Phylogenomics, convergent evolution, and the Anthropocene

Bastien Boussau

Université Claude Bernard - Lyon 1
Laboratoire de Biométrie et Biologie Évolutive

Phylogenomics, convergent evolution, and the Anthropocene

Phylogénomique, évolution convergente, et l'Anthropocène

Bastien BOUSSAU

Habilitation à diriger les recherches
Soutenue publiquement le 21 mars 2019

Jury :

Frédéric Delsuc	ISEM, Montpellier	Rapporteur
Christophe Douady	LEHNA, Lyon	Examinateur
Hélène Frérot	EEP, Lille	Rapportrice
Hervé Philippe	SETE, Moulis	Rapporteur
Philippe Vandenkoornhuyse	ECOBIO, Rennes	Examinateur

Acknowledgements

I would like to thank the three rapporteurs, Hélène Frérot, Frédéric Delsuc and Hervé Philippe, and the two examinateurs Philippe Vandenkoornhuyse and Christophe Douady, for taking the time to evaluate my HDR.

I would like to thank all my colleages at the LBBE, and the BPGE team in particular. BPGE is the main reason why I came back to Lyon. People there are smart and interesting, but they are also very nice and supportive. It was not easy for my colleagues and I to make the decision to create Le Cocon and I will do my best to keep strong ties with the members of my former team and department.

I would also like to thank my collaborators, without whom all my work would have been even worse. In particular I would like to thank the students and postdocs I have been allowed to advise. I know they have taught me much, and I hope they learned a few things in the process.

I would like to thank Vincent and Eric, who share my office, for they have taught me to question why I choose to question what I question. They like to question a lot of things, all the time.

I would like to thank Philippe who knows when to tell me the truth or when to tell me that I did well.

I would like to thank Anouk for her decision to opt for Enercoop despite a hostile environment.

I would like to thank M&M, for they make my life so sweet. This manuscript is dedicated to them.

Plan

Introduction

Throughout my career, I have developed bioinformatic methods to study genomic sequences and learn about their evolution and their function. The methods I have developed are typically based on probabilistic models, which I use to perform statistical inference in the Maximum Likelihood or Bayesian frameworks. They involve sophisticated algorithms and often rely on parallel computing on supercomputers to cope with the vast amounts of data contained in whole genome sequences. By using these methods on simulated and empirical data, I have been able to show their accuracy and to uncover new biological insights.

In this manuscript, I will be focusing on three aspects of my work. Here it should be understood that when I write “my work”, I mean work done by my colleagues and myself. Research work is highly collaborative, and relies on the combined efforts of many scientists with complementary expertises. All the work I will be presenting benefitted from synergies between several brains, including the brains of reviewers of published papers. Recently my close collaborators include, in reverse alphabetical order, Tom Williams, Philippe Veber, Eric Tannier, Gergely Szöllősi, Marie Sémon, Céline Scornavacca, Nicolas Lartillot, Sebastian Höhna, Laurent Guéguen, Marlène Dreux, Vincent Daubin. They also include students and postdoctoral researchers such as Adrian Davín, Carine Rey, Peter Markov, Kassian Kobert,Vincent Lanore.

Firstly I will present some of my work on gene tree-species tree models. These models attempt to capture the processes that make the histories of genomic segments differ among themselves, and differ from the species tree. By doing so they can provide information on the history of the species and of their genomes. I will include 7 publications on this topic in this manuscript.

Secondly, I will present my efforts to study the genomic roots of convergent phenotypic evolution. Faced with similar environmental constraints, living organisms have repeatedly come up with similar solutions. Examples include the convergent adaptations to life underground in Aselloidea, a clade of Crustaceans, and the convergent adaptations of rodents to life in arid environments. Undeground crustaceans have all developed different stages of the troglomorphic syndrome, *i.e.* have less pigmentation, a reduced metabolic rate, and enlarged antennas. Rodents living in arid environments have kidneys that are very good at saving water. I will include 3 publications on this topic in this manuscript and describe a research project on the adaptation of plants to abiotic stressors such as drought, heat, or soil salinity. These stressors are expected to become more intense and widespread as climate change unfolds.

Thirdly, I will outline a series of disparate efforts to engage with the society beyond academia. These efforts include outreach work, but also attempts to reduce the environmental footprint of research.

I will conclude with a short exposé of my views on the position of researcher in the Anthropocene.

But first, let me briefly describe some work I chose not to detail in this manuscript.

Brief summaries of additional work

There are some parts of my work that I have chosen not to present. These include work I have done on RevBayes, on experimental evolution of viruses in cell cultures, some outreach work, and some teaching work.

- RevBayes

RevBayes is a piece of software for Bayesian statistical inference, in particular in phylogenetics. During my postdoctoral stay at UC Berkeley in the lab of John Huelsenbeck, I became a core member of the development team of the software RevBayes. Since then I have maintained this involvement. RevBayes is a program for performing Bayesian analyses, in particular in phylogenetics. It is designed to be extremely general and malleable and yet is efficient, and is used through a dedicated R-like language (Höhna *et al*., 2016). I have been part of the initial design discussions and have contributed code to the software. I have also participated in workshops, hackathons and tutorials on RevBayes taught in the USA, and organized a hackathon in Lyon during the summer 2018.

- Experimental viral evolution

I have also chosen not to present work I have done on experimental evolution of viruses in cell cultures, in collaboration with Marlène Dreux and her team, at the Centre International de Recherche en Infectiologie. Standard approaches to understanding how viruses work involve painstaking experiments in molecular and cellular biology. Each experiment is intensive in time and resources, and must be devoted to testing solid hypotheses on the function of specific sequences in the genome of a virus. To generate such hypotheses, experimental evolution has often been used. By evolving a virus in a controlled environment for several replications, and studying its genome before and after evolution, one can identify changes in the genome that belie adaptation to the experimental conditions.

We undertook experimental evolution assays in the riboriviruses Zika and Dengue viruses. These RNA viruses evolve fast, are both transmitted by mosquitoes, infect more than 400 million people yearly, and expand their geographical range as climate change unfolds. We studied how these viruses respond to the innate immunity of their host, humans or mosquitoes, by passaging viruses in cell cultures in which specific innate responses had been tuned up or down. In all experiments performed in human cell lines, the viruses adapted to the experimental conditions and replicated faster at the end than at the beginning of the experiment. We then extracted the viral genomes at different time points and sequenced them at a high depth with a particular sequencing protocol to identify low frequency variants while ensuring a low sequencing error rate. I advised two postdoctoral researchers on this project to build the needed computational tools and to analyze the sequences. In particular, a read mapper specific to the sequencing approach we used was implemented. Our rationale was to identify the sequence variants that increase in frequency during the experiment, *i.e.* those that caused adaptation of the virus to the experimental conditions. Two candidate sites were identified and tested in vitro: they were found to

explain most of the improvement in viral replication time. A manuscript is currently being written up.

- Outreach

In 2017, I contacted the “Transports en Commun Lyonnais” (TCL) with an outreach idea. The TCL have a private TV channel showing programs on their public transportation network. My idea was to produce little videos on evolutionary biology, showing that all living species are coming from a *common* ancestor (pun with the word “commun” from “transports en commun”). The goal was to discuss what they have in *common:* their phenotypic characteristics are either inherited from their ancestors, or have evolved convergently. The TCL liked this idea and financed the production of 20 videos that were designed in collaboration with Damien de Vienne and Sylvain Charlat. These videos, currently on display on the TCL channel, can be seen by the 800.000 people traveling daily on the TCL network. They are also available here: http://pointscommuns.fr/ , along with more information for those who would like to learn more on the subject.

- Teaching

Finally I will not be discussing my teaching, which has been fairly limited. I have been teaching courses to master, PhD students and researchers. Beyond courses on RevBayes as explained above, I have been contributing for 3 years to a course on Bayesian inference organized by Marie-Laure Delignette-Muller, Fabien Subtil and Nicolas Lartillot in the LBBE in Lyon for researchers and PhD students. In 2017 and 2018 I have been invited to teach phylogenetic inference in the “Computational Molecular Evolution” EMBO course organized by Nick Goldman (EBI Cambridge, UK), Ziheng Yang (Imperial College, UK) and Alexis Stamatakis (HITS, Germany), which took place in Cambridge (UK) and in Crete. I will be teaching this course again in 2019, in Cambridge.

Gene tree-species tree models

My work on gene tree-species tree models fits within a larger scheme defined in a manuscript I wrote with Vincent Daubin in 2010 (Boussau and Daubin, 2010). We argued there that computational genomics had much to gain by rethinking the analysis pipeline of genomic sequences which brings raw reads to annotated genomes and functional and evolutionary insights by a series of inferential steps. In this pipeline, every step is performed in almost complete isolation from other steps (Figure 1, left). In particular, such a pipeline ignores the uncertainty associated with early inferrential steps when performing later steps. Every mistake made at an early step is therefore propagated to later steps, which can result in incorrect inferences on the function of genes, or on their evolution (Figure 2). This point of view certainly did not come as a surprise to praticians of evolutionary genomics, who know all too well that careful manual “cleaning” of the data is often necessary to obtain reliable conclusions.

Gene tree-species tree models attempt to make the inferrential pipeline more robust by inferring jointly gene trees and species trees. With their help, the uncertainty about gene trees is kept when inferring the species tree, and the species tree can be used to improve the inference of gene trees, in particular when there is not much information in gene sequences. We have targeted one of the last steps of the pipeline because it was a way for us to learn about the processes of gene duplication, gene transfer and gene losses. I believe we have indeed made progress on those questions, as the following papers should demonstrate. But we still rely entirely on early steps of the pipeline, whose uncertainty is discarded. For this reason I suspect many mistakes coming from those early steps probably plague our inferences. Improving those early steps of the pipeline still represents a whole research agenda for the next decade. I expect that it would be helpful to use the information from the species tree to improve the clustering of homologous genes and to improve gene alignments.

Now I address gene tree-species tree models in more details. Genomes contain thousands of individual genes, which we will here define loosely as sequences whose history we can reconstruct across species boundaries. Throughout their history, these genes have tracked the history of speciations their host species went through, but they have also undergone events of duplications, transfers and losses. As a result, phylogenies reconstructed based on gene sequences are often more complicated than, and disagree with, the species phylogeny (Fig. 1) (Boussau and Daubin, 2010). While all nodes in a species phylogeny correspond to speciation events, nodes in a gene phylogeny can correspond to a variety of events: speciations, duplications, or transfers. I have developed methods to reconstruct accurate gene phylogenies and to robustly annotate their nodes (Szöllősi *et al*. 2015a).

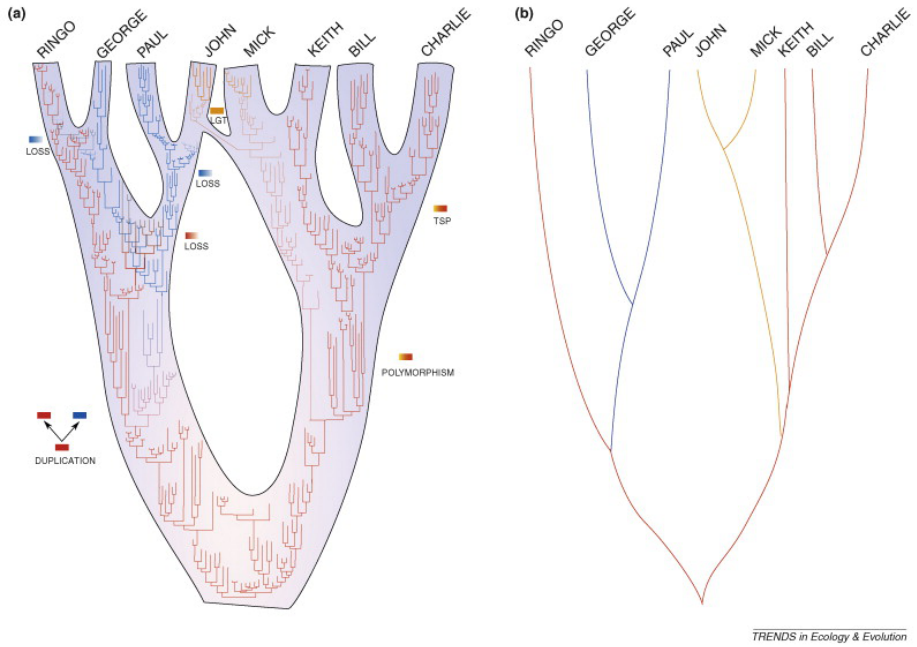


Figure 1: Various processes can generate discordance between species and gene trees. (a) A tree depicting the relationships of eight species. Ringo and George, Paul and John are on one side of the root, and Mick and Keith, Bill and Charlie on the other. The history of a gene family is depicted within the bounds of this species tree, and processes acting at the genome level (duplication, loss, gene transfer) as well as population level (polymorphism) are shown. (b) The gene tree reconstructed from this gene family shows a topology that conflicts with the species tree. Following a duplication and losses, George and Paul are grouped together, a gene transfer groups John and Mick, and trans-specific polymorphism leads to Keith being clustered with Bill and Charlie. From Boussau and Daubin (2010).

Inferring accurate gene histories is important to link gene and species histories. Firstly, without sorting speciation nodes from other types of nodes in gene phylogenies, one runs the risk of inferring incorrect species trees. This was the motivation behind one of my first forays into gene tree-species tree models, with the development of Phyldog, a program to jointly infer gene trees and species trees in the presence of gene duplications and losses (Boussau *et al*. 2013). With Phyldog I demonstrated that thousands of gene trees and a species trees could be inferred jointly and accurately for dozens of genomes, on a supercomputer. Secondly, events of gene duplication, transfer or loss provide functional information: a gene that got duplicated or transferred and fixed in a species probably was functionally important when the event took place; a gene that got lost probably became functionally less important than it was before the loss. My colleagues and I have shown that we can robustly identify such events for instance to reveal highways of gene transfers in Fungi (Szöllősi *et al*. 2015b), root species trees (Szöllősi *et al*. 2012), or study genome evolution and reconstruct ancestral gene contents in Archaea (Williams *et al*. 2018). I have also collaborated with American colleagues on species tree reconstruction based on orthologous sequences in the presence of incomplete lineage sorting (Mirarab *et al*. 2014, for instance).

Beyond information on the topology of species trees and on the past functionality of genes, gene tree-species tree models can extract information about the age of ancient speciations. In the past, species phylogenies have been dated thanks to a combination of fossil information and a model of the rate of sequence evolution across lineages. This approach is limited because models of rate evolution are inaccurate, because some clades contain very few fossils, and because early life barely left any fossil at all. However, a PhD student I co-supervised with Vincent Daubin showed that gene transfers contain reliable and consistent information about the relative order of speciation nodes in clades of Bacteria, in Archaea, and in Fungi (Fig. 2) (Davín *et al*., 2018). This entirely new source of information is extremely precious to date the history of life, because it is abundant in clades where fossils are rare.

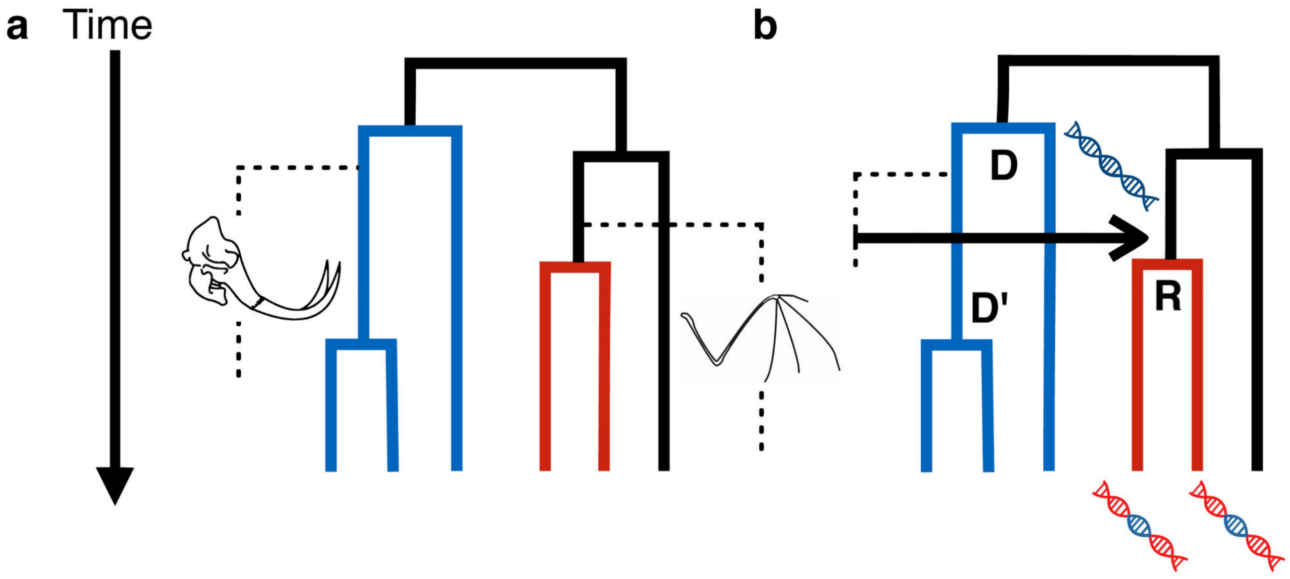


Figure 2: Gene transfers, like fossils, carry information on the timing of species divergence. a) The geological record provides the only source of information concerning absolute time: the age of the oldest fossil representative of a clade provides direct evidence on its minimum age, but inferring maximum age constraints (e.g. dashed line for the red clade), and by extension the relative age of speciation nodes, must rely on indirect evidence on the absence of fossils in the geological record. b) Gene transfers, in contrast, do not carry information on absolute time, but they do define relative node age constraints by providing direct evidence for the relative age of speciation events: the gene transfer depicted by the black arrow implies that the diversification of the blue donor clade predates the diversification of the red clade (i.e. node D is necessarily older than node R).

Selected articles on gene tree-species tree models:

- Method development
    - Syst Biol 2015
    - Genome Research
    - Syst Biol amalgamation
    - Science 2013
- Learning about genome evolution
    - Fungi Phil Trans 2015
    - PNAS Archaea
- Dating trees
    - NEE

Perspectives

The cocon team and reconciling three levels of evolution

Perspective: dating in RevBayes

Method development

- Syst Biol 2015

In 2015 my colleagues and I wrote a review on gene tree-species tree models. It focuses on methods based on probabilistic models, and spends less time on methods based on parsimony. This choice reveals our preference for probabilistic approaches.

Both approaches attempt to find solutions to a problem. Parsimony methods attempt to find solutions that involve the minimum number of events. For instance, when reconstructing phylogeny on a sequence alignment, parsimony methods would favour trees that require the lowest numbers of substitutions to explain a given alignment. This criterion reflects a particular philosophical stance, that nature operates with parsimony, meaning it tends to use few events to achieve a result.

Probabilistic models reflect a different viewpoint, and allows a range of solutions, including non-parsimonious ones. Adopting this framework does not necessarily mean that we believe the world is probabilistic ; it may well be deterministic, but with so many variables that we cannot include them all in our models. In that case, operating probabilistically implicitly integrates them out.

I personnally favour probabilistic models because they lend themselves to statistical inferential frameworks, such as Maximum Likelihood (ML) and Bayesian Inference (BI). When using these frameworks, one sits on a series of theoretical results that prove that inference is consistent, *i.e.* it is correct on infinitely large amounts of data generated according to the inferential model. Such results are of course of little practical use, because we never have infinite amounts of data, and because on empirical data we never use for inference the model that generated the data. But they are reassuring: the problems we are trying to solve are so difficult from a modelling or an algorithmic point of view that I take solace in the idea that, were I able to do things right, with large amounts of data, I might have a chance at getting a correct answer. Parsimony does not have this property and is therefore less appealing to me. Yet in lots of cases it is faster than probabilistic approaches and fairly accurate, which means it can be used as a convenient heuristic to help speed up inferential algorithms that operate in the ML or BI frameworks.

SYST BIOL 2015 ARTICLE INCLUDED HERE

- Genome Research 2013

I got interested in gene tree-species tree models in 2007, after reading an article by Scott Edwards that inferred gene trees and species trees with the multispecies coalescent (Edwards et al., 2007). My colleagues and I started working on it, and it took me 6 years to publish my first first-author paper on the subject. Our aim was to develop a program to jointly reconstruct gene trees and species trees with a model of gene duplication and loss.

We made several choices that differed from Edwards et al.'s. They had a two step process: first they inferred gene trees using a commonly used Bayesian method (mrBayes). Then they used the reconstructed tree distributions as an input to their main program, which inferred a posterior distribution of species trees and branch-wise parameters. They also obtained from this second step a reweighting of their input gene tree distributions. This reweighting amounted to reconstructing gene trees not only based on the gene sequence information, but also on the reconstructed species tree distribution.

We did not want to have a two-step approach: we felt it was more elegant to jointly infer gene trees and species trees. Our method can thus start from gene alignments, and reconstruct jointly the gene trees and the species tree. The method alternates between optimizing gene trees based on the sequences and a current proposed species tree, and individual topological moves on the current species tree. Contrary to Edwards et al., I chose to use an optimization approach that returns a single species tree and single gene trees per alignment rather than a Bayesian approach that returns a sample of the posterior distribution. The motivation for this choice was in part the hope that this would make the whole algorithm faster, but more importantly I had little knowledge of Bayesian methods at the time and felt more comfortable with optimization algorithms.

The resulting program, PHYLDOG, is still unique in its ability to jointly infer gene trees and species trees. The computational challenge associated to this aim is still sizeable, and it is not clear that the gains in accuracy are worth the effort. Instead I would recommend a simpler iterative approach, where one first reconstructs gene trees based on gene alignments, and second infers a species tree based on the fixed gene trees. These two steps could be repeated a couple of times. This should be more efficient than PHYLDOG's algorithm, which optimizes gene trees entirely every time a single topological move on the species tree is performed.

GENOME RESEARCH 2013 PAPER HERE

- Syst Biol Amalgamation

PHYLDOG taught us that joint inference of gene trees and species tree was computationally challenging. In Szollosi et al. 2012, we presented another piece of software that used fixed gene

trees and focused on inferring the species tree with a model of gene duplication, transfer and loss (DTL model). This DTL model is computationally intensive in and of itself, so it seemed much more reasonable to bypass gene tree inference and instead start from input gene trees reconstructed by another method.

However we were not entirely pleased with our software, and knew that the input gene trees contained many mistakes: after all, gene alignments are short and contain a limited amount of information, probably insufficient in a lot of cases to choose among the vast number of possible binary tree topologies.

We liked the idea of relying on input gene tree distributions as in Edwards et al. (2007): by doing so, one can integrate over the uncertainty associated with each gene tree. However this meant multiplying the computational load by the number of gene trees included in the gene tree distribution. This was prohibitive for the DTL model, which was too slow.

We thought of a solution to this problem after reading the supplementary material to David and Alm (2011). The authors had come up with an idea they called *amalgamation* to consider several gene trees at a time in a parsimony framework and save some computing time. The idea was simple: trees from a tree distribution will share some subtrees. One can thus make subtree-specific computations only once and use the result in all trees that contain this subtree. The speed improvement can be massive, especially for tree distributions that are not too diverse, i.e. in which a limited number of subtrees are found many times.

The DL and DTL models can be associated naturally to the amalgamation idea: they are based on dynamic programming algorithms, which means they rely on subtree-specific computations. We could thus develop a DL or DTL dynamic programming algorithm that worked on a tree distribution instead of on a single tree. One additional thing we had to do to extend the David and Alm amalgamation idea to the probabilistic framework, was to keep track of subtree frequencies and subtree conditional frequencies. This way, with only minor approximations, our method can take as input a posterior tree distribution from software like mrBayes or RevBayes and yield the same results as if it had started from the gene alignments.

There is some irony in the fact that it took us several years and detours to come up with a solution that is similar to Edwards' 2007 algorithm. However, our solution to the problem is more efficient than Edwards', but also more accurate. Because it considers subtrees, amalgamation has the property that it generalises the input tree distribution: it offers estimates of the probability of trees never observed in the tree distribution, but made of subtrees that have been observed in the tree distribution.

SYST BIOL AMALGAMATION PAPER HERE

- Science 2013

In parallel to our work on DL and DTL models, several researchers were actively developing fast methods for inferring species trees with the multispecies coalescent. They had abandoned the idea of inferring better gene trees by taking into account the species tree, because it was computationally too costly. Instead, they were developing “summary methods”, which take individual gene trees as input. The methods became very popular, but suffered from that problem I mentioned above: there is only a limited amount of information in gene alignments, so point estimates of the gene trees necessarily contain mistakes.

Users of these methods noticed this problem, and some researchers became highly critical of all summary methods. Since each gene alignment contains so little information that you can’t trust the gene trees, why reconstruct gene trees? Can’t we just concatenate all gene alignments and infer one species tree? Supporters of summary methods pointed out that the concatenation approach can be positively misleading: in some part of parameter space, the multispecies coalescent would generate gene alignments that would lead the concatenation approach to infer an incorrect species tree, even with an infinite amount of data. To the outside observer, it was apparent that we had a fight between a method that did not work in practice and a method that could not work in theory.

Unfortunately, the amalgamation trick could not be used here, because subtrees cannot be treated independently of each other. Further, the community of phylogeneticists was eager to use larger and larger data sets, which prohibited the use of methods that jointly estimate gene trees and species trees. My colleagues Siavash Mirarab, Tandy Warnow and myself came up with a trick.

The trick was to improve gene tree estimation without relying on the species tree, and is very simple. It acknowledges that single gene alignments do not contain enough information for reconstructing accurate gene trees, but considers that there may be enough information in a concatenate of a few gene alignments that share similar phylogenetic signal. It operates as follows: first, reconstruct gene trees based on individual alignments; second; concatenate together gene alignments whose gene trees agree enough, based on a support threshold value; third, repeat the first two steps as long as some concatenation can be done; fourth, infer a species tree using a summary method from the trees inferred from the concatenates. This simple trick performed really well on simulations and was used to infer a species tree for a bird phylogenomic project (Jarvis et al., 2014).

This trick improved the poor empirical performance of summary methods, but at an important cost: we had lost the theoretical guarantees that summary methods should yield the correct species tree if they were given as input the true gene trees. We addressed this problem in another manuscript by slightly modifying the algorithm: instead of running the summary method directly on the trees inferred from the concatenates, we reweigh the concatenates by the number of genes they contain (Bayzid et al., 2015). This little modification was enough for us to prove that the whole method, when run on an infinite number of true gene trees generated according to the multispecies coalescent, would return the correct species tree.

SCIENCE 2013 PAPER

Gene tree-species tree models to learn about genome evolution

One important motivation behind the development of PHYLDOG was to study the dynamics of gene duplications and losses in mammals. Could we observe more duplications on branches of the species tree where the effective population size was smaller? Could we identify cases of neofunctionalization or subfunctionalization after gene duplications? I hoped that the improved accuracy of gene trees reconstructed with DL models could bring a new light on those questions.

It turns out that these expectations had been too naive. Despite efforts by Ensembl, mammalian genomes were at the time pretty poor. Firstly, many genes were missing from the genome assemblies because genome coverage for some of the species was very low: some species had 2X coverage, meaning on average each site of the genome was covered by only 2 reads. Secondly, gene annotation was not very reliable. In some gene families, different isoforms had been chosen for different species. The result was that gene trees clustered together closely related isoforms instead of closely related genes. Some gene trees were spectacularly messy, and could group together sequences from a pig, a cat and a primate on one side, and from a cow, a dog and another primate on another side.

Despite my efforts, I could not find a way to study the biological questions that had motivated me with these data. Frustrated by our attempts to study genomes from multicellular Eukaryotes, we turned our interests to prokaryotes and unicellular Eukaryotes. Their genomes are much smaller, which means they are easy to sequence entirely, so we don’t have to worry about missing data so much. And they have no or very few introns, so we don’t have to worry about gene annotation so much.

One study we performed dealt with genome evolution in Cyanobacteria and in a subclade of Fungi. We were curious to see whether our DTL model would detect different amounts of gene transfer between these two clades. We found that, after normalization, the rates of gene transfer were similar between Cyanobacteria and Fungi. However, gene transfer in Fungi was highly heterogeneous: some clades had lots, and others, very little. It would be interesting to investigate the determinants of these heterogeneities: they could notably be linked to differences in the lifestyles of these fungi, or to differences in their molecular machineries.

DTL models provide an interesting window into inferring ancient lifestyles, because they allow reconstructing ancestral gene contents. Since they reconstruct scenarios of gene family evolution, by placing events of gene duplication, transfer, and loss, they also reconstruct for each gene family the number of genes in each ancestral genome. From these ancestral genomes, assuming ancestral genes had functions similar to their extant relatives, one could infer ancestral

lifestyles. This is something I had played with as an undergrad with simple parsimony methods (Boussau et al., 2004), but we still have to exploit fully this idea with better ancestral genomes based on DTL models.

This work resulted in the first paper of Adrian Davín, a PhD student I co-advised with Vincent Daubin.

PAPIER FUNGI PHIL TRANS 2015

- PNAS Archaea

Our DTL model is fairly unique in its capabilities. It caught the eye of a few colleagues around the world. In particular Tom Williams, from Bristol University, expressed interest in our methods. He had been trying with little success to root Archaea using branch-heterogeneous models of sequence evolution. These models, which formed the heart of my PhD work, are non-reversible, meaning they provide different likelihoods to phylogenetic trees rooted on different branches. Unfortunately so far they have not yielded convincing results on their ability to root empirical data, even though they can work on simulated data. Tom contacted us with the hope that DTL models could help him root Archaea.

DTL models can in principle root species trees, because different rootings of the species tree impose different scenarios of gene duplications, transfers and losses. Our hope was that, when run on thousands of gene families, DTL models would have enough statistical power to distinguish between candidate roots.

We applied our DTL model on Tom's data and indeed found that they favoured a particular root of the Archaeal tree. We also took a look at the ancestral genomes that DTL models could reconstruct, and inferred the lifestyles of ancient Archaea that lives hundreds or thousands of million years ago. There again, much improvement coud be done on the way we interpreted ancestral gene contents. In particular, ancestral gene contents are inferred one gene family at a time, independently of other gene families. There is much to gain by considering several gene families together, because many genes work in collaboration with other genes and can't do much on their own. Using this functional information should improve individual gene histories and thus ancestral gene contents.

PNAS ARCHAEA HERE.

Gene tree-species tree models to date species phylogenies

So far we have seen that gene tree-species tree models could be used to reconstruct better gene trees, species trees, and to study genome evolution. In particular, DTL models allow inferring ancient lateral gene transfers. These transfers in turn tell tales of ancient proximities.

For a transfer to occur between two species, the two species have to be in close proximity, both in time and in space. Gene transfers can occur through direct contact between two cells, or through egoist elements in which they get inserted, or through transformation, *i.e.* the capture of free-floating DNA coming from a cell that has died. In all those cases, there cannot be a lot of space or a lot of time separating the two cells, because DNA and egoist elements have a limited half-life outside of living cells. Therefore, any transfer we detect is a testimony of an ancient proximity between the donor and the recipient cells.

One could use the ancient transfers that DTL models detect to infer ancient spatial proximities between species. Species that exchange a lot of genes probably shared an ecological niche. We could then confront such predictions to predictions based on gene content. We have not pursued this type of research yet. Instead, we have focused on what ancient transfers tell us about time.

If we were able to detect donor and recipient species of individual gene transfers, we could directly infer that those two species lived at the exact same time. Unfortunately, we do not have access to individual gene transfers, because we are limited by our taxonomic sampling. Many species that lived at the time of the transfers we detect have left no descendants among the extant species we have sampled. Their lineages may have become extinct, or we may have failed to sequence their extant descendants. It is thus very likely that we detect composite transfers, i.e. transfers that actually represent a series of evolutionary events, including speciations, duplications, or transfers. An example of a possible composite transfer would be a gene that originated in a donor lineage whose descendants we have sampled, then was transferred to some ancient species whose descendants we have not sampled, evolved in those species for some time, and finally was transferred into some species whose descendants we have sampled. This is what Figure 3b shows: a descendant of species D transferred a gene to some species that has left no sampled descendant, but in turn transferred a gene to an ancestor of species R.

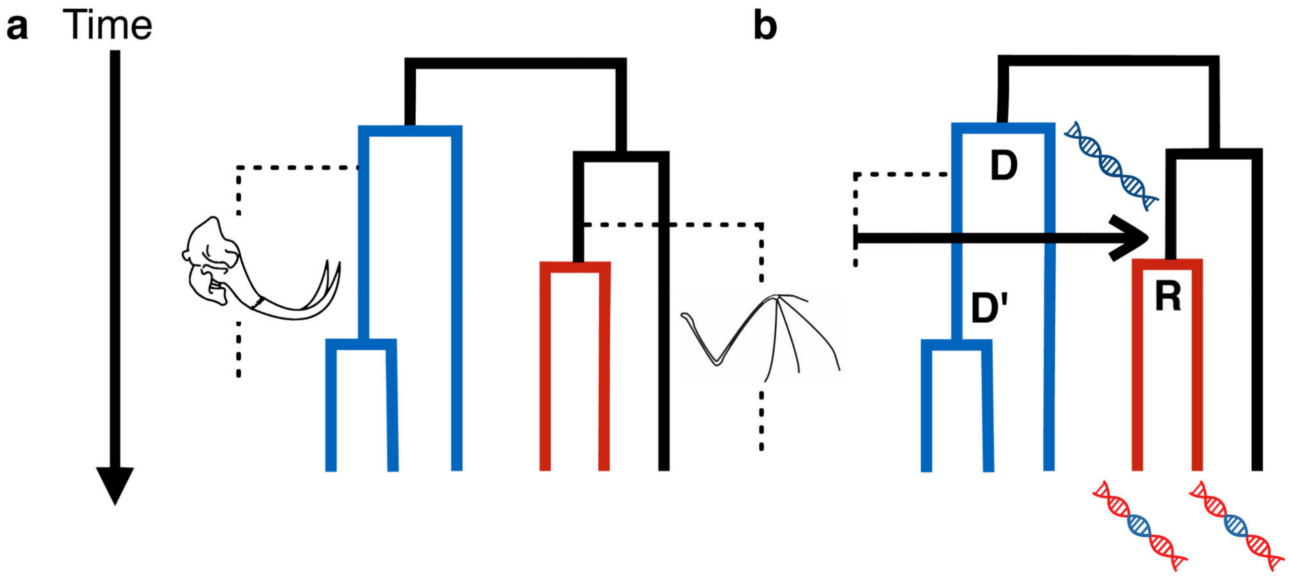


Figure 3 (2 again!): Gene transfers, like fossils, carry information on the timing of species divergence. a) The geological record provides the only source of information concerning absolute time: the age of the oldest fossil representative of a clade provides direct evidence on its minimum age, but inferring maximum age constraints (e.g. dashed line for the red clade), and by extension the relative age of speciation nodes, must rely on indirect evidence on the absence of fossils in the geological record. b) Gene transfers, in contrast, do not carry information on absolute time, but they do define relative node age constraints by providing direct evidence for the relative age of speciation events: the gene transfer depicted by the black arrow implies that the diversification of the blue donor clade predates the diversification of the red clade (i.e. node D is necessarily older than node R).

Another composite transfer is illustrated in Figure 4. Figure 4b shows a composite transfer that could be detected by our DTL models, between an ancestor of species F and E and an ancestor of species B. What actually occurred is a speciation between a lineage that led to species E and F, and an extinct or unsampled lineage leading to D, and then an individual transfer between the lineage leading to D and an ancestor of species B.

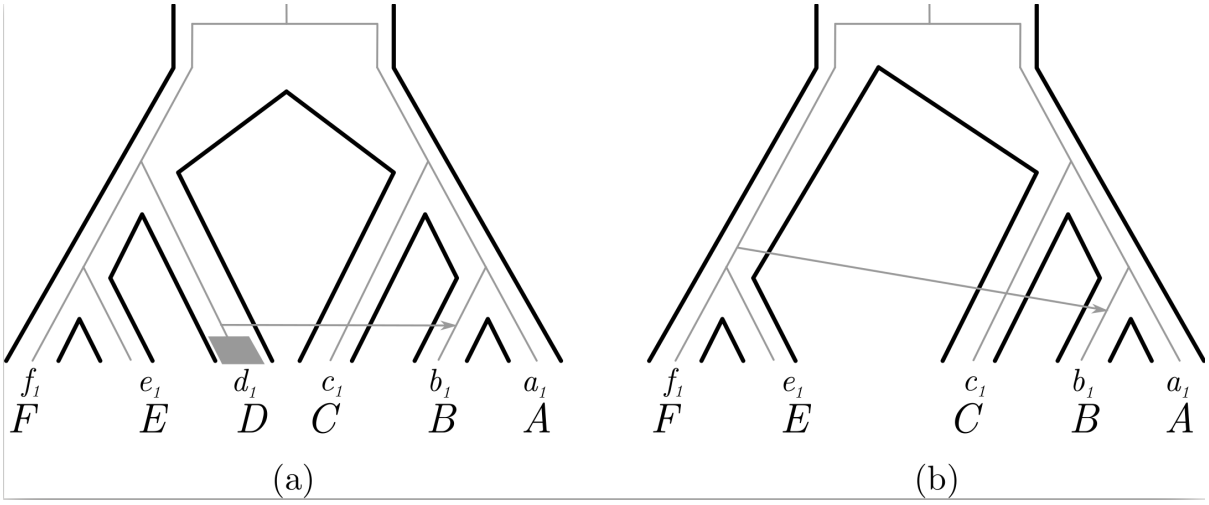


Figure 4: DTL models can detect composite gene transfers, not individual transfers. The true evolutionary scenario is depicted in a). What DTL models can capture at best is the transfer in b), because we have not sampled D, or D has become extinct. This transfer is a composite transfer, that combines a speciation event, evolution along an unsampled lineage, and a transfer. Figure borrowed from Boussau and Scornavacca, under review.

Despite our inability to find individual gene transfers that would prove that two ancient species where contemporaneous, composite gene transfers detected by DTL models still carry dating information: they tell us that the ancestor of the donor species is necessarily older than the descendant of the recipient species: D is older than R in Figure 3b.

My colleagues and I have developed methods to exploit this dating information from gene transfers. In particular we have developed methods to order nodes of a species tree thanks to the thousands of gene transfers that DTL models infer from genome-wide analyses (Chauve et al., 2017; Davín et al., 2017). This is a project that was led by Adrian Davín, a PhD student I co-supervised with Vincent Daubin. What was really hard about this project was to convince ourselves that the dating information contained in gene transfers was trustworthy. The main evidence we have that they contain *bona fide* dating information is that they agree with relaxed molecular clocks. We were worried that perhaps well-known phylogenetic artifacts such as long branch attraction due to accelerated rates of evolution in some lineages would create artifactual agreement between transfer-based dating and relaxed clocks. The many tests we did never exposed such a bias, so we ended up trusting our results.

We argue that transfers contain *bona fide* dating information in Fungi, Archaea, and Cyanobacteria. We believe this is an important result because there is little dating information available to date phylogenies across the tree of life: fossils are rare, sometimes difficult to associate to a particular ancient lineage, and they are limited to species that fossilize well. In particular the fossil record for microbial organisms is nearly non-existent. My colleagues and I are currently working to apply transfer-based dating across the tree of life.

PAPER NEE HERE

Perspectives

- Perspective 1: dating with transfers and relaxed clocks in RevBayes

In Davín et al. 2017 we were able to use the dating information contained in gene transfers to order the nodes of a species phylogeny relative to each other. We found that this order agreed well with relaxed molecular clock estimates of dated phylogenies. A natural extension of this work is to combine relaxed molecular clock and relative constraints in a probabilistic method to date species trees.

I have implemented such a method in RevBayes. We have tested this implementation and are currently in the final phases of its testing on simulations and on empirical data. Its principle is very simple: it samples dated trees during the MCMC algorithm but rejects all trees that disagree with transfer-based constraints. An improved version of this method, that still needs to be developed, would associate a probability to each transfer-based constraint.

- Perspective 2: team Le cocon

My colleagues Eric Tannier, Annabelle Haudry, Laurent Guéguén, Damien de Vienne and Vincent Daubin and I have decided to create a new research team called Le cocon in the LBBE, which became active in January 2019. One of the central themes of this team will be gene tree-species tree models. In particular, we will continue exploring the ability of DTL models to date species trees and to illuminate the ancient history of life. However, we will also extend the range of DTL models to tackle host-parasite or host-symbiont relationships. This will be an opportunity for us to collaborate more closely with colleagues of the LBBE that specifically study symbioses, notably on insect model systems.

Another defining theme in Le cocon will be our desire to engage with society at large, beyond the academic world. We all have the desire to communicate about our work, but we also feel that we can and perhaps should tie closer relationships with our fellow citizens. In part three of this manuscript, I will elaborate on my attempts to do so by focusing on environmental issues.

Convergent evolution

Students of historical sciences face a replicability problem: it is often difficult to replicate processes that unfold over long time scales, under a set of conditions that may be unique. To circumvent this problem, one has to turn to “pseudo-replicates”: events that resemble each other so closely that we can hope to learn from their commonalities much like an experimental scientist would learn from her replicates. In evolutionary biology, cases of convergent evolution have occurred during such pseudo-replicates: confronted with similar selective pressures, different species have evolved similar phenotypes. I study such cases of convergent evolution to learn

about the link between the genotype and the phenotype: when convergent phenotypes have appeared, have they been caused by convergent changes in the genomes? Were identical substitutions at particular sites involved? If not, were some homologous genes repeatedly recruited? If not, were different genes from a particular pathway repeatedly recruited? We can then learn if there are few or many different ways to produce a particular phenotype. We can also investigate whether confounding factors such as mutation biases or changes in effective population sizes can generate non-adaptive convergent genomic evolution.

Convergenomix

I have been the coordinator on the ANR grant "Convergenomix" that aims at studying the genomic underpinnings of convergent phenotypic evolution in 3 clades of animals. This project gathers 4 teams from 4 different labs from Lyon, the LBBE, the LBMC with Sophie Pantalacci and Marie Sémon, the LEHNA with Tristan Lefébure and Christophe Douady, and the IGFL with Abderrahman Khila. For each clade, we have been generating transcriptomic data sets with similar sequencing and bioinformatic tools to make sure that the results we obtain can be compared across clades. The sampling in the clades was also meant to be similar across clades, as shown in Figure 5.

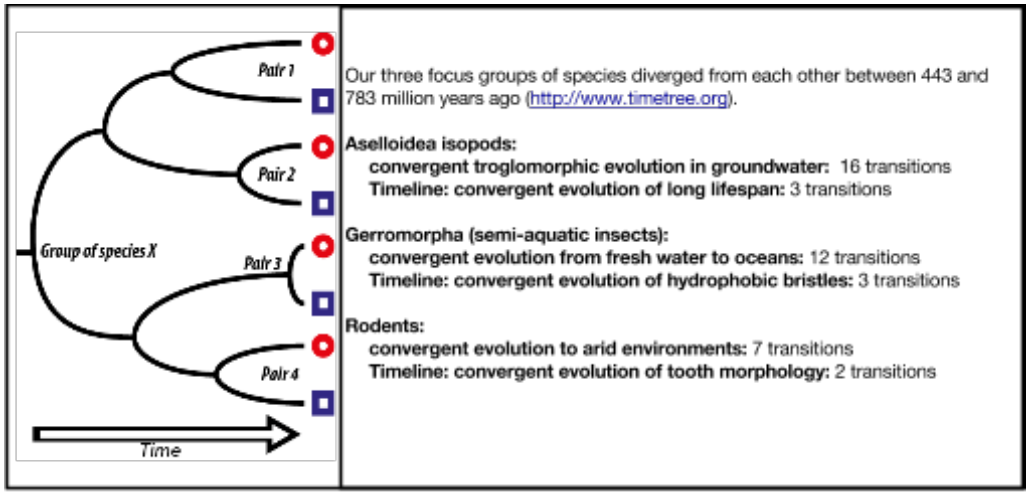


Figure 5: Experimental design: in each of the 3 groups, species pairs showing convergent phenotypic changes (here between red circle and blue square) are studied for convergent evolution at different evolutionary depths (from 7 to 16 species pairs per group; here only 4 pairs are shown).

My role in this consortium has notably been to oversee the development of bioinformatic methods for data set assembly as well as for sequence analysis. This has resulted in two publications by Carine Rey, a PhD student that I am co-supervising with Marie Sémon (LBMC). The first one presents a new method, CAARS, for assembling genes from sequencing data with the help of assembly information from related species, including distantly related species, taking ad-

vantage of their phylogenetic context (Rey *et al*., 2018b). The second one presents a new probabilistic method, PCOC, to identify convergent substitutions (Rey *et al*., 2018a). In both cases, we compared our approaches to state-of-the art methods, and we find that our developments have improved markedly upon the competition. A postdoctoral researcher I co-advise with Philippe Veber (LBBE), Vincent Lanore, has also worked on another method, Diffsel, originally developed by Nicolas Lartillot (LBBE), to detect convergent genomic evolution. This method is similar in spirit to PCOC, but uses the Bayesian framework instead of the ML framework. Vincent Lanore is a computer scientist by training and has done tremendous amounts of work to re-implement diffsel to make it faster but also easier to modify, and to parallelize it. Vincent and Carine are co-first authors on a manuscript in its second review round where we analyze the ability to detect convergent genomic evolution in the presence of changes in effective population size, for all published methods, including diffsel and PCOC (Rey et al., under review). We find that model-based approaches, ours in particular, are more robust to this confounding factor.

We have also started applying our tools to empirical data sets. To this end, we obtained 5.5 million hours of computing time on the OCCIGEN supercomputer. In particular, with the team of Marie Sémon and Sophie Pantalacci (LBMC), we are writing up a manuscript presenting analyses of convergent evolution to arid environments in rodents. We notably found one interesting candidate gene that harbours several convergent substitutions. Importantly, this gene is expressed in the kidney with expression levels that convergently change between species living in arid environments and species living in mesic environments.

Building and annotating families of homologous genes from transcriptomes

One motivation for the Convergenomix project was to develop a bioinformatic pipeline to generate high quality families of homogous genes and identify paralogs and orthologs. These are necessary steps for a vast array of analyses, but it seems that each and every team has their own way of doing things. At least within Convergenomix, we wanted to have a standardized approach to building families so that we could compare the results across taxa and perhaps learn something general about convergent evolution in animals.

We felt we could improve upon commonly used variants of such pipelines by taking a comparative approach. In particular, it is usual to assemble genes from a species just based on the sequencing reads of that species, in isolation from the assembled genes of other species. This seemed like an important waste of potentially useful information. In addition, we wanted to annotate paralogs and orthologs based on reconciled gene phylogenies built with a gene tree-species tree model.

Carine Rey thus implemented CAARS, a program to build and annotate gene sequences from sequencing reads. CAARS uses annotated gene sequences from user-chosen species to help

assemble the transcripts of a target species, and places the assembled transcripts within families of homologous genes. Then it uses Phyldog to build reconciled gene trees and identify orthologs and paralogs.

Our tests show that CAARS is more accurate than a commonly used pipeline, and we have been using it in the context of Convergenomix. However, there is still much room for improvement, in particular in terms of computational efficiency. The accuracy of CAARS comes at a high computational cost, and it would be extremely useful to be able to reduce it.

CAARS PAPER HERE

A model-based method to detect convergent amino acid substitutions

One of the first objectives of Convergenomix was to identify convergent substitutions, *i.e.* substitutions that subtend convergent phenotypes. Several methods had been proposed in the past, but they all had shortcomings we wanted to avoid. Besides, when used on the same data sets, they provided different answers.

We decided to develop a probabilistic model to detect convergent amino acid substitutions. First, we had to define what would constitute an interesting convergent substitution. Here, "interesting" must be understood as "interesting enough that we would want to follow up on this substitution with experimental validation in the lab". We agreed on the following definition, in layman's terms: an interesting convergent substitution would be such that on every branch where the phenotype changes from its ancestral state to the convergent state, the site undergoes a substitution towards some particular type of amino acid. We then developed a probabilistic model to embody this definition within a program called PCOC.

The model behind PCOC works with amino acid profiles. An amino acid profile is a vector of 20 frequencies, one per amino acid state. It describes the equilibrium frequencies of a model of sequence evolution: if we let evolution run for an infinite amount of time, the probabilities that we end in each state will be those 20 frequencies. We interpret those as a proxy to the amino acid fitnesses at a particular site. We use this profile idea in our definition of a convergent site, which now becomes: an interesting convergent substitution would be such that on every branch where the phenotype changes from its ancestral state to the convergent state, the site undergoes a substitution towards a particular amino acid profile that differs from the ancestral amino acid profile. Figure 6 illustrates our definition.

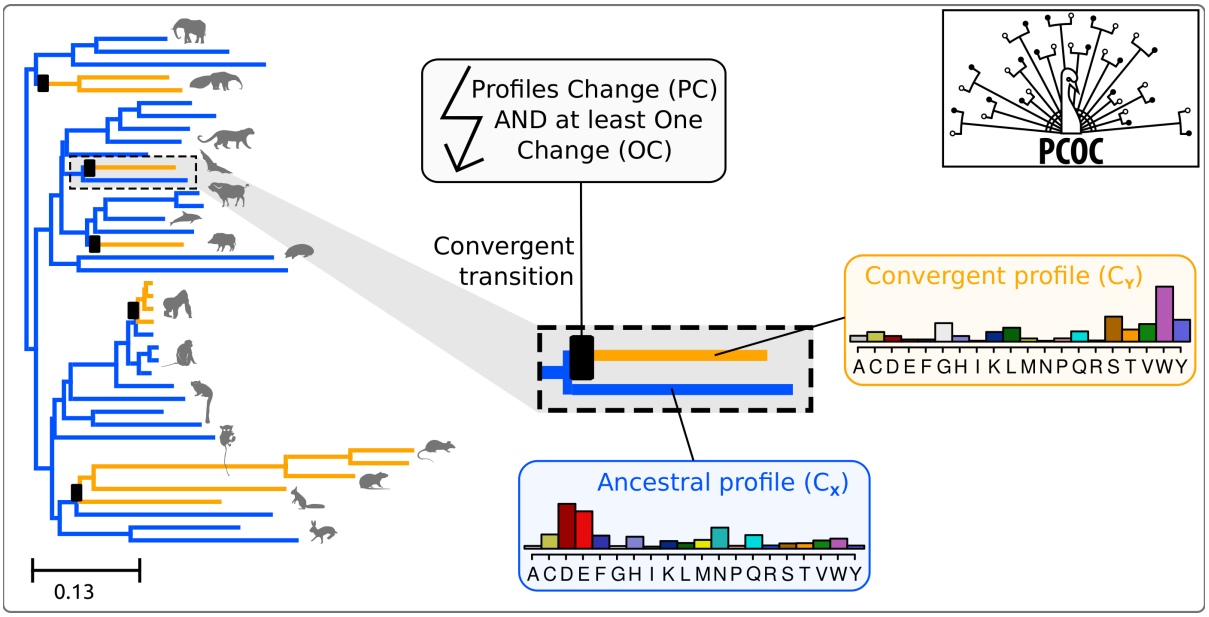


Figure 6: The definition of a convergent substitution in PCOC. Lineages with the convergent phenotype are in orange, those with the ancestral phenotype in blue. Bar plots represent amino acid profiles.

To detect convergent amino acid evolution with PCOC, we compare the likelihood of two models: one where we assume the site is convergent, and one where we assume that the site evolves according to a single profile on all branches, without necessarily undergoing a substitution on branches where the phenotype changes from ancestral to convergent. Convergent sites should significantly prefer the former model over the latter.

The article below provides more details on how PCOC works. We found that it could recover previously identified candidate sites on empirical data, and that it was performing very well on simulations that fit its definition. This latter statement is not very surprising, but illustrates a shortcoming of computational research. We tend to validate methods on simulated data, and we often simulate data with models very similar to those we use for inference. In the present case, we tried to make a mismatch between the reconstruction model and the simulation model, by using different sets of amino acid profiles. But the structure of the models remained the same. We were not able to come up with a better simulation protocol.

PCOC PAPER HERE

A comparison of methods to detect convergent substitutions

We decided to pursue our research started with the article above. Firstly, not all published methods to detect convergent substitutions had been compared in the article above. In particular, two model-based approaches, diffsel and tdg09, had not been included.

Secondly, the simulations had been performed under the PCOC model, in particular assuming that there had to be a substitution on every branch where the phenotype changed. This modelling choice made sense for detecting a very specific pattern, but was less natural for simulation. Instead, we wanted to try a more mechanistic simulation model, where phenotype changes would be associated to amino acid fitness profile changes. Under this model, patterns similar to those generated by PCOC, with substitutions on all branches where the phenotype changed, are still possible, in the cases where the differences between the ancestral and the convergent profiles are very pronounced. But subtler and more realistic patterns, with substitutions only on a subset of the branches where the phenotype changed, can also occur.

Thirdly, we wanted to explore some confounding factors that could generate spurious convergent substitutions, which could mislead detection methods. In particular, we explored the impact of changes in the efficacy of selection on the ability of detection methods to identify *bona fide* convergent substitutions.

Carine Rey and Vincent Lanore, a postdoctoral researcher co-advised by Philippe Veber and myself, did most of the analyses. The resulting paper is currently in its second round of reviews in *Philosophical Transactions of the Royal Society B.*

PAPER PHIL TRANS HERE.

Perspectives

A flexible framework for constructing Bayesian models to detect convergent molecular evolution

We were fortunate in the Convergenomix project to hire Vincent Lanore, a computer scientist by training, who specializes in software engineering and parallel computing. Early on we decided he would develop methods to detect convergent genomic evolution starting from code by Nicolas Lartillot, in the Bayesian framework. Since then, Vincent has deeply changed the way probabilistic models are implemented in this code base. It is now easier to develop complex Bayesian models that are both efficient and correct, and that can run in parallel on clusters of computers. We are currently using this code base for our analyses. Looking forward, we will also rely on it to develop more sophisticated models to improve our ability to detect *bona fide* convergent substitutions.

Studying convergent adaptation to abiotic stresses in plants

The Convergenomix project will end in 2020. We have produced several tools that can be used at the genomic scale to identify sites likely to subtend convergent phenotypes, and we are now applying them to animals in collaboration with the teams involved in the project. However, convergent evolution is not limited to animal genomes, and other clades appear to be very interesting.

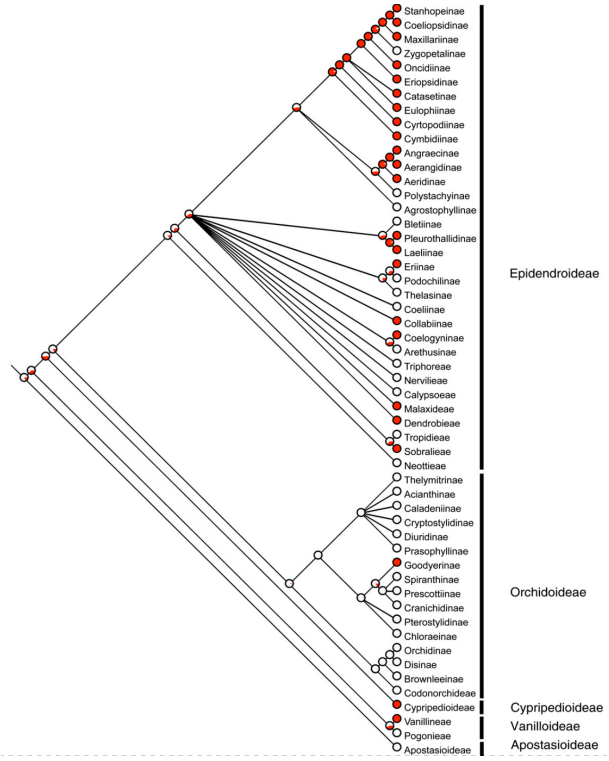
I plan to build upon the Convergenomix project to study convergent genomic evolution in plants that can tolerate thermal, salinity and drought stresses, arguably the most important abiotic stresses. These stresses are expected to become more widespread and intense as climate change unfolds. For instance, focusing on crops, recent episodes of drought have had severe effects: in 2010 wheat yield decreased by 33% in Russia, and in 2003 corn yield decreased by 30% in France (van der Velde *et al*., 2003). Further climatic projections suggest that higher heat will result in decreases of up to 30% on crop yields, even without taking extreme events into account (Gammans *et al*., 2017). Climate change will also alter precipitation regimes, which could increase salinity in agricultural fields: in particular, lower precipitations will lead to more irrigation, which increases salinity. Overall, it will be very useful to understand how plants can adapt to thermal, drought and salinity stresses; it may even be important for securing crop yields in future years.

There have been several instances of convergent evolution to abiotic stresses in plants. For instance the CAM metabolism, which provides tolerance to drought, has evolved many times in independent lineages, about a dozen times in Orchidaceae alone (Silvera et al. 2009, Fig. 7), and the ability to grow in the presence of high salinity has evolved about 70 times in grasses alone (Bromham 2015, Fig. 8). The convergenomix project has delivered solid methodological foundations for benefiting from these pseudo-replicates and identify in the genomes of plants the genes and the sites involved in the tolerance to abiotic stresses.

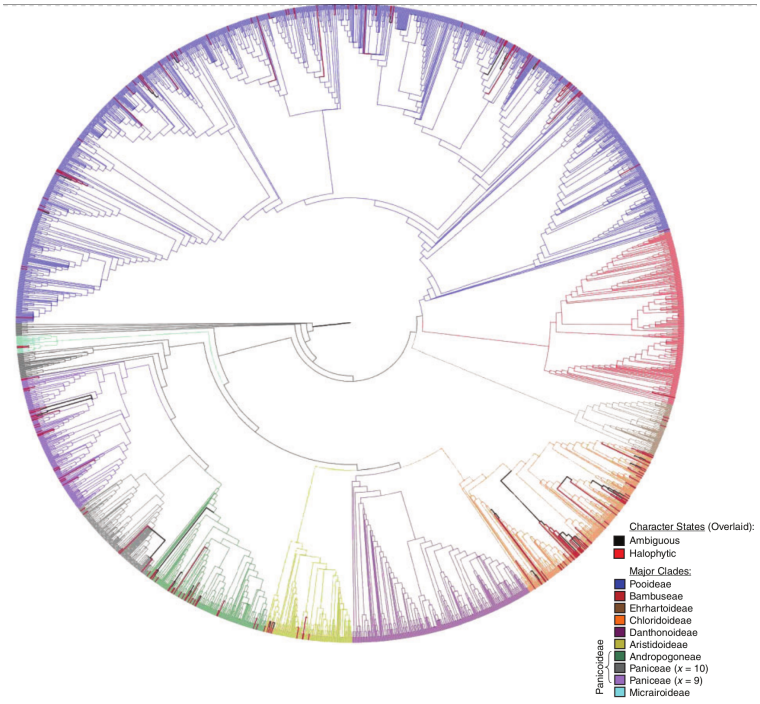
I will build upon the PCOC and Diffsel software and extend them to identify convergent adaptations that are linked to abiotic stresses. One useful extension will be to go beyond discrete phenotypic variables. In Convergenomix I have focused on situations where there are two states: the ancestral state, and the convergent state. I will extend the models to handle continuous variables (e.g. summer temperatures or the amount of precipitation in the area where the plant has been sampled), or hierarchically ordered discrete variables. The former would be methodologically similar to work done by Lartillot and Poujol (2010). The latter would be useful to use e.g. phenotype ontology databases (for instance the planteome database) or to specify the different types of photosynthetic metabolisms found in plants. They are generally clustered in 3 groups: C3, C4 and CAM metabolisms. C4 and CAM may be considered as more closely related, because they both work by concentrating CO2 around the RuBisCO enzyme for performing photosynthesis in difficult growing conditions: C4 does so through spatial compartmentalization, and CAM does so through temporal compartmentalization. One might expect that some genes show more similarities between the C4 and the CAM metabolisms than between either of those

and C3. Another extension of the models that would be crucial is the ability to consider several overlapping phenotypes together: some plants can be both heat and salinity tolerant for instance, and have a CAM metabolism, while others can be C3 heat tolerant plants. Considering several phenotypes together would be especially useful since some genes may be involved in the tolerance to any abiotic stress, and so could show signals of convergent evolution in plants adapted to heat, drought, and salinity. Considering several stresses at a time will allow distinguishing genes that are involved in the tolerance to a particular stress from those that are involved in the tolerance to abiotic stresses. Finally, given the evolutionary distances considered, it would be useful to relax the assumption that all taxa in a given condition share the same profile at a convergent site. Instead, it would be useful to consider that closely related taxa are likely to develop a particular phenotype in a similar way, while distantly related taxa are less likely to do so. All those developments can be included naturally in a Bayesian framework as used in Diffsel, which will allow me to benefit from its parallel framework to handle whole genome data. In the end, this project will identify candidate sites subtending tolerance to important abiotic stresses. These sites can then be validated by *in vitro* and *in vivo* experiments. Validated sites might then be ultimately used for selection and breeding of cultivars or varieties in which they are found as natural variants, or for introduction into plants by genomic engineering, as a last resort.

Data for such a project will be easy to come by, because plant genomic data is accumulating fast. The 1kP project, which aimed to generate 1 000 plant transcriptomes, is now being followed by the 10kP project, which plans on sequencing 6 000 plant genomes in the next couple of years. Phenotypic data should not be a problem either, because phenotypic databases for plants exist (e.g. the TRY database), and the phenotypes I am interested in are easy to assess or can be gathered from environmental variables of the distribution area of a plant.



7: Phylogenetic tree of 53 subtribes of Orchidaceae showing plants with CAM metabolism in red. About a dozen convergent transitions to CAM appear to have taken place in this clade (Silvera et al. 2009).



8: Phylogeny of grasses from Bromham 2015, in which halophytic plants have been colored in red. About 70 independent origins of halophily have been found in grasses.

Science in the Anthropocene

It has become undeniable that climate change is unfolding right now and that we as a society must act quickly and radically to avoid its worst effects. For instance, as a rule of thumb, it is commonly considered that the average French person should reduce his.her CO2 footprint 4 or 5 fold to stay in line with the Paris Agreement, which is extremely challenging to do. Among other things, it basically requires not owning a car, having a very well isolated apartment and not heating it much, avoiding planes, repairing and reusing things instead of buying them new, and eating much less meat. Such massive changes have to occur across the board, in all aspects of our lives, including in our professional lives. Yet the scientific community so far has not taken measures to reduce its environmental footprint. This is particularly problematic since the scientific community is overwhelmingly convinced that climate change is real and caused by humans. If those who are most convinced of climate change do not act to mitigate it, who will? This past year I have started working on this situation. I have started to investigate the environmental footprint of academic work, and I have drafted a plan to reduce the environmental footprint of scientific research. Finally, as I became more knowledgeable about the environmental crises, I decided to give seminars on the environmental crisis to the lay audience.

Groupe de Travail Empreinte Ecologique

All human activities have an environmental footprint. However, it is still very unclear how much of an impact academic research has, and it would be very useful to know, among our activities, which have the largest impacts and which have the lowest. To investigate this, I have started the “Groupe de Travail Empreinte Ecologique” in the LBBE. One of its primary goals is to measure and analyse the detailed carbon footprint of the whole lab. This laboratory is quite interesting because it harbours several different activities whose footprint needs to be assessed: field work in France or abroad, lab work, computational work, and of course conferences and

seminars. The direction of the lab supports this initiative, several members of the lab have joined me in this endeavour, and we are currently collecting the data.

Reducing the environmental footprint of research

Beyond diagnosing the environmental footprint of research, I consider that we also need to decrease our footprint as a community. I have been drafting a plan to mitigate the environmental footprint of research by targetting funding agencies. I have named this plan “plan D”, D as in Decarbonization, to mirror “plan S” launched earlier this year by European research agencies and science funders to favour open access to scientific research. They all agreed on a set of principles that research they fund must follow. This has been perceived as an important step forward for open access (Else 2018).

I propose that scientists should agree on a path to decrease their professional carbon footprint, with explicit targets for greenhouse gas emissions in 2030, 2040 and 2050. To achieve this, research agencies and science funders will need to add a carbon footprint criterion to their calls for grant proposals, so that each applicant will be required to provide an estimate of the carbon footprint of their project. Each funder will get a yearly carbon budget to split between grants. Projects would still be evaluated based on their scientific merit, in the limits of the total carbon and financial budgets of the funder.

The challenges associated to such a project can all be overcome. Firstly it would just add another criterion on grant proposals; some funders already include gender balance or the amount of permanent/temporary contracts as non-budgetary criteria in their grants. Secondly, the estimation of the carbon footprint can be automated so that applicants would only need to fill out a spreadsheet to get the footprint of the proposed research. Thirdly, as shown by the “plan S”, research agencies can agree on important principles and take action. Here they would need to pledge to reduce their footprint to meet targets agreed-upon during an international summit. A yearly global footprint would be split between funders based on their yearly footprint at the time of the agreement. Every five years, the distribution of the emission rights would be reassessed to balance rights between developed and developing countries notably. An international committee would be in charge of managing the agreement and checking that emissions do not exceed the rights.

I am still currently reflecting over the details of the plan and its proposed implementation, but I have had a draft read by a dozen colleagues, who all gave positive feedback. I will present it to the entire research community in the next few months, and hope to see it discussed and implemented.

Outreach

In late 2018, a colleague of mine discussed for a few hours with three M2 students who were in a state of distress. They had come to him with the following question: “what is the point

of studying and doing research if our societies are on the verge of collapse?". This anecdote illustrates that the idea of an imminent collapse of our society has recently become widespread in France, with several books, popular youtube videos and magazines discussing about it. The general idea behind collapse is that climate change, finite resources, political and economical instabilities all conspire to make the current state of our societies unsustainable so that, sooner or later, our society will collapse. Such a point of view is notably publicized by Pablo Servigne and Yves Cochet, a former minister of ecology in France. They argue that 10 years from now, our lives will have changed drastically. This obviously clashes with more optimistic views of the future, in particular those that expect that the current state of affair will keep on improving as technological progress continues.

Given this context, I have decided to examine the primary scientific literature to summarize what science has to say on this matter, and to communicate about the result of this research. Since September 2018, in collaboration with Mathilde Paris, I have given two talks to the lay audience, one during the "Fête de la Science", and one during the "Rencontres régionales de l'éducation et de la promotion de la santé et de l'environnement" which was followed by a round table (https://goo.gl/eaCy5L and http://goo.gl/GvFtqj, respectively). The feedback I received after these two talks confirms that the lay audience is in demand for such information, and motivates me to continue to give talks on this topic in the next few years.

Conclusion: a scientist in the Anthropocene

Warning: work in progress and scattered thoughts. These paragraphs are the result of ongoing discussions with my colleagues from the team Le Cocon, but I take full responsibility for the content.

Last week-end, friends and I rented a car to go outside of Lyon. We had taken the touristic road on the way out, but on the way back we decided to take the highway. There were some "Gilets Jaunes" on a roundabout close to the tollgate. The toll barriers had been lifted. Yet, I decided to pay. It just did not feel right not to pay. This shows how well brought up I am. Maybe a bit conservative, or docile, too.

It is commonly argued by researchers that scientific research can transform society. A project to study Dengue may argue that it might make people stop worrying about mosquitoes. A project in machine learning may argue that it might make car accidents a thing of the past. A project on the evolution of life may argue that it might change the way people think about themselves.

Another argument we commonly hear from scientists is the following one: "Politics should stay out of the lab; what we're doing is science, and science aims to be objective.". Isn't there a contradiction here? How can we propose to change society and claim at the same time

that we are not doing politics? It could instead be argued that the choice that a researcher makes to work on a specific topic is necessarily politic. Researchers have a lot of freedom to choose the work they want to do. To some extent, they get to decide how some amount of public money –their salary and possibly some grant money– gets spent. Every day when they decide to work on project A instead of project B, they get to nudge the society in a way that they have chosen. Every day they make a political decision.

It took a while for me to realise that; it is difficult to question commonly held views when your natural urge is to pay for tolls when they are free. I was convinced that, because I was doing fundamental research, my activities had no impact on society, were limited to the realm of ideas. It is probably true that as an individual I did not do a huge amount of harm; but now I'd like to try to do some good. Choosing what "good" means is obviously difficult, but I would argue that it is part of my responsibility as a researcher to articulate why society should devote my time and skills to project A instead of project B.

I still believe the methods of scientific research are very efficient at avoiding errors to better understand the world around us. It is important to try to falsify hypotheses, to acknowledge one's own biases, to subject one's work to critical reviews. Through the years, I've learned some of those methods, which I intend to put to use in the projects I have outlined in this manuscript. I intend to teach those methods to students that I get to advise, but I hope I will also motivate them to think more critically about their work than I did when I was a student myself.

References

My name appears underlined and in bold, and with an asterisk when I have last author status.

Bayzid MS, Mirarab S, **Boussau B**, Warnow T. Weighted Statistical Binning: Enabling Statistically Consistent Genome-Scale Phylogenetic Analyses. *PLoS One.* (2015) 10 ppe0129183.

Boussau B, Karlberg EO, Frank AC, Legault BA, Andersson SG. Computational inference of scenarios for alpha-proteobacterial genome evolution. *Proc Natl Acad Sci U S A.* (2004) 101:9722-9727.

Boussau B, Daubin V. Genomes as documents of evolutionary history. *Trends Ecol Evol.* (2010) 25 pp224-32.

Boussau B, Szöllosi GJ, Duret L, Gouy M, Tannier E, Daubin V. Genome-scale coestimation of species and gene trees. *Genome Res.* (2013) 23 pp323-30.

Boussau B, Scornavacca C. Reconciling gene trees with species trees. *Book chapter*, under review.

Bromham L. Macroevolutionary patterns of salt tolerance in angiosperms. *Annals of Botany*. (2015) 115 pp 333-341.

Chauve C, Rafiey A, Davín AA, Scornavacca C, Veber P, **Boussau B**, Szöllosi GJ, Daubin V, Tannier E. MaxTiC: Fast ranking of a phylogenetic tree by Maximum Time Consistency with lateral gene transfers. *Recommended by PCI Evol Biol* (2017)

David LA, Alm EJ. Rapid evolutionary innovation during an Archaean genetic expansion. *Nature*. (2011) 469:93-96.

Davín AA, Tannier E, Williams TA, **Boussau B**, Daubin V, Szöllősi GJ. Gene transfers can date the tree of life. *Nat Ecol Evol.* (2018) 2 pp904-909.

Edwards SV, Liu L, Pearl DK. High-resolution species trees without concatenation. *Proc Natl Acad Sci U S A*. (2007) 104 pp5936-5941.

Else H. Radical open-access plan could spell end to journal subscriptions. *Nature* (2018) **561**, pp17-18.

Gammans M, Mérel P, Ortiz-Bobea A. Negative impacts of climate change on cereal yields: statistical evidence from France. *Environmental Research Letters* (2017) pp054007.

Höhna S, Landis MJ, Heath TA, **Boussau B**, Lartillot N, Moore BR, Huelsenbeck JP, Ronquist F. RevBayes: Bayesian Phylogenetic Inferrence Using Graphical Models and an Interactive Model-Specification Language. Syst Biol. (2016) 65 pp726-36.

Jarvis ED, Mirarab S, Aberer AJ, Li B, Houde P, Li C, Ho SY, Faircloth BC, Nabholz B, Howard JT, Suh A, Weber CC, da Fonseca RR, Li J, Zhang F, Li H, Zhou L, Narula N, Liu L, Ganapathy G, **Boussau B**, Bayzid MS, Zavidovych V, Subramanian S, Gabaldón T, Capella-Gutiérrez S, Huerta-Cepas J, Rekepalli B, Munch K, Schierup M, Lindow B, Warren WC, Ray D, Green RE, Bruford MW, Zhan X, Dixon A, Li S, Li N, Huang Y, Derryberry EP, Bertelsen MF, Sheldon FH, Brumfield RT, Mello CV, Lovell PV, Wirthlin M, Schneider MP, Prosdocimi F, Samaniego JA, Vargas Velazquez AM, Alfaro-Núñez A, Campos PF, Petersen B, Sicheritz-Ponten T, Pas A, Bailey T, Scofield P, Bunce M, Lambert DM, Zhou Q, Perelman P, Driskell AC, Shapiro B, Xiong Z, Zeng Y, Liu S, Li Z, Liu B, Wu K, Xiao J, Yinqi X, Zheng Q, Zhang Y, Yang H, Wang J, Smeds L, Rheindt FE, Braun M, Fjeldsa J, Orlando L, Barker FK, Jønsson KA, Johnson W, Koepfli KP, O'Brien S, Haussler D, Ryder OA, Rahbek C, Willerslev E, Graves GR, Glenn TC, McCormack J, Burt D, Ellegren H, Alström P, Edwards SV, Stamatakis A, Mindell DP, Cracraft J, Braun EL, Warnow T, Jun W, Gilbert MT, Zhang G. Whole-genome analyses resolve early branches in the tree of life of modern birds. *Science*. (2014) 346 pp1320-1331.

Lartillot N, Poujol R. A phylogenetic model for investigating correlated evolution of substitution rates and continuous phenotypic characters. *Mol Biol Evol.* (2011) pp729-744.

Mirarab S, Bayzid MS, **Boussau B**, Warnow T. Statistical binning enables an accurate coalescent-based estimation of the avian tree. *Science.* (2014) 346 pp1250463.

(2018a) Rey C, Guéguen L, Sémon M, **Boussau B*.** Accurate detection of convergent amino-acid evolution with PCOC. *Mol Biol Evol.* (2018).

(2018b) Rey C, Veber P, **Boussau B***, Sémon M. CAARS: comparative assembly and annotation of RNA-Seq data. *Bioinformatics*. (2018).

Rey C, Lanore V, Veber P, Gueguen L, Lartillot N, Semon M, **Boussau B***. Detecting convergent adaptive amino acid evolution. *Under review*. http://biorxiv.org/cgi/content/short/513010v1

Silvera K, Santiago LS, Cushman JC, Winter K. Crassulacean Acid Metabolism and Epiphytism Linked to Adaptive Radiations in the Orchidaceae. *Plant physiology.* (2009) pp1838-1847.

Szöllosi GJ, **Boussau B**, Abby SS, Tannier E, Daubin V. Phylogenetic modeling of lateral gene transfer reconstructs the pattern and relative timing of speciations. *Proc Natl Acad Sci U S A.* (2012) 109:17513-8.

(2015a) Szöllősi GJ, Tannier E, Daubin V, **Boussau B***. The inferrence of gene trees with species trees. *Syst Biol.* (2015) 64 pp e42-62.

(2015b) Szöllősi GJ, Davín AA, Tannier E, Daubin V, **Boussau B***. Genome-scale phylogenetic analysis finds extensive gene transfer among fungi. *Philos Trans R Soc Lond B Biol Sci.* (2015) 370 pp20140335

van der Velde M, Tubiello F, Vrieling A, Bouraoui F. Impacts of extreme weather on wheat and maize in France: evaluating regional crop simulations against observed data. Climatic Change (2010) pp751-765.