June 15-19, 2014

MATHEMATICAL AND COMPUTATIONAL EVOLUTIONARY BIOLOGY

Labex NUMEV

Institut de biologie computationnelle

CeMEB

GdR BiM

UNIVERSITÉ MONTPELLIER 2 SCIENCES ET TECHNIQUES

Institut des Sciences de l'Évolution-Montpellier

la Région Languedoc Roussillon

Laboratoire Informatique Robotique Microélectronique Montpellier

cnrs

INFORMATIONS

Meeting Point

Bus station, Parking du grand Saint-Jean
(Close to the train station and served by the tram)
A « Bancarel » bus will leave Montpellier on Monday at **16H30**.

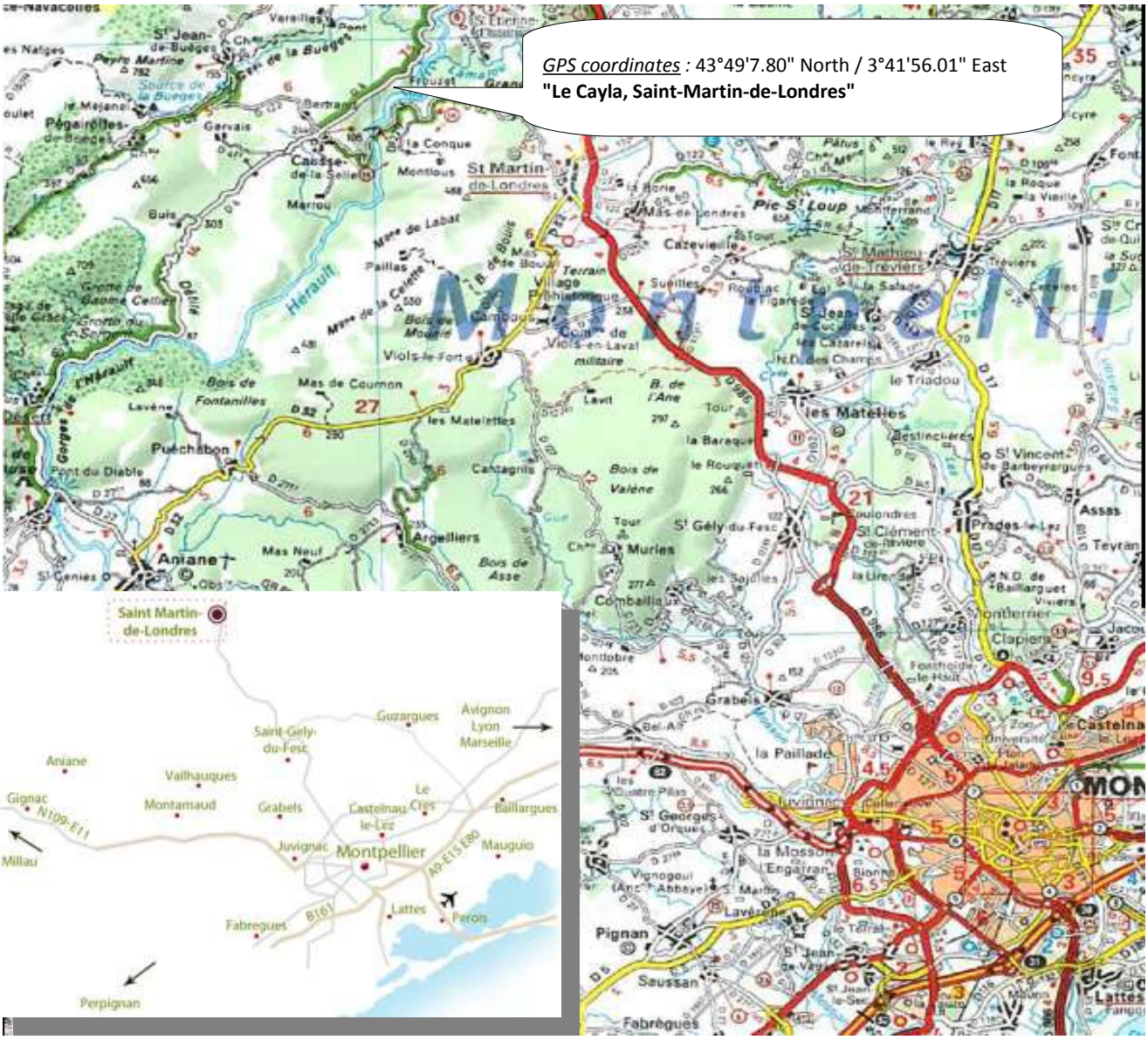
In case of problems :

Olivier Gascuel : 33 (0) 06 17 80 37 57

Sylvain Milanesi : 33 (0) 06 74 13 05 58

Location

The conference will be held at the ***Hameau de l'Etoile***,
a hamlet dedicated to seminars and conferences,
located at about 25 km north of Montpellier (south of France).



Practical informations

Domaine Le Hameau de l'Etoile
Route de Frouzet
34380 ST-MARTIN-DE-LONDRES
Tél (+33) **04 67 55 75 73**
Fax (+33) 04 67 55 09 10

Taxi :

Taxi de St Martin de Londres
Call first « Mr Bernard Valette » at **06 81 16 93 75**
Rates : In week, tram station Saint-Roch Train Station = 55 € / Airport=70 €
Week - end / Night : + 20 euros

Hotels in Montpellier :

Hôtel d'Aragon *** 10, rue Baudin 34000 MONTPELLIER Tél : 33 (0)4 67 10 70 00 fax : 33 (0)4 67 10 70 01	Hôtel d'Angleterre ** 7, rue Maguelone 34000 MONTPELLIER Tél : 33 (0)4 67 58 59 50 fax : 33 (0)4 67 58 29 52	Hôtel le Mistral ** 25, rue Boussairolles 34000 MONTPELLIER Tél : 33 (0)4 67 58 45 25 / 33 (0)6 60 53 73 40 fax : 33 (0)4 67 58 23 95
Hôtel Le Guilhem *** 18, rue Jean Jacques Rousseau 34000 MONTPELLIER Tél : 33 (0)4 67 52 90 90 fax : 33 (0)4 67 60 67 67	Hôtel des Arceaux ** 33/35, boulevard des Arceaux 34000 MONTPELLIER Tél : 33 (0)4 67 92 03 03 fax : 33 (0)4 67 92 05 09	Hôtel Nova ** 8, rue Richelieu 34000 MONTPELLIER Tél : 33 (0)4 67 60 79 85 fax : 33 (0)4 67 60 89 06
Newhotel du Midi *** 22, boulevard Victor Hugo 34000 MONTPELLIER Tél : 33 (0)4 67 92 69 61 fax : 33 (0)4 67 92 73 63	Hôtel des Arts ** 6, boulevard Victor Hugo 34000 MONTPELLIER Tél : 33 (0)4 67 58 69 20 fax : 33 (0)4 67 58 85 82	Hôtel du Palais ** 3, rue du Palais 34000 MONTPELLIER Tél : 33 (0)4 67 60 47 38 fax : 33 (0)4 67 60 40 23
Royal Hotel *** 8, rue Maguelone 34000 MONTPELLIER Tél : 33 (0)4 67 92 13 36 fax : 33 (0)4 67 92 59 80	Hôtel Colisée Verdun ** 33, rue de Verdun 34000 MONTPELLIER Tél : 33 (0)4 67 58 42 63 fax : 33 (0)4 67 58 98 27	Hôtel du Parc ** 8, rue Achille Bégé 34000 MONTPELLIER Tél : 33 (0)4 67 41 16 49 fax : 33 (0)4 67 54 10 05
Hôtel Acapulco ** 445, rue Auguste Broussonnet 34090 MONTPELLIER Tél : 33 (0)4 67 54 12 21 fax : 33 (0)4 67 52 26 10	Hôtel François de Lapeyronie ** 80, rue des Pétètes 34090 MONTPELLIER Tél : 33 (0)4 67 52 52 20 fax : 33 (0)4 67 63 56 65	Hôtel Les Troenes ** 17, avenue Emile Bertin Sans 34040 MONTPELLIER Tél : 33 (0)4 67 04 07 76 / 33 (0)6 29 02 31 17 fax : 33 (0)4 67 61 04 43
Hôtel Les Alizés ** 14, rue Jules Ferry 34000 MONTPELLIER Tél : 33 (0)4 67 12 85 35 fax : 33 (0)4 67 12 85 30	Hôtel Littoral ** 3, Impasse Saint Sauveur 34000 MONTPELLIER Tél : 33 (0)4 67 92 28 10 fax : 33 (0)4 67 92 72 20	Hôtel les Fauvettes * 8, rue Bonnard 34000 MONTPELLIER Tél : 33 (0)4 67 63 17 60 / 33 (0)6 89 26 63 58

PROGRAM

Sunday, June 15

> 16h30 : Bus from Montpellier (see « Meeting point » - p2)

> 20:00 : Apéritif

> 20h30 : Dinner

Monday, June 16

> 09h30 – 11h00 : *KEYNOTE* *p.18*

Rasmus Nielsen (University of California, Berkeley, US)
« Theory and methods developments in Population Genetics »

> 11h : Coffee break

> 11h30 – 13h00 : 4x 20 min *TALKS* (including questions)

Simon BOITARD (Laboratoire TIMC-IMAG, FR) *p.10*
« Inferring the past dynamics of effective population size using genome wide molecular data : an ABC approach »

Sophie LÈBRE (Laboratoire ICube, Université de Strasbourg, FR) *p.12*
« Deriving the degrees of freedom in the genetic code of overlapping genes »

Laurent GUÉGUEN (Lab. de Biométrie et Biologie Evolutive (LBBE), Villeurbanne, FR) *p.11*
« Using stochastic mapping to estimate non-homogeneous models »

Stephan SCHIFFELS (Auckland University of Technology,NZ) *p.14*
« Inferring human population size and separation history from multiple genome sequences »

> 13h00 - 14h30 : Lunch & Swimmingpool

> 14h30 – 15h00 : 1x20 min *TALK* (including questions)

Steffen KLAERE (Laboratoire TIMC-IMAG, FR) *p.12*
« Phylogenetic Inference with real confidence »

> 15h00 – 16h30 : *KEYNOTE* *p.8*

Arne MOOERS (Simon Fraser University, CA)
« Considering the Tree of Life in Conservation »

> 19h00 : *POSTERS* *p.16*
Vin, apéritif et discussions

> 20h30 : Dinner

Tuesday, June 17

> 09h30 – 11h00 : *KEYNOTE* *p.9*

Adam SIEPEL (Cornell University, Ithaca, US)
« Genome-wide inference of ancestral recombination graphs »

> 11h00 : Coffee break

> 11h30 – 13h00 : *KEYNOTE* *p.8*

Hélène MORLON (Ecole Polytechnique, FR)
« Macroevolution; macroecology; biodiversity research »

> 13h00 : Lunch

> 14h00 - 20h30: Free afternoon (hiking, canoe, etc.)

> 20h30 : Dinner

Wenesday, June 18

> 09h30 – 11h00 : *KEYNOTE* *p.7*

Daniel HUSON (Center for Bioinformatics ZBIT, Tuebingen University, DE)
« Environmental metagenomics »

> 11h00 : Coffee break

> 11h30 – 13h00 : 4x 20 min *TALKS* (including questions)

Catherine MATIAS (Laboratoire Statistique et Génome , FR) *p.12*
« Co-phylogeny reconstruction via an approximate Bayesian computation »

Sebastian NOVAK (Institute of Science and Technology Austria, Klosterneuburg,AT) *p.13*
« Dispersal Evolution: Bridging the gap between the two H's »

Srdjan SARIKAS (Institute of Science and Technology Austria, Klosterneuburg,AT) *p.14*
« Path ensembles in population genetics »

Frédéric DELSUC (Institut des Sciences de l'Evolution, Montpellier, FR) *p.10*
« Exploring topological incongruence for detecting contaminations in phylogenomic datasets »

> 13h00 - 14h30 : Lunch & Swimmingpool

> 14h30 – 15h00 : 1x20 min *TALK* (including questions)

Fabio PARDI (LIRMM, Institut de Biologie Computationelle, Montpellier, FR) *p.12*
« Identifiability of phylogenetic networks: do not distinguish the indistinguishable »

> 15h00 – 16h30 : *KEYNOTE* *p.7*

Rampal ETIENNE (University of Groningen, NL)
« Evolutionary community ecology »

19h00 : *POSTERS* *p.16*
Vin, apéritif et discussions

> 20h30 : Dinner

Thursday, June 19

> 09h30 – 11h00 : *KEYNOTE* *p.9*

Mike STEEL (University of Canterbury, NZ)
« What can we reconstruct from the past? »

> 11h00 : Coffee break

> 11h30 – 13h00 : 4x 20 min *TALKS* (including questions)

Priya MOORJANI (Columbia University, US) *p.12*
« Estimating the generation time in human evolution »

Eric TANNIER (INRIA, LBBE, Université de Lyon 1, FR) *p.15*
« Ancient and ancestral genomes »

Louis DU PLESSIS (ETH Zürich, CH) *p.10*
« Detecting diversity dependent speciation using a birth-death model in BEAST 2 »

Renaud VITALIS (Institut des Sciences de l'Evolution, Montpellier, FR) *p.15*
« Detecting and measuring selection from gene frequency data »

> 13h00 - 14h30 : Lunch & Swimmingpool

> 14h30 – 16h00 : *KEYNOTE* *p.7*

Nicolas LARTILLOT (Lab. de Biométrie et Biologie Evolutive (LBBE), Villeurbanne, FR)
« The molecular comparative method: Bayesian integrative models of macro-evolutionary processes »

> 16h00 : Bus to Montpellier (train station ~17:30, airport ~18:00)

KEYNOTE SPEAKERS

> Rampal Etienne
University of Groningen, NL

Evolutionary community ecology
Species assemblages or ecological communities are collections of species that occur together in a particular habitat. There is often an amazing diversity of species in such communities which makes one wonder how such communities come about, and how they are maintained, and what determines what species are present.
In this lecture I will give a brief overview of the development of the field of community ecology in the seventies and eighties, to move on to more modern views on community assembly that take into account the biogeographic and macroevolutionary history of ecologicla communities, including the neutral theory of biodiversity and biogeography and phylogenetic community ecology. I will end with two ongoing projects that I am involved in: a dynamic model for phylogenetic community ecology and a macro-evolutionary theory of island biogeography.

> Daniel Huson
Center for Bioinformatics (ZBIT), Department of Computer Science, Tuebingen University, DE

Environmental metagenomics
Metagenomics, the study of uncultured microbes using DNA sequencing, adds a whole new dimension to many different fields of research in the life sciences, from ecology to medicine. There is currently much interest in using this type of approach (including metatranscriptomics and other meta- analyses, as well).
In the first part of this talk, we give an introduction to the field of metagenomics, describe the types of problems that are studied and discuss a number of key papers in the area. High-throughput sequencing requires high-through analysis methods to enable researchers to explore metagenomic data in detail. In the second part of the talk, we present recent work on one of the computationally most challenging steps in metagenome analysis, the alignment of reads against a protein reference database such as NCBI-NR. A typical problem size is to align 5 billion DNA reads against 30 million protein reference sequences. Our new tool called DIAMOND (Buchfink, Xie & Huson, under review) solves this problem, offering a 16,000 fold speedup over BLASTX.In the final part of the talk, we present MetaScope (Huson, Buchfink & Xie, in preparation), our winning entry to the one-million dollar DTRA challenge “Identifying Organisms from a Stream of DNA” in 2013.

> Nicolas Lartillot
Laboratoire de Biologie et Biométrie Évolutive, Lyon, FR

The molecular comparative method: Bayesian integrative models of macro-evolutionary processes
Estimating divergence times, understanding molecular evolutionary mechanisms, or testing macroevolutionary hypotheses about patterns of diversification and morphological evolution, are usually considered as separate research questions, adressed by distinct, although overlapping, scientific communities. Yet, many connections would deserve to be made between these various topics in evolutionary sciences. Over the last years, several attempts at integrating some of these various aspects of macro-evolutionary sciences have been made, using hierarchical modeling approaches. After reviewing them, I will more specifically present a Bayesian framework for modeling the macroevolutionary process in an integrative manner. This framework can be seen as a fusion between the classical comparative method and methods for divergence time estimation. Taking as an input a multiple sequence alignment, data about life-history traits of extant species, and fossil calibrations, it then jointly estimates divergence times, life-history evolution and correlations between substitution patterns and quantitative traits. Application of the method to placental mammals reveals extensive correlations between life-history and molecular evolution, providing stimulating observations for testing macroevolutionary hypotheses.

> Arne Mooers
Simon Fraser University, CA

Considering the Tree of Life in Conservation
Phylogenetic information has informed conservation prioritization for decades. For example, the US Endangered Species Act Regulations from 1983 suggests higher funding priority be given to endangered monotypic species than endangered subspecies. This notion of maximizing the sum of the edge weights on a tree when choosing taxa for conservation attention was formalized by Vane-Wright et al. (1991) and Faith (1992). The max-sum approach can be extended to include the probabilities of persistence of leaves (see, e.g. Witting & Loeschcke 1995) and the costs of their conservation (see, e.g., Weitzman, 1998), which together is often considered the "Noah's Ark Problem", and to network representations of diversity (Minh et al, 2009). The max-sum approach has also been approximated using leaf-level metrics of average marginal gain (see, e.g., Redding et al., 2008).
It may be useful to think more broadly about the remit of conservation phylogenetics. In biodiversity science and in ecology, the diversity of a set of taxa (e.g. of a local community) is often characterized by its richness, by measures of average divergence and by the evenness of those divergences (see, e.g., Mason et al., 2005). Pavoine and Bonsall (2011) have projected these dimensions onto a phylogenetic framework by considering a sum (e.g. of edge weights connecting a set of leaves), a mean (e.g. of pair-wise patristic distances), and a variance (e.g. of pair-wise patristic distances) on a tree. While the max-sum approach in conservation focuses on the first measure, it might be interesting to explore when the other two might be relevant.
Importantly, all these perspectives assume that edge weights (and so measures of patristic distances) map onto conservation-relevant variation. Different models of how this variation is captured by the edge weights may change both the utility of a phylogenetic approach to conservation and what approaches are preferred (see, e.g., Bordewich et al., 2008). Indeed, the field likely needs robust, field-tested models of trait evolution in order to advance much further.

> Hélène Morlon
Ecole polytechnique, FR

Macroevolution; macroecology; biodiversity research
Why are some groups of species richer than others? Why are some regions on the planet richer than others? And what explains phenotypic diversity across taxonomic groups, ecological communities, and geographic regions? In this lecture i will give a brief overview of how phylogenetic comparative methods can help us to answer these keys questions in ecology and evolutionary biology. I will review the existing models that can be used to evaluate support for various diversification scenarios and estimate rates of speciation and extinction. I will then present results of one of the first attempt to compute likelihoods of phylogenetic trees for an individual-based model of diversification, inspired from the neutral theory of biodiversity and biogeography. I will end by illustrating how the use of these phylogenetic comparative methods on various empirical systems can help us to understand the various extrinsic and intrinsic factors that influence how species and morphological diversity are distributed in space and time.

> Rasmus Nielsen
University of California, Berkeley, US

Theory and methods developments in Population Genetics
I will give a general overview of methods for analyzing population genomic data. Full likelihood functions are usually are intractable, and researchers are therefore often instead using approximate methods, including composite likelihood methods and Approximate Bayesian Computation methods, when analyzing large population genomic data sets. An alternative approaches is to use data reduction, i.e. focusing on just a few markers with special properties such as low recombination rates. Finally, many methods rely on approximate models, where the population genetic model is modified in one way or another. This often involves enforcing a Markov condition on the process generating ancestral recombination graphs.
In the second part of the talk I will discuss joint work with my student Mason Liang on the properties of admixture proportions and admixture fragment lengths in a population. We show that

admixture fragments often are far from iid exponentially distributed as otherwise assumed in all previous inferential models. We also obtain analytical expressions for the distribution of admixture proportions within a population for some simple models, which can be used to estimate admixture times.

> Adam Siepel
University of California, Santa Cruz, US

Genome-wide inference of ancestral recombination graphs

I will discuss recent progress by my research group on the long-standing problem of inferring an "ancestral recombination graph" (ARG) from sequence data. The ARG provides a complete characterization of the correlation structure of a collection of sequences sampled from a population, and, in principle, fast, high-quality ARG inference could enable many improvements in population genomic analysis. However, the available methods for ARG inference are either extremely computationally intensive, depend on fairly crude approximations, or are limited to very small numbers of samples, and, as a consequence, they are rarely used in applied population genomics. I will present a new method for ARG inference, called ARGweaver, that is efficient enough to be applied on the scale of dozens of complete mammalian genomes. Experiments with simulated data indicate that ARGweaver converges rapidly to the true posterior distribution and is effective in recovering various features of the ARG, for twenty or more megabase-long sequences generated under realistic parameters for human populations. We have begun to apply our methods to high-coverage individual human genome sequences from Complete Genomics, and I will show that signatures of selective sweeps, background selection, recombination hot spots, and other features are all evident from properties of the inferred ARGs.

> Mike Steel
University of Canterbury, UK

What can we reconstruct from the past?

The famous physicist Niels Bohr once quipped "prediction is very difficult, especially about the future". Yet even predicting the pass can be hard. It is a general property of any irreducible Markov process that the information that present observations provide about a past state decays exponentially with time. This sets fundamental limits on the ability to confidently infer phylogenetic branching events and ancestral states in the distant past. These limits hold for any inference method (though it is also instructive to also compare different reconstruction methods to see how they perform relative to each other). Biological data is also subject to processes that can obscure this already weakened tree-like signal, such as lateral gene transfer or lineage sorting. In this talk, I offer an overview of the mathematical aspects of information loss in evolution. This, in turn, provides hints as to how to improve accuracy, for example by using more (or occasionally less) data, different types of data, or alternative inference methods. Moreover, we will see that mixtures of Markov processes (such as models of DNA evolution that allow certain distributions of rates across sites) can escape the curse of the exponential loss of information with time. And some recent results are presented that show how simple stochastic models of lateral gene transfer allow a remarkable amount of transfer events before the tree signal is lost.

TALKS

> Simon Boitard [1,2]
[1] UMR 7205 OSEB (EPHE - MNHN - CNRS), Paris, France.
[2] UMR 1313 GABI (INRA - AgroParisTech), Jouy En Josas, France.

Inferring the past dynamics of effective population size using genome wide molecular data : an ABC approach

Effective population size is a key notion in population genetics and is often used as a summary of population genetic diversity. Inferring the effective size of a population, and its eventual expansions or reductions in the past, from genetic data, also provides important knowledge about the demographic history of this population. Until a few years, methods allowing to infer past effective population size were designed for data sets including a small number of independent markers or non recombining DNA sequences. However, the spectacular progress of genotyping and sequencing technologies during the last decade has enabled the production of high density genome wide data in many species, so new statistical methods are needed to take benefit of this new type of data. Promising approaches have recently been proposed (Li and Durbin, 2011; Song et al, 2013), but so far they are still limited to small sample sizes or small DNA segments. Due to the complexity and the high dimension of the mathematical models related to this question, improving these methods will necessarely involve approximations of the models and intensive numerical computations.

In this study I present an Approximate Bayesian Computaion (ABC) approach for inferring the past effective size of a single population. This approach is based on coalescent simulations and on the use of a large number of summary statistics related to allele frequencies and linkage disequilibrium. I illustrate the performance of this approach using cross validation, comparing several estimation strategies and discussing the influence of the different summary statistics. I finally apply this method to a set of 25 bovine sequences from the Holstein breed and compare the results with those obtained by the Pairwise Sequentially Markovian Coalescent approach of Li and Durbin (2011).

> Frédéric Delsuc [1]; Khalid Belkhir [1]; Céline Scornavacca [1]
[1] Institut Des Sciences De L'Evolution, CNRS Université Montpellier 2, France

Exploring topological incongruence for detecting contaminations in phylogenomic datasets

Molecular phylogenetic studies now routinely rely on multigene data sets that are often automatically constructed using bioinformatics pipelines. This process generally involves gathering sequences from public databases to combine with newly generated data for the taxonomic group under focus. Such a procedure is error-prone because of misidentified sequences in the databases, but also frequent contaminations in next-generation sequencing data, leading to phylogenetic artifacts. Here, we propose computational solutions aimed at avoiding these problems by exploring topological incongruence based on the distribution of bipartitions among gene trees and multigene concatenations. We show that this allows the reliable identification of mislabeled sequences and contaminations, which can negatively impact phylogenetic inference as exemplified by a case study on placental mammal phylogenetics

> Louis Du Plessis [1,2]; Tanja Stadler [2];
[1] Theoretical Biology, Institute Of Integrative Biology, Department Of Environmental Systems Science, ETH Zürich, Zürich Switzerland;
[2] Computational Evolution, Department Of Biosystems Science And Evolution, ETH Zürich, Basel, Switzerland;

Detecting diversity dependent speciation using a birth-death model in BEAST 2

The amazing biodiversity of life on earth is thanks to the constant evolution of new species. Although diversification is a constant process, the rates at which new species arise and others die out are far from constant. Speciation and extinction rates are affected by external factors, such as climate change and habitat restructuring, and also by events such as adaptive radiations, triggered by a key innovation or colonization of a new environment [1]. This results in many available niches, which leads to an initial increase in the speciation rate. As the available niches are filled by novel

species the speciation rate subsequently slows down again. We model diversification as a birth-death process where speciations correspond to births and extinctions to deaths. The model we use is an extension of the model described by Stadler [2]. In addition to allowing for shifts in speciation and extinction rates through time [2], we allow for an explicit dependence between the speciation rate and the number of species. This diversity dependence is achieved by bounding the clade-level carrying capacity, which is identical to limiting the number of niches that new species can colonize. Furthermore, the model also makes provision for changes to the carrying capacity over time. We implement the model in the BEAST 2 framework. Diversity dependence is implemented in a similar fashion to density dependence in the epidemiological model Leventhal et al. [3] used to describe SIS-dynamics. By decoupling the intrinsic speciation rate and the clade-level carrying capacity we can investigate if diversity dependence explains observed shifts in the speciation rate. Moreover, identifying shifts in the rates and the number of niches allows us to determine times during which adaptive radiations or mass extinctions occurred. Furthermore, as the method is implemented in BEAST 2, it can be used as a tree-prior in conjunction with a model of molecular evolution to directly infer new phylogenies. Currently, shifts in the model parameters can only occur across the whole phylogeny. However, it is a simple extension to look for shifts in only some subclades, which will allow us to test for radiations within a particular subclade. Furthermore, fossil data can also easily be incorporated in the model to distinguish between shifts in the speciation rate and mass extinctions.

References:
[1] Etienne RS, Haegeman B, "A conceptual and statistical framework for adaptive radiations with a key role for diversity dependence", The American Naturalist 180(4), (2012), pp E75–E89.
[2] Stadler T, "Mammalian phylogeny reveals recent diversification rate shifts", PNAS, 108(15), (2011), pp 6187–6192.
[3] Leventhal GE, Günthard HF, Bonhoeffer S, Stadler T, "Using an epidemiological model for phylogenetic inference reveals density dependence in HIV transmission", MBE 31(1), (2014), pp 6–17.

> Laurent Guéguen
Lab. De Biométrie Et Biologie Evolutive (LBBE), Villeurbanne, France

Using stochastic mapping to estimate non-homogeneous models

Probabilistic modelling is a powerful way to study molecular evolution, and many efficient statistical approaches exist to infer models on an evolutionary process along a tree, given a set of sequences. However, when the evolutionary process is heterogeneous among branches, models are difficult to infer, mostly because of the large number of parameters involved. Stochastic mapping is a way to tackle this difficulty by computing some relevant features of the process on the branches, such as the count of substitutions of a given type [1, 4, 5]. However, counts of stochastic mapping depend on the composition of ancestral sequences, and up to now this method has been used as an intermediate tool to provide data augmentation [3, 6] or to cluster branches, in bayesian statistics or maximum likelihood approaches [2, 7].

I propose a way to take explicitly into account the composition of ancestral sequences in the counts of stochastic mappings, to compute directly and independently on all the branches informative features such as the expected dN or dS, or the GC-equilibrium frequency of the process. The goal is to use mapping to infer directly models and parameters in the context of full heterogeneous modelling.

References:
[1] F. Ball and R.K. Milne. Simple derivations of properties of counting processes associated wih markov renewal processes. J. Appl. Prob., 42(4):1031-1043, 2005.
[2] Julien Y. Dutheil, Nicolas Galtier, Jonathan Romiguier, Emmanuel J.P. Douzery, Vincent Ranwez, and Bastien Boussau. Efficient selection of branch-specific models of sequence evolution. Molecular Biology and Evolution, 29(7):1861-1874, 2012.
[3] I. Holmes and G.M. Rubin. An expectation maximization algorithm for training hidden substitution models. J. Mol. Biol., 317:753-764, 2002.
[4] V.N. Minin and M.A. Suchard. Fast, accurate and simulation-free stochastic mapping. Phil. Trans. Roy. Soc. B, 363:3985-3995, 2008.
[5] R. Nielsen. Mapping mutations on phylogenies. Syst. Biol., 51(5):729-739, 2002.
[6] N. Rodrigue, H. Philippe, and N. Lartillot. Uniformization for sampling realizations of Markov processses: applications to bayesian implementations of codon substitution models. Bioinformatics, 24(1):56-62, 2008.
[7] Jonathan Romiguier, Emeric Figuet, Nicolas Galtier, Emmanuel J. P. Douzery, Bastien Boussau, Julien Y. Dutheil, and Vincent Ranwez. Fast and robust characterization of time-heterogeneous sequence evolutionary processes using substitution mapping. Plos One, 7:1-10, 2012.

> Steffen Klaere [1], Barbara Holland [2], Michael Charleston [3], Stephane Guindon [1]
[1] Department Of Statistics, University Of Auckland, NZ
[2] Department Of Mathematics And Physics, University Of Tasmania, Hobart, AU
[3] Department Of Computer Science, University Of Sydne, AU

Phylogenetic inference with real confidence

There is an abundance of statistical methods for phylogenetic inference, but only few of these answer the biologists' most pressing question: Did I get the right tree? A successful test combines statistical accuracy with applicability. The tests most widely used nowadays have many shortcomings but remain popular because they are easily applied. In this talk we present novel statistically accurate tests to help biologists asking their most pressing question.

> Sophie Lèbre [1], Olivier Gascuel [2]
[1] Laboratoire ICube, Université De Strasbourg, France
[2] Institut de Biologie Computationnelle, LIRMM, CNRS – Univ. de Montpellier, France

Deriving the degrees of freedom in the genetic code of overlapping genes

In 1977 it was discovered that a single DNA sequence may code for several overlapping genes: one gene is coded in a reference reading frame, whereas a second gene is read with a frame shift of length 1 or 2 (frame +1 or +2). Overlapping genes may also be coded in the opposite sense on the complementary DNA strand, with three possible shifts (frame 0, -1 or -2). Overlapping genes were first found in non-viral species and were suggested to have multiple functions, such as regulation of gene expression, translational coupling and genome imprinting. It recently came out that the number of overlapping genes could be greater than expected, especially in the virus world, where they seem to be privileged therapeutic targets. Indeed, such double-coding DNA sequences should be highly conserved due to strong associated functional and structural constraints, which should prevent from the rapid adaptation of viruses and fast appearance of resistance mutations. Preliminary analyses of this ‘double coding’ were published in the 1980s. Basing on information theory, Smith and Waterman (1981) have shown that the five possible overlapping reading-frame configurations differ significantly in their coding flexibility and thus in their global information content. Applications to detect new overlapping genes remained limited but several routes for improvement do exist. Although the genetic code is degenerate (64 codon codes for 20 amino acids plus the stop signal), the case of overlapping genes is obviously over-constrained, and thus a central question is, how many degrees of freedom remain and what the constraints imply at the protein level? Our recent studies based on an algebraic approach allowed us to make explicit basic constraints, in terms of amino acid or polypeptide frequencies. Let us consider amino acids and denote by N1=([A1],[C1],...) and N2=([A2],[C2],...), the vectors giving the number of occurrences of the 20 amino acids and the stop signal, in the first and second reading frames respectively. Vector Q of size 4^4 gives the number of occurrences of ‘quadri-nucleotides’ or ‘quadons’ in the first reading frame. Each quadon codes exactly for one amino acid in each reading frame. This is written for any reading frame f in [-2,2] as N=Mf.Q, where N=(N1,N2) is the vector of amino acid and stop frequencies in both reading frames and Mf is the correspondence 42x256, 0/1 matrix. Mf gives a complete description of the deterministic constraints imposed by the genetic code, and allows for the derivation of the frequency constraints by solving a linear system. It turns out that the number and contents of constraints differ considerably among reading frames; for example, frame -2 has 10 linear constraints in coding sequences (space of dimension 40), such as [Ala1]=[Ala2] or [His1]+[Gln1]=[Cys2]+[Trp2]. These explicit constraints (at the amino-acid but also polypeptide levels) should help to improve the current methods to analyze and detect overlapping genes.

> C. Matias [4], C. Baudet [1][2], B. Donati [3], B. Sinaimeri [1][2]; P. Crescenzi[3], C. Gautier[1][2], M-F. Sagot [1][2]
[1] INRIA Grenoble Rhône-Alpes, 38330 Montbonnot Saint-Martin, France;
[2] Université De Lyon, F-69000 Lyon; Université Lyon 1; CNRS, UMR5558, Laboratoire De Biométrie Et Biologie Evolutive, F-69622 Villeurbanne, France;
[3] Universita Di Firenze, Dipartimento Di Sistemi E Informatica, I-50134 Firenze, Italy;
[4] Laboratoire Statistique Et Génome, UMR CNRS 8071 & USC INRA, Université D’Evry, France;

Co-phylogeny reconstruction via an approximate Bayesian computation

Despite an increasingly vaster literature on co-phylogenetic reconstructions for studying host-parasite associations, understanding the common evolutionary history of such systems remains a problem that is far from being solved. Most algorithms for host-parasite reconciliation use an event-

based model, where the events include in general (a subset of) co-speciation, duplication, loss, and host-switch. All known event-based methods then assign a cost to each type of event in order to find a reconstruction of minimum cost. The main problem with this approach is that the cost of the events strongly influence the reconciliation obtained. To deal with this problem, we developed an algorithm, called Coala, for estimating the frequency of the events based on an approximate Bayesian computation approach.
The benefits of this method are twofold: (1) it provides more confidence in the set of costs to be used in a reconciliation, and (2) it allows to estimate the frequency of the events in cases where the dataset consists of trees with a large number of taxa. We evaluate our method on simulated and on real datasets. We show that in both cases, for a same pair of host and parasite trees, different sets of frequencies for the events constitute equally probable solutions. Moreover, sometimes these sets lead to different parsimonious optimal reconciliations, in the sense of presenting a different number of the events. For this reason, it appears crucial to take this into account before attempting any further biological interpretation of such reconciliations. More generally, we also show that the set of frequencies can vary widely depending on the input host and parasite trees. Indiscriminately applying a standard vector of costs may thus not be a good strategy.

> Priya Moorjani [1][4], Minyoung Wyman[1][4], Ziyue Gao[2], And Molly Przeworski [1][3]
[1] Department Of Biological Sciences, Columbia University, New York, USA;
[2] Committee On Genetics, Genomics And Systems Biology, University Of Chicago, Chicago, USA;
[3] Department Of Systems Biology, Columbia University, New York, USA;
[4] Contributed Equally

Estimating the generation time in human evolution
Recent next-generation sequencing studies in human pedigrees yields mutation rate per year that is significantly lower than the one obtained from phylogenetic methods, introducing substantial uncertainty about the chronology of human evolution. This discrepancy would be resolved if changes in the generation time had led to a rapid evolution of the mutation rate in hominoids. Here, we test this hypothesis by estimating the generation time in humans based on the insight that distinct types of mutations arise through different mechanisms and hence have different dependencies on generation time. Notably, transitions at CpG sites are thought to occur primarily through spontaneous deamination, implying that their mutation rate should depend largely on absolute time and thus should be relatively insensitive to the generation time. In contrast, replication-driven mutations at non-CpG sites occur in higher numbers following male puberty, and yearly mutation rates therefore reflect generation times. We characterize these different time dependencies from human pedigree data. By relating them to the observed average pairwise diversity values at CpG and non-CpG sites, we then obtain joint estimates of the mean coalescence time and the generation time in years. Our method provides a novel, population genetics based estimator of the generation time, applicable to any species with polymorphism data and mutations with different time dependencies. We illustrate the method by estimating the historical generation time for modern human populations as well as for archaic populations of Neandertal and Denisova and discuss the role of generation time in shaping human mutation rate.

> Sebastian Novak [1]
[1] Institute Of Science And Technology Austria, Klosterneuburg, Austria

Dispersal Evolution: Bridging the gap between the two H's.
Dispersal is a universal trait of any natural population that embeds species into their geographical environment. Understanding the evolution of dispersal is therefore crucial for understanding the dynamics of spatially structured populations. Two main driving forces of dispersal evolution have been identified: Spatio-temporal variability of the habitat and relatedness between individuals. In my talk, I establish a link between the effects of habitat heterogeneity and relatedness on the evolution of dispersal. I briefly introduce consequences of a heterogeneous habitat for dispersal evolution (dating back to Hastings, 1983). By identifying a class of evolutionarily stable dispersal strategies I show that the classical findings are reflected in my model. However, within this class of dispersal strategies, other forces come into play as higher-order effects. These constitute an alternative approach to finding effects of relatedness and thus create a link to the body of dispersal evolution literature based on renowned work by Hamilton and May (1977). As a side effect, a limitation to the use of continuous differential equations in modelling natural processes is illustrated.

> Fabio Pardi [1], Celine Scornavacca [2]
[1] *Institut de Biologie Computationnelle, LIRMM, CNRS – Univ. de Montpellier, France*
[2] Institut Des Sciences De L'Evolution, CNRS Université Montpellier 2, France

Identifiability of phylogenetic networks: do not distinguish the indistinguishable
Phylogenies are almost invariably represented as trees. Although in many cases this is reasonable, in many others phylogenies should be represented as networks (more precisely directed acyclic graphs). This is due to a number of biological phenomena collectively known as reticulation events, whereby a species or a gene inherits genetic material from more than one parent organism. This may be caused by events such as hybrid speciation, introgression, horizontal gene transfer, or recombination. Phylogenetic network inference methods are in their infancy, but they are almost invariably based on the following idea: the goodness of a candidate network is evaluated on the basis of how well the trees it contains fit the data. This poses a problem: different networks may contain exactly the same set of trees, meaning that these networks will be considered "indistinguishable" by most network inference methods, no matter the input data. We propose a novel definition of what constitutes a "uniquely reconstructible" network: for each class of indistinguishable networks, we define a canonical form. Under mild assumptions, the canonical form is unique. Given data coming from any phylogenetic network, only its canonical equivalent can be uniquely reconstructed. This is a fundamental limitation that implies a drastic reduction of the solution space in phylogenetic network inference

> Srdjan Sarikas, Harald Ringbauer, Nick Barton
Institute Of Science And Technology (IST) Austria, Am Campus 1, 3400 Klosterneuburg, Austria

Path ensembles in population genetics
The most straightforward way for investigating stochastic processes in population genetics is by direct simulations. Intuitive and bias-free as it may be, this approach often is not the most efficient. The diffusion approximation partially alleviates this deficiency, yet it also often requires further simplifications to yield analytic results, like restriction to asymptotic cases. Rather then considering ensembles of propagating stochastic particles, we propose their alternative description that stems from considering ensembles of paths, i.e. histories, the diffusing particle undergoes. Assigning a probability to each history, we allow population genetics to use ideas and results previously developed in physics and applied mathematics. For example, we can sample this path ensemble by the Markov Chain Monte Carlo methods or calculate the rate of rare processes using large deviation theory. We present the main ideas, the formalism and give simple examples of usefulness of this concept. The formalism seems to be more natural in some cases, e.g. for longitudinal studies and can be readily applied in cases that are difficult to solve in the diffusion approximation. The goal is to devise a new CPU-time efficient technique for parameter inference (population size, selection coefficient), applicable also for multidimensional problems involving recombination.

> Stephan Schiffels
Wellcome Trust Sanger Institute, UK

Inferring human population size and separation history from multiple genome sequences
The availability of complete human genome sequences from populations across the world has given rise to new population genetic inference methods that explicitly model their ancestral relationship under recombination and mutation. So far, application of these methods to evolutionary history more recent than 20-30 thousand years ago and to population separations has been limited. Here we present a new method that overcomes these shortcomings. The Multiple Sequentially Markovian Coalescent (MSMC) analyses the observed pattern of mutations in multiple individuals, focusing on the first coalescence between any two individuals. Results from applying MSMC to genome sequences from nine populations across the world suggest that the genetic separation of non-African ancestors from African Yoruban ancestors started long before 50,000 years ago, and give information about human population history as recently as 2,000 years ago, including the bottleneck in the peopling of the Americas, and separations within Africa, East Asia and Europe.

> Eric Tannier [1], Cedric Chauve [2]
[1] INRIA, LBBE, Université De Lyon 1, France;
[2] Simon Fraser University, Vancouver, Canada

Ancient and ancestral genomes

Paleogenomics, or the reconstitution of extinct species macromolecule sequences, can be understood in at least two different ways: ancient genome sequencing on one side, or computational prediction of an ancestral genome from the comparison of its descendants on the other. Ancient whole bacterial genomes have been released in the last years, while ancestral genome reconstruction methods are now able to propose relatively reliable solutions thanks to integrative evolutionary models. We propose to explore the ways these two approaches can feed each other. For example computational ancestral reconstructions can help the design of baits to capture ancient DNA better than a single extant genome. Or, ancient sequences can be used as a validation for comparative methods. For several bacterial clades for which a close ancien relative has been sequenced, we reconstruct the full genome sequences of all ancestors in a phylogeny. This includes the full organization in chromosomes and plasmids, and ancestral sequence at the nucleotide level. For this we mix phylogeny estimation, ancestral sequence reconstruction, and ancestral whole genome reconstructions at the organisational level. We compare the obtained sequences to the ancient sequenced reads and also use the predicted sequence to guide the ancient DNA capture. We also propose to assemble fragmented ancient genomes with a comparative method, using informations from several descendants or relatives. We derive insights into the pestis or vibrio recent molecular evolution.

> Renaud Vitalis [1,2], Mathieu Gautier [1,2], Kevin J Dawson [3], Mark A Beaumont [4]
[1] Centre De Biologie Pour La Gestion Des Populations (CBGP), INRA Montpellier, France;
[2] Institut De Biologie Computationnelle (IBC), Montpellier, France;
[3] Cancer Genome Project, The Wellcome Trust Sanger Institute, Hinxton, UK;
[4] Department Of Mathematics And School Of Biological Sciences, University Of Bristol, UK.

Detecting and measuring selection from gene frequency data

The recent advent of high throughput sequencing and genotyping technologies makes it possible to produce, easily and cost-effectively, large amounts of detailed data on the genotype composition of populations. Detecting locus-specific effects may help identify those genes that have been, or are currently, targeted by natural selection. How best to identify these selected regions, loci or single nucleotides remains a challenging issue. Here, we introduce a new model-based method, called SelEstim (Vitalis et al. 2014), to distinguish putative selected polymorphisms from the background of neutral (or nearly neutral) ones, and to estimate the intensity of selection at the former. The underlying population genetic model is a difusion approximation for the distribution of allele frequency in a population subdivided into a number of demes that exchange migrants. We use a Monte Carlo Markov Chain algorithm for sampling from the joint posterior distribution of the model parameters, in a hierarchical Bayesian framework. We present evidence from stochastic simulations, which demonstrates the good power of SelEstim to identify loci targeted by selection and to estimate the strength of selection acting on these loci, within each deme. We also re-analyze a subset of SNP data from the Stanford HGDP-CEPH Human Genome Diversity Cell Line Panel to illustrate the performance of SelEstim on real data. In agreement with previous studies, our analyses point to a very strong signal of positive selection upstream of the LCT gene, which encodes for the enzyme lactase-phlorizin hydrolase and is associated with adult-type hypolactasia. The geographical distribution of the strength of positive selection across the Old World matches the interpolated map of lactase persistence phenotype frequencies, with the strongest selection coeficients in Europe and in the Indus Valley.

Vitalis R., Gautier M., Dawson K.J. and Beaumont M.A. (2014) Detecting and measuring selection from gene frequency data. Genetics, in press

POSTERS

Poster 1

> **Sarah Bastkowski** [1], Danial Mapleson [1], Andreas Spillner [2], David Swarbreck [1], Vincent Moulton [3]
[1] The Genome Analysis Centre (TGAC), Norwich, UK;
[2] University of Greifswald, Greifswald, Germany;
[3] University of East Anglia, Norwich, UK

SPECTRE : A Suite of Phylogenetic Tools for Reticulate Evolution

We present SPECTRE (Suite of PhylogEnetiC Tools for Reticulate Evolution), a new open source software package that contains a number of useful tools, data structures and algorithms for inferring and visualising evolutionary patterns associated with recticulate evolution.

The current version of SPECTRE is focused on tools that either create or use split networks, such as SuperQ [1], FlatNJ [2], NetME [3] and several NeighborNet variants [4,5]. Furthermore we provide a tool for visualising the produced trees and networks. The tools are designed to have consistent graphical and command line interfaces and are efficient, reliable and portable to all common desktop platforms and high performance computing environments. Key data structures and algorithms, such as splits, trees, networks, distances and quartets, shared between tools are stored in a central library. This library also contains code for handling commonly used phylogenetics file formats such as Nexus, Phylip and Newick. Finally, the library is designed to be easily integrated into future tools or into external community projects. Our aim is to make phylogenetic tools accessible and useful to three groups.

First, for biologists, SPECTRE is easy to install and easy to use. Second, bioinformaticians will appreciate the efficiency and reliability so they can construct robust automated pipelines for performing phylogenetic inference. Finally, developers have a well architected codebase, which is easy to extend and integrate, allowing them to build their own tools using the shared library.

[1] S. Grünewald, A. Spillner, S. Bastkowski, A. Bögershausen, V. Moulton (2013) SuperQ: Computing Supernetworks from Quartets. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 10(1):151-60.
[2] M. Balvočiūtė, A. Spillner, V. Moulton (2014) FlatNJ: A novel network-based approach to visualize evolutionary and biogeographical relationships. Systematic Biology.
[3] S. Bastkowski, A. Spillner, V. Moulton (2014) Fishing for minimum evolution trees with Neighbor-Nets. Information Processing Letters, 114(1–2): 13–18.
[4] D. Levy, L. Pachter (2010) The neighbor-net algorithm. Advances in Applied Mathematics, 47: 240- 258.
[5] D. Bryant, V. Moulton (2004) Neighbor-Net: An Agglomerative Method for the Construction of Phylogenetic Networks. Molecular Biology and Evolution, 21 (2): 255-265.

Poster 2

> **Filip Bielejec** [1], Philippe Lemey [1], Luiz Max Carvalho[2], Guy Baele [1], Andrew Rambaut[3], Marc A. Suchard [4,5]
[1] Department Of Microbiology And Immunology, Rega Institute, KU Leuven, Leuven, Belgium
[2] Program For Scientific Computing (PROCC), Fundaçao Oswaldo Cruz, Rio De Janeiro, Brazil;
[3] Institute Of Evolutionary Biology, University Of Edinburgh, Edinburgh, United Kingdom
[4] Department Of Biomathematics and human genetics, D. Geffen school of medecine at UCLA, University Of California, Los Angeles, CA, USA;
[5] Department of Biostatistics, UCLA Fielding School of Public Health, University of California, Los Angeles, CA, USA

piBUSS: a parallel BEAST/BEAGLE utility for sequence simulation under complex evolutionary scenarios

Background: Simulated nucleotide or amino acid sequences are frequently used to assess the performance of phylogenetic reconstruction methods. BEAST, a Bayesian statistical framework that focuses on reconstructing time-calibrated molecular evolutionary processes, supports a wide array of evolutionary models, but lacked matching machinery for simulation of character evolution along phylogenies.

Results: We present a flexible Monte Carlo simulation tool, called piBUSS, that employs the BEAGLE high performance library for phylogenetic computations within BEAST to rapidly generate large sequence alignments under complex evolutionary models. piBUSS sports a user-friendly graphical user interface (GUI) that allows combining a rich array of models across an arbitrary number of partitions. A command-line interface mirrors the options available through the GUI and

facilitates scripting in large-scale simulation studies. Analogous to BEAST model and analysis setup, more advanced simulation options are supported through an extensible markup language (XML) specification, which in addition to generating sequence output, also allows users to combine simulation and analysis in a single BEAST run.
Conclusions: piBUSS offers a unique combination of flexibility and ease-of-use for sequence simulation under realistic evolutionary scenarios. Through different interfaces, piBUSS supports simulation studies ranging from modest endeavors for illustrative purposes to complex and large-scale assessments of evolutionary inference procedures. The software aims at implementing new models and data types that are continuously being developed as part of BEAST/BEAGLE.

Poster 3

> **Katarina Bodova** [1], Nick Barton [1], Gasper Tkacik [1]
[1] Institute of Science and Technology, Austria

A hybrid discrete--continuous maximum entropy method in quantitative trait dynamics of a single locus

Evolutionary processes including selection, mutation and random drift affect the dynamics of allele frequencies and consequently of quantitative traits. While the macroscopic dynamics of quantitative traits can be measured, the allele frequencies are typically unknown. Without knowing these microscopic processes the key question remains: How do the macroscopic observables respond to changes in evolutionary forces? The problem has previously been studied with the help of statistical mechanics principles [Barton, de Vladar, 2009]. This approach describes the microscopic stationary distribution of allele frequencies using a principle of maximal entropy with a particular entropy function representing the evolutionary forces. The method also allows study of the temporal changes of macroscopic quantities, i.e., trait mean and genetic variability in response to changes in evolutionary forces, and gives very accurate predictions, particularly in a regime of a strong mutation. However, in the regime of weak mutation the method breaks down due to a qualitative change in the steady state allele frequency distribution. We propose and study an extension of the maximum entropy method to a case of a single locus in the regime of small mutation rate.

Poster 4

> **Olga Chernomor** [1,2], Bui Quang Minh [1], Arndt Von Haeseler [1,2]
[1] Center for Integrative Bioinformatics Vienna (CIBIV), Max F. Perutz Laboratories (MFPL), Vienna, Austria;
[2] Bioinformatics and Computational Biology, Faculty of Computer Science, University of Vienna, Vienna, Austria;

Viable taxon selection under Split Diversity

Phylogenetic Diversity (PD) is a measure of biodiversity based on the evolutionary history of species (Faith 1992). Here we discuss optimization problems related to the use of PD, and the more general measure split diversity, in conservation prioritization (Spillner et al. 2008, Minh et al. 2009). Depending on the conservation goal and the information available about species, one can construct optimization routines that incorporate different biological constraints. Viable taxon selection problem was introduced by Moulton et al. (2007). Here, the prioritization is constrained by dependencies between species in the community. In practice, incorporating such constraints has the potential to prevent the use of limited resources on specialist taxa unless a sufficient resource base to support them is also preserved. In our work we extend the viable taxon selection to account for weighted dependency networks. Further, we show how to model such optimization problems in Integer Programming (IP) parlance, which allows available IP software packages to solve them. We demonstrate the usage of viability constraints in conservation prioritization on Caribbean Coral Reef Community (Opitz 1996). Here, predator-prey relationships between species in the community are used to define the constraints and the dependency network used is a weighted food web. Such food webs allow us to analyze an entire set of species as an interaction network rather than as isolated units.

References :
Faith, D.P. (1992) Conservation evaluation and phylogenetic diversity. Biological Conservation, 61, 1-10. Minh, B.Q., Klaere, S. & von Haeseler, A. (2009)
Taxon selection under split diversity. Systematic Biology, 58, 586-594. Moulton, V., Semple, C. & Steel, M. (2007)
Optimizing phylogenetic diversity under constraints. Journal of Theoretical Biology, 246, 186-194. Opitz, S. (1996)
Trophic interactions in Caribbean coral reefs. ICLARM Technical Reports, Makati City, Philippines. Spillner, A., Nguyen, B.T.

& Moulton, V. (2008)
Computing phylogenetic diversity for split systems. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 5, 235-244.

Poster 5

> **Giulio Valentino Dalla Rova** And Others
Biomathematics Research Centre, University of Canterbury, Christchurch, New Zealand

The Web Traits and the Tree

Food webs exhibit a phylogenetic signal: similar species play similar roles wherever they are found – i.e., they have similar trophic level, motifs distribution, etc. Modelling food webs as random dot-product graphs, we can represent predation relationships between species as a stochastic event: the probability of an interaction between two species depends on their vector of traits. Using spectral embedding techniques we can estimate, directly from the observed food webs, the vectors of vulnerability and foraging traits. This may represent a direct link between network based diversity measures and phylogenetic diversity measures. In conclusion, we show how a generalization of a field-of-bullets model of extinction based on species traits can be implemented.

Poster 6

> **Riet De Smet[**1,2], Keith Adams[1,3], Klaas Vandepoele[1,2], Steven Maere[1,2] & Yves Van De Peer[1,2,4]
[1] Department of Plant Systems Biology, VIB, Technologiepark 927, 9052 Ghent, Belgium;
[2] Department of Plant Biotechnology and Bioinformatics, Ghent University, Technologiepark 927, 9052 Ghent, Belgium
[3] Department of Botany, University of British Columbia, 6270 University Blvd, Vancouver, BC, V6T 1Z4, Canada
[4] Genomics Research Institute (GRI), University of Pretoria, Private bag X20, Pretoria, 0028, South Africa

Convergent Gene Loss Following Gene And Genome Duplications Creates Single-Copy Families In Flowering Plants

The importance of gene gain through duplication has been appreciated for a long time. Contrary, the importance of gene loss has only recently attracted attention. Indeed, studies in organisms ranging from plants to worms and humans suggest that duplication of some genes might be better tolerated than that of others. Here we have undertaken a large-scale study to investigate the existence of duplication-resistant genes in the sequenced genomes of 20 flowering plants. We demonstrate that there is a large set of genes that is convergently restored to single-copy status following multiple genome-wide and smaller-scale duplication events. We rule out the possibility that such a pattern could be explained by random gene loss only and therefore propose that there is selection pressure to preserve such genes as singletons. This is further substantiated by the observation that angiosperm single-copy genes do not comprise a random fraction of the genome, but instead are often involved in essential housekeeping functions that are highly conserved across all eukaryotes. Furthermore, single-copy genes are generally expressed more highly and in more tissues than non-single-copy genes, and they exhibit higher sequence conservation. Finally, we propose different hypotheses to explain their resistance against duplication.

Poster 7

> **Ludovic Duvaux**[1]; Mark Beaumont[2]; Martin Hinsch[1]; Roger K Butlin[1]
[1] University of Sheffield, Sheffield, United-Kingdom
[2] University of Bristol, Bristol, United-Kingdom

An ABC framework to investigate the role of natural selection in divergence between aphid host races

Speciation is central to evolutionary biology. Recent attention has focused on divergent natural selection where populations evolve different adaptations in distinct habitats. However, natural selection may be opposed by the movement of genes between populations leading to the so called 'speciation with gene flow' (Nosil 2012). Therefore, a major route to further understanding is to identify the genes that respond to divergent selection and so to determine how they influence the movement of other genes between populations. There have recently been spectacular advances in

the technology of DNA sequencing. These advances have allowed many groups to generate large data sets covering many genes, sometimes even whole genome sequences, in animals and plants subject to divergent natural selection. In principle, it should be possible to search these data for genes that are unusually divergent between populations, so identifying the direct targets of natural selection. Unfortunately, this is difficult because the history of populations is unknown and events in the past such as strong reductions in population size cause variation in the levels of divergence of genes, regardless of whether they are under selection or not (Crisci et al. 2012). There is a major need for better methods of analysis that will facilitate the use of the flood of new data to solve this difficult problem: first to find the best model for the history of the study populations and then to use that model to test for genes that are under selection. The pea aphid, a species whose genome has recently been sequenced and which is a model for many aspects of aphid biology, is a fascinating example of progress towards speciation as a result of divergent natural selection. In Europe, the pea aphid has 11 races, each adapted to a different host plant species and forming a continuum of genomic divergence. We have available several very large data sets from these host races, using several of the latest sequencing technologies. Here, we present the first results of an approximate Bayesian framework (Beaumont et al. 2002, Peter et al. 2012) designed to understand the setting up of the genomic architecture of reproductive isolation by accounting both for demography and selection. We will present results from simulations of a wide range of scenarios (including violations of the model assumptions) as well as from applications to our wide range of pea aphid data.

References :
Beaumont MA, Zhang W, Balding DJ (2002) Approximate Bayesian computation in population genetics. Genetics, 162, 2025–35.
Crisci JL et al. (2012) Recent Progress in Polymorphism-Based Population Genetic Inference. The Journal of heredity, 1–10.
Nosil P (2012) Ecological Speciation. Oxford University Press, Oxford.
Peter BM et al. (2012) Distinguishing between selective sweeps from standing variation and from a de novo mutation. PLoS genetics, 8, e1003011.

Poster 8

> **Fanny Gascuel** [1], Régis Ferrière [2,3], Robin Aguilée [4], Amaury Lambert [1,5]
[1] Center for Interdisciplinary Research in Biology, CNRS UMR 7241, Collège de France, France;
[2] Laboratoire Ecology et Evolution, CNRS UMR 7625, Ecole Normale Supérieure, Paris, France;
[3] Department of Ecology and Evolutionary Biology, University of Arizona, Tucson, USA;
[4] Laboratoire Evolution et Diversité Biologique, CNRS UMR 5174, Université Paul Sabatier, Toulouse, France;
[5] Laboratoire Probabilités et Modèles Aléatoires, CNRS UMR 7599, UPMC Université Paris 06, Paris, France

Ecology affects phylogenetic tree shape: insights from an individual-based model of eco-evolutionary diversification

How and why species diversity changes across multiple scales of time and space is a perennial and controversial problem in the study of life. The role of biotic factors, such as competition, and that of abiotic factors, such as climatic or geological changes, have been debated largely independently. Here our goal is to advance our understanding of the combined roles of biotic and abiotic factors in driving the history of species diversity. To this end, we developed a general individual-based model of adaptive radiation, which explicitly takes into account biotic ecological processes (intra and inter-specific competition for resources) and abiotic environmental changes, through time (landscape dynamics and local catastrophic extinctions) and space (landscape asymmetry). We use this framework to advance fundamental questions in evolutionary biology: how do biotic and abiotic factors influence speciation and extinction rates, as well as their heterogeneity among clades? do these factors leave an imprint on the shape (branching tempo and topology) of phylogenies reconstructed from extant species? Our results underline the role of both the biotic and abiotic factors in shaping thepatterns of diversification of life at macroevolutionary time scales, and provides some clues about the processes which may be responsible for the shapes of empirical phylogenies.

Poster 9

> **M. Gautier** [1,2], A. Cruaud [1], M. Galan [1], J. Foucaud [1], L. Sauné [1], G. Genson [1] , E. Dubois [3], S. Nidelet [3], T. Deuve [4] And J.-y. Rasplus [1]
[1] INRA, UMR1062 CBGP, Montferrier-sur-Lez, France
[2] Institut de Biologie Computationnelle, 95 rue de la Galéra, 34095 Montpellier, France
[3] Montpellier GenomiX, c/o Institut de Génomique Fonctionnelle, Montpellier, France
[4] MNHN, UMR7205 OSEB, Muséum National d'Histoire Naturelle, Paris, France

Empirical assessment of RAD sequencing for interspecific phylogeny

Restriction-site-associated DNA sequencing (RAD-seq) has been used to infer the recent evolutionary history (<3 My) of few organisms. However, it was repeatedly emphasized that with increasing genetic distances, mutations in the restriction sites will dramatically reduce the number of orthologous loci, making RAD-seq unsuitable to infer relationships between more distant taxa. Conversely, recent in silico studies suggested that a sufficient number of markers could be obtained from distant species. During our talk we will report on the first empirical test of this prediction. We compared the power of Sanger and RAD sequencing approaches to resolve relationships between 18 nonmodel species of ground beetles (Carabus), whose divergences ranged from 1.2 to 17 My. For three times the price of our Sanger experiment (three mitochondrial and six nuclear markers), our homemade RAD library produced 400 times more sites (25,425 loci; alignment length = 2,262,825 bp; more than 132,000 informative sites) and fully resolved relationships between the species of beetles. This one-shot study demonstrates the feasibility, affordability and unprecedented power of RAD-seq to infer relationships within many groups of Eukaryotes.

Poster 10

> **Carla Giner-delgado** [1,2], David Castellano [1], Magdalena Gayà-vidal [1], Sergi Villatoro [1], Mario Cáceres [1,3];
[1] Institut de Biotecnologia i de Biomedicina, Universitat Autònoma de Barcelona, Bellaterra (Barcelona), Spain;
[2] Departament de Genètica i de Microbiologia, Universitat Autònoma de Barcelona, Bellaterra (Barcelona), Spain;
[3] Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Spain;

Evolutionary trajectories of human polymorphic inversions

Chromosomal inversions have been a paradigm for evolutionary biology for decades. A key effect of inversions is that they suppress recombination as heterozygotes. Because of this property, they have been proposed as key factors in processes like local adaptation, evolution of sex chromosomes, and speciation. However, very little is known about the evolutionary dynamics of these mutations, especially in humans. Here, we explore evolutionary trajectories of human polymorphic inversions to better understand the forces acting on them. This study is framed within an exhaustive project towards the characterization of inversions in the human genome (INVFEST). In particular, we took advantage of the large-scale genotyping effort of 41 inversions in ~550 individuals from 7 HapMap populations, and the information provided about the global frequencies and differences among populations. Additionally, by combining the inversion genotypes with the 1000 Genomes Project data, we have also been able to explore the nucleotide variation patterns associated to the inversions. This analysis revealed that a high proportion of those mediated by inverted repeats are recurrent, while those with clean breakpoints seem to have a unique origin. Focusing in the inversions derived from a unique inversion event, we have applied and adapted methods to estimate their age from both nucleotide variation and frequency data. Taking into account the uncertainties of each type of information, we have explored the different demographic and selective scenarios that could lead to the observed results. Our findings suggest that some candidate inversions may have been favored by selection and deserve further molecular and phenotypic characterization.

Poster 11

> **Florian Massip** [1,2], Michael Sheinman [2], Peter Arndt [2]
[1] MIG - INRA Jouy en Josas Cedex France;
[2] Max Planck Institute for Molecular Genetic, Berlin, Germany

Statistical Properties of Pairwise Distances on a Random Yule Tree

A Yule tree is the result of a process with constant branching and elimination rates. Such a process serves as an instructive null model of many real systems. Often the only available information is the distribution of pairwise distances between a small fraction of the leaves in a tree. Motivated by this, we studied statistical properties of the pairwise distances. Using a method based on a recursion,we derived an exact, analytic and compact formula for the expected number of pairs separated by a certain time distance, t. The last turns out to be a simple increasing exponential function in t. Our method can be further used to derive other quantities. We demonstrate it by calculatingthe expected number of n-most closely related pairs of leaves and number of cherries, separated by a certain time distance. To make our results more useful for realistic scenarios we also take intoaccount incomplete sampling of the leaves in a tree.

Poster 12

> **Elina Numminen**, Jukka Kohonen, Jukka Corander
PL 68 (Gustaf Hällströmin katu 2b) 00014 Helsingin yliopisto

When relationships get complex: a general unsupervised approach for the inference on ecological interactions between species

The dynamics of ecological communities are often well described by the interactions the species have on each other. Such interactions give rise to interesting emergent properties of the whole system, sometimes promoting species diversity and persistence, and sometimes the opposite. Statistical inference on the nature of species interactions from observational studies is complicated by the large number of co-occurring species in a habitat, making the number of hypothesis on possible interactions even larger. Also, observational data might not contain enough observations of the desired type, so that all the interesting hypothesis could be carefully assessed. We present a very general unsupervised method for studying species interactions from longitudinal observational studies on ecological communities. We assume the system state, i.e. the co-occurrence of species at certain time in a habitat, evolves as a Markov process and we cluster the state space based on the frequencies of transitions made. The obtained clusters reveal whether there exists important interactions between species, and also which of the species are interacting with whom.
Therefore, our approach could serve as an important exploratory tool, for assessing firstly to what extent the data is informative on the interactions and secondly which interactions seem more evident and could be studied in more detail. The advantage of this approach is that it requires minimal amount of prior information on interactions or species demography in general. Since the method unravels the interaction network instead of the strength of each single interaction, it also suits better for systems with several species. We present our approach and the motivation to it by applying it to co-occurrence data on 5 bacterial species sharing the same habitat, human nasopharynx.

Poster 13

> **Murray Patterson** [1,2,3], Gergely Szollozi [2,4], Vincent Daubin [2] And Eric Tannier [1,2]
[1] INRIA Rhone-Alpes, 655 avenue de l'Europe, F-38344 Montbonnot, France;
[2] Laboratoire de Biometrie et Biologie Évolutive, CNRS and Université de Lyon 1, 43 boulevard du 11 novembre 1918, F-69622 Villeurbanne, France;
[3] Centrum Wiskunde & Informatica, Science Park 123, 1098 XG, Amsterdam, The Netherlands; [4] ELTE-MTA "Lendület" Biophysics Research Group 1117 Bp., Pázmány P. stny. 1A., Budapest, Hungary

Lateral Gene Transfer, Rearrangement, Reconciliation

Background. Models of ancestral gene order reconstruction have progressively integrated different evolutionary patterns and processes such as unequal gene content, gene duplications, and implicitly sequence evolution via reconciled gene trees. These models have so far ignored lateral gene transfer, even though in unicellular organisms it can have an important confounding effect, and can be a rich source of information on the function of genes through the detection of transfers of clusters of genes.

Result. We report an algorithm together with its implementation, DeCoLT, that reconstructs ancestral genome organization based on reconciled gene trees which summarize information on sequence evolution, gene origination, duplication, loss, and lateral transfer. DeCoLT optimizes in polynomial time on the number of rearrangements, computed as the number of gains and breakages of adjacencies between pairs of genes. We apply DeCoLT to 1099 gene families from 36 cyanobacteria genomes.
Conclusion. DeCoLT is able to reconstruct adjacencies in 35 ancestral bacterial genomes with a thousand gene families in a few hours, and detects clusters of co-transferred genes. DeCoLT may also be used with any relationship between genes instead of adjacencies, to reconstruct ancestral interactions, functions or complexes.
Availability. http://pbil.univ-lyon1.fr/software/DeCoLT/

Poster 14

> **Willy Rodriguez**
Institut National des Sciences Appliquées de Toulouse, Toulouse, France

Exploring the estimation quality of PSMC using different sizes of input sequences

Since its publication in 2011, the Pairwise Sequentially Markovian Model (PSMC) method [1] has been widely used in order to infer the demographic history of different species. [2][3][4]. In all these studies, the authors could rely on full genome information, and the PSMC has been shown to perform well in these cases. However, it is not always possible to obtain a whole diploid genome, due to the high sequencing costs. For several non model species only unordered (or partially ordered) scaffolds are available [5] and it is not fully clear how this information could be used in order to infer the demographic history of these species. It is well understood that the quality of PSMC estimates is related to the number of recombinations events that are expected to occur in each defined time window (i.e. the time periods during which the effective population size is estimated and assumed to be constant). However, there is still uncertainty on the “minimal” length of the input sequences for getting statistically acceptable results.
In this study, I will present an analysis of how the size of the input sequence changes the quality of the demographic inference by using Goodness of Fit and Variance measures associated to the inference. The data used for this purpose have been generated by simulations using software ms [6].

References
[1] Inference of human population history from individual whole-genome sequences. Li H, Durbin R. (2011)
[2] Polar and brown bear genomes reveal ancient admixture and demographic footprints of past climate change. Miller et al. (2012)
[3] Baiji genomes reveal low genetic variability and new insights into secondary aquatic adaptations. Zhou et al. (2013)
[4] Whole-genome sequencing and analysis of the Malaysian cynomolgus macaque (Macaca fascicularis) genome. Higashino et al. (2012)
[5] Aye-aye population genomic analyses highlight an important center of endemism in northern Madagascar . Perry et al. (2013)
[6] msHOT: modifying Hudson's ms simulator to incorporate crossover and gene conversion hotspots. Garrett Hellenthal and Matthew Stephens (2007)

Poster 15

> **Magali Semeria** [1], Laurent Guéguen [1], Eric Tannier [1,2]
[1] Laboratoire de Biométrie et Biologie Évolutive (LBBE), Université Lyon 1, Villeurbanne, France;
[2] INRIA Grenoble Rhône-Alpes, Montbonnot, France

Evolution of gene relationships

Phylogenetic reconstruction methods are traditionally based on models of nucleotide or amino-acid sequence evolution. But the evolution of genomes can be studied at different scales: the gene level, accounting for gains and losses, and the genome level, accounting for rearrangements of chromosome organization. An integrative method has been developed that reconstructs species and genes' histories simultaneously and takes into account sequence evolution, gene birth, duplication, transfer and loss [1]. Because it accounts for gene level and sequence level signal, this method provides far more reliable species trees and gene trees for the inference of evolutionary histories. However, as most existing methods do, this method neglects that genes evolve in a structured environment, such as gene regulation pathways or genomic structure. In fact, genes share relationships with other genes when they participate in a common function or because of physical proximity. As genes undergo evolutionary events such as duplication and rearrangement, the relationships they share can be kept, gained or lost. The relationship between two evolving

genes can then be considered as an evolutionary object itself [2]. We model the evolution of relationships between genes and show that we can then inform gene tree reconstruction [3]. We argue that our method is a first step toward the integration of co-evolution information in gene tree reconstruction methods.

[1] B. Boussau, G. J. Szöllosi, L. Duret, M. Gouy, E. Tannier, and V. Daubin, “Genome-scale coestimation of species and gene trees.,” Genome Res., vol. 23, no. 2, pp. 323–30, Feb. 2013.
[2] S. Bérard, C. Gallien, B. Boussau, G. J. Szöllősi, V. Daubin, and E. Tannier, “Evolution of gene neighborhoods within reconciled phylogenies.,” Bioinformatics, vol. 28, no. 18, pp. i382–i388, Sep. 2012.
[3] C. Chauve, N. El-Mabrouk, L. Guéguen, M. Semeria, and E. Tannier, “Duplication, Rearrangement and Reconciliation: A Follow-Up 13 Years Later,” in in Models and Algorithms for Genome Evolution, C. Chauve, N. El-Mabrouk, and E. Tannier, Eds. Springer, 2013, pp. 47–62.

Poster 16

> **Bertrand Servin** [1], Simon Boitard [2,3], Maria-Ines Fariello [1], Florence Phocas [3], Magali San Cristobal [1]
[1] Génétique, Physiologie et Systèmes d'Élevage (Genphyse), INRA, Toulouse, France;
[2] Museum National d'Histoire Naturelle, Paris, France;
[3] Génétique Animale et Biologie Intégrative (GABI), INRA, Jouy-en-Josas, France

On the advantages of a multipoint approach for the detection of selection signatures: Lessons from the 1000 Bull Genomes data.

Genome scans for selection can be used to identifiy genes and mutations underlying adaptive traits. The classical FST statistic does not account for complex population history. As a single SNP statistic, it is also not modelling linkage disequilibirum. We present two new methods aimed at taking these phenomenons into account. We will illustrate application of these methods on two datasets in cattle, a resequencing dataset on four international breeds and a high-density genotyping dataset of 20 french breeds. In particular, we will show how these methods can be used to reveal the selective history of a QTL in cattle populations.

Poster 17

> **Jakub Truszkowski**
EBI Goldman Group and Cancer Research UK (S.Tavare), University of Cambridge, UK

Using hashing to speed up phylogenetic inference and placement

Due to rapid expansion in sequence databases, very fast phylogenetic reconstruction algorithms are becoming necessary. Large sequence alignments can contain up to hundreds of thousands of sequences, making traditional methods, such as Neighbor Joining, computationally prohibitive. We present LSHTree, the first sub-quadratic time algorithm with mathematical accuracy guarantees under a Markov model of sequence evolution. Our new algorithm runs in O(n^{1+\gamma(g)} log^2 n) time, where \gamma is an increasing function of an upper bound on the mutation rate along any branch in the phylogeny, and \gamma(g) < 1 for all g. For phylogenies with very short branches, the running time of our algorithm is close to linear. In experiments, our prototype implementation was more accurate than the current fast algorithms, while being comparably fast. Our current work focuses on applying the LSH framework to the problem of phylogenetic placement of environmental sequence reads. This work appeared in the conference proceedings of WABI 2012.

ATTENDEES

Bastkowski	Sarah	The Genome Analysis Centre,GB
Berry	Vincent	LIRMM, IBC,FR
Bielejec	Filip	KU Leuven,BE
Binet	Manuel	LIRMM IBC,FR
Bodova	Katarina	Institute Of Science And Technology,AT
Boitard	Simon	INRA / MNHN,FR
Cassan	Elodie	LIRMM,FR
Chernomor	Olga	CIBIV, University Of Vienna (UniVie),AT
Chevenet	François	IRD/IBC,FR
Dalla Riva	Giulio	University Of Canterbury,NZ
De Smet	Riet	VIB/Ghent University,BE
Delsuc	Frederic	ISEM, CNRS Université Montpellier 2,FR
Du Plessis	Louis	Theoretical Biology, Institute Of Integrative Biology, ETH Zürich,CH
Duvaux	Ludovic	University Of Sheffield,GB
Etienne	Rampal	University Of Groningen,NL
Gascuel	Olivier	IBC, LIRMM, CNRS – Univ. de Montpellier,FR
Gascuel	Fanny	Collège De France,FR
Gautier	Mathieu	INRA,FR
Gavruskin	Alex	Auckland University Of Technology,NZ
Giner-Delgado	Carla	Universitat Autònoma De Barcelona,ES
Guéguen	Laurent	LBBE Université Lyon 1,FR
Hoscheit	Patrick	INRA,FR
Huson	Daniel	ZBIT, Department Of Computer Science, Tuebingen University,DE
Klaere	Steffen	Auckland University,NZ
Ladret	Veronique	Institut De Mathématiques Et Modélisation De Montpellier (I3M),FR
Lartillot	Nicolas	Laboratoire De Biologie Et Biométrie Évolutive, Lyon,FR
Latrille	Thibault	ENS Lyon,FR
Leblois	Raphael	INRA,FR
Lefort	Vincent	CNRS,FR
Lèbre	Sophie	Laboratoire ICube, Université De Strasbourg.,FR
Marin	Jean-Michel	University Montpellier 2,FR
Massip	Florian	MIG-INRA,FR
Matias	Catherine	CNRS, LaMME, Evry,FR
Merle	Coralie	UM2/INRA, I3M/CBGP,FR
Mooers	Arne	Simon Fraser University,CA
Moorjani	Priya	Columbia University,US
Morlon	Hélène	Ecole Polytechnique,FR
Mourad	Raphael	LIRMM,FR
Nielsen	Rasmus	University Of California, Berkeley,US
Novak	Sebastian	IST Austria,AT
Numminen	Elina	Department Of Mathematics And Statistics, University Of Helsinki,FI
Palero	Ferran	UVEG,ES
Pardi	Fabio	IBC, LIRMM, CNRS – Univ. de Montpellier,FR
Patterson	Murray	Centrum Wiskunde ,NL
Pouyet	Fanny	LBBE-UMR5558,FR
Pudlo	Pierre	I3M - Université De Montpellier 2,FR
Rodriguez Valcarce	Willy	Institut National Des Sciences Appliquées De Toulouse,FR
Rousset	Francois	Institut Des Sciences De L'Evolution, Montpellier,FR
Sarikas	Srdjan	IST Austria,AT
Saulnier	Emma	LIRMM,FR
Schiffels	Stephan	Auckland University Of Technology,NZ

Scornavacca	Céline	ISE-M,FR
Semeria	Magali	LBBE - UCB Lyon 1,FR
Servin	Bertrand	INRA,FR
Siepel	Adam	Cornell University,US
Steel	Mike	University Of Canterbury,NZ
Swenson	Krister	IBC, LIRMM, CNRS – Univ. de Montpellier,FR
Tannier	Eric	INRIA, LBBE, Université Lyon 1, France,FR
Truszkowski	Jakub	Auckland University Of Technology,NZ
Van Iersel	Leo	CWI,NL
Vitalis	Renaud	INRA,FR
Zhang	Chi	Swedish Museum Of Natural History,SE