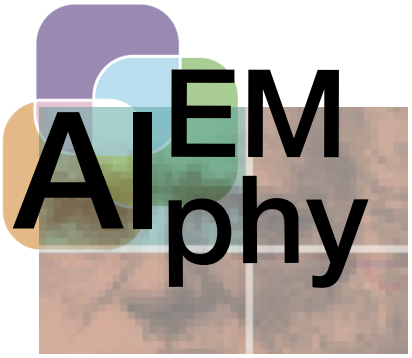
2023 Alphy & AIEM Joint Meeting

Amphitheater Boucherle, Université Grenoble Alpes

23-25 January 2023



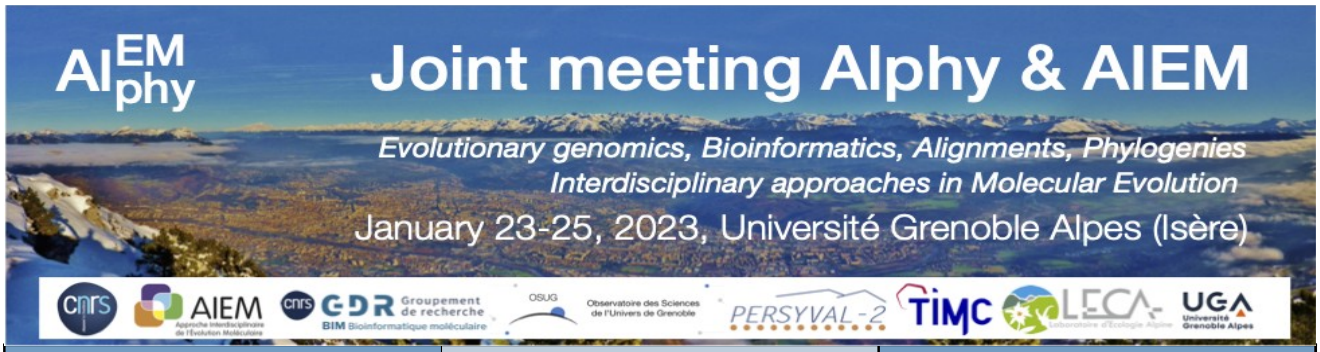
Financial support

ORGANIZING COMMITTEE

Sophie Abby (TIMC, CNRS, Université Grenoble Alpes)
Frédéric Brunet (IGFL, ENS de Lyon)
Magali Brunet (TIMC, Université Grenoble Alpes)
Laurence Després (LECA, Université Grenoble Alpes)
Antoine Frenoy (TIMC, Grenoble-INP, Université Grenoble Alpes)
Nelle Varoquaux (TIMC, CNRS, Université Grenoble Alpes)
Flora Gaudillière (TIMC, Université Grenoble Alpes)
Margaux Jullien (TIMC, CNRS, Université Grenoble Alpes)
Morgane Roger-Margueritat (TIMC, Université Grenoble Alpes)
Sophie-Carole Chobert (TIMC, CNRS, Université Grenoble Alpes)

SCIENTIFIC COMMITTEE

Sophie Abby (TIMC, CNRS, Université Grenoble Alpes)
Guillaume Achaz (MNHM, Université Paris-Cité; CS AIEM, GDR AIEM chair)
Frédéric Brunet (IGFL, ENS de Lyon; CS AIEM)
Laurence Després (LECA, Université Grenoble Alpes)
Laurent Duret (LBBE, CNRS, UCBL; GDR BiM, Alphy chair)
Antoine Frenoy (TIMC, Grenoble-INP, Université Grenoble Alpes)
Céline Scornavacca (ISEM, CNRS, Université de Montpellier; GDR BiM, Alphy chair)
Nelle Varoquaux (TIMC, CNRS, Université Grenoble Alpes)



Monday, January 23rd 2023

13h00	Welcome coffee and registration
13h30	Introduction
13h45	**Invited speaker - François Parcy** - Evolution of the floral regulator LEAFY
14h30	Paul Zaharias - Robustness of Felsenstein's versus Transfer Bootstrap Supports with respect to Taxon Sampling
14h50	Nisha Dwivedi - Resolving the evolutionary history of Nesoenas picturata in the Mascarenes
15h10	Clément Gain - A quantitative theory for genomic offset statistics
15h30	Cofffee break
16h00	Marie Raynaud - Population genomics reveal PRDM9-dependent recombination landscapes in salmonids
16h20	Thibault Latrille - Up to 25% of beneficial mutations in protein sequences are not adaptive innovations in mammals
16h40	Julien Joseph - Recombination and selection efficiency in humans
17h00	Antoine Taupenot - Social polymorphism and supergene in the ant Myrmecina graminicola: insights from population genomics
17h20	Poster session & beers
19h00	End of first day

Tuesday, January 24th 2023

8h30	Opening
9h00	**Invited speaker - Anne-Florence Bitbol** - Optimization and historical contingency in protein sequences
9h45	Ignacio Bravo - Transcriptomic, proteomic and functional cis- and trans-acting effects in human cells of gene expression with varying codon usage preferences, in a long-term selection experiment
10h05	Luca Nesterenko - Phyloformer: Towards fast and accurate phylogeny reconstruction with self-attention networks
10h25	Coffee break
10h55	Florian Bénitière - Investigating the impact of random genetic drift on synonymous codon usage in metazoans
11h15	Margaux Jullien – COCOATree: benchmarking coevolution methods for sector identification
11h35	Romain Feron - Reproducible workflows to study conservation of genomic sequence using multispecies whole genome alignments
11h55	Rémi-Vinh Coudert - MPS-Sampling (Multi Proteins Similarity Sampling) to select evolutionary significant representative genomes from large databases
12h05	Lea Bou Dagher - Evolutionary signal captured from topological properties of proteins via persistent homology
12h25	Lunch break
14h00	**Invited speaker - Mathieu Groussin** - Evolution of host-gut microbiome interactions in the context of industrialization
14h45	Flora Gaudillière - Understanding the evolutionary dynamics of insertion sequences in prokaryotic genomes
15h05	Fanny Mazzamurro - Evolution of natural transformation in bacteria
15h25	Sophie-Carole Chobert - Unraveling the relative emergence of quinones biosynthetic pathways
15h45	Coffee break
16h15	Guillaume Louvel - Causes of discord in eukaryotic protein domains inherited from Archaea
16h35	Anton Crombach - Cell type diversification across paleo– and neocortex revealed by single cell multiomics analysis
16h55	Jasmine Gamblin - Beyond one-gain models for pangenome evolution
17h15	Charles Coluzzi - Epistatic interactions between genetic background and antibiotic resistances genes (and vice versa)
17h35	Mélodie Bastian - Bridging the gap between population genomic and phylogenetic approaches by the study of the effective population size
20h00	Gala dinner @Bouillon A

Wednesday, January 25th 2023

8h30	Opening
9h00	Jean-Baptiste Grodwohl - Opening Motoo Kimura's archives: on the history of molecular evolution and the neutralist school
9h20	Leonardo Trujillo - Adaptive walks don't do walks on hypercubes
9h40	Thibaut Capblancq - In search of islands of speciation in the genomes of two Coenonympha butterfly sister species
10h00	Elise Gay - Mitonuclear discordance in the great white shark (Carcharodon carcharias): sex biased dispersal, mitonuclear incompatibility, or both?
10h20	Coffee break
10h50	Alia Abbara - Spatially structured populations on graphs beyond update rules
11h10	Thomas Forest - Birds demography inference based on genomic data
11h30	Riccardo Poloni - Genetics and genomics help understanding the colour polymorphism in the invasive Box Tree Moth
11h50	Matthieu Joron - Subtle signals of adaptive introgression in the late stages of the speciation continuum
12h20	Conclusion
12h30	Lunch (to go)

Program

Oral Presentations

1

Population Genetics I

Evolution of the floral regulator LEAFY

Francois Parcy*[1]

[1]Physiologie Cellulaire et Végétale - Grenoble – CNRS : UMR5168, Commissariat à l'Énergie Atomique et aux Énergies Alternatives (CEA) - Grenoble, Institut national de la recherche agronomique (INRA), , Universitey of Grenoble Alpes (UGA) – France

Résumé

Angiosperms represent a successful group of plants characterised by the presence of flowers. Molecular genetics in angiosperm monde species identified the LEAFY transcription factor as a master regulator of flower development. However, LEAFY is also present in algae, mosses, ferns and gymnosperms and we wondered what are the evolutionary steps that lead LEAFY to become a key floral regulator. For this, we (and others) study both the evolution of the LEAFY role and the evolution of the structural and biochemical properties of the LEAFY protein. I will describe recent progress in both axes and discuss about what we learned about the origin of flowers.

*Intervenant

Robustness of Felsenstein's versus Transfer Bootstrap Supports with respect to Taxon Sampling

Paul Zaharias*[1], Frédéric Lemoine[2], and Olivier Gascuel†[3]

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

[2]Institut Pasteur – Hub Bioinformatique et Biostatistique – C3BI, USR 3756 IP CNRS – Institut Pasteur de Paris – 25-28 Rue du Docteur Roux, 75015 Paris, France

[3]Institut de Systématique, Evolution, Biodiversité – Museum National d'Histoire Naturelle, Ecole Pratique des Hautes Etudes, Sorbonne Universite, Centre National de la Recherche Scientifique : UMR7205, Université des Antilles – France

Résumé

The bootstrap method is based on resampling alignments and reestimating trees. Felsenstein's bootstrap proportions (FBP; Felsenstein 1985) is the most common approach to assess the reliability and robustness of sequence-based phylogenies. However, when increasing taxon-sampling (i.e., the number of sequences) to hundreds or thousands of taxa, FBP tends to return low supports for deep branches. The Transfer Bootstrap Expectation (TBE; Lemoine et al. 2018) has been recently suggested as an alternative to FBP. TBE is measured using a continuous transfer index in (0,1) for each bootstrap tree, instead of the {0,1} index used in FBP to measure the presence/absence of the branch of interest. TBE has been shown to yield higher and more informative supports, without inducing falsely supported branches. Nonetheless, it has been argued that TBE must be used with care due to sampling issues, especially in datasets with high number of closely related taxa. In this study, we conduct multiple experiments by varying taxon sampling and comparing FBP and TBE support values on different phylogenetic depth, using empirical datasets. Our results show that the main critic of TBE stands in extreme cases with highly unbalanced sampling among clades, but that TBE is still robust in most cases, while FBP is inescapably negatively impacted by high taxon sampling. We suggest guidelines and good practices in TBE (and FBP) computing and interpretation.

*Intervenant

†Auteur correspondant: olivier.gascuel@mnhn.fr

Resolving the evolutionary history of Nesoenas picturata in the Mascarenes

Nisha Dwivedi*[1], Ben Warren[2], and Stefano Mona[3,4]

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

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

[3]Ecole Pratique des Hautes Etudes (EPHE) – Ecole Pratique des Hautes Etudes – PSL Research University, Paris, France

[4]Institut de Systématique, Evolution, Biodiversité (ISYEB) – Museum National d'Histoire Naturelle, Université Pierre et Marie Curie - Paris 6, Ecole Pratique des Hautes Etudes : UMR7205, Centre National de la Recherche Scientifique – 57 rue Cuvier, CP39, France

Résumé

Many island species are threatened with extinction due to competition from introduced species. Knowing whether a species is native or introduced is therefore critical for effective conservation action which, on islands, commonly involves the culling of non-native individuals. The colonization history of *Nesoenas picturata* throughout the Mascarenes is uncertain. Though known to be native to Madagascar, its status in Réunion and Mauritius is unclear. Moreover, while *N.picturata* is protected in Réunion, it is culled in Mauritius. Fossil records of *N. picturata* in Mauritius suggest that the extant population could be native, the result of hybridization between native and introduced individuals or fully introduced after the extinction of the native population. To investigate the colonization history of *N.picturata* in the Mascarenes we created a genetic time series by whole genome sequencing historical and modern samples. Population structure analyses suggest that the *N. picturata* populations of Madagascar, Reunion and Mauritius form three separate clades. Historical demographic inferences show a similar trajectory in Reunion and Mauritius, but with discrepancies in the effective population size and the timing of demographic events. Such results are suggestive of two independent colonization events originating in Madagascar, the more ancient to Réunion and the more the more recent to Mauritius. By harnessing the power of temporal series, we are currently testing for hybridization and full modeling inter-island migration using coalescence simulations. This allows us to better understand the routes and timing of colonization to both Réunion and Mauritius, as well as refining the more recent changes in effective population size that occurred after human arrival.

*Intervenant

A quantitative theory for genomic offset statistics

Clément Gain*[1] and Olivier François[2]

[1]Techniques de lÍngénierie Médicale et de la Complexité - Informatique, Mathématiques et Applications, Grenoble - UMR 5525 – VetAgro Sup - Institut national d'enseignement supérieur et de recherche en alimentation, santé animale, sciences agronomiques et de l'environnement, Centre National de la Recherche Scientifique : UMR5525, Université Grenoble Alpes, Institut polytechnique de Grenoble - Grenoble Institute of Technology – France

[2]UMR 5525 TIMC-IMAG – Université Grenoble-Alpes – France

Résumé

Genomic offset statistics assess the maladaptation of populations to rapid habitat alteration based on association of genotypes with environmental variation. Despite substantial evidence for empirical validity, genomic offset statistics have well identified limitations, and lack a theory that would facilitate interpretations of predicted values. Here, we clarified the theoretical relationships between genomic offset statistics and adaptive traits, and proposed a new measure to predict fitness after rapid change in local environment. The predictions of our theory were verified in computer simulations and in empirical data on African pearl millet (Cenchrus americanus) obtained from a common garden experiment. Our results proposed a unified perspective on genomic offset statistics, and outlined their importance for conservation management in the face of environmental change.

*Intervenant

Population genomics reveal PRDM9-dependent recombination landscapes in salmonids

Marie Raynaud*[1], Pierre-Alexandre Gagnaire[2], and Nicolas Galtier[3]

[1]Institut des Sciences de lÉvolution de Montpellier – Centre de Coopération Internationale en Recherche Agronomique pour le Développement : UMR116, Ecole Pratique des Hautes Etudes, Université de Montpellier, Institut de recherche pour le développement [IRD] : UR226, Centre National de la Recherche Scientifique : UMR5554 – France

[2]Institut des Sciences de l'Évolution de Montpellier – Centre National de la Recherche Scientifique, Université de Montpellier, Institut de recherche pour le développement [IRD] : UR226 – France

[3]ISEM – CNRS, UMR 5554 Institut des Sciences de l'Evolution (Université de Montpellier) – France

Résumé

In most mammals, the genomic localization of recombination hotspots is directed by the PRDM9 protein. The complete or partial loss of PRDM9 in other taxa such as birds, some teleost fishes or invertebrates, results in a different localization of hotspots and an apparently more stable dynamics of recombination landscapes than in mammals. The existence of mammalian-like recombination landscapes in taxa that have retained a functional PRDM9 remains unclear, raising questions about the ancestral role of PRDM9 in animals. Salmonid fishes have a full-copy of the PRDM9 gene - Is it involved in the regulation of recombination as in humans and mice? To address this question, we used linkage disequilibrium information from whole-genome polymorphism data to build fine-scale population-based recombination maps in three Salmonid species: Oncorhynchus kisutch, O. mykiss and Salmo salar. The three species showed recombination rate variation at both broad and fine scales, with a tendency for hotspots to be localized away from transcription start sites, similarly to mammals. Moreover, the comparison of recombination landscapes among species of salmonids revealed a rapid evolution of the recombination rate distribution. These findings strongly suggest that PRDM9 has a function of directing DNA double strand breaks in salmonids, arguing in favor of an ancestrally PRDM9-mediated recombination process in vertebrates. Confirmation of these results via ChipSeq and binding motif analysis is on its way.

*Intervenant

Up to 25% of beneficial mutations in protein sequences are not adaptive innovations in mammals

Thibault Latrille*[1], Julien Joseph[2], and Nicolas Salamin[1]

[1]Département de biologie computationnelle - Université de Lausanne – Suisse
[2]Laboratoire de Biométrie et Biologie Evolutive – Université de Lyon, Université Lyon 1 – France

Résumé

In this work based on genome-wide studies across species and populations, we estimated the proportion of beneficial mutations in protein coding sequences that are not adaptive innovations.
Our study is based on the premise that slightly deleterious mutations scattered across the genome are reaching fixation due to genetic drift. These mutations are then subsequently reverted by beneficial back-mutations, generating a balance at which genomes are constantly both damaged and repaired simultaneously at different loci. Even though the existence of these back-mutations is predicted by the nearly neutral theory, they have been largely overlooked, and positive selection has been countlessly interpreted as adaptation to changing environments. In this work, we integrated datasets across the entire exome of 96 species at the mammalian scale, with polymorphism for 28 populations from 6 genera (Equus, Bos, Capra, Ovis, Chlorocebus and Homo). We then estimated selective effects of mutations inside mammalian protein coding sequences, under a model assuming no adaptation at the phylogenetic scale. We finally estimated the proportion of beneficial mutations that are not adaptive innovations among all beneficial mutations at the population scale. Our work confirms that deleterious substitutions have accumulated in mammals and are currently being eliminated. In modern humans, it results in around 25% of beneficial mutations that are not adaptive innovations, but instead are repairing previous deleterious changes.

*Intervenant

Recombination and selection efficiency in humans

Julien Joseph*†[1], Thibault Latrille[2], Laurent Duret[1], and Nicolas Lartillot[1]

[1]Laboratoire de Biométrie et Biologie Evolutive - UMR 5558 – Université Claude Bernard Lyon 1, Institut National de Recherche en Informatique et en Automatique, VetAgro Sup - Institut national d'enseignement supérieur et de recherche en alimentation, santé animale, sciences agronomiques et de l'environnement, Centre National de la Recherche Scientifique : UMR5558 – France
[2]Université de Lausanne = University of Lausanne – Suisse

Résumé

Meiotic recombination is a major force driving genome evolution in eukaryotes. It allows the independent segregation of alleles with different selective value hereby enhancing the efficiency of selection. However, it has also been shown in mammals and especially in humans that recombination events can induce a transmission bias of GC alleles call GC-biased gene conversion, which can increase the genetic load. In the end recombination can have both a beneficial and a deleterious effect on the efficiency of selection, but it is not clear which effect outweights the other, and if an increase of the recombination rate is overall beneficial or deleterious in humans. In this study, we leverage an estimation of the fitness landscape at the mammalian scale to test wether highly recombining protein are more or less fit than lowly recombining ones in humans. We conclude that recombination and GC-biased gene conversion is responsible for around a third of the fixed genetic load, showing that even relatively low recombination rates are already deleterious because of GC-biased gene conversion. We discuss the reason of this recombination-induced load in the light of intragenomic conflicts involving fertility and genome integrity.

*Intervenant
†Auteur correspondant: julien.joseph@ens-lyon.fr

Social polymorphism and supergene in the ant Myrmecina graminicola: insights from population genomics

Antoine Taupenot*[1], Elise Gay[2], Claudie Doums†[3], Mathieu Molet‡[1], and Stefano Mona§[3]

[1]Institut d´cologie et des sciences de lénvironnement de Paris – Institut de Recherche pour le Développement : UMR242, Sorbonne Universite : UMR113, Université Paris-Est Créteil Val-de-Marne - Paris 12 : UMR7618, Centre National de la Recherche Scientifique : UMR7618, Institut National de Recherche pour l'Agriculture, l'Alimentation et l'Environnement : UMR1392 – France
[2]École pratique des hautes études – Université Paris sciences et lettres – France
[3]Institut de Systématique, Evolution, Biodiversité – Museum National d'Histoire Naturelle, Ecole Pratique des Hautes Etudes, Sorbonne Universite, Centre National de la Recherche Scientifique : UMR7205, Université des Antilles – France

Résumé

Supergenes are genomic regions of various length inherited as a single block due to the suppression of recombination. They lock set of genes determining complex (adaptive) phenotypes. They are well known for being responsible of some spectacular polymorphisms like sex chromosomes, mimetic morphs in butterflies or colony social organization in ants. Up to now, it has been shown that five distantly related ant species harbor a supergene responsible for the number of queens per colony. This social polymorphism has been also linked to morphological, chemical, and behavioral variations involved in dispersal abilities. However, its origin and maintenance remain to be clearly elucidated and new test cases are warranted. In this spirit, we selected the ant species *Myrmecina graminicola* for which a social polymorphism (i.e., the occurrence of both monogynous and polygynous colonies) has been described and was shown to have a simple mendelian inheritance. To test for the presence of the supergene and its relationship with the social polymorphism we whole genome sequenced 24 queens from the Fontainebleau Forest (Ile-de-France) belonging to the two colony types. Genomic scan based on Fst, linkage disequilibrium and the site frequency spectrum identified 6 scaffolds (for a total of ˜11 Mb) rich in coding genes (˜400) differentiating monogynous from polygynous colonies, clearly identifying the presence of a social supergene. Syntenic analyses further highlight that the genomic organization of *M. graminicola* supergene differs from those of the other ant lineages, suggesting the existence of a novel evolutionary trajectory leading to the social phenotype. Further analyses are ongoing to study the genetic structure of the two types of colonies and to verify that the identified scaffolds constitute a single genomic region. This will allow us to better characterize the origin and maintenance of this supergene.

*Intervenant
†Auteur correspondant: claudie.doums@ephe.sorbonne.fr
‡Auteur correspondant: mathieu.molet@gmail.com
§Auteur correspondant: mona@mnhn.fr

2

Sequence Analysis

Optimization and historical contingency in protein sequences

Anne-Florence Bitbol*†[1]

[1]EPFL – Suisse

Résumé

Protein sequences are shaped by functional optimization on the one hand and by evolutionary history, i.e. phylogeny, on the other hand. A multiple sequence alignment of homologous proteins contains sequences which evolved from the same ancestral sequence and have similar structure and function. In such an alignment, correlations in amino-acid usage at different sites can arise from structural and functional constraints due to coevolution, but also from historical contingency.

Correlations arising from phylogeny often confound coevolution signal from functional or structural optimization, impairing the inference of structural contacts from sequences. However, inferred Potts models are more robust than local statistics to these effects, which may explain their success (1). Dedicated corrections can further increase this robustness (2). Moreover, phylogenetic correlations can in fact provide useful information for some inference tasks, especially to infer interaction partners from sequences among the paralogs of two protein families. In this case, signal from phylogeny and signal from constraints combine constructively (3), and explicitly exploiting both further improves inference performance (4).

Protein language models have recently been applied to sequence data, greatly advancing structure, function and mutational effect prediction. Language models trained on multiple sequence alignments capture coevolution and structural contacts, but also phylogenetic relationships (5). They are able to disentangle signal from structural constraints and from phylogeny more efficiently than Potts models (5), and they have promising generative properties (6).
(1) Dietler N, Lupo U, Bitbol A-F (2022) "Impact of phylogeny on structural contact inference from protein sequence data", https://arxiv.org/abs/2209.13045

(2) Colavin A, Atolia E, Bitbol A-F, Huang KC (2022) "Extracting phylogenetic dimensions of coevolution reveals hidden functional signals", Scientific Reports 12(1):820

(3) Gerardos A, Dietler N, Bitbol A-F (2022) "Correlations from structure and phylogeny combine constructively in the inference of protein partners from sequences", PLoS Computational Biology 18(5): e1010147

(4) Gandarilla-Perez CA, Pinilla S, Bitbol A-F, Weigt M (2022) "Combining phylogeny and coevolution improves the inference of interaction partners among paralogous proteins",

*Intervenant

†Auteur correspondant: anne-florence.bitbol@epfl.ch

https://arxiv.org/abs/2208.11626

(5) Lupo U, Sgarbossa D, Bitbol A-F (2022) "Protein language models trained on multiple sequence alignments learn phylogenetic relationships", Nature Communications 13: 6298 (2022)
(6) Sgarbossa D, Lupo U, Bitbol A-F (2022) "Generative power of a protein language model trained on multiple sequence alignments", https://arxiv.org/abs/2204.07110

Transcriptomic, proteomic and functional cis- and trans-acting effects in human cells of gene expression with varying codon usage preferences, in a long-term selection experiment.

Ignacio Bravo*[1], Picard Marion[2], Jallet Arthur[2], Borvetö Fanni[2], and Decourcelle Mathilde[3]

[1]CNRS – MIVEGEC – CNRS – France
[2]CNRS – MIVEGEC – CNRS – France
[3]Pôle Protéomique Montpellier (FPP) – CNRS : UMRUAR3426 BioCampus – France

Résumé

Codon Usage Bias (CUBias) of a gene strongly modulates its own expression (cis-acting effects) and may lead to important changes in the cellular homeostasis (trans-acting effects). Integrative studies quantifying the phenotypic impact of CUBias across molecular and cellular levels are lacking, especially in human cells.
We generated and independently expressed six synonymous versions of the *shble* antibiotic resistance gene fused to a fluorescent reporter, and performed experimental evolution on human cultured cells over hundred generations under three different selection regimes. Our experimental design focused on differences during translation elongation. Multiscale phenotype was assessed by means of: i) overall transcriptomic and proteomic analysis; ii) cellular fluorescence, as a proxy for single-cell level expression; and iii) real-time cell proliferation, as a proxy for cell fitness.

Differences in CUBias strongly impact the molecular and cellular phenotypes: i) for the focal gene, they result in large differences in mRNA and in protein levels; ii) they introduce non-predicted splicing events; iii) they lead to reproducible phenotypic heterogeneity; iv) they lead to a trade-off between the benefit of antibiotic resistance and the burden of heterologous expression; v) the global transcriptome undergoes substantial changes, not associated with the direction of the CUBias of the heterologous sequence; vi) the global proteome presented only modest changes.

Upon evolution, we communicate a remarkable and rapid convergence towards similar expression phenotypes, independently of the CUBias of the heterologous gene and without any synonymous mutation in the sequence of the focal gene, suggesting that cis-mediated effects of CUBias can be leveraged by evolution and/or plasticity. In the absence of selection pressure, gene silencing through different mechanisms evolves in certain populations, possibly benefiting of a fitness advantage by limiting overexpression. We identify significant changes in the transcriptome over time, but leading to very limited impact on the proteome, suggesting that immediate trans-acting effects of CUBias impose only a moderate effect on human

*Intervenant

cells in culture.
In human cells in culture, changes in CUBias can lead to important *cis*- acting effects in gene expression, leading to differences in protein levels and eventually eliciting phenotypic differences, but cellular homeostasis can buffer the phenotypic impact of overexpression of heterologous genes with extreme CUBias.

Phyloformer: Towards fast and accurate phylogeny reconstruction with self-attention networks

Luca Nesterenko*[1,2,3]

[1]Luca Nesterenko – Univ Lyon, Université Claude Bernard Lyon 1, CNRS UMR 5558, LBBE, F-69100, Villeurbanne, France – France

[2]Bastien Boussau – Univ Lyon, Université Claude Bernard Lyon 1, CNRS UMR 5558, LBBE, F-69100, Villeurbanne, France – France

[3]Laurent Jacob – Univ Lyon, Université Claude Bernard Lyon 1, CNRS UMR 5558, LBBE, F-69100, Villeurbanne, France – France

Résumé

An important problem in molecular evolution is that of phylogenetic reconstruction, that is, given a set of sequences descending from a common ancestor, the reconstruction of the binary tree describing their evolution from the latter. State-of-the-art methods for the task, namely Maximum likelihood and Bayesian inference, have a high computational cost, which limits their usability on large datasets. Recently researchers have begun investigating deep learning approaches to the problem but so far these attempts have been limited to the reconstruction of quartet tree topologies, addressing phylogenetic reconstruction as a classification problem. We present here a radically different approach with a transformer-based network architecture that, given a multiple sequence alignment, predicts all the pairwise evolutionary distances between the sequences, which in turn allow us to accurately reconstruct the tree topology with standard distance-based algorithms. The architecture and its high degree of parameter sharing allow us to apply the same network to alignments of arbitrary size, both in the number of sequences and in their length. We evaluate our network Phyloformer on several types of simulations and find that its accuracy competes with that of a Maximum Likelihood method while being significantly faster.

In our work, we rely on a supervised learning based paradigm for phylogenetic reconstruction which exploits the fact that sampling data under probabilistic models of sequence evolution is computationally cheap, even in cases where maximizing the likelihood under these models is expensive. Accordingly, large numbers of pairs of phylogenetic trees and multiple sequence alignments (MSAs) evolved along these trees can be generated, and used, via a supervised learning approach, to learn a function which takes an MSA as input and outputs the phylogenetic tree. Learning this function can be a costly process, but once done the function can be used to reconstruct a tree from an MSA very rapidly, regardless of the complexity of the model of sequence evolution that was used to generate the training examples.

In order to parameterize this function, we propose a self-attention network that progressively updates via learnable functions, the representation of each pair of sequences in the input alignment, alternatively focusing on different pairs and on different sites and thus

*Intervenant

getting, through several layers, an increasingly predictive estimate of all the pairwise evolutionary distances. The interactions across different sequence pairs enable the network to build context aware representations allowing to exploit all the information contained in the input MSA and overcoming, with this joint prediction, the drawback of classical distance-based methods which usually take in input independently computed distance estimates.

Investigating the impact of random genetic drift on synonymous codon usage in metazoans

Florian Bénitière*†[1], Laurent Duret‡[2], and Tristan Lefébure[3]

[1]Laboratoire de Biométrie et Biologie Evolutive - UMR 5558 – Université Claude Bernard Lyon 1, Institut National de Recherche en Informatique et en Automatique, VetAgro Sup - Institut national d'enseignement supérieur et de recherche en alimentation, santé animale, sciences agronomiques et de l'environnement, Centre National de la Recherche Scientifique : UMR5558 – France

[2]Laboratoire de Biométrie et Biologie Evolutive (LBBE) – CNRS : UMR5558, Université Claude Bernard - Lyon I, INRIA – 43 Bld du 11 Novembre 1918 69622 VILLEURBANNE CEDEX, France

[3]Laboratoire d'Ecologie des Hydrosystèmes Naturels et Anthropisés (LEHNA) – CNRS : UMR5023, Université Claude Bernard - Lyon I – France

Résumé

The genetic code contains 61 codons for 20 amino acids. The different synonymous codons that encode the same amino-acid are not uniformly used. These biases in synonymous codons usage (SCU) may result from the combination of two types of processes : i) non-adaptive processes, such as mutational biases, that affect the nucleotide composition of the entire genome, and ii) translational selection, *i.e.* selection favoring codons that are decoded by the most abundant tRNAs, thereby optimizing the speed and the accuracy of the translation. The intensity of translational selection increases with gene expression : the more a codon is translated, the more its optimization will impact the fitness of the organism. Population genetics principles state that the ability of selection to promote beneficial mutations or eliminate deleterious mutations depends on the intensity of selection (s) relative to the power of random genetic drift (defined by the effective population size, Ne). Thus, random drift is expected to set an upper limit on the capacity of selection to favor usage of optimal synonymous codons. To test this prediction, we analyzed the relationship between variations in synonymous codon usage and drift across metazoans. For this, we selected 262 metazoans spanning a wide range of Ne for which a complete genome assembly and gene annotations were published. In each species, we quantified gene expression level based on RNAseq data, and we annotated tRNA genes in the genome assembly. It has been shown in human and drosophila that the number of tRNA genes in the genome is a good proxy of their abundance within cells. In our dataset, we observed that 94% of species show a significant positive correlation between the number of tRNA genes in their genome and the abundance of the corresponding amino-acids in their proteome (weighted by gene expression level). This implies that in metazoans, there is generally a strong constraint on tRNA abundance to match the demand in amino-acid usage. For each amino-acid, we determined which of its synonymous codons are predicted to be optimal, based on the relative abundance of the corresponding tRNAs. We then quantified the intensity of selection on SCU

*Intervenant

†Auteur correspondant: florian.benitiere@univ-lyon1.fr

‡Auteur correspondant: Laurent.Duret@univ-lyon1.fr

within each species, by analyzing variation in the frequency of optimal codons according to gene expression level. We controlled for possible variation in mutational biases, by analyzing the frequency of corresponding triplets in introns. We found that signature of translational selection vary widely across metazoans, being very strong in nematodes and dipteran insects, very weak in vertebrates. The intensity of selection on SCU appears to be positively correlated with different proxies of Ne. However, we noted a strong phylogenetic inertia on SCU, which precludes drawing clear conclusion regarding the significance of these correlations. We will discuss different factors that may drive the strong variation in SCU across metazoans.

COCOATree: benchmarking coevolution methods for sector identification

Margaux Jullien*[1], Sophie Abby[1], Nelle Varoquaux[1], and Junier Ivan[1]

[1]Translational microbial Evolution and Engineering – Translational Innovation in Medicine and Complexity / Recherche Translationnelle et Innovation en Médecine et Complexité - UMR 5525 – France

Résumé

Progress in sequencing has caused an explosion of available genome sequences, allowing the study of large multiple sequence alignments (MSAs) of families of homologous proteins. These studies provide key information about the function and structure of proteins. In particular, coevolution information between pairs of residues can help not only predicting structural contacts but also identifying functional domains known as protein sectors (Halabi et al. 2009).
Several metrics have been used to quantify coevolution trend between pairs of residues, including covariance methods and mutual information. The precise identification of coevolving residues is also known to be limited by at least three factors: (1) heterogeneity of residue conservation, (2) phylogeny bias due to the over-representation of similar sequences in the MSA, and (3) background noise bias due to a low number of sequences. Various corrections have thus been developed to counter these biases.

In this context, we introduce COCOATree, a Python package which encompasses different coevolution metrics and corrections, thus allowing for a comprehensive benchmarking of the different methods proposed to identify protein sectors. In addition, COCOATree aims at systematically mapping sector composition on phylogenetic trees, establishing a link between phylogenetics and statistical approaches of residue coevolution. The package will also include bootstrap procedures in order to quantify the relevancy of results.

As an application, we will show results obtained for the family of flavin-containing monooxygenases involved in the biosynthesis of ubiquinone (UQ). This family of hydroxylases is characterized by a broad diversity of regioselectivities, with enzymes capable of hydroxylating one, two or three positions of the UQ aromatic cycle. An analysis coupling sector identification and phylogenetic information should thus provide insights dictating the functioning and evolution of these enzymes.

Keywords: coevolution, protein sectors, method benchmarking, enzyme evolution, ubiquinone pathway

References:
Halabi, Najeeb, Olivier Rivoire, Stanislas Leibler, and Rama Ranganathan. 2009. "Protein Sectors: Evolutionary Units of Three-Dimensional Structure." *Cell* 138 (4): 774–86. https://doi.org/10.1016/j.cell.2009.07.038.

*Intervenant

Reproducible workflows to study conservation of genomic sequence using multispecies whole genome alignments

Romain Feron*[1,2]

[1]Université de Lausanne – Suisse
[2]Swiss Institute of Bioinformatics – Suisse

Résumé

Thanks to ambitious initiatives like the Earth BioGenome Project and the Darwin Tree of Life aiming to coordinate the sequencing of all identified eukaryotic species, a growing number of high-quality assemblies is continuously filling the current taxonomic gaps in available genome sequences. These assemblies enable the study of sequence conservation across entire genomes in progressively more diverse taxa, at different evolutionary scales, and with an increasingly finer resolution. Conserved sequences are likely to be under selective constraints and therefore to play key functional and biological roles in their respective taxa. In practice, these sequences can be identified from multispecies whole genome alignments (MWGA), which have been successfully used to uncover new functional elements and to identify signatures of selection in several species. Here, we present a workflow to generate a MWGA from a set of assemblies and perform standard analyses on the resulting alignment, including computing basic alignment metrics, PhastCons and PhyloCSF scores, identifying conserved elements, and generating a GenomeBrowser track hub. This workflow implements the steps used to compute the majority of MWGAs published so far (e.g. alignments available at the UCSC Genome Browser) in an open, reproducible, portable, and scalable manner by leveraging the many features of the workflow management language and engine Snakemake. With this workflow, we generated MWGAs for several arthropod clades and we are now using these alignments to quantify nucleotide-level sequence conservation and explore which features and functions are associated with high or low levels of conservation. In the future, we plan to generate alignments and sequence conservation metrics for an increasing number of arthropod clades. In order to select high quality assemblies to include in the alignments, and to be able to integrate new and updated high-quality assemblies over time, we developed an assembly quality assessment catalog, the A3Cat. This resource provides BUSCO scores and NCBI GenBank metadata for all available Arthropod assemblies and is updated regularly thanks to a collection of clade-agnostic workflows. We hope that the resources and reproducible workflows we developed will help cataloging the current state of available assemblies to guide sequencing efforts, and to leverage the influx of high-quality assemblies to perform large scale comparative analyses.

*Intervenant

MPS-Sampling (Multi Proteins Similarity Sampling) to select evolutionary significant representative genomes from large databases

Rémi-Vinh Coudert*[1,2], Céline Brochier[1], Jean-Pierre Flandrois[3], Jean-Philippe Charrier[2], and Frédéric Jauffrit[4]

[1]Laboratoire de Biométrie et Biologie Evolutive - UMR 5558 – Université Claude Bernard Lyon 1, Institut National de Recherche en Informatique et en Automatique, VetAgro Sup - Institut national d'enseignement supérieur et de recherche en alimentation, santé animale, sciences agronomiques et de l'environnement, Centre National de la Recherche Scientifique : UMR5558 – France

[2]Microbiology Research Department, bioMérieux S.A. – BIOMERIEUX – France

[3]Laboratoire de Biométrie et Biologie Evolutive – CNRS : UMR5558, Université Claude Bernard - Lyon I (UCBL), INRIA – France

[4]Microbiology Research Department, bioMérieux S.A. – BIOMERIEUX – 376 Chemin de l'Orme, Marcy l'Etoile, France

Résumé

The recent burst of genome sequencing provides a wealth of data and an ever increasing access to genetic information. However, available genomic data is growing in an unbalanced way, both in terms of quality, as most sequenced genomes are in fact draft assemblies, and in terms of diversity, with the over representation of a few species and taxonomic groups, reflecting medical, agronomic or industrial considerations. This huge amount of sequence data makes exhaustive analyzes complicated, costly in computation time, or even technically impossible. For these reasons, most studies rely on subsamples of available genomic data. Current approaches can be classified into three main categories: taxonomy based, whole genome similarity based and phylogeny based. However, none of these approaches is suitable to select representative genomes from very large datasets. Taxonomy-based methods are highly sensitive to incomplete or invalid affiliation, whole genome similarity-based methods are efficient at the genus/species level but not above, while the quality of the phylogenetic trees inferred with large datasets is usually very low.
To address these limitations, we present MPS Sampling (Multiple Protein Similarity Sampling): a fast, scalable, and efficient method for the selection of reliable representative genomes from large datasets. MPS Sampling is ab initio and does not rely on environmental, academic, historical or cultural criteria. Using a set of single copy protein families (e.g. core proteins, ribosomal proteins), MPS-Sampling performs three successive steps of clustering to constitute homogenous groups of genomes . Then, a representative genome is chosen per group based on quality and centrality criteria.

MPS Sampling was tested on 158,027 bacterial genomes using 47 ribosomal proteins available in RiboDB, to obtain samples of variable density. The resulting samples have been

*Intervenant

investigated and compared with taxonomy and phylogeny, both globally (Bacteria) and at a lower scale on three major taxonomic families (Lactobacillaceae, Bacillaceae, Enterobacteriaceae). The genomes chosen by MPS Sampling were always coherent with both taxonomic and phylogenetic diversity, demonstrating that MPS Sampling was particularly suitable to obtain reliable samples of representative genomes.

Evolutionary signal captured from topological properties of proteins via persistent homology

Lea Bou Dagher*[1], Céline Brochier-Armanet†[2], and Philippe Malbos‡[3]

[1]Boudagher – Univ Lyon, Université Claude Bernard Lyon 1, CNRS UMR 5208, Institut Camille Jordan, 43 blvd. du 11 novembre 1918, F-6962 Villeurbanne Cedex, France-0, CNRS, VetAgro Sup, Laboratoire de Biométrie et Biologie Evolutive (LBBE), UMR CNRS 5558, F-69100, Villeurbanne – France

[2]Brochier-Armanet – CNRS, VetAgro Sup, Laboratoire de Biométrie et Biologie Evolutive (LBBE), UMR CNRS 5558, F-69100, Villeurbanne – France

[3]Malbos – Univ Lyon, Universite Claude Bernard Lyon 1, CNRS UMR 5208, Institut Camille Jordan, 43 blvd. du 11 novembre 1918, F-6962 Villeurbanne Cedex, France – France

Résumé

Proteins are essential components of biological processes. Furthermore, their primary sequences carry an historical (phylogenetic) signal is routinely used to infer the evolutionary history of proteins and organisms. However, phylogenetic approaches based on the analysis of primary sequences have limitations (1). Here, we explore the potential of methods from topological data analysis (TDA) (2) to analyze the information contained in protein 3D structures. Specifically, we apply persistent homology (PH) (3), a tool of TDA, to describe and compare the topological properties of protein 3D structures. Our working hypothesis is that variations in the topological characteristics of proteins could provide information on the evolution of proteins, both quantitatively and qualitatively. This approach is original because until now, most applications of PH in biology have been used to classify/identify biological objects (proteins, tumors...) (4), but for evolutionary studies. To test our hypothesis, we have investigated a large collection of protein 3D structures using three types of topological distances. We show that these distances are strongly correlated with phylogenetic distances estimated from the primary protein sequences, which means that they could represent a new way to catch the evolutionary signal contained in proteins.

(1): Phylogenomics and the reconstruction of the tree of life. Delsuc F., Brinkmann H., Philippe H. Nat Rev Genet. 2005 May; 6(5):361-75. doi: 10.1038/nrg1603. PMID: 15861208

(2): Topology and data. Carlsson, G. Bull. Amer. Math. Soc. (N.S.). 2009 April; 46(2):255–308.

(3): Topological persistence and simplification. Edelsbrunner H., Letscher D., and Zomorodian A. Discrete Comput. Geom. 2002 Nov; 28(4):511–533.

(4): Topology based data analysis identifies a subgroup of breast cancers with a unique

*Intervenant

†Auteur correspondant: celine.brochier-armanet@univ-lyon1.fr

‡Auteur correspondant: malbos@math.univ-lyon.fr

mutational profile and excellent survival. Nicolau M., Levine A., Carlsson G. PNAS. 2011 Feb; 108(17):7265–7270.

3

Genome Evolution

Evolution of host-gut microbiome interactions in the context of industrialization

Mathieu Groussin*†[1], Mathilde Poyet[2], and The Gmbc Consortium[3]

[1]Kiel University – Allemagne
[2]Kiel University – Allemagne
[3]OpenBiome – États-Unis

Résumé

Concurrent with industrialization, the human gut microbiome has drastically decreased in diversity and shifted in composition. However, to what extent transitioning from hunter-gatherer to industrialized lifestyles impacted host-microbiome interactions and host physiology is unknown. Here, we generate paired human and gut microbiome multi omics data associated with host physiology and metadata from dozens of populations worldwide, ranging from hunter-gatherers to fully industrialized groups, and show that microbiome compositions, microbial functions, intestinal inflammation, humoral immune response and patterns of horizontal gene transfers (HGT) between bacteria strongly changed with industrialization. Overall, our results suggest that industrialization perturbed our gut ecosystem and homeostasis on many levels, which could contribute to the rise of chronic inflammation diseases observed worldwide.

*Intervenant
†Auteur correspondant: m.groussin@ikmb.uni-kiel.de

Understanding the evolutionary dynamics of insertion sequences in prokaryotic genomes

Flora Gaudilliere*†[1], Sophie Abby[1], Thomas Hindré[1], and Ivan Junier‡[1]

[1]Techniques de lÍngénierie Médicale et de la Complexité - Informatique, Mathématiques et Applications, Grenoble - UMR 5525 – Institut Polytechnique de Grenoble - Grenoble Institute of Technology, Centre National de la Recherche Scientifique : UMR5525, Université Grenoble Alpes – France

Résumé

Insertion sequences (IS) are the simplest prokaryotic transposable elements: they usually only code for a transposase, the enzyme that catalyses their movement within genomes. The transposition of IS elements within a genome can have drastic effects on gene expression that are often deleterious for the cell's fitness: as a consequence, IS are mostly considered as genomic parasites. However, IS transposition can also be a vector of adaptation. For instance, the acquisition of antibiotic resistance genes is often linked to IS movements. Our project focuses on the evolutionary dynamics of insertion sequences in prokaryotic genomes. We use automated detection tools to look for insertion sequences in a wide range of prokaryotic species and study the distribution of IS families across phyla. We then try to model these distributions to see which minimal factors can be invoked to explain them. One of the key questions is to determine whether purifying selection plays a role in the evolutionary dynamics of insertion sequences, or if a neutral dynamic is enough to recapitulate their distribution.

*Intervenant

†Auteur correspondant: flora.gaudilliere@univ-grenoble-alpes.fr

‡Auteur correspondant: ivan.junier@univ-grenoble-alpes.fr

Evolution of natural transformation in bacteria

Fanny Mazzamurro*[1,2], Jason Chirakadavil[3], Christophe Ginevra[4], Sophie Jarraud[4], Xavier Charpentier[3], and Eduardo P. C. Rocha†[1]

[1]Microbial Evolutionary Genomics – Institut Pasteur, Université de Paris Cité, CNRS UMR3525, F-75015 Paris, France – France

[2]Collège Doctoral – Sorbonne Université, F-75005 Paris, France – France

[3]CIRI, Centre International de Recherche en Infectiologie – Inserm, U1111, Université Claude Bernard Lyon 1, CNRS, UMR5308, École Normale Supérieure de Lyon, Univ Lyon, 69100, Villeurbanne, France – France

[4]Centre national de Référence des Légionelles – Centre de biologie Nord, 103 grande-rue de la Croix-Rousse 69317 LYON Cedex 04, France – France

Résumé

Natural transformation is a key mechanism of gene exchange between bacteria. While it was the mechanism that allowed to demonstrate that DNA is the support of genetic information almost a century ago, its existence and evolutionary benefits remain debated. Hypothesis for the latter include acquisition of novel functions, DNA repair, and removal of mobile genetic elements. In the light of the latter hypothesis, mobile genetic elements may respond and defend themselves by impairing transformation. Despite the competence genes being conserved in the bacterial core genomes and transformation being widely spread among bacteria, very large within-species variations in natural transformation frequencies have been observed in several species. The source of these variations remains unknown and could provide important clues on the benefits and costs of natural transformation. We built two representative collections of hundreds of genomes from two naturally transformable bacterial species, which present important within-species variations in their experimentally measured transformation rates. We analyzed these genomes, built core and pangenomes, and inferred phylogenetic trees with and without recombination. We then used the trees to test different evolutionary models that might explain the observed distribution of transformation frequencies. Our results are in favor of evolution by jumps, since extant strains show sudden and recent acquisition of genetic determinants of transformation. We searched to identify these determinants using a variety of computational methods. We first characterized the mobile genetic elements and bacterial defense systems susceptible to interact with them. We then conducted unitig based GWAS on these elements to identify potential transformation-associated genetic elements. Our results highlight the intimate link between mobile genetic elements and the variation of transformation rates.

*Intervenant

†Auteur correspondant: eduardo.rocha@pasteur.fr

Unraveling the relative emergence of quinones biosynthetic pathways

Sophie-Carole Chobert*[1], Ivan Junier[1], Fabien Pierrel[1], and Sophie Abby[1]

[1]Laboratoire TIMC - Equipe TrEE – Univ. Grenoble Alpes, CNRS, UMR 5525, VetAgro Sup, Grenoble INP, TIMC, 38000 Grenoble, France – France

Résumé

The Great Oxidation Event is thought to have had a profound impact on the evolution of bioenergetics on Earth. We aim at getting more insight regarding how bacteria adapted their energetic metabolism to this massive change. We focus on a family of molecules with a key role in respiratory chains of most living organisms: the isoprenoid quinones (hereafter simply referred to as quinones). In respiratory chains, quinones shuttle electrons and protons between proteins in the membrane. According to their mid-point redox potential, quinones can be classified as low potential (LP) such as menaquinone (MK) or high potential (HP) quinones such as ubiquinone (UQ). The traditional view is to consider that LP quinones are involved in anaerobic processes and that HP quinones emerged to cope with rising O2 levels. We believe that quinone biosynthetic pathways have captured capital information regarding the evolution of energy metabolisms. Several pathways can lead to the production of the same quinone, and enzymes from different pathways often share homologies. Our group recently showed that UQ, a HP quinone only present in Proteobacteria and Eukaryotes, is synthesized via two biosynthetic pathways: the classical O2-dependent one, and the O2-independent that uses a source of oxygen other than O2 (1).

To reconstitute the evolutionary history of the quinone pathways, we combined several approaches: 1) analysis of the pathways and enzymes distribution; 2) follow up of the dynamics of genetic architecture of the pathways; 3) phylogenetic approaches. Our first observations in Proteobacteria showed a widespread distribution of the O2-independent pathway across this phyla which indicates that it could have existed in the common ancestor of UQ-producing Proteobacteria. This would suggest that high O2 levels in the atmosphere were not a prerequisite for the emergence of UQ (1). In addition, we recently showed that UQ produced by the O2-independent pathway can be used for anaerobic respiration (nitrate respiration in *Pseudomonas aeruginosa)* (2). Our systematic investigation of quinone biosynthetic pathways across bacterial genomes allowed us to gather evidence to decipher their relative order of appearance and to get a better understanding of their respective role. Altogether, our study questions the relative timing between the appearance of HP quinones in the context of rising O2 levels.

(1) L. Pelosi, et al., Ubiquinone Biosynthesis over the Entire O2 Range: Characterization of a Conserved O2-Independent Pathway, MBio. 10 (2019) 21.
(2) C.-D.-T. Vo, et al., The O2-independent pathway of ubiquinone biosynthesis is essential for denitrification in Pseudomonas aeruginosa, Journal of Biological Chemistry. 295 (2020) 9021–9032. https://doi.org/10.1074/jbc.RA120.013748.

*Intervenant

Causes of discord in eukaryotic protein domains inherited from Archaea.

Guillaume Louvel*[1] and Laura Eme[2]

[1]Ecologie Systématique et Evolution – AgroParisTech, Université Paris-Saclay, Centre National de la Recherche Scientifique : UMR8079 – France

[2]Ecologie Systématique et Evolution – Université Paris-Sud - Paris 11, AgroParisTech, Centre National de la Recherche Scientifique : UMR8079 – bat. 360 91405 ORSAY CEDEX, France

Résumé

Eukaryotes descend from at least two prokaryotic ancestors: the alphaprotebacterium endosymbiont that gave rise to mitochondria, and an archaeal parent, that donated most components of the genetic machinery. However, over the last several decades, phylogenetic analyses gave conflicting results regarding the precise placement of eukaryotes with regard to Archaea.
The development of more complex evolutionary models –better able to model evolutionary processes over extremely long periods of time–, as well as the discovery of novel archaeal lineages, led to a stronger consensus in the last few years. In particular, Asgard archaea, a superphylum identified primarily through metagenomics, appear to represent the closest lineage to eukaryotes in recent phylogenomic analyses.

However, in the sequence-based strategy for inferring species relationships, discord among individual genes is pervasive: different genes support different trees. This can be due to many reasons: real incongruence caused by horizontal gene transfer, hybridization and incomplete lineage sorting, or by sequence data contamination, or, more commonly, because of methodological artefacts, such as hidden patterns of duplications and losses, incorrect ortholog grouping, statistical uncertainty (erosion of phylogenetic signal), or inadequate sequence evolution model.

In this project, we aim to identify causes for incongruence in a set of eukaryotic protein domains originating from either TACK or Asgard archaea, as identified by a previous study (1). In particular, we aim to identify characteristics of the proteins (function, length, aminoacid composition...), alignments (degree of conservation, length...), and tree topologies (branch length, tree length, branch support...) that correlate with the nature of the sister-group of eukaryotes supported by each marker.
(1): Vosseberg et al 2021

*Intervenant

Cell type diversification across paleo– and neocortex revealed by single cell multiomics analysis

Sara Zeppilli*[1], Robin Attey[1], Alonso Ortega Gurrola[2], Pinar Demetci[3], Ritambhara Singh[3], Maria Antonietta Tosches[2], Anton Crombach†‡[4], and Alexander Fleischmann§[3]

[1]Brown University – États-Unis
[2]Columbia University – États-Unis
[3]Brown University – États-Unis
[4]INRIA – Univ Lyon, ENS de Lyon, CNRS, Inria, UCB Lyon 1, IXXI, LIP, 69342 Lyon, France – France

Résumé

The mammalian cortex comprises ancient paleocortical structures, exemplified by the olfactory cortex, as well as more recently evolved neocortical areas. Neurons in the mouse olfactory cortex differ from neurons in neocortex in their developmental origin, morphology, functional properties and molecular identities. However, the mechanisms driving cell type diversification across these distinct cortical traits remain unknown.
Here, we characterize the gene regulatory network (GRN) that defines cell type identity in the mouse olfactory cortex, and we compare GRN activity across paleo-, peripaleo- and neocortical areas. Using single cell ATAC and RNA sequencing, we reveal enhancer-gene interactions and identify combinatorial transcription factor activity as a key mechanism for cell type diversification. Finally, by comparing mouse paleo- and neocortical neurons to single cell data from turtle, lizard, and salamander, we propose semilunar cells of the olfactory cortex as the ancestral neuronal cell type of the mammalian cortex.
Our data provide the first comprehensive molecular description of cell types in the mouse olfactory cortex and identify regulatory mechanisms underlying cell type diversification during evolution.

*Auteur correspondant: sara_zeppilli@brown.edu
†Intervenant
‡Auteur correspondant: anton.crombach@inria.fr
§Auteur correspondant: alexander_fleischmann@brown.edu

Beyond one-gain models for pangenome evolution

Jasmine Gamblin*[1], François Blanquart[1,2], and Amaury Lambert[1,3]

[1]Centre interdisciplinaire de recherche en biologie – Labex MemoLife, Collège de France, Centre National de la Recherche Scientifique : UMR7241, Institut National de la Santé et de la Recherche Médicale : U1050 – France

[2]Infection, Anti-microbiens, Modélisation, Evolution – Institut National de la Santé et de la Recherche Médicale : U1137, Université Paris Cité : $UMR_{S1}137, Université Sorbonne Paris nord - - France$

[3]Institut de Biologie de l'ENS Paris – Département de Biologie - ENS Paris, Institut National de la Santé et de la Recherche Médicale : U1024, Centre National de la Recherche Scientifique : UMR8197 – France

Résumé

A species pangenome is the set of all genes carried by at least one representant of the species. In bacteriae, pangenomes can be much larger than the set of genes carried by one individual. Many questions remain unanswered regarding the evolutive forces shaping these bacterial pangenomes. One of them is to explain the U-shape of the gene frequency spectrum: there are more genes present in very few or almost all genomes than at intermediate frequencies. Two papers from 2012 (Baumdicker et al. and Heageman and Weitz) explained this distribution with stochastic models allowing genes to be gained only once along the species phylogeny. However the importance of intra-specific horizontal gene transfer (HGT) in many bacterial species calls for more complex models. Using a dataset of 436 commensal E.coli genomes, we show that a model with only one gain per gene is not able to reproduce the patterns of presence/absence of genes at the leaves of the phylogeny. We thus introduce a new model of pangenome evolution including a category of genes that can be gained and lost in the phylogeny multiple times, interpreted as genes undergoing frequent HGT. Both the gene frequency spectrum and the presence/absence patterns are reproduced more accurately.

*Intervenant

Epistatic interactions between genetic background and antibiotic resistances genes (and vice versa)

Charles Coluzzi*[1] and Eduardo Rocha[1]

[1]Génomique évolutive des Microbes / Microbial Evolutionary Genomics – Institut Pasteur [Paris], Centre National de la Recherche Scientifique : UMR3525, Université Paris Cité – France

Résumé

Bacterial populations can adapt quickly to environmental challenges. This process is often concomitant with genome modifications, which can be endogenous (e.g. mutations) or exogenous (e.g. horizontal gene transfer). Antibiotic resistance is a perfect example of this process since it can arise in both ways. However, if we often attribute the antibiotic resistance phenotype to well described genetic markers, these are actually the result of a subsequent, stepwise evolutionary process. Furthermore, epidemiological data show that some lineages are more prone to acquire such resistances than others, suggesting that adaptation requires adequate genetic backgrounds.
Here, we inferred the correlated evolutionary events leading to antibiotic resistance in two different contexts using a likelihood framework. In a first analysis of 1600 E. coli genomes, we tested if specific genes were horizontally acquired prior the acquisition of point mutations leading to fluoroquinolone resistance. We found hundreds of genes frequently acquired prior the acquisition of the fluoroquinolone resistance. In a second analysis of 757 ST410 *E. coli* genomes, we searched for point mutations acquired prior the horizontal acquisition of gene-conferring beta-lactam resistance. We found that some point mutations were systematically acquired before the beta-lactam genes.
This study shows how the acquisition of horizontally acquired genes potentiates future phenotypic evolution by point mutation. But also, how the evolution of the genetic background by simple point mutations favors the acquisition of new functions by horizontal gene transfer.

*Intervenant

Bridging the gap between population genomic and phylogenetic approaches by the study of the effective population size

Mélodie Bastian*†[1] and Nicolas Lartillot[2]

[1]Bioinformatique, phylogénie et génomique évolutive – Département PEGASE [LBBE] – France
[2]Laboratoire de Biométrie et Biologie Evolutive - UMR 5558 – Université Claude Bernard Lyon 1, Université de Lyon, Institut National de Recherche en Informatique et en Automatique, VetAgro Sup - Institut national d'enseignement supérieur et de recherche en alimentation, santé animale, sciences agronomiques et de l'environnement, Centre National de la Recherche Scientifique : UMR5558 – France

Résumé

The study of the different evolutionary forces, their mechanisms, and their impact on the shape of genomes is a vast research field. Among these different forces, a very interesting one is genetic drift, which is inversely proportional to the number of breeders in an ideal population, i.e. the effective size (Ne). Genetic drift can impact both short and long-term evolutionary processes (estimated from polymorphism and divergence data, respectively), which concern different disciplines of study with different proper methods.
The availability of an increasing amount of genetic data allows us to contrast these intraspecific and interspecific data to bridge the gap between population genomic and phylogenetic studies.

In this study, I estimated variations in Ne between 180 species based on synonymous polymorphism estimated from a set of 7200 orthologous genes, correcting for variations in mutation rate (μ), along the mammalian phylogeny in order to study correlations between Ne, ecological traits, and molecular traits such as selection intensity (both long-term, based on dN/dS, and short-term, based on pN/pS).
For this purpose, I devised a pipeline from the recovery of orthologous gene sequences, ecological and polymorphism data to Bayesian integrative analysis, aimed at reconstructing Ne using a multivariate process (1).
As found previously by other Authors, I recover a positive correlation between dN/dS and life history traits, which has been classically interpreted as a negative correlation of both variables with long-term Ne. I also observe a correlation between synonymous polymorphism and pN/pS, which is compatible with an impact of Ne on the variation in selection efficiency at the population scale. On the other hand, I did not find a significant correlation between synonymous polymorphism and life history traits or dN/dS. This observation suggests the possibility that short-term Ne (from population genomic data) and long-term Ne (from phylogenomic data), might be partially decoupled, as also discussed in (2).

*Intervenant
†Auteur correspondant: melodie.bastian@univ-lyon1.fr

1.Brevet, M. & Lartillot, N. Reconstructing the history of variation in effective population size along phylogenies. 793059 https://www.biorxiv.org/content/10.1101/793059v4 (2021) doi:10.1101/793059.
2.Müller, R., Kaj, I. & Mugal, C. F. A Nearly-Neutral Model of Molecular Signatures of Natural Selection after Change in Population Size. Genome Biology and Evolution evac058 (2022) doi:10.1093/gbe/evac058.

4

Population Genetics II

Opening Motoo Kimura's archives: on the history of molecular evolution and the neutralist school

Jean-Baptiste Grodwohl*[1]

[1]Laboratoire SPHERE UMR 7219 – CNRS : UMR7219 – France

Résumé

This talk considers the history of molecular evolution from the vantage point of a historian of science. I have recently consulted Motoo Kimura's archive, which shed much insight into the birth of the neutral theory. Using these sources, I will discuss the following questions: (1) Why did Kimura take interest in molecular evolution in the 1960s and why did he develop a neutralist hypothesis? (2) Who were the first members of the neutralist school and what were their relationships? (3) What were the major difficulties faced by this hypothesis over the 1970s and how did neutralists cope with them? I will then briefly compare the respective fates of neutralists and of neutral models in the 1980s.

*Intervenant

Adaptive walks don't do walks on hypercubes

Leonardo Trujillo*†[1], Paul Banse[2], and Guillaume Beslon[2]

[1]Inria Lyon – Institut National de Recherche en Informatique et en Automatique – France
[2]Inria Lyon – Institut National des Sciences Appliquées (INSA) - Lyon – France

Résumé

Molecular evolution is often conceptualised as adaptive walks on rugged fitness landscapes, driven by mutations and constrained by incremental fitness selection. The defining property of fitness landscapes are the genotype spaces in which they reside. This space is generally thought of as a hypercube in which each genotype is a vertex and the edges connect neighbouring sequences that differ by one point of mutation. In this talk, in addition to the widely used point mutations, we present a minimal model of sequence inversions to simulate adaptive walks. We use the well known NK model to instantiate rugged landscapes. We show that adaptive walks can reach higher fitness values through inversion mutations, which, compared to point mutations, allows the evolutionary process to escape local fitness peaks. To elucidate the effects of this chromosome rearrangement, we use a theoretical graph representation of accessible mutants and show how new evolutionary pathways are uncovered. Thus we show that adaptive walks don't do walks on hypercubes but in more connected and complex networks. Our model suggests a simple mechanistic rationale to analyse escapes from local fitness peaks in molecular evolution driven by (intragenic) structural inversions and reveals some consequences of the limits of point mutations for simulations of molecular evolution.

*Intervenant
†Auteur correspondant: leonardo.trujillo-lugo@inria.fr

In search of islands of speciation in the genomes of two Coenonympha butterfly sister species.

Thibaut Capblancq*[1] and Laurence Despres[2]

[1]Laboratoire d'écologie alpine (LECA) – CNRS : UMR5553, Université Joseph Fourier - Grenoble I, Université de Savoie – bat. D - Biologie 2233 Rue de la piscine - BP 53 38041 GRENOBLE CEDEX 9, France

[2]Laboratoire d'Ecologie Alpine (LECA) – Université Joseph Fourier - Grenoble I – 2233 rue de la piscine 38400 Saint-Martin d'Heres, France

Résumé

In this study we took advantage of a newly produced *Coenonympha* reference genome and whole genome re-sequencing data to explore the genomic landscape of speciation between the two butterfly species *C. arcania* and *C. gardetta*. Knowing that these two species diverged relatively recently, still hybridized occasionally in the wild and occupied two very different ecological niches, we wanted to understand if their isolation relied on a genome-wide resistance to gene flow or was dependent on a few highly divergent genomic regions that particularly drove reproductive isolation. We started our investigations by re-inferring the demographic history of speciation between the two sister species and formally tested if gene flow accompanied their divergence. We then screened the genomes of the species in search for barrier loci. We inferred multiple population genetics statistics to look for genomic regions of high differentiation, low diversity and/or high linkage among loci. In parallel, we also used coalescent inferences to locate genomic regions of reduced gene flow. All analyses were performed at the genome level but with a particular attention given to comparing the autosomes to the Z chromosome, with the goal of assessing the role of the sex chromosomes in this speciation event. Our results highlight several autosomal regions with strong evidence of selection and suggest an important role of the Z chromosome in the divergence of the two species.

*Intervenant

Mitonuclear discordance in the great white shark (Carcharodon carcharias): sex biased dispersal, mitonuclear incompatibility, or both?

Elise J Gay*[1,2], Romuald Laso-Jadart†[2], Shanon Corrigan‡[3], Lei Yang§[3], Pierre Lesturgie¶[2], Gavin J P Naylor‖[3], and Stefano Mona**[2,4]

[1]EPHE, Ecole Pratique des Hautes Etudes, Sorbonne Universités, Paris, France – EPHE – France
[2]Museum Nationnal d'Histoire Naturelle de Paris – Muséum National d'Histoire Naturelle (MNHN) – France
[3]Florida Museum of Natural History, University of Florida, Gainesville, FL, USA – États-Unis
[4]EPHE, Ecole Pratique des Hautes Etudes, Sorbonne Universités, Paris, France – EPHE – France

Résumé

The great white shark (*Carcharodon carcharias*) is an ocean-wide distributed species that lives in temperate and coastal waters. A first attempt to investigate the population genetic structure of this species highlighted a marked difference between mitochondrial and microsatellite nuclear markers, with the former being characterized by larger *Fst*values. This was suggested to be the consequence of female philopatry, a pattern which is believed to be common in many shark species. Recently, the great white shark nuclear genome has been fully assembled at the chromosome level, including sex chromosomes (unveiling a XY sex determination system). To shed more light on the mitonuclear discordance, we sequenced the whole nuclear and mitochondrial genome of 21 (including 12 males and 9 females) and 283 individuals respectively. With these data, we first refined the global population structure and inferred the most likely demographic scenario based on the nuclear genomic variation. Then, we predicted through coalescent simulations mitochondrial and Y-chromosome variation under this scenario. We found that the observed mtDNA variation could never be reproduced even when integrating female philopatry, while the observed Y-chromosome was compatible with simulated data. Finally, we scanned the nuclear genome following a sliding window approach to identify regions correlated to mtDNA variation. We found 24 candidate regions, one of which harboring OXPHOS genes, which are responsible for mitonuclear cross-talk. Our results suggests that female philopatry alone cannot explain the observed mitonuclear discordance, and we conclude that mitonuclear incompatibility plays a major role in determining the observed pattern. This calls for a reappraisal of the origin of such discordance also in other shark species.

*Intervenant
†Auteur correspondant: romuald.laso-jadart@mnhn.fr
‡Auteur correspondant: shancorrigan1@gmail.com
§Auteur correspondant: leiyangslu@gmail.com
¶Auteur correspondant: pierre.lesturgie@mnhn.fr
‖Auteur correspondant: gjpnaylor@gmail.com
**Auteur correspondant: stefano.mona@mnhn.fr

Spatially structured populations on graphs beyond update rules

Alia Abbara*[1] and Anne-Florence Bitbol[1]

[1]EPFL – Suisse

Résumé

In a well-mixed microbial population, all individuals are equally in competition, but original evolutionary dynamics might appear in a spatially structured population where competition only holds locally. In particular, a complex structure can affect the fixation of a mutant (1). Here, we consider a population on a graph, where each node holds a homogeneous sub-population, and exchanges can happen along the edges. We suggest a model in which population dynamics evolve following cycles of two steps. First, a phase of local exponential growth on each node, during which node sizes become very large. Then, a dilution step samples a given number of individuals from each node and migrations occur along edges according to migration rates. At the end of these two steps, node sizes reach their bottleneck values. Our model allows to bypass the choice of an update rule (such as birth-death or death-birth with the Moran model on graphs (2)) and provides a more universal description, encompassing large migration regimes which were not accessible in previous models where nodes had limited carrying capacities (3). We investigate the impact of the star graph, a structure known to amplify selection in birth-death models with a single individual per node (4), on the fixation of a mutant. We draw a parallel with the Wright-Fisher model to provide an analytical prediction in the regimes of extremely rare migrations. For frequent migrations, we use a branching process approximation to predict the extinction probability and the average extinction time. Our results have excellent agreement with simulations. We find that not only graph structure, but also parameter values, greatly impact the fixation probability. In particular, the star graph can display both amplification or suppression of natural selection in different migration regimes. It can also be an accelerator of evolutionary dynamics, as the initial mutant can reach fixation or extinction faster on average than in a well-mixed population with the same total size. (1) Lieberman E, Hauert C, Nowak MA. Evolutionary dynamics on graphs, Nature, 433(7023):312–316, 2005.
(2) Yagoobi S, Traulsen A. Fixation probabilities in network structured meta-populations, Sci Rep 11, 17979, 2021.
(3) Marrec L, Lamberti I, Bitbol AF. Toward a universal model for spatially structured populations, Phys. Rev. Lett. 127, 218102, 2021.
(4) Allen B, Sample C, Jencks R, Withers J, Steinhagen P, Brizuela L, et al. Transient amplifiers of selection
and reducers of fixation for death-Birth updating on graphs, PLoS Comput Biol., 16(1):e1007529, 2020.

*Intervenant

Birds demography inference based on genomic data

Thomas Forest*[1,2,3], Guillaume Achaz[4,5,6], Jérôme Fuchs[7], and Amaury Lambert[5,8]

[1]Centre interdisciplinaire de recherche en biologie – Collège de France : SMILEGroup, Institut National de la Santé et de la Recherche Médicale : U1050, Centre National de la Recherche Scientifique : UMR7241 – France

[2]Éco-Anthropologie – Museum National d'Histoire Naturelle, Centre National de la Recherche Scientifique : UMR7206 – France

[3]ISYEB (Institut de Systématique, Évolution, Biodiversité) – Museum National d'Histoire Naturelle - MNHN (FRANCE) : UMR7205 – France

[4]Atelier de BioInformatique – MNHN, CNRS, UPMC, EPHE – France

[5]Centre interdisciplinaire de recherche en biologie – Inserm : U1050, CNRS : UMR7241, Collège de France – France

[6]Musée de l'Homme – Muséum National d'Histoire Naturelle (MNHN) – France

[7]ISYEB7205 (MNHN) – Museum National d'Histoire Naturelle - MNHN (FRANCE) – France

[8]Institut de biologie de l'ÉNS Paris (UMR 8197/1024) – École normale supérieure - Paris, Université Paris sciences et lettres, Centre National de la Recherche Scientifique, Institut National de la Santé et de la Recherche Médicale : U1024, Centre National de la Recherche Scientifique : UMR8197 – France

Résumé

The quantification of demographic change is a common, yet tricky question in population genetics. Here, techniques based on genetic information (allele frequency spectra) are experienced on a diverse set of bird species from diverse ecozones (western palearctic, nearctic, afrotropical...), which present different types of demographic structure, history (decline, steady...) at different time scales. Some progress on species using genetic data will be presented, including based on the genome of the green woodpecker, *Picus viridis*, that we assembled recently. Convergences or inconsistencies with IUCN Red List status will also be discussed through these examples.

*Intervenant

Genetics and genomics help understanding the colour polymorphism in the invasive Box Tree Moth

Riccardo Poloni*[1], Charles Perrier[2], and Mathieu Joron[3]

[1]CEFE, Univ Montpellier, CNRS, EPHE, IRD, Montpellier, France, 1919 route de Mende, 34090 Montpellier, France – CEFE, Univ Montpellier, CNRS, EPHE, IRD, Montpellier, France, 1919 route de Mende, 34090 Montpellier, France, CEFE UMR 5175 – France

[2]MR CBGP, INRAE, CIRAD, IRD, Institut Agro, Univ Montpellier, Montpellier, France – Institut de Recherche pour le développement UMR CBGP – France

[3]CEFE, Univ Montpellier, CNRS, EPHE, IRD, Montpellier, France, 1919 route de Mende, 34090 Montpellier, France – CEFE UMR 5175 – France

Résumé

Polymorphism under selection allows studying phenotype-fitness relationships in a controlled environment. In some species, polymorphism plays a pivotal role in their ecology with complex co-occurring evolutionary forces. The Box Tree Moth, *Cydalima perspectalis* (Crambidae), is an Asian moth invading Europe and North America, displaying a colour polymorphism, with a white form and a melanic form. It is a serious concern for habitats where Box-Tree is a keystone species. However, the genetic and ecological determinants of this polymorphism are not known.

We investigated the molecular mechanism of colour polymorphism using a) mendelian crosses, b) GWAS, c) identification of structural variants. As a first step, we performed 84 controlled crosses to highlight dominance patterns and obtain individuals with known genotype for sequencing. Then, we produced a reference genome for a black individual and used Pool-Seq and GWAS to highlight the region controlling the white/melanic switch. We also produced two chromosome-scale assemblies, one per each colour morph, and whole-genome resequencing to detect structural variants associated with the polymorphism.

The white morph is recessive, whereas the melanic one is dominant with a very low homozygous ratio. GWAS highlighted a very clear association in chromosome 11 of our assembly, in the same region of the gene cortex, known to control polymorphism in other species of Lepidoptera, placing further back in the past the role as wing colour gene.

*Intervenant

Subtle signals of adaptive introgression in the late stages of the speciation continuum

Quentin Rougemont*[1], Barbara Huber†[2,3], and Mathieu Joron‡§[1]

[1]Centre d'Ecologie Fonctionnelle et Evolutive – Université Paul-Valéry - Montpellier 3 : UMR5175, Ecole Pratique des Hautes Etudes : UMR5175, Centre National de la Recherche Scientifique : UMR5175, Institut de Recherche pour le Développement : UMR5175, Institut National de Recherche pour l'Agriculture, l'Alimentation et l'Environnement : UMR5175, Institut Agro Montpellier, Université de Montpellier : UMR5175 – France

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

[3]Universidad de Mérida – Venezuela

Résumé

Quantifying gene flow between lineages at different stages of the speciation continuum is central to understanding of how distinct lineages emerge. *Heliconius* butterflies have undergone an adaptive radiation in wing colour patterns driven in part by natural selection for local mimicry. They are also well known for assortative mating based on wing pattern. Therefore, wing patterns are sometimes considered "magic traits" facilitating speciation. Here, we take a multifaceted approach to explore speciation and species boundaries between closely-related species *H. hecale* and *H. ismenius*. We focus our research in geographic regions where the two are mimetic and contrast this with a geographic region where the similarly co-occur but do not mimic each other. To examine population history and patterns of inter- and intraspecific gene flow, we developed a 4-population model accounting for linked selection. This model suggests that the two species have remained isolated for the majority of their history, yet with a small amount of gene exchange between the co-mimics. Accordingly, local signatures of genomic introgression were small and dispersed except at a major wing pattern allele. We combine this with behavioural assays suggesting that other cues determine strong sexual isolation. Finally, tests for chemical differentiation between species identified major differences in putative pheromones which likely mediate species recognition. Our results show that strong divergence and adaptive introgression at wing pattern loci may not reveal the whole speciation history. In a clade where assortative mating is readily triggered by Wing pattern differences may trigger assortative mating, but our results show that the accumulation of other barriers to gene flow are important in the completion of speciation

*Auteur correspondant: quentinrougemont@orange.fr

†Auteur correspondant: babahuber@gmail.com

‡Intervenant

§Auteur correspondant: mathieu.joron@cefe.cnrs.fr

Posters

Exploration of Myriapoda genomes

Dorine Merlat*†[1], Gemma Collins[2], Clément Schneider[2], Arnaud Kress[1], Peter Decker[2], Ricarda Lehmitz[2], Miklos Balint[2], and Odile Lecompte[1]

[1]Laboratoire des sciences de l'ingénieur, de l'informatique et de l'imagerie (ICube) – ENGEES, Institut National des Sciences Appliquées [INSA] - Strasbourg, université de Strasbourg, CNRS : UMR7357 – 300 bd Sébastien Brant - BP 10413 - F-67412 Illkirch Cedex, France

[2]LOEWE Centre for Translational Biodiversity Genomics (LOEWE-TBG), Senckenberg Research Institute, Frankfurt am Main, Germany – Allemagne

Résumé

The soil is one of the main reservoirs of biodiversity, with 40% of terrestrial species being associated with soil at some point in their cycle. In this ecosystem, invertebrates are important actors involved in many processes such as nutrient cycling or water regulation but they are directly threatened by anthropic actions and their consequences (1). The diversity of these soil invertebrates is still largely unknown, it is estimated that only one soil invertebrate specie out of 20 is known today. Among them, myriapods are particularly poorly studied at the genomic level with only 7 genome available. This group of small arthropods present in all terrestrial environments is divided into four classes: chilopods (centipedes), diplopods (millipedes), symphylans, and pauropods. The MetaInvert project was initiated to explore the diversity of soil invertebrates with the objective of establishing reference genomes and identifying correlations between the genomic and ecological traits of species. It allowed the sequencing of almost 300 genomes, including 50 myriapod genomes. We selected 20 myriapod genomes (10 chilopods and 10 diplopods) with a BUSCO-Complete score higher than 70% for annotation and comparative genomic analysis. To do so, we developed a tool called EXOGAP (Exotic Organism Genome Annotation Pipeline), an automated annotation pipeline adapted to non-model species with little available data. Using EXOGAP, we predicted protein-coding genes, non-coding genes, pseudogenes, and repetitive elements. Preliminary results showed that repetitive elements have a determining impact on genome size in myriapods with a marked difference in evolutionary dynamics between chilopods and diplopods. Our analysis also revealed a strong differentiation of protein-coding gene repertoires between chilopods and diplopods. Further comparative and evolutionary genomics studies will allow us to functionally explore the core genome of myriapods as well as specific genes related to the diverse ecological niches adopted by the different species. In addition, we will leverage these new annotated genomes by conducting a phylogenomic analysis to resolve still debated phylogenetic relationships within myriapods but also with other groups of arthropods. (1) State of knowledge of soil biodiversity - Status, challenges and potentialities. FAO

*Intervenant

†Auteur correspondant: odile.lecompte@unistra.fr

The influence of genetic dosage on PRDM9-dependent evolutionary dynamics of meiotic recombination

Alice Genestier*†[1], Nicolas Lartillot‡[2], and Laurent Duret[3]

[1]Département PEGASE [LBBE] – Laboratoire de Biométrie et Biologie Evolutive - UMR 5558 – France
[2]Laboratoire de Biométrie et Biologie Evolutive - UMR 5558 – Université Claude Bernard Lyon 1, Université de Lyon, Institut National de Recherche en Informatique et en Automatique, VetAgro Sup - Institut national d'enseignement supérieur et de recherche en alimentation, santé animale, sciences agronomiques et de l'environnement, Centre National de la Recherche Scientifique : UMR5558 – France
[3]Laboratoire de Biométrie et Biologie Evolutive (LBBE) – CNRS : UMR5558, Université Claude Bernard - Lyon I, INRIA – 43 Bld du 11 Novembre 1918 69622 VILLEURBANNE CEDEX, France

Résumé

Meiosis is an important step in the eukaryotic life cycle during which recombination and proper chromosome segregation takes place. In mammals, recombination is regulated by the Prdm9 gene. This gene, which possesses a double function (recruitment of the double strand break machinery and facilitation of the pairing of homologous chromosomes), induces an intra-genomic Red Queen resulting from the opposition of two antagonistic forces : erosion of the recombination landscape by biased gene conversion and positive selection on Prdm9. This Red Queen was previously modeled, but without taking into account the role of PRDM9 as a pairing facilitator. Accordingly, I developed a mechanistic model taking into account the dual role of PRDM9. This modeling work gives important insights into the Red Queen mechanism, thus completing previous studies. In particular, it reveals that positive selection of new PRDM9 alleles is due to the reduced symmetrical binding caused by the loss of high affinity binding sites and, on the other hand, it demonstrate the influence of the genetic dosage of PRDM9 on the dynamics of the Red Queen, which can result in negative selection on new PRDM9 alleles entering the population.

*Intervenant
†Auteur correspondant: alice.genestier@univ-lyon1.fr
‡Auteur correspondant: nicolas.lartillot@univ-lyon1.fr

PhylteR, a tool for analyzing, visualizing and filtering phylogenomics datasets

Damien M $_{d}eVienne^{*1}$

[1]Laboratoire de Biométrie et Biologie Evolutive - UMR 5558 – Université Claude Bernard Lyon 1, Centre National de la Recherche Scientifique, Centre National de la Recherche Scientifique : UMR5558 – France

Résumé

In phylogenomics, incongruences between gene trees, resulting from both artifactual and biological reasons, are known to decrease the signal-to-noise ratio and complicate species tree inference. The amount of data analysed nowadays in classical phylogenomics analyses rules out manual detection and deletion of errors but a standard for the quality control of phylogenomics datasets is still lacking.
Here we present PhylteR, a method that allows a rapid and accurate detection of outliers in phylogenomics datasets, i.e. species from individual gene trees that do not follow the general trend. PhylteR relies on Distatis, an extension of multidimensional scaling to 3 dimensions to compare multiple distance matrices at once. Distance matrices are obtained either directly from multiple sequence alignments, or are extracted from individual gene phylogenies.
Using PhylteR on biological datasets, we show (i) that it identifies as outliers sequences that can be considered as such by other means, (ii) that the removal of these sequences improves the concordance between the gene trees and the quality of the species tree reconstructed afterwards, and (iii) that the identifcation of outliers is much more accurate than other available software. Thanks to the generation of numerous graphical outputs, PhylteR also allows the rapid and easy characterisation of the dataset at hand, thus aiding in the precise identification of errors.

*Intervenant

Predicting interaction partners using masked language modeling

Umberto Lupo*†[1], Damiano Sgarbossa‡[1], and Anne-Florence Bitbol§[1]

[1]EPFL – Suisse

Résumé

Determining which proteins interact together from their amino acid sequences is an important task. In particular, even if an interaction is known to exist in some species between members of two protein families, determining which other members of these families are interaction partners can be tricky. Indeed, it requires identifying which paralogs interact together. Various methods have been proposed to this end. Here, we present a new one, which relies on a protein language model trained on multiple sequence alignments and directly exploits the fact that this model was trained to fill in masked amino acids. We obtain promising results on two different benchmark pairs of interacting protein families where partners are known. In particular, performance is good even for shallow alignments, while previous coevolution-based methods require deep ones. Performance is also found to quickly improve by giving the model correct examples of interacting sequences.

*Intervenant

†Auteur correspondant: umberto.lupo@epfl.ch

‡Auteur correspondant: damiano.sgarbossa@epfl.ch

§Auteur correspondant: anne-florence.bitbol@epfl.ch

sciencesconf.org:alphy-aiem-2023:435933

Using whole-genome data to unravel the evolutionary history of a commercially important species: demographic history and chromosomal inversions in the Thorny Skate (Amblyraja radiata)

Pierre Lesturgie*[1], John Denton[2], Jeff Kneebone[3], Romuald Laso-Jadart[4], Lei Yang[2], Stefano Mona[5], and Gavin Naylor[2]

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

[2]Florida Museum of Natural History [Gainesville] – États-Unis

[3]New England Aquarium – États-Unis

[4]Institut de Systématique, Evolution, Biodiversité (ISYEB) – Muséum National d'Histoire Naturelle (MNHN) – France

[5]Institut de Systématique, Evolution, Biodiversité – Museum National d'Histoire Naturelle, Université Pierre et Marie Curie - Paris 6, Ecole Pratique des Hautes Etudes : UMR7205, Centre National de la Recherche Scientifique – France

Résumé

The thorny skate (*Amblyraja radiata*) is a vulnerable species of commercial interests living in the North Atlantic Ocean. To date, mitochondrial data have suggested that the species is weakly structured or even panmictic across its range, which is in contrast with ecological evidence of strict residency. In addition, individuals with size and lifecycle differences have been observed in the Gulf of Maine, raising the question of the existence of cryptic species. By harnessing the power of whole-genome sequencing data from 50 individuals sampled throughout its range, we aim here to re-evaluate population structure and investigate the demographic history and size difference pattern in Gulf of Maine individuals. Clustering and pairwise FST analyses showed strong genetic differentiation between the eastern (from Norway up to East Greenland) and western (from the Gulf of Maine to Canada) part of the range (FST_˜0.20), while limited genetic structure within each group and no genetic differentiation within the Gulf of Maine (FST_˜0.003) were observed. Coalescence-based inference of variation in the effective size through time (PSMC, stairwayplot) suggested an independent demographic history of the two clusters for the last 20,000 generations. This was confirmed by fitting an isolation with migration (IM) model to the pairwise two-dimensional site frequency spectrum, thus reversing the previous evidence of panmixia suggested by the mtDNA data. Sliding windows of FST, linkage disequilibrium, and allele frequency identified the presence of two large inversions, occurring in two different chromosomes and characterized by independent demographic trajectory. Preliminary results suggest that one of the inversions may be related to the size pattern observed in the Gulf of Maine cluster, but

*Intervenant

additional data and analyses are needed to properly detect the phenotypic consequences of both inversions. This study highlights the power of whole genome sequencing data to uncover the population structure and historical demography of commercially important species, and points to the relevance of chromosomal inversions in the evolutionary history of a species.

Urban population genomics: dispersal and adaptation of the brown rat (Rattus norvegicus) in Paris

Romuald Laso-Jadart*†[1], Elise Gay[1,2], Thierry Feuillet[3], Benoît Pisanu[4], Bertrand Bed'hom[1], Aude Lalis‡[1], and Stefano Mona§[1,5]

[1]Institut de Systématique, Evolution, Biodiversité (ISYEB) – Muséum National d'Histoire Naturelle (MNHN) – France

[2]Ecole pratique des hautes études – Université Paris sciences et lettres, École Pratique des Hautes Études [EPHE] – France

[3]Caen University, UMR IDEES CNRS, Caen, France – CNRS – France

[4]UMS 2006 Patrimoine Naturel, AFB, MNHN, CNRS – UMS 2006 Patrimoine Naturel, AFB, MNHN, CNRS – 36 rue Geoffroy Saint-Hilaire, 75005, Paris, France

[5]Ecole pratique des hautes études – Université Paris sciences et lettres, École Pratique des Hautes Études [EPHE] – France

Résumé

Urban environments are prone to drastically shape species evolution. The proximity of human populations modifies food availability and quality and induces habitat fragmentation. With the expansion of urbanization, understanding how species disperse within cities and adapt to the novel habitat is becoming crucial for both ecological and public health issues. Here we focus on the brown rat (*Rattus norvegicus*), which is among the most abundant species living in urban environments, but it remains poorly studied in this ecological context. As the lack of knowledge on their dispersal ability and adaptation is detrimental to a wise control of their populations in large urban centers, we whole genome sequenced 23 specimens in Paris to investigate local population structure and search for putative signature of selection. We found that Parisian rats present weak population structure at the city level, with the Seine acting as the main barrier to gene flow. To model dispersal at the fine scale, we built a raster map of Paris including various environmental features (buildings, parks, streets, the Seine and the bridges among others). Based on this raster, we computed distances between sampling sites using both least-cost and isolation-by-resistance algorithms. The permeability of each urban feature was estimated by maximizing the correlation between the geographic distances and the matrix of pairwise genetic relatedness computed from genome wide autosomal data, Y-chromosome, and mtDNA. We found signatures of sex-biased dispersal, with males being the main vector of city-wide connectivity. Finally, we compared the genomes of the Parisian rats to twelve Chinese rats to find footprints of natural selection. We discovered a large inversion characteristic of some Parisian rats and several other candidate regions potentially involved in urban adaptation.
This work, by providing insights in the dispersal and adaptation dynamics, will help to improve the monitoring and the management of the brown rat in Paris.

*Intervenant

†Auteur correspondant: romuald.laso-jadart@mnhn.fr

‡Auteur correspondant: aude.lalis@mnhn.fr

§Auteur correspondant: mona@mnhn.fr

Inference of the Cultural Transmission of Reproductive Success from human genomic data: ABC and machine learning methods

Arnaud Quelin*[1,2], Jérémy Guez[1,2], Ferdinand Petit[1,2], Frédéric Austerlitz[1], and Flora Jay[2]

[1]Éco-anthropologie – UMR 7206, Muséum national d'Histoire naturelle, CNRS, Université de Paris, Musée de l'Homme 17 Place du Trocadéro 75016 Paris, France – France

[2]Laboratoire Interdisciplinaire des Sciences du Numérique – CentraleSupélec, Université Paris-Saclay, Centre National de la Recherche Scientifique : UMR9015 – France

Résumé

The Cultural Transmission of Reproductive Success (CTRS) is one of the various cultural processes that can impact human genetic evolution. In this process, individuals from large families have more children on average. Here, we develop and evaluate methods to infer this process from genomic data, using two approaches: (1) Approximate Bayesian computation, which uses summary statistics computed on inferred genealogies from genomic data and (2) deep neural networks, which are directly trained on genomic data. These methods rely on large simulated datasets incorporating varying levels of CTRS. Both competing approaches show a good ability to infer CTRS on genomic data and worth investigating under more complex evolutionary histories.

*Intervenant

Thirdkind : Drawing reconciled phylogenetic trees up to 3 reconciliation levels

Simon Penel*[1], Hugo Menet[1], Théo Tricou[1], Vincent Daubin[1], and Eric Tannier[1,2]

[1]Laboratoire de Biométrie et Biologie Evolutive - UMR 5558 – Centre National de la Recherche Scientifique : UMR5558 – France

[2]INRIA Rhône-Alpes (INRIA Grenoble Rhône-Alpes) – INRIA – ZIRST 655 Avenue de l'Europe Montbonnot 38334 Saint Ismier cedex, France

Résumé

The history of a species is related to the history of its genes. Associating genome evolution to the evolution of its genes is a way to describe this relationship. Reconciliating gene tree with species tree consists into mapping the nodes of the gene tree and the associated events to the nodes of the species tree. Reconciliation can as well be used to map the history of a parasite with the history of a host, or to map the history of a protein domain with the history of a sequence.
Phylogenetic reconciliations can visualised with various programs and interfaces as NOTUNG (1), SylvX (2), Treerecs (3), Jane (4), eMPRess (5) and Capybara (6). However at the exception of SylvX, all are integrated in a specific reconciliation software and cannot visualise reconciliations produced by others. None of these software is handling recPhyloXML (7), a XML format proposed as a standard to describe phylogenetic reconciliations, and none of them is generic to any kind of reconciliation nor can handle multiple horizontal transfer and the consideration of numerous possible scenarios.

Furthermore there is no software able to combine two nested reconciliations i.e. to get in a single representation the gene/symbiont reconciliation and the symbiont/host reconciliation.

Here we present Thirdkind (8) a command-line software allowing to easily generate graphical output from one or several recphyloXML files with a large choice of options (orientation, police size, branch length, multiple gene trees, multiples species trees, multiple files, redundant transfers handling, etc.) and to handle the display of two nested reconciliations (displaying a gene/symbiont/host reconciliation for example).

Home page : https://github.com/simonpenel/thirdkind/wiki

References

(1) K Chen *et al.* NOTUNG: a program for dating gene duplications and optimizing gene family trees. *J. Comput. Biol.*, (7):429–447, 2010.

(2) F Chevenet *et al.* SylvX: a viewer for phylogenetic tree reconciliations. *Bioinformatics*,

*Intervenant

(32):608–610, 2016.

(3) N Comte *et al.* Treerecs: an integrated phylogenetic tool, from sequences to reconciliations. *Bioinformatics*, (36):4822–4824, 2020.

(4) C Conow *et al.* Jane: a new tool for the cophylogeny reconstruction problem. *Algorithms Mol Biol.*, DOI: 10.1186/1748-7188-5-16, 2010.

(5) S Santichaivekin *et al.* eMPRess: a systematic cophylogeny reconciliation tool. *Bioinformatics*, (37):2481–2482, 2021.

(6) Y Wang *et al.* Capybara: equivalence ClAss enumeration of coPhylogenY event-BAsed ReconciliAtions. *Bioinformatics*, (36):4197–4199, 2021.

(7) W Duchemin *et al.* RecPhyloXML: a format for reconciled gene trees. *Bioinformatics*, (34):3646–3652, 2018.
(8) S Penel *et al.* Thirdkind: displaying phylogenetic encounters beyond 2-level reconciliation. Bioinformatics, https://doi.org/10.1093/bioinformatics/btac062, 2022.

Towards alife-generated benchmarks for phylogeny

Marco Foley[1], Jonathan Rouzaud-Cornabas[1,2], Eric Tannier[1,3], and Guillaume Beslon*[1,2]

[1]BEAGLE – INRIA, Institut National des Sciences Appliquées [INSA] - Lyon – France
[2]Laboratoire d'Informatique en Image et Systèmes d'information – Université Lumière - Lyon 2, Ecole Centrale de Lyon, Université Claude Bernard Lyon 1, Centre National de la Recherche Scientifique : UMR5205, Institut National des Sciences Appliquées de Lyon – France
[3]Laboratoire de Biométrie et Biologie Evolutive (LBBE) – CNRS : UMR5558, Université Claude Bernard - Lyon I (UCBL), INRIA – 43 Bld du 11 Novembre 1918 69622 VILLEURBANNE CEDEX, France

Résumé

Inspired by the double-blind principle that governs testing in science, we propose a new way to test methods for molecular evolution with computer simulations. Here, two teams (the INRIA Beagle Team, specialized in computational evolution, and the CNRS/LBBE Le Cocon Team, specialized in phylogeny) worked concurrently, Beagle producing evolutionary simulations – without information about the analysis tools – while Le Cocon tested phylogenomic tools on the simulated data without information on the way they have been generated.
Blind sequence generation: The Beagle team adapted its Aevol platform to allow for the simulation of 4-bases sequences (while the original platform uses binary sequences). This allows analyzing the simulated genomes with on-the-shelf phylogenomic tools. Using this new version, we let a population evolve for 800.000 generations. Then, we simulated a random branching process and simulated evolution along the branches up to generation 1.000.000. This results in 40 different populations that evolved for the same duration in the same conditions but that diverged in their past at random times. We extracted the genome of the 40 best final organism and sent them to Le Cocon for "double-blind" analysis.

Blind phylogenomic reconstruction: A first attempt to align the 40 sequences with MAFFT – handling, as most alignment softwares, local mutations (substitution, InDels) – gave no satisfying results, which convinced the inference team that it was necessary to account for rearrangements (duplication, inversion, translocation). We then used the Mauve sequence aligner, which segments the genomes into aligned pieces. The aligned pieces, scattered across all initial genomes, were concatenated to produce 40 aligned virtual sequences, which are each rearranged segments of the initial sequences. This alignment was given as input to IQtree with a "model test" option to let the program choose the inference model, resulting in an inferred tree. Importantly, none of the tools used integrate knowledge about the simulation software and the simulations have been produced without a priori knowledge about the tools operated by the inference team.
Comparison of the inferred tree with the ground-truth showed that its shape matched almost exactly, with three differences that correspond to the lower branch supports of the inferred tree. As far as we know, this is the first time an artificial life simulation software produced

*Intervenant

sequences that could be directly analyzed with on-the-shelf tools. Although the simulated sequences evolved in the simplest possible setting (all species evolving in the same conditions and differing only by their branching times), this first result is promising. Indeed, it opens the way to simulating more complex trees with e.g. variations in mutation rates or population sizes (including bottlenecks) in order to challenge phylogenomic methods.

MacSyFinder v2: An improved search engine to model and identify molecular systems in genomes

Bertrand Néron*[1], Denise Remi†[2], Charles Coluzzi[3], Marie Touchon[3], Eduardo Rocha[3], and Sophie Abby[4]

[1]Institut Pasteur – Hub Bioinformatique et Biostatistique – C3BI, USR 3756 IP CNRS – Institut Pasteur de Paris – 25-28 Rue du Docteur Roux, 75015 Paris, France

[2]APC Microbiome Ireland School of Microbiology, University College Cork, Cork – Irlande

[3]Génomique évolutive des Microbes / Microbial Evolutionary Genomics – Institut Pasteur [Paris], Centre National de la Recherche Scientifique : UMR3525, Université Paris Cité – France

[4]Techniques de lÍngénierie Médicale et de la Complexité - Informatique, Mathématiques et Applications [Grenoble] (TIMC-IMAG) – Centre National de la Recherche Scientifique : UMR5525, Université Grenoble Alpes – Domaine de la Merci - 38706 La Tronche, France

Résumé

Complex cellular functions are most often encoded by a set of genes rather than individual ones. Furthermore, the genes in such "systems" are often encoded nearby in microbial genomes. MacSyFinder uses these properties to model and then accurately annotate cellular functions in microbial genomes at the system-level rather than at the individual-gene level. We hereby present a major release of MacSyFinder (1), MacSyFinder version 2 (v2). This new version is coded in Python 3 ($> = 3.7$). The code was improved and rationalized to enable higher maintainability over time. Several new features were added to allow more flexible modeling of the systems. We introduce a more intuitive and comprehensive search engine to identify all the best candidate systems and sub-optimal ones that still respect the models' constraints. We also present the novel macsydata companion tool that enables the easy installation and broad distribution of the models developed for MacSyFinder (macsy-models) from GitHub repositories. Finally, we have updated, improved, and made available MacSyFinder popular models to this novel version: TXSScan and TFF-SF, CONJscan, and CasFinder. MacSyFinder v2 can be found at this URL: https://github.com/gem-pasteur/macsyfinder

References

1. Sophie S Abby, Bertrand Néron, Hervé Ménager, Marie Touchon, Eduardo PC Rocha. MacSyFinder: A Program to Mine Genomes for Molecular Systems with an Application to CRISPR-Cas Systems. PLOS ONE, https://doi.org/10.1371/journal.pone.0110726, 2014.

*Auteur correspondant: bertrand.neron@pasteur.fr

†Intervenant

Towards creating longer genetic sequences with Generative Adversarial Networks

Antoine Szatkownik*[1], Burak Yelmen[2], Flora Jay†[2], and Guillaume Charpiat[3]

[1]LISN – Laboratoire Interdisciplinaire des Sciences du Numérique (LISN) – France
[2]LISN – Université Paris-Sud - Université Paris-Saclay – France
[3]INRIA – Université Paris-Sud - Université Paris-Saclay – France

Résumé

Deep Learning has leveraged the large quantities of data produced by high-throughput sequencing to improve our understanding of complex biological processes like evolution.The generation of omic data, in particular proteomic data stemming from the protein design community, has been extensively studied in the last years, yet little progress has been made toward genome-wide scale DNA sequence generation.
Public biobanks of human genomes such as the 1000G project and HapMap are biased toward the over-representation of certain populations, lack the specificity that can be found in private datasets oriented in the study of diseases, or simply contain too few samples which make models prone to overfitting. This raises the need to augment public databases with artificial genomes, possibly designed to have desired properties, that will serve as proxies for non-accessible, sensitive data.

Here, we report on current developments for creating artificial human genomes. Since we are focusing on bi-allelic polymorphism markers, encoded as binary, we aim to generate binary data. To do so, rather than applying a simple rounding scheme (1), we learn the binarization, that is we use, in the last layer, a sign function in the forward pass and approximate the sign function in the backward pass (2).
Previous generative neural networks for genomic data are not suitable for genome-wide sequences as they are based on dense layers, thus exploding the number of parameters needed when the sequence length increases. To scale up to longer DNA sequences, we perform the learning of a Wasserstein Generative Adversarial Network (WGAN) in a dimensionally reduced space, like the one obtained by applying PCA, which allows generating new points in the PCs space and then projecting the generated PC scores in the original genomic space. The WGAN is a variant of GAN where the discriminator is replaced by a critic that no longer assigns

*Intervenant
†Auteur correspondant: flora.jay@lri.fr

Gene conversion drives allelic dimorphism in two paralogous surface antigens of the malaria parasite, P. falciparum

Brice Letcher*[1] and Zamin Iqbal[1]

[1]EMBL-EBI – Royaume-Uni

Résumé

In the malaria parasite, *P. falciparum*, a number of genes display exactly two, highly-diverged sequence forms at high-frequencies across malaria-endemic countries, presumably maintained by balancing selection for immune escape.
Why exactly two forms exist, and not more, and why these forms have not been shuffled by recombination - a phenomenon called allelic dimorphism in the literature - has not been resolved, however. Further, fully characterising the gene sequences from whole-genome sequencing data has proved difficult, as all available data for 1,000s of global P. falciparum field samples is Illumina data.

Here, we developed a new genotyping pipeline to comprehensively characterise two highly-variable and dimorphic malaria cell-surface antigens, DBLMSP and DBLMSP2. The genes are paralogs, sharing multiple protein domains, and located only _˜16kbp apart. Our pipeline leveraged multiple approaches, including genome-graph-based genotyping to alleviate reference bias, and outperformed a widely-used pipeline for *P. falciparum* based on GATK, a single-reference-based variant caller.
Using our newly-resolved sequences, we confirmed dimorphism in each gene, and further identified that one of the two diverged forms in each gene is shared between the genes. We found clear evidence this is driven by mitotic gene conversion between the two paralogs. Based on these data we propose a new model for allelic dimorphism, in which a strong population bottleneck in P. falciparum (known from previous studies), followed by gene conversion between diverging paralogs, can create allelic dimorphism.

*Intervenant

Conservation of Structured Populations : Insights from the Structured Coalescent

Alexane Jouniaux*†[1], Lounés Chikhi[2], and Olivier Mazet[1]

[1]Institut de Mathématiques de Toulouse – Institut National des Sciences Appliquées (INSA) - Toulouse – France

[2]Evolution et Diversité Biologique – Institut de Recherche pour le Développement, Université Toulouse III - Paul Sabatier, Centre National de la Recherche Scientifique, Centre National de la Recherche Scientifique : UMR5174 – France

Résumé

Genetic data are increasingly used to inform conservation actions. It is then crucial to predict the genetic consequences of conservation decisions. This is particularly important given that natural habitats are increasingly fragmented with limited gene flow between populations that used to be connected. An example of theoretical development in this field is the work of Alcala et al. (2019) (1). In this paper, authors study the genetic diversity of populations subdivided in one up to four demes and how changes in the number of connections or in the number of demes affect it. They use the expected within-subpopulation nucleotide diversity as a measure of genetic diversity, which is an indicator based on the expected coalescence times of samples of two genes taken in the same deme. Their results suggest that some configurations are able to maintain high levels of genetic diversity and that when some connections are lost, the within-subpopulation nucleotide diversity can decrease significantly.
However, by using only the within-subpopulation nucleotide diversity, the authors did not make use of all the genetic information available through coalescence time distributions. Using the structured coalescent theory (cf. Herbots (1994) (2), Mazet et al. (2016) (3), Rodriguez et al. (2018) (4)), we can, for instance, compute the probability of having a certain number of differences between two sequences. By taking the probability of having at least one difference, we obtain a new indicator of genetic diversity. It allows us to consider the entire information provided by coalescence times and the sampling of two genes in disconnected subpopulations, that are not taken into account in the article.

In the present work, we compute the probability of having a certain number of differences between two sequences for all the population configurations of (1) and prove that we can retrieve their results with this indicator. We also varied some parameters of this measure, such as the mutation rate of the population or the number of differences considered, to see how it affects genetic diversity.

(1) N. Alcala, A. Goldberg, U. Ramakrishnan, and N. A. Rosenberg, "Coalescent Theory of Migration Network Motifs," *Mol. Biol. Evol.*, vol. 36, no. 10, pp. 2358–2374, Oct. 2019.

*Intervenant

†Auteur correspondant: jouniaux@insa-toulouse.fr

(2) H. M. J. D. Herbots, "Stochastic Models in Population Genetics: Genealogy and Genetic Differentiation in Structured Populations," 1994.

(3) O. Mazet, W. Rodríguez, S. Grusea, S. Boitard, and L. Chikhi, "On the importance of being structured: Instantaneous coalescence rates and human evolution-lessons for ancestral population size inference?," *Heredity (Edinb).*, vol. 116, no. 4, pp. 362–371, 2016.

(4) W. Rodríguez *et al.*, "The IICR and the non-stationary structured coalescent: towards demographic inference with arbitrary changes in population structure," *Heredity (Edinb).*, vol. 121, no. 6, pp. 663–678, 2018.

Bayesian inference of the origin of ancient individuals.

Guillaume Laval*[1]

[1]Institut Pasteur [Paris] – Institut Pasteur [Paris] – 25-28, rue du docteur Roux, 75724 Paris cedex 15, France

Résumé

A number of studies in humans showed that ancient DNA data (aDNA) provide valuable information to increase our understanding of the recent evolutionary past of human populations. However, the origin of ancient individuals is often unclear. A study ignoring the genetic discontinuity due to waves of migration that may occur in recent past estimated that no ancient samples represent direct ancestors of modern Europeans (Schraiber, Genetics, 2018). Using more realistic models of the recent evolution of European populations we implemented new Approximate Bayesian Computation (ABC) and machine learning methods to analyze 2,654 ancient European genomes covering multiple epochs over the last 35,000 years. The methods show that models which explicitly consider the recent waves of migration in Europe, i.e., the arrival of Anatolian farmers _˜8,500 years ago (ya) and that of populations associated with the Yamnaya culture around _˜4,500 ya, allows to properly infer the origin of ancient individuals. In addition, through the use of new Deep Learning algorithms (Convolutional Neural Networks) we also show that ancient DNA data contains enough information to provide estimations of the age of each ancient sample.

*Intervenant

Genome size variation in animals: impact of effecive population size and transposable elements

Alba Marino*[1], Benoit Nabholz[1], Annabelle Haudry[2], and Anna-Sophie Fiston-Lavier[1]

[1]Institut des Sciences de l'Evolution de Montpellier – Université de Montpellier – France
[2]Laboratoire de Biométrie et Biologie Evolutive - UMR 5558 – Université Claude Bernard Lyon 1 – France

Résumé

Animals display a remarkable variation in genome size (GS) that covers four orders of magnitude, spanning from 20 Mb in the nematode *Pratylenchus coffeae* to 120 Gb in the african lungfish *Protopterus aethiopicus*. While it is now acknowledged that the major quantitative factor affecting GS is non-coding DNA, the evolutionary drivers of such variation are still debated. Lynch & Conery (2003) observed that organisms with small effective population size (*Ne*) show bigger GS than organisms with high *Ne*: they proposed that strong genetic drift - measured as small *Ne* - allows slightly deleterious mutations such TEs to more easily reach fixation in the population and contribute to GS increase, while the genomes of organisms with high *Ne* remain more streamlined as a consequence of a more effective selection against TE proliferation. While this hypothesis was proposed to explain patterns of genome architecture variation across all living organisms, it is still not clear whether its framework also applies at shorter phylogenetic scale. We leaverage a dataset of 808 animal species, including 189 insects, 148 ray-finned fish, 260 birds, 183 mammalians and 28 molluscs, to systematically test this hypothesis. In particular, we aim at (1) confirming whether TE content affects and explains GS variation across animals as a general trend, and (2) assessing the goodness of *Ne* as predictor of the variations of GS and TE density. In order to bypass the issue of repeated sequences causing fragmented and missing information in genome assemblies, we use the c-values of a species subset to systematically correct GS estimations from assembly size; moreover, we estimate the TE content from low coverage sequencing reads. Finally, we use both non-synonymous over synonymous substitution rate and life history traits to model the relationships of *Ne* with GS and TE content.

*Intervenant

An early origin of Iron-Sulfur cluster biosynthesis machineries before Earth oxygenation

Pierre Garcia*[1], Francesca D'angelo[1], Sandrine Ollagnier De Choudens[2], Macha Dussouchaud[1], Emmanuelle Bouveret[1], Simonetta Gribaldo[1], and Frédéric Barras[1]

[1]Institut Pasteur – Université Paris Cité, Centre national de la recherche scientifique - CNRS (France) – France

[2]Université Grenoble Alpes – Centre national de la recherche scientifique - CNRS (France), Commissariat à l'Énergie Atomique et aux Énergies Alternatives (CEA) - Grenoble – France

Résumé

Iron-sulfur (Fe-S) clusters are ubiquitous co-factors essential for life. It is largely thought that the emergence of oxygenic photosynthesis and progressive oxygenation of the atmosphere led to the origin of multiprotein machineries (ISC, NIF, and SUF) assisting Fe-S cluster synthesis in the presence of oxidative stress and shortage of bioavailable iron. However, previous analyses have left unclear the origin and evolution of these systems. Here, we combine exhaustive homology searches with genomic context analysis and phylogeny to precisely identify Fe-S cluster biogenesis systems in over 10,000 archaeal and bacterial genomes. We highlight the existence of two additional and clearly distinct "minimal" Fe-S cluster assembly machineries, MIS and SMS, which we infer in the Last Universal Common Ancestor (LUCA), and we experimentally validate SMS as a bona fide Fe-S cluster biogenesis system. These ancestral systems were kept in Archaea whereas they went through stepwise complexification in Bacteria to incorporate additional functions for higher Fe-S cluster synthesis efficiency leading to SUF, ISC, and NIF. Horizontal gene transfers and losses then shaped the current distribution of these systems, driving ecological adaptations such as the emergence of aerobic lifestyles in archaea. Our results show that dedicated machineries were in place early in evolution to assist Fe-S cluster biogenesis, and that their origin is not directly linked to Earth oxygenation.

*Intervenant