provided algorithms and tools that could be applied to solve efficiently minimum episodes clustering problems. Our comparative study helps to identify which model is the most reasonable choice in inferring genomic duplication events.

[1] R. Guigo et al., “Reconstruction of ancient molecular phylogeny,” Mol Phylogenet Evol, vol. 6, no. 2, pp. 189–213, OCT 1996.
[2] J. Paszek and P. Gorecki, “Genomic duplication problems for unrooted gene trees,” BMC Genomics, vol. 17, no. 1, pp. 165–175, 2016.
[3] M. Fellows et al., “On the multiple gene duplication problem,” LNCS 1533, 1998, pp. 347–356.
[4] M. S. Bansal and O. Eulenstein, “The multiple gene duplication problem revisited,” Bioinformatics, vol. 24, no. 13, pp. i132–8, 2008.
[5] C.-W. Luo et al.,“Linear-time algorithms for the multiple gene duplication problems,” IEEE/ACM Trans Comput Biol Bioinform, vol. 8, no. 1, pp. 260–265, Jan. 2011.
[6] Rabier CE, Ta T, Ané C. Detecting and locating whole genome duplications on a phylogeny: a probabilistic approach. Molecular biology and evolution. 2014;31(3):750–62.

Poster 30

> **Swati Patel** [1]; Reinhard Burger [1]
[1] University of Vienna

The impacts of eco-evolutionary feedbacks on a two-locus three-species model

In recent years, there has been a lot of empirical evidence that feedbacks between ecological processes, such as community dynamics, and evolutionary processes, such as trait evolution, can drastically impact the qualitative dynamics of both processes. This evidence has spurred much theoretical investigation into when and how eco-evolutionary feedbacks alter dynamics. By investigating a two-locus model with linkage and recombination, we explore the role the genetic architecture of evolving traits may have on the impact of eco-evolutionary dynamics. In particular, we investigate an eco-evolutionary model of a predator evolving in a trait controlled by two potentially-linked loci, which determine its interactions with two independent prey species. Through analytical and computational analyses, we ask how do eco-evolutionary feedbacks affect the maintenance of genetic variation. Previous theoretical work on the pure evolutionary dynamics of this scenario has shown that both alleles at both loci are maintained for low recombination rates but not for intermediate to high recombination rates, since the latter prevents linkage disequilibrium. Here, we show that eco-evolutionary feedbacks enable the maintenance of genetic variation even for intermediate recombination rates. Interestingly, the coupling of ecological and evolutionary processes leads to cycling of population densities, allele frequencies, and linkage disequilibrium at these intermediate rates of recombination. Our work highlights that the aspects of the genetics of evolving traits play an important role in the overall effects of eco-evolutionary feedbacks.

Poster 31

> **Umberto Perron** [1]; Iain Moal[1]; Nick Goldman[1]
[1] European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Trust Campus, Hinxton, United Kingdom

Modelling structural constraints on protein evolution via side-chain rotamers

Very few models of sequence evolution have incorporated parameters that describe protein structure. We know, however, that structure is consistently more conserved than protein sequence [1] and that it is essential for protein functions. The increasing availability of structural information opens up the possibility of creating structurally aware substitution models. A promising candidate feature for this is the rotameric state of side-chains: each residue can adopt one or more of a discrete set of rotamers, which is a geometric pattern of atomic positions defined by the dihedral angles between covalently linked atoms. Rotamer states depend on backbone conformation and play a role in protein folding and docking mechanisms by determining side-chain positioning. Furthermore, they are a feature of the constituent residues of a sequence and therefore fit well with existing approaches to modelling sequence evolution. We are able to assign a rotamer state and its corresponding probability to each residue of proteins with known structure, starting from X-ray crystallography data on side chain atoms' position and using states defined in the Dunbrack rotamer library [2]. Here we present a survey of rotamer state conservation across a number of protein families, highlighting evolutionary patterns for these features. Our work aims ultimately to generate a structurally aware Dayhoff-like model employing an expanded state set (alphabet) composed of symbols containing both residue type and rotamer state information. This has the potential to improve our understanding of the relationships between protein sequence structure and evolution and aid in obtaining better tree topology and sequence divergence estimates, and may also allow reconstruction of ancestral rotameric states and potentially full modelling of ancestral structures given the observed structures at leaf nodes. Chotia C, Lesk AM. (1986) The relation between the divergence of sequence and structure in proteins. EMBO J., 5:823-826. Shapovalov MV, Dunbrack RL. (2011) A smoothed backbone-dependent rotamer library for proteins derived from adaptive kernel density estimates and regressions. Structure, 19:844-858.

Poster 32

> **Sylvain PULICANI**[1,2,3]; Krister SWENSON[1,2]; Eric RIVALS[1,2]; Giacomo CAVALLI[3]
[1] Laboratoire d'Informatique, de Robotique et de Microélectronique de Montpellier (LIRMM), Montpellier, FRANCE;
[2] Institut de Biologie Computationnelle (IBC), Montpellier, FRANCE;
[3] Institut de Génétique Humaine (IGH), Montpellier, FRANCE

Linking Large Scale Chromosomal Rearrangements to 3D Chromatin Structure in Drosophila

Genomic rearragements like inversion, duplication or transposition of a chromosomic segments are initiated by double-strand break in the DNA, delimiting the involved segment. Inversions are known in Drosophila since 1910, and have been linked to aberrant phenotypes, e.g. female-only offspring. Rearrangements can prevent fertile inter-species cross or disturb genes regulation. We study the constraints that are linked to the apparition and the fixation of such rearrangements. In Eukaryotes, chromosomes are organized in complexes polymers named chromatin. The structure of the chromatin has been shown to be organized using technologies like Hi-C. This spatial organization has a functional importance, and is conserved along evolution. However, what is the influence of that structure on the position of the rearrangement breakpoints? To respond to that question, we compute evolutionary scenarios based on the chromosomes structure. We design algorithms that infer such scenarios, maximizing the proximity of involved breakpoints. We apply our work on Drosophila datasets.

Poster 33

> **Adam Rohrlach**[1][2]; Barbara Holland[3]; Nigel Bean[1][2]; Jonathan Tuke[1][2]
[1] The University of Adelaide, Adelaide, Australia;
[2] ARC Centre of Excellence for Mathematical and Statistical Frontiers, Melbourne, Australia;
[3] The University of Tasmania, Hobart, Australia

The Detection of Demographic Structure and Migration in mtDNA

The detection of demographic structure can be of key importance for assessing the modelling assumptions of a phylogenetic analysis, and may be the very question of interest when analyzing sequence data. Methods exist for the exploration of autosomal DNA, however, these methods can not be applied to non-autosomal DNA, such as mtDNA and Y-chromosomal DNA. In this talk I will introduce a method for calculating dissimilarity distances between individuals in an alignment based on chi-square distances via Multiple Correspondence Analysis. Using these distances I will show that individuals can be tested for correlations with quantitative supplementary variables, such as geographic location, through an updated Mantel test. Finally I will describe a method for detecting possible migration routes for individuals

in an alignment. To illustrate the power of these methods I will analyze the mtDNA of Aboriginal Australians obtained from hair samples with provenance data predating European arrival in Australia.

Poster 34

> **Dominik Schrempf**[1,2,3]; Asger Hobolth[4]; Bui Quang Minh[5]; Arndt von Haeseler[5,6]; Carolin Kosiol[1,2]
[1] School of Biology, University of St Andrews, United Kingdom;
[2] Institut für Populationsgenetik, Vetmeduni Vienna, Austria;
[3] Vienna Graduate School of Population Genetics, Vienna, Austria;
[4] Bioinformatics Research Center, Aarhus University, Denmark;
[5] Center for Integrative Bioinformatics Vienna, Max F. Perutz Laboratories, University of Vienna, Medical University of Vienna, Austria;
[6] Bioinformatics and Computational Biology, University of Vienna, Austria

Discrete multivariate boundary mutation models and their application to tree inference

The multivariate Wright-Fisher and Moran models are standard models in population genetics. However, analytical solutions are usually intractable and numerical treatment is cumbersome if population sizes are large. We present a simple and intuitive derivation of the stationary distribution of these models for general rate matrices under the assumption of low mutation rates (Schrempf and Hobolth 2017). The derivation is based on three key ingredients. First, the decoupled Moran model is used to describe genetic drift. Second, mutations are limited to monomorphic states (boundary mutation). Third, the rate matrix is separated into a time-reversible part and a part that describes the probability flux around closed loops in the allelic state space. Besides theoretical advantages, our result is a valuable tool for inference on population data because the expected sampling distribution approximately matches the stationary distribution. In particular, we investigate the application of our model to a Polymorphism-aware phylogenetic Model (PoMo, De Maio et al., 2015; Schrempf et al., 2016). Thereby we combine population genetic approaches with phylogenetics. PoMo allows for parameter estimation and species tree reconstruction for up to a hundred species with many individuals per species. The inference from small and large evolutionary time scales is combined by employing population data such as site frequency spectra. This is especially useful because vast amounts of genomic data from closely related species and from individuals within species are now available. The analysis of closely related species may provide clues towards understanding the dynamics of speciation. However, complex processes such as incomplete lineage sorting pose a challenge. Incomplete lineage sorting results in incongruences between gene trees and leads to biases in species tree reconstruction. The cause of these problems is ancestral variation which is captured by our model. Consequently, simulation studies show that PoMo has high accuracy in estimating tree topologies, divergence times and mutation rates for species whose evolutionary history exhibits incomplete lineage sorting. We apply PoMo to exome-wide alignments of population data from six Baboons populations. We demonstrate that PoMo is a valuable tool for dating speciation events and for large-scale and genome-wide species tree reconstruction using population data.
De Maio N, Schrempf D, and Kosiol C 2015. PoMo: An Allele Frequency-Based Approach for Species Tree Estimation. Syst. Biol. 64(6):1018–1031. Schrempf D and Hobolth A 2017. An alternative derivation of the stationary distribution of the multivariate neutral Wright-Fisher model for low mutation rates with a view towards mutation rate estimation from site frequency data. Theor. Popul. Biol., 114:88–94. Schrempf D et al. 2016. Reversible polymorphism-aware phylogenetic models and their application to tree inference. J. Theor. Biol., 407:362–370.

Poster 35

> **Daniah Tahir**[1]; Sylvain Glemin[2]; Martin Lascoux[1]; Ingemar Kaj[1]
[1] Uppsala University, Uppsala, Sweden;
[2] Institut des Sciences de l'Evolution de Montpellier, Montpellier, France.

Modeling Trait-Dependent Evolution on a Random Species Tree

We present a probabilistic modeling framework for binary trait random species trees, in which the number of species and their traits are given by a two-type, continuous time Markov branching process. In addition, depending on their trait, the species in the phylogenetic tree accumulate mutations over time. A typical area of application of this model lies within mating systems in plant species and in this work, we have considered outcrossing and selfing species as the underlying binary traits. We also carry out a study of dN/dS, the ratio of non-synonymous to synonymous substitutions. A methodology is introduced which enables us to match model parameters with parameter estimates from phylogenetic tree data. The properties obtained from the model, applied to outcrossing and selfing species trees in the Geraniaceae and Solanaceae family, allow us to investigate not only the branching tree rates, but also the mutation rates and the intensity of selection.
References: [1] Athreya K.B., Ney P.E. 1972. Branching processes. New York: Springer Verlag Berlin Heidelberg.
[2] Galtier N., Gouy M., Gautier C. 1996. SEAVIEW and PHYLO-WIN: two graphic tools for sequence alignment and molecular

phylogeny. Comput. Appl. Biosci. 12:543–548.
[3] Glemin S., Muyle A. 2014. Mating systems and selection efficacy: a test using chloroplastic sequence data in angiosperms. J. Evol. Biol. 27:1386–1399.
[4] Goldberg E.E., Kohn J.R., Lande R., Robertson K.A., Smith S.A., Igic B. 2010. Species selection maintains self-incompatibility. Science 330:493–495.
[5] Igic B., Bohs L., Kohn J.R. 2003. Historical inferences from the self-incompatibility locus. New Phytol. 161:97–105.
[6] Yang Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. Comput. Appl. Biosci. 13:555–556.

Poster 36

> Magnus Bordewich[1]; Charles Semple[2]; **Nihan Tokac**[3]
[1] Durham University, Durham, UK;
[2] University of Canterbury, Christchurch, New Zealand;
[3] Yildiz Technical University, Istanbul, Turkey

Constructing tree-child networks from distance matrices

we show that a tree-child network on x with a positive real-valued weighting of its edges is essentially determined by the path-length distances between elements in x.

Poster 37

> **Meike J. Wittmann**[1]; Sylvain Mousset[1]; Joachim Hermisson[1]
[1] Mathematics and BioSciences Group, University of Vienna, Vienna, Austria

Stable polymorphisms due to seasonally fluctuating selection and their genetic footprint

For organisms with several generations per year, seasonally fluctuating selection can be a powerful mechanism to maintain genetic polymorphism. For example, an allele favored during summer may stably coexist with an allele favored during winter, a form of balancing selection. Despite intense debate over decades, it is still unclear how much of the variation observed in the genomes of natural populations is due to balancing selection. In recent years, evolutionary biologists have started scanning genomes for genetic footprints of balancing selection (e.g. regions of increased diversity and positive Tajima's D). However, these scans have generally assumed the simplest form of balancing selection where alleles are maintained at constant frequencies over time. There is currently insufficient theory to tell us what genetic footprint to expect under seasonally fluctuating selection, and how to distinguish it from neutrality but also from other forms of balancing selection. Here we use coalescent theory and stochastic simulations to characterize the genetic footprint of seasonally fluctuating selection acting at one locus or multiple linked loci. We find that seasonally fluctuating selection generally leads to an increase in genetic diversity close to the selected site, but reduces diversity further away from the selected site, often even leading to a net decrease in diversity on the scale of the chromosome. In addition, fluctuations affect haplotype structure and site-frequency spectra. For two or more linked loci under seasonally fluctuating selection, qualitatively different behaviors emerge depending on whether the genetic distances between loci are below or above a certain threshold distance. Above the threshold distance, recombination breaks up linkage disequilibria faster than selection generates them. Thus all loci are in complete linkage equilibrium and fluctuate in the same way as in the single-locus case. If loci are closer together than the threshold distance, recombination is not strong enough and linkage disequilibrium builds up. Depending on the initial conditions, the population reaches one of two or more alternatively stable attractors with different patterns of linkage disequilibrium, and qualitatively different seasonal dynamics and genetic footprints. These results might contribute to explaining empirical observations of long-range linkage disequilibria in natural populations.

ATTENDEES

ALLMAN	Elizabeth	Dept of Mathematics and Statistics, University of Alaska Fairbanks,USA
ARBISSER	Ilana	Stanford University, USA
AYABINA	Diepreye	Imperial College London, UK
BAELE	Guy	Rega Institute / KU Leuven - Evolutionary and Computational Virology Section, BE
BARIDO-SOTTANI	Joëlle	ETH Zurich, CH
BASTIDE	Paul	UMR MIA-Paris, AgroParisTech, INRA, Université Paris-Saclay, Paris, FR
BASTKOWSKI	Sarah	The Earlham Institute, Norwich, UK
BECHELER	Arnaud	IRD-CNRS, FR
BEERLI	Peter	Florida State University, USA
BOITARD	Simon	INRA, GenPhySE, Toulouse, FR
CLEMENTE	Florian	Institut de Biologie Computationnelle, FR
COLIJN	Caroline	Imperial College London, UK
CROTTY	Stephen	University of Vienna, AT
CULSHAW	Victoria	Real Jardín Botánico, Madrid, CSIC, SP
DE MAIO	Nicola	University of Oxford, UK
DE VIENNE	Damien	Laboratoire de Biométrie et Biologie Evolutive (LBBE), Lyon, FR
DENISE	Rémi	Microbial Evolutionary Genomics, Institut Pasteur and CNRS, UMR3525, FR
DEPAULIS	Frantz	Ecole Normale Supérieure, FR
DIB	Linda	Université de Lausanne, CH
DRAY	Stéphane	Laboratoire de Biométrie et Biologie Evolutive (LBBE), Lyon, FR
FREUND	Fabian	Institute of Plant Breeding, Seed Science and Population Genetics, Univ. of Hohenheim, DE
FUJISAWA	Tomochika	Kyoto University, JP
FUTSCHIK	Andreas	Johannes Kepler University Linz, AT
GASCUEL	Olivier	C3BI USR 3756 Institut Pasteur – CNRS, FR
GAVRYUSHKIN	Alex	ETH Zürich, CH
GAVRYUSHKINA	Alexandra	ETH Zurich, CH
GUEGUEN	Laurent	LBBE - Université Lyon 1, FR
GUILLOT	Elsa	DEE, Université de Lausanne, CH
GUINDON	Stéphane	CNRS, FR
HERMANN	Philipp	Johannes Kepler University Linz, AU
HIVERT	Valentin	INRA UMR CBGP, FR
HOLLAND	Barbara	Theoretical Phylogenetics Group, School of Mathematics and Physics, Univ. of Tasmania, AU
ISHIKAWA	Sohta	Institut Pasteur, FR
JOHN	Sona	Section of Population Genetics, Technische Universität München, DE
KALAGHATGI	Prabhav	Max Planck Institute for Informatics, DE
LEBLOIS	Raphael	INRA, FR
LEMOINE	Frederic	Institut Pasteur, FR
LEVKIVSKYI	Ivan	Institute of Ecology and Evolution, Biology, Bern, CH
LINARD	Benjamin	LIRMM-CNRS, FR
LOUVEL	Guillaume	Institut de Biologie de l'École Normale Supérieure, FR

MALASPINAS Anna-sapfo Institute of Ecology and Evolution, University of Bern, CH
MANCEAU Marc Collège de France, FR
MATIAS Catherine CNRS - Univ. Pierre et Marie Curie, FR
MATUSZEWSKI Sebastian EPFL Lausanne, CH
MAYROSE Itay Tel Aviv University, IS
MEHTA Rohan Stanford University, USA
MERLE Coralie Université de Montpellier, FR
METZIG Cornelia Imperial College London, UK
MOSLONKA Mathieu Institut Pasteur, FR
MOURAD Raphaël IBCG, Toulouse, FR
MUGAL Carina Uppsala University, SW
PARDI Fabio LIRMM – CNRS, FR
PASZEK Jaroslaw University of Warsaw, PL
PATEL Swati University of Vienna, AT
PERRON Umberto EMBL-EBI, UK
PULICANI Sylvain IBC – LIRMM, FR
RHODES John University of Alaska Fairbanks, USA
RODRIGUEZ Willy INRA - Le Moulon, FR
ROHRLACH Adam The University of Adelaide, AU
SAULNIER Emma UM, IBC, MIVEGEC, FR
SCHREMPF Dominik School of Biology, University of St Andrews, UK
SCORNAVACCA Céline ISE-M, CNRS, FR
SONG Yun Calabi-Simons Chair in Mathematics and Biology, Dep. of Math. and Biol., Univ. of Pennsylvania, USA
SUCHARD Marc David Geffen School of Medicine at UCLA, Dep. of Biomath., Biostat. and Human Genetics, USA
SWENSON Krister LIRMM, CNRS, FR
TAHIR Daniah Department of Mathematics, Uppsala University, SW
THOMPSON Robin University of Oxford, UK
TOKAC Nihan Yildiz Technical University, TR
TRUSZKOWSKI Jakub EMBL-EBI and Cancer Research UK, University of Cambridge, UK
VOLZ Erik Imperial College London, UK
WANG Liangliang Simon Fraser University, DE
WELLER Mathias IBC, LIRMM, FR
WITTMANN Meike University of Vienna, Mathematics and BioSciences Group, AU
ZHUKOVA Anna Institut Pasteur, FR