AAPG2019	EVOLUTHON		PRC
Coordinated by:	Eric TANNIER	48 months	298 keuros
Mathématiques et sciences du numérique pour la biologie et la santé			

Evoluthon: Artificial life as a benchmark for Molecular Evolutionary studies

Summary table of persons involved in the project:

Partner	Name	First name	Current position	Role & responsibilities in the project (4 lines max)	Involvement
Beagle-Inria	*Tannier*	*Eric*	*Researcher, Inria*	*Project coordinator, Principal Investigator of partner 1 Principal Investigator, Task 1*	*18p.month*
LBBE	*Daubin*	*Vincent*	*Research Director, CNRS*	*Principal investigator of partner 2, responsible for Task 3, revealing how evolutionary methods work on artificial life*	*9.6p.month*
Beagle-Inria	*Beslon*	*Guillaume*	*Professor INSA*	*Principal investigator Task 2, upgrading Aevol to produce benchmarks*	*6p.month*
LBBE	*Boussau*	*Bastien*	*Researcher CNRS*	*Evaluating methods for detecting convergent genomic evolution Task 3*	*4.8p.month*
Beagle-Inria	*Rouzaud-Cornabas*	*Jonathan*	*Assistant Professor, INSA*	*High performance computing Task 2*	*9.6p.month*
Beagle-Inria	*Parsons*	*David*	*Research Engineer, Inria*	*Development management and Software quality*	*5p.month*
Beagle-Inria	*To be hired*		*Ph-D*	*Adaptation of the Aevol model to produce benchmarks*	*36p.month*
LBBE	*To be hired*		*Post-doc*	*Methods for multiscale co-evolution and test campain on benchmarks*	*24p.month*
				Total	*110.8 p.m*

I. Proposal's context, positioning and objective(s)

a. Objectives and research hypothesis

Context.
Methods in molecular evolution, genomics and phylogenetics are applied widely across Biological sciences. For example, they are used for uncovering the functional importance of genes for species of interest (Liu et al, 2015), predicting seasonal viral strains against which a vaccine needs to be developed (Luksza et al, 2014), understanding human migrations on earth (Slatkin et al, 2016), managing agro-systems (Thrall et al, 2011), annotating medically-relevant variants (Cooper and Shendure, 2011), or even in judicial inquiries (Scaduto et al, 2010). Because these methods perform inferences of a historical nature, **they face a validation issue**: it is not possible to travel in time and verify hypotheses and predictions, which concern events that can be up to 4 billion years old. A possible experimental validation can be to evolve organisms in the lab (Randall et al, 2016), but experiments are short term, costly and have never lead to instances able to discriminate among different methods. Cross validation can be performed by comparing with the fossil (Szöllősi et al, 2012, Romiguier et al, 2012) or ancient DNA record (Duchemin et al, 2015), but samples are rare, in particular in the microbial world, and ancient DNA is not preserved above one million years. Predicted ancestral proteins can be synthesized to verify that they are still functional (Groussin et al, 2017), but even simplistic methods with known shortcomings seem to produce functional ancestral proteins. Theoretical considerations about the models and methods can also help to choose among competing approaches (e.g. statistical consistency, computational complexity), and there are ways to assess the robustness of the results (e.g. by data re-sampling), but those say nothing about the validity of the underlying modeling choices (Felsenstein, 2003). Throughout the scientific literature the most popular validation approach remains **computer simulations**. Genome evolution can be simulated *in silico* for a much higher number of generations than in experimental evolution, at a much lower cost. Then the results of simulations can be used as instances of inference methods.

Performing simulations for validation requires epistemological and organizational thinking. Indeed very often an individual method is tested with an ***ad hoc*** simulation, *i.e.* a simulation made on purpose to test it. In that situation some elements of the method are inevitably integrated in the simulator, which is then likely to generate only easy instances for this method and has no chance to reach the **complexity** of real data. Even when simulations are based on a general software that has not been designed for a specific study (Dalquen et al, 2011, Sjostrand et al, 2013, Arenas et al, 2014, Mallo et al, 2015, Edgar et al, 2018), some important underlying principles remain, shared between the simulation and inference methods simply because they are widely accepted (and often implicit) in the bioinformatics community. These principles are the "Natural Interpretations" of the community (Feyerabend, 1975): for example genes are considered as evolutionary units and intragenic rearrangements are neglected; simulations are performed at the inter-specific level, and ignore population-level processes where mutation, drift and selection occur; sites evolve independently from each other and independently from higher-level (e.g. structural) constraints. Extinct or unsampled species are ignored and not simulated. In such a situation, methods are only tested in a world designed for them, which does not assess their efficiency in the real world.

There is a need for a cooperative effort to organize and standardize benchmarks, as acknowledged for example by the addition of a section in PLoS Computational Biology dedicated to benchmarking, or the upcoming edition in 2019 of a special issue of Genome Biology on benchmarking studies.

Proposition.
We propose an original, principled way of benchmarking models and methods for molecular evolution studies with computer simulations. We are inspired by the "double blind" principle that governs test studies in science in general, and also some software development techniques (Pugh, 2011), where development and test teams are separated and work independently. The principles are that:

(1) Inference methods and simulated benchmarks should not be built by the same team. Moreover, the benchmark and inference teams, while having a common biological culture, should be

"**methodologically blind**" to each other, meaning that principles from inference methods should not be included in simulations, and the other way around, principles specific to simulations should not be used by inference methods. To this aim, the simulation and the inference methods should be produced by teams belonging to **different scientific communities**.

(2) The simulated benchmarks are produced by a model which **has not been designed to be used** as a benchmarking tool. While this seems hardly doable and somewhat contradictory, we argue that this is the way of approaching the double blind principle, and that it is possible for molecular evolution because of the existence of disjoint scientific communities around the modeling of genome evolution.

(3) As much as possible, **processes, and not patterns, should be simulated**. This means that instead of tuning parameters to resemble empirical data in some arbitrary sense, we should uncover the processes that produce these empirical data and implement them into a mechanistic model. Although it is desirable to produce simulated data that looks like empirical data, the definitions of the similarity measures can themselves be *ad hoc* design choices, dependent on a particular inference method.

We will implement these principles by gathering two teams from two different backgrounds, and by organizing an original mode of collaboration between the two. The first is the Inria "Beagle" team, specialized in *in silico* experimental evolution, and the second is the CNRS "Cocoon" team from the Biometry and Evolutionary Biology Lab (LBBE) of the University of Lyon, specialized in molecular evolution inference methods. The scientific background of Beagle is in bio-inspired computer science: complex systems, genetic algorithms, genetic programming and multi-agent models. The scientific background of Cocoon is molecular biology, evolutionary biology and bioinformatics. The Beagle team has developed several *in silico* evolutionary platforms. Importantly they were not devised with the aim to be used as benchmarks for inference methods, which paradoxically creates an ideal situation to use them for this purpose.

The two teams will work in close collaboration, exchanging results, expectations and challenges, but, importantly, not exchanging ideas on computational models. The Beagle team will construct a benchmark useful for **a large variety of bioinformatics models and methods** including clustering of genes into homologous families, orthologous gene detection, reconstruction of multiple alignments, phylogenies, ancestral genomes, demographic history, and detection of selection, adaptation and convergent genomic evolution. The Cocoon team will organize the application of state-of-the-art methods on these benchmarks, by proposing to scientific labs around the world to participate to a benchmarking challenge. It will also propose improvements based on the results, particularly on its area of expertise: multi-scale interactions in evolution –the nano-scale for genes, the micro-scale with microbes, the visible scale with animals or plants, and a macro-scale, with the global environment. Indeed, the team has a renown record in integrative phylogenetics at several scales and part of the benchmarking activity will tend towards modeling processes of such interactions.

The results of this benchmarking will go far beyond the two involved teams. Indeed we will maintain an open access to all data and expect the benchmark to become an international reference for validating a model or a method. We will organize an "Evoluthon" contest to compare different methods. All teams in molecular evolution will be welcome to submit their method, and we hope to gain the reputation of a standard benchmark for any new method susceptible to be tested by standard simulations.

b. Position of the project as it relates to the state of the art

State of the Art in genetic and genomic simulation.
Computer simulations are used in many contexts in genomics. They are used for assessing model adequacy through posterior predictive simulations, for inference using Approximate Bayesian Computation, or for calibrating machine learning approaches (Chan *et al.*, 2018). For what concerns the validation of evolutionary inference models and methods one can identify two kinds of simulators.

- **Ad-hoc simulations**. The most frequent situation is the *a posteriori* programming of a simulator, aimed at validating a precise method. This is what we call an *ad-hoc* simulator. These *ad-hoc* simulations are not useless. They address identifiability problems, they help to measure the range of parameters in which, under some model, the inference method is efficient. They sometimes invalidate the method of a concurrent team. However they cannot be taken for a validation, and not even for a serious test of the limits of a model. Hundreds of references could be cited here, where a method is supposedly but misleadingly "validated" by an *ad-hoc* simulation. For example, methods to detect genetic introgression, the exchange of genes between different species via hybrids (Rosenzweig *et al*, 2016) are systematically tested with simulations including introgression between ancestors of sampled species. This is what the inference methods can detect, but this is severely misleading to include it in simulations because it makes the absurd hypothesis that introgression in the past only happens between species whose descendants are sampled millions of years later (Szollosi *et al*, 2013). This kind of hypothesis is representative of an *ad-hoc* simulation, importing the principles of an inference method, even if this import is incompatible with the theory of Evolution. A simulation without this import could well invalidate the results, as we recently reported it for the *Anopheles* phylogeny (Davin et al, 2018). Another example is the reconstruction of phylogenetic trees in domains of the biodiversity where speciation events are suspected to have occurred close to each other. It has been advocated that some summary methods that use reconstructed gene trees as input were superior to competing approaches, mostly for two reasons: the method is statistically consistent, *i.e.* given an infinite number of true gene trees it returns the correct species tree, and the method works beautifully on simulations (Liu and Edwards, 2010). Such results however raised a lot of heated discussion, centered on the impact of errors in gene trees (Song et al., 2012; Gatesy and Springer, 2013; Mirarab et al., 2014), which had been completely overlooked in the initial simulations. Finally, another example of an *ad hoc* simulation can be found in Nelson-Sathi et al. (2015), where a simulation was used to test the ability of a method developed by the authors to assess the similarity of two sets of trees. Their simulation involved randomly swapping branches in each tree of a set, and checking that their method could indeed see that the original set and the set with swapped branches were different. We pointed out that this *ad hoc* simulation was specially devised to advertise a method, at the risk of using particular and unrealistic conditions (Groussin et al. 2016).

We could multiply such examples *ad libitum*. The common point between them is that methods have been tested on an *ad-hoc* simulator by the team that developed the method itself. It is a common requisite for the publication of a method that it is first tested on some simulations, with no precise requirement on the type of simulations, except that they should be "realistic", which is not very prescriptive. In *ad-hoc* simulations the virtual organisms obey to laws taken not from the observation of Nature but from the inference method themselves. In summary they are often designed to sell a method and a result more than to really test it.

- **Generalist simulations**. Some simulation programs are designed for a wider use. They are not limited to the test of a single method, but address a general problem on which a few methods can be tested. For example, Seq-Gen (Strope et al, 2009) produces simulated sequences along a phylogenetic tree; BottleSim (Kuo et al, 2003) simulates the process of population bottlenecks; Simphy (Mallo et al, 2015) produces gene trees with a process of duplication, loss, transfers, incomplete lineage sorting. The National Cancer Institute of the NIH has a (non-exhaustive) website classifying 143 published computer simulation programs for genetic studies (phylogeny, population genetics, RNA or protein folding, next generation sequencing simulation…)[1]. Carvajal-Rodríguez (2008) chooses 25 of them and compares their properties. The requirements of what should contain a general simulation method is discussed in Carvajal-Rodríguez (2010), where a "simulation requirements document" is proposed. It is advised to include specifications of simulation software just as in any software development project. It is still not among the requirements that they should avoid to construct artificial worlds that are specifically designed for inference methods. That is, even if these generalist simulators are designed by teams different from the ones using them, the teams belong to the same scientific community. As a result, simplifying hypotheses generally included in inference methods are

1 https://popmodels.cancercontrol.cancer.gov/gsr/

transposed to simulations. One of the most striking examples of a universal simplifying hypothesis common to inference and simulation is the gene taken as an evolutionary unit. While it is known that during evolution genes are combined, fused, fissioned, cut, extended, diversely transcribed, these events are neglected in simulators because they are absent from models used for inference. An empirical definition of the gene used to cluster them into families is transposed into an essential definition of *a priori* families, which is a way to include knowledge of the future (genes will be clustered into families) in ancestral genomes.

A particular interesting case is the software **Evolver** (Edgar et al, 2018). Evolver is thought to evolve whole genomes at the nucleotide scale, in order to produce histories that resemble as much as possible the supposed history of mammalian genomes. It is not dedicated to test a method in particular. It can be used as a benchmark for diverse methods, as gene finding, gene clustering, multiple alignment, genome rearrangements, phylogeny. As such it is close to our proposition and is probably our closest competitor. However being developed by authors of famous inference methods, they –intentionally or not– include the same simplifying assumptions than in these methods: inside genes, no duplication and no rearrangement is allowed, mirroring the fact that in multiple alignment programs, these are not handled. Population effects, including selection, are averaged and not simulated, because they are not directly used by the methods to be tested. Sites evolve without structural constraints. As such, methods are still tested on a world designed for them, without much space for the unexpected, which is the real resemblance to the real world that we should try to simulate.

The use of artificial life.

Compared to this state of the art, we propose a **completely novel and original approach**. Acknowledging that biases of simulation conception are partly unconscious, we propose that simulations for benchmarking obey to certain principles that may seem obvious when formulated but have nevertheless never been implemented. The main principle is that simulations should not be designed by the same teams than inference methods. They should even belong to different scientific communities. This situation is made possible by the existence of **artificial life** and, within this community, of *in silico* experimental evolution (Batut et al, 2013), also called digital genetics (Adami, 2006). There are many digital genetics platforms, and their important common property is that, as life has not evolved to be studied by evolutionists, they have not been designed to produce benchmarks. As such, they are paradoxically better positioned to produce interesting benchmarks. Moreover, being grounded in a completely different field (mostly biophysics or bioinspired computation) their developers were only scarcely connected to bioinformatics. Most digital genetics platforms are not readily usable for benchmarking evolutionary studies because they evolve objects too far from biological sequences (see Hindré et al., (2012) for a review). For instance the well-known platform Avida (Wilke et al., 2001; Lenski et al, 2003; Adami 2006) evolves pseudo-assembler code. Others evolve *e.g.* digital electronic circuits (Kashtan and Alon, 2005), graphs or networks (Crombach and Hogeweg, 2008). Such models proved useful to decipher macro-evolutionary rules but they cannot be directly used to generate benchmarking data.

Aevol (Knibbe et al, 2007), oppositely, occupies an interesting position. In Aevol the structure of the fitness landscape of an evolving population is strongly determined by the structure of the biological information coding. Hence, Aevol precisely mimics the biological genomic structure and the organization of biological genotype-to-phenotype mapping. As a direct consequence, in Aevol the evolved objects are purposely realistic, in the sense that genomes are sequences of nucleotides, that these sequences are transcribed into mRNA carrying genes, and organisms are in Darwinian competition with regard to the comparison of a non trivial phenotype and an environment. Interestingly in Aevol very few of the usual simplifications imported by simulators from inference methods are present. For example genes are fully evolving entities which can combine, overlap, undergo rearrangements, partial duplications, just like in reality. When observing evolution in action within Aevol, intragenic rearrangements are largely counter-selected. However, it do regularly happen that they are eventually fixed in the lineage.

Importantly, and contrary to most simulators used to validate bioinformatics methods, Aevol does not simulate the evolution of a sequence, but of populations of virtual organisms encoded by sequences. Although the difference may seems purely semantic, it leads to very different simulation principles: (1) Aevol does not simulate a single lineage; It simulates a population and the fixed lineage is recovered thereafter. (2) Aevol does not simulate only fixed mutations (which requires a substitution model); It simulates the biophysical roots of random DNA copy errors and the selection/drifts process acting at the population level leads (or not) to their fixation. A direct consequence is that the substitution model is not given; it is an observable that emerges from the complex interactions between the biophysical model of mutations and the selection process that drives their spreading in the population.

Finally, Aevol has not been designed for benchmarking but has interesting properties to be used as such, in particular it saves a perfect fossil record of all events at each evolutionary step and gives the ability to reconstruct the genes trees (Knibbe and Parsons, 2014). Moreover, it can be used to simulate several kinds of genomes, by varying their sizes (virus-like very compact and small genomes, or bacteria-like less dense ones), which is done by tuning the mechanisms of mutations.

Proofs of principle.
We have recently shown that using Aevol to produce a benchmark for **comparative genomics** methods was efficient to uncover unexpected pitfalls in usual inference methods. Lehman et al (2018) claim that a noticeable characteristic of *in silico* experimental evolution, is that artificial organisms produce **unexpected and surprising** behaviors. That is, the systems are complicated enough to be largely unpredictable. The crossed influence of many objects or processes (genes, RNA, proteins, mutation, selection, phenotype, environment,...) makes these systems much richer than other types of simulations addressing a precise question. This is particularly interesting for us, because simulations for testing inference methods are interesting mainly if they point at unexpected behaviors of the method.
We have experienced this unexpectedness, and showed that oppositely to ad-hoc simulators, using Aevol could reveal pitfalls of inference methods and help to construct better methods. A typical example of such a case study was about the **chromosome inversion problem**. It consists in comparing the gene order of chromosomes in two different species, and estimating the number of chromosome inversions that have occurred during evolution in the lineages of these species, since their last common ancestor. It is a very studied problem, with combinatorial and statistical solutions to the estimation of this genomic distance (see surveys by Eriksson, 2004, and Fertin, *et al*, 2009). All models translate gene orders on chromosome into permutations, used to estimate the number of inversions. In particular intergenic sequences were not used by statistical estimators. *Ad-hoc* or generalist simulations were used to validate the estimators, in which intergenic sequences are systematically neglected. They are not even mentioned, their presence is simply never imagined. Indeed if a parameter is not used by inference methods, simulation designers do not even think of integrating it, even if it can interfere. For the chromosome inversion problem **we tested a dozen of statistical estimators** using a benchmark generated by Aevol. Aevol includes intergenic sequences because it is agnostic to inference methods. So it generates a lot of features *a priori* not used by statistical estimators. But not used does not mean useless, because unused feature can interfere with used ones. In this case the results were dramatically different than tests with *ad-hoc* simulators (Biller et al, 2016a). No method reached half of the performance claimed from *ad-hoc* simulations. We interpreted this failure by the interference with intergenic sequences, and designed new, better, statistical estimators (Biller et al, 2016b), which perform an order of magnitude better on Aevol data as well as on *ad-hoc* simulators. This outcome was unexpected before the simulations: the influence of the intergene size was not suspected beforehand and was discovered thanks to the blind procedure. Intergenic sequences in Aevol were not included in order to test their effect on inversion distances but simply because they were necessary to model DNA double-strand breaks which repair leads to inversions.

Apart of this proof of principle, such procedure of involving different scientific communities for testing with a certain degree of blindness between them **has never been attempted**. The consortium we gather is a unique and original configuration that allows to attempt to improve validation

procedures in evolutionary studies. This warrants a big impact from the outcome of this project, even if part of what we expect is by essence unexpected.

c. Methodology and risk management

Methodology: The blind interdisciplinarity.

Our methodology is based on several counter-intuitive principles: Benchmarks have to be produced by tools which have not been designed to produce benchmarks; While we gather an interdisciplinary group we promote some extent of “blindness” in the collaboration, which requires a thoughtful way of cooperating.

These principles can be interpreted as risks and need to be properly managed. This is the objective of Task 1 (Management, cooperation, competition and the Evoluthon challenge). The project coordination has a special importance compared to usual multidisciplinary projects, because it includes protocols of communications that organize the use of information exchanged between the teams. Of course we won’t forbid people from different teams to communicate, to exchange knowledge, to present methods and results to the others. All these will be encouraged as in any other collaborative project. However the design of the artificial world relies on the absence, or at least scarcity of features that would be specially designed for inference methods. Conversely, we expect inference methods to avoid integrating information specific to artificial worlds. We will organize an emulative arm race between the teams. Artificial life designers will challenge the evolutionists by adding unexpected difficulties taken from biological processes. Biologists will challenge the artificial benchmark team to find and implement mechanisms able to discriminate among methods.

The key aspect of this collaboration is that the goal of both teams won’t be to validate a method, but oppositely to find its limits. The usual mode of publication biases the results towards validation and overselling of the capacities of a precise method. This collaboration mode will reverse this tendency and promote better reliability.

Risk 1. The design of methods tuned for artificial life.

We expect to produce benchmarks that will be universally used to test and compare evolutionary inference methods. As we won’t avoid artificial organisms to have some specificities that are not found in the real world, the danger will be to include in the design of these tested methods some features specially dedicated to pass the artificial life test, without correlation with their actual efficiency for biology. It is a very general problem of all benchmarks. For example engineers at Volkswagen had tuned the car’s software to react differently to the pollution benchmarks than in reality. Every time a measure of performance is constructed (like, in Research, the impact factor and the h-index), there is a risk that cheaters will try to fit the measure instead of being efficient. All the more since we will be totally transparent regarding the algorithm generating the benchmarks: it is open source, and the parameters are public. To circumvent this risk, first our own phylogeny team will not participate to the international competition we will organize. Second, we will diversify the simulations, with different parameters, so that it will be difficult to tune a method. Third, if the algorithms generating the benchmarks are open, the measures to define success to the challenges will be kept secret until the submission deadline. Eventually we count on the unexpected behavior of the digital organisms, that use to surprise even their developers (Lehman et al, 2018). They should be all the more complex to cheat with for users.

Risk 2. About the realism.

In a short novel by Jose Luis Borges, an imaginary author, Pierre Ménard, reproduces part of the Quixote by Cervantes. He has two options to reproduce it. One is to become Cervantes, so he would easily reproduce his text with the same mechanisms. The other, which he finds more difficult and more interesting, is to write exactly the same text from a completely different context. Borges then compares some extracts from the text by Cervantes and Ménard. They are completely identical, but have very different explanations because they were written at different periods, with different contexts, by

different authors. Borges addresses here a literature variant of statistical identifiability: different parameters can produce the same results, which, back to Evolution, can mean that there is more than one solution to an inference problem. Pierre Ménard can also be seen as an attempt to simulate with precision a world without knowing the exact processes that have shaped it. This is a possible caveat of "realistic" simulations. Tuning the parameters of a simulator to reproduce an aspect of extant data, while having simplified a lot of other ones, can orient simulations in an unrealistic narrow process. For example, in Evolver (Edgar et al, 2018), rearrangement rates are tuned to obtain a comparative genomic landscape of mouse and human which is indistinguishable from measures on the real genomes. This way of simulating is validated by **adversarial learning**: a simulation is labeled sufficiently realistic when learning algorithms cannot make the difference between artificial and real data.

Without denying the merits of such approaches, we will take another path. We aim at being realistic in the processes and not necessarily in the patterns. We would rather become Cervantes, at the risk of producing another book because writing is not necessarily a deterministic process, than being Ménard and resembling reality. We don't expect to pass adversarial learning easily, because we will put efforts on realistic processes more than in realistic patterns. This is a risk concerning the communication with the scientific community, for which pattern realism is often a criterion. We will advertise our way of proceeding with solid arguments, privileging a certain universality in the simulated data than a resemblance with specific organisms. That said, any discrepancy between patterns in the data between artificial and real organisms will be addressed to the team producing benchmarks. The latter cannot implement artificial tricks to fit the pattern, but search for the processes responsible for that pattern that have not been implemented, and find a way to include them.

Risk 3. The difficulty to anticipate the unexpectedness.

More than in any other scientific project, we count on unexpected results. Part of the project is to provoke serendipity. This is why we do not describe into details in this project, and in particular in Task 3, which improvements will be brought to phylogenetic methods. The essence of the project is to unearth unknown caveats of current methods, and possibly to correct them. We are confident that it will happen because it happened in all the attempts we made (see Proof of principle above), and surprising results is a constant of digital evolution, as witnessed by a recent paper specially dedicated to this aspect (Lehman, 2018). The post-doc hired for Task 3 will have a good knowledge of probabilistic models for evolutionary inference, and will first be given the task of detecting unknown artifacts. This limits that risk because the evaluation procedure itself will be valorized.

Risk 4. The limits of computation resources.

Realistic Computer simulations require significant amounts of resources. Recently we have improved a lot the algorithms and code design of Aevol, and it is able to evolve millions of generation for populations of thousands of individuals, in a few days. Adding the possibility to evolve several independent lineages through speciation will keep the same order of magnitude for the simulations. When the lineages will show some dependencies due to horizontal gene transfer, hybridization or migrations, we will face a computational issue. Overcoming it will be one of the main challenges of Task 1, helped by specialists in high performance computing and software engineering in the team.

Note however that we are not engaged in a simulation that requires evaluating parameters, and in consequence exploring a wide range of them. A few combinations will be sufficient to produce a few dozens of conditions on which to test inference methods. We are not at all required to be exhaustive, so the limits of computation resources is not a severe risk.

Risk 5. The limits of blindness and validation in science.

Our proposition of a protocol close to the "double blind" principle has of course its limits. We cannot totally exclude common points between simulations and inference methods simply because both use computers and a common mathematical background. Both are computational models, while the real world is not. Both use Markov processes, and our interpretation of randomness via the theory of

probabilities. We cannot propose a way to abstract ourselves from current science and technology. A way to go further would be to use in vitro or in vivo experimental evolution (Barrick et al, 2009, Randall et al, 2016). This has been attempted at limited scales, and future projects can be to do it at a large scale. This would however necessitate an amount of time and money that we cannot afford presently. Even pushing the limits of the blindness, it is evident that the stricto sensu “validation” of a method will forever stay out of reach. As any scientific theory, evolutionary methods can be invalidated in some cases, but their validation will always be synonymous of an ability to pass many tests aimed at invalidating it. With this project, we aim at providing difficult instances that will help to improve methods and have an improved confidence on our evolutionary predictions. We propose to counterbalance the tendency to produce easy instances as selling arguments for a method, pushed by the current habits of the scientific community to publish striking results. Nonetheless we expect this original procedure to have a significant impact and be published in general journals.

II. Organisation and implementation of the project

a. Scientific coordinator and its consortium / its team

Implication of the scientific coordinator and partner’s scientific leader in on-going project(s)

Name of the researcher	Person.month	Call, funding agency, grant allocated	Project’s title	Name of the scientific coordinator	Start - End
E Tannier	7,2	ANR	Sthoriz	Damien de Vienne	2018-2020
E Tannier	7,2	ANR	LncEvoSys	Anamaria Necsulea	2017-2020
V Daubin	7.2	ANR	Horizon	Sylvain Charlat	2018-2020
V Daubin	7.2	ANR	Dasire	Nicolas Lartillot	2016-2019

The scientific coordinator Eric Tannier has a key position in this consortium. He is a member of both involved teams. He has long lasting collaborations with the other members of Inria Beagle as well as LBBE, including co-supervisions and co-publications. From a scientific point of view he also occupies a hinge position. His background is in Discrete Mathematics, and Theoretical Computer Science, he is employed by Inria, the national computer science research institute, and he has been inside an evolutionary biology lab for more than ten years. He has a renowned research record both in Computer Science and in Biology. He recently started several projects involving large collaborative or participative initiatives, including the present one.

Eric Tannier has a long experience in multidisciplinarity. He published in scientific journals as diverse as “Theoretical Computer Science”, “Genome Research”, “PNAS”, “Nature Ecology and Evolution”, “Discrete Applied Mathematics”, “Trends in Plant Science”. He has national and international collaborations in biology, in diverse aspects of computer science, and in bioinformatics. Moreover, he teaches scientific ethics, research integrity, epistemology, in addition to mathematics and computational biology, to diverse publics, from Master students to general audiences.

Quantitatively, Eric Tannier is the co-author of 3 published books, 5 book chapter, co-editor of 2 books, and the author of more than 60 internationally peer-reviewed scientific papers. He is also the author of several licensed software. He was in the executive committee of the Ancestrome project, funded 2.2 million euros by the “Investissement d’Avenir” programme in 2012, and he is the leader of a technological action programme on phylogeny, funded 120 keuros so has also the experience of leading research projects.

Moreover he is engaged in several actions promoting social and environmental responsibilities of researchers. In particular he is a member of the national open science committee of the ministry of education and research, and of the ethics platform of university of Lyon. He will then be able to implement the requirements of the ANR in terms of integrity and open science, in terms of practice, open access of publication and data, reproducibility and good science practice.

The role of the coordinator in this project will consist in the usual managing tasks, plus the management of the particular interdisciplinary protocol inherent to this project. This necessitate a mix between full collaboration between partners and a sort of competitive emulation. While the two partner are not directly in competition because they develop different aspects of the project, they will constantly try to play with the limits of the other teams, in order to progressively improve the results of both. The coordinator will also organize an international competition between evolutionary methods and a dedicated workshop. So he will be in charge of the valorization and publicity of this project.

The composition of the consortium is the spirit of the project itself. The project is indissociable from this consortium. Indeed the leading principle of this project is that the benchmarks should be designed by a team which is not involved in evolutionary inference methods, while an inference team should ensure that the benchmarks can be used by current inference methods. An important point is that the two teams have different scientific backgrounds, one is bio-inspired computing, the other in bioinformatics.

The project leader Eric Tannier is the proposed PI for the Beagle team from the Inria, which has developed *in silico* experimental evolution platforms for more than ten years. Importantly the developers of artificial life in the team (Guillaume Beslon, Jonathan Rouzaud-Cornabas, David Parsons) are only scarcely engaged in collaborations with bioinformatics or phylogeny teams but they maintain strong collaborations with renowned groups in experimental evolution through its participation to the "Laboratoire International Associé" EvoAct (together with D. Schneider team in Grenoble and R. Lenski and C. Ofria team at Michigan State University, US). Guillaume Beslon, head of the Beagle team and responsible for Task 2 of the project, is the historical designer of Aevol, which is now widely known in the artificial life community, and found applications in medicine and teaching (Beslon & Schneider, 2017; Beslon et al., 2013). He organized in 2017 the yearly international conference on artificial life ECAL in Lyon and was project leader of the 2013-2016 “EvoEvo” European Project that gathered research groups in experimental evolution, computational biology, artificial life and software development. Hence he has a strong expertise of interdisciplinary projects. High performance computing will be decisive in this project to reach credible size benchmarks with a complex system. Jonathan Rouzaud-Cornabas, who is a specialist of high performance computing, and also worked on Aevol to scale it up, will work on this aspect.

The Cocoon team from the CNRS, University of Lyon lab LBBE (Biometrics and Evolutionary Biology) is an emerging team of young researchers internationally already renowned in molecular evolution. Bastien Boussau, participant to the project, and Vincent Daubin, the responsible for this partner in the project, are both leading researchers in phylogeny, with international top level publication records. They are both experts in methodology for phylogeny and ancestral genome reconstructions, with applications in all domains of life. The team in engaged in novel methodologies for science, as participative protocols or crowdsourcing, which will be implemented for this project.

b. Implemented and requested resources to reach the objectives

We organize the work program into three tasks. Two (Tasks 2 and 3) are dedicated to works within each involved team, and one organizing task (Task 1) deals with the management and communication between the two, and the organization of an international competition based on the benchmarks. This task will also cover the aspects of project management and monitoring. The way the two teams will cooperate and challenge each other will be decisive for the success of this project.

TASK 1 - Management, cooperation, competition and the Evoluthon challenge
HEAD: Eric Tannier, Inria
PARTICIPANTS: All partners
STARTING: M1 **ENDING**: M48
OBJECTIVES. The goals of this task are:
- to organize the cooperation between the two teams of the consortium
- to promote the approach and organize the Evoluthon contest
- project management and monitoring

DESCRIPTION OF THE WORK.

I. We will organize regular meetings of the consortium, as in any other scientific project, where researchers from both teams will freely and transparently speak about their progresses. However, we will promote a particular way of information exchange and a way to address questions to the other team on the challenge mode.
On the one hand, the researchers from Inria Beagle, will receive requests from the Cocoon to integrate some features in the artificial world, so that some particular issues can be tested. How to integrate these features however should not be inspired by the inference methods, but by mere biological and biophysical knowledge. The challenge for Inria Beagle will be to construct mechanisms responsible to a pattern while avoiding direct implementation of the pattern.
On the other hand, the LBBE members will be challenged to reconstruct some aspects of the artificial history produced by the Beagle team, who will design difficult instances on purpose. That is, we won't organize the collaboration so that it is tempting to cheat by adapting each team's work to the other, but oppositely on the challenge mode, trying to refute the other team's work. Note that this mode of cooperation is not a competitive one, because the two teams work on different fields. It is more on the "proof and refutation" mode described by Lakatos (1976).

II. We will organize a contest for a wide range of evolutionary methods, based on the produced benchmarks. We will follow the model of Quest for Orthologs (Dessimoz et al, 2012), Alignathon (Earl et al, 2014) or Assemblathon (Bradnam et al, 2013), which gather several teams around challenging instances where the ground truth is known. The contest will be called "Evoluthon" and will encourage many researcher from the whole world to test their methods on our benchmarks. Our inference team The Cocoon won't be a challenger for this contest because of the privileged collaboration it will have had with the producers of the benchmarks. Lastly, we will all participate to the organization and to the definition of the measures to define the success to rank participants. These measures will be kept secret to all participants until the submission deadline. Then they will be released.
We will organize a workshop to celebrate this challenge, as a satellite of one of the big annual conferences in Evolution or Computational Biology. Several members of the consortium have participated a lot in the organization of these conferences, as organizing chair, PC chair, PC member or organizing committee member (RECOMB, RECOMB Comparative Genomics, SMBE, ECAL, ALIFE, ISMB, ECCB, …). We will contact future organizers to propose a keynote presentation to the winner of the challenge.
A book will be edited for the proceedings of this workshop, describing the purpose of the project, the mode of collaboration we choose, how it actually happened, and the contribution of all participants.

III. Eric Tannier, as coordinator will have a global overview of all tasks through regular contacts with all project members. This is made easy since both partners are on the same campus. Based on the contract information approved (present document), he will support the collaboration to keep Evoluthon on track according to deadlines and resources allocated for each task. His coordination role will benefit from his strong interdisciplinary experience, he is be able to discuss/understand all scientific elements of the project.
Semester meetings will be organized to directly assess global and individual progress, discuss planning and future engagements. These meetings will also be means of internal dissemination to warrant sharing of knowledge between all partners (in the respect of the double-blind methodology proposed). Meeting draft agenda will be proposed 4 weeks ahead and final version sent out 2 weeks before the event. Minutes of the events will be drafted by the coordinator and available in the following 2 weeks. External experts may be invited to the meetings depending on the needs of the consortium and agreement of both partners. Monitoring will assess work progress towards initial objectives, deliverable preparation, writing and finalization after validation by task leaders and project leader. Pre-defined milestones will help to assess regular progress and give green light for go-ahead. In case of delay fall-back corrections will be discussed between the two partners.

DELIVERABLE D1.1. Set of rules for the contest. Submission date: M30
DELIVERABLE D1.2. Set of measures for ranking participants' propositions to the contest. M40
DELIVERABLE D1.3. Workshop on the Evoluthon contest proceedings. Edited book. M48
DELIVERABLE D1.4. Project advancement reports for ANR M6,M12,M24,M36,M48
MILESTONE M1.1. Call for participation to the Evoluthon contest. Expected date: M30
MILESTONE M1.2. Evoluthon contest deadline. Expected date: M40
MILESTONE M1.3. Analyses of the results Expected date: M44
MILESTONE M1.4. Detailed program of the workshop dedicated to the contest, Expected date: M46

TASK 2 - Producing a blind benchmark for evolutionary studies
HEAD: Guillaume Beslon, Inria
PARTICIPANTS: Beagle, Inria (Eric Tannier, Jonathan Rouzaud-Cornabas, Ph-D to hire)
STARTING: M1 **ENDING**: M40
OBJECTIVES. We will adapt the Aevol software so that it can produce instances of different levels of difficulty for a wide range of comparative methods in Molecular Evolution.
DESCRIPTION OF THE WORK. Aevol is already usable for some comparative genomics tests. We have used it with success in particular to test statistical estimators for the chromosome inversion distance problem (Biller et al, 2016a). In theory it can be readily used to test selection detection, gene clustering, multiple alignment and gene phylogeny for some limited cases, because genes diversify in the artificial organisms, and it is possible to try to retrieve this diversification. However this is limited by the absence of key features. This weakness is precisely what make the force of our project: Aevol has not been designed to be used for benchmarking. So in order to adapt it, we have to implement the following features:

I. Genetic code. For sake of simplicity and computational efficiency, Aevol currently uses a simplified genetic code. The first step will thus be to use a more complex genetic code. This will enable to introduce different biases in the model, e.g. codon usage bias or mutation biases. Note that an alpha-version of this modification is currently under test in the Beagle Team (Liard et al., 2017).
II. Polyploidy. In current version of the model organisms are haploid with circular chromosomes. We will develop a diploid version (including cross-over and dedicated recombination operators) and, if possible, a polyploid one. An option will allow choosing between linear and circular chromosomes.
III. Speciation. For the moment Aevol simulates evolution of a single population at a time. We have to include a diversification of population to be able to test species phylogeny reconstruction methods, which are one of our main targets. The speciation model will include the possibility of variable population size, between the species or along time within a same specie (*e.g.* bottleneck). This will enable producing benchmarks to test ancient demography methods.
IV. Extinctions. Little attention has been put on the extinction of a population for the moment, because an experiment stops when the population goes extinct. However if speciation is included, species extinction have a particular importance.
V. Horizontal transfers. Transfers are currently possible as recombining events between individuals of a population, but if speciation is included, two different species should be able to exchange genes.
VI. Mutation rates and transposable elements. It is known that selection can act on the mutation kind and rate, and that the variations along the history and among organisms is also a difficulty for phylogenetic software. Including evolving mutation rates is a challenge, because the determinants of the mutations have to be included in the genotype, whereas for the moment it is outside as a parameter. On a same idea, we will add Transposable Elements (TE) to the model. This will be beneficial not only to produce difficult benchmarks, but also in Evolution in general, to observe the behavior of evolving mutation rates or TE dynamics.

This list is not exhaustive. One of the key aspect of Task 1 is to allow Cocoon members to suggest modification of the model without choosing the implementation themselves. To this aim, once a

modification has been proposed and collectively accepted, Beagle members will not question Cocoon members about the instantiation of the modification. They will rather use their acquaintances in evolutionary biology and in experimental evolution to design the model. Of course these guided modifications of the model might seem contradictory with our principle that benchmarks should be produced by software not designed for benchmarking. Indeed "Aevol_BM" will be made to produce benchmarks. We have two answers for that. First the engine of the software does not change, and includes a complex system of unexpected interactions, independently from the modifications (those being actually limited compared to the whole system). Second, every additional feature will be added indirectly through a modification of the mechanistic process, and not "phenomenologically", *i.e.* as a pattern. The pattern is what is measured by the inference method, for example, the species phylogeny, or the heterogeneous mutation rate among sites of a protein. We should avoid implementing a phylogeny or an heterogeneous mutation rate, but we will constantly think of the processes generating these patterns, for example diversification of species or population bottlenecks.

In order to generate multiple benchmarks (and, beforehand, to test the model), we will need to run simulations lasting for millions of generations. with the current state of Aevol, it would necessitate months of computation (in a recent test of the model we simulated 10 millions of generation for a population of bacteria-like organisms. The simulation lasted 1,5 months on a 32 core computer). This performance is already enough to produce decent benchmarks, but for some features like varying demography, we will need to optimize the numerical and computational problems behind Aevol. This optimization will be done on two aspects: First new methods and/or optimizations will be proposed for the numerical methods underlying the biological model, taking into account the specificities of *in silico* experimental evolution. Second, the numerical methods will be ported on HPC platforms and new computational approaches and/or optimization for the specificities of *in silico* experimental evolution. Furthermore, we will propose an easy way (*e.g.* frameworks, domain specific language, graphical interface) to generate new benchmarks. We will design a simulation workflow that will automatically compose a benchmark from simulations specifications (e.g., species tree structure) expressed in a few dozens lines.

As timely modifications of the software are mandatory for the progress of the project, they will be strictly monitored through a list of Milestones (see below). Any deviation from these milestones will immediately be discussed within the consortium to propose correction measures.

DELIVERABLE D2.1. Benchmark for a phylogenetic pipeline — Submission date: M18
DELIVERABLE D2.2. Benchmark for a population genetics pipeline — Submission date: M24
DELIVERABLE D2.3. Set of benchmark for the Evoluthon contest — Submission date: M30
DELIVERABLE D2.4. Aevol_BM, a special release of the software Aevol — Submission date: M40
MILESTONE M2.1. Genetic code and diploidy implemented — Expected date: M10
MILESTONE M2.2. Speciation implemented — Expected date: M12
MILESTONE M2.3. Variable population size implemented — Expected date: M18
MILESTONE M2.4. Horizontal transfers implemented — Expected date: M24
MILESTONE M2.5. Evolvable mutations and transposable elements implemented — M30

TASK 3 - Evaluating evolutionary inference methods on artificial benchmarks
HEAD: Vincent Daubin, CNRS
PARTICIPANTS: LBBE, CNRS, Univ Lyon (Bastien Boussau, Eric Tannier, hired post-doc)
STARTING: M12
ENDING: M48

OBJECTIVES. We will implement the proof of principle of the use of artificial life for benchmarking by applying on the results of Task 2 standard methodologies in phylogeny, as well as modern methods developed in the team, in particular multiscale co-evolution. Thus we will promote this method so that it can be used by a wide range of methodologists in evolutionary biology.

DESCRIPTION OF THE WORK.

Our main interest will be to test and improve the current approaches used in phylogenomic pipelines, from the recognition of homologous sequences to the reconstruction of the history of species, their

genomes and their interactions. This is why we will start this Task after Milestone 2 of Task 2. A typical phylogenomic pipeline considers the following steps, usually independently: gene annotation, gene clustering into families, multiple sequence alignment of each family, filtering sites according to their mutational pattern in the multiple alignment, gene tree reconstruction and species tree reconstruction. The fact that each step is computed independently from the previous one is a well known problem (Boussau and Daubin, 2010), but the extent to which errors accumulate throughout the pipeline is not well understood. The Cocoon team has done an extensive work on how to account for dependencies between several steps (gene trees / species trees inference), and proposed models to improve reconstruction at both levels, while inferring evolutionary events such as duplication, transfer, loss or deep coalescence of genes (Szollosi et al., 2012, 2013). This is achieved by modeling processes at several scales, genes and species, which is transposable to other interactions, like host-symbiont. In the context of artificial life simulations, one can measure errors at each step of the pipeline and the impact on the final outcome., This radically differs from usual quality assessments with simulations, where a single step of the pipeline is tested, often assuming that the others have been completed without error.

The extent to which we will be able to produce reliable phylogenies will inform us on our abilities to reconstruct the history of biodiversity as we find it today. Moreover, it will point at the steps of the procedure which are the bottleneck for such a reconstruction. Currently these bottlenecks are not known with certainty. Teams working on each step are often disjoint, and they complexify and improve their part without knowing exactly the influence of an improvement when the other parts have been neglected. For example, models for gene phylogenies from sequences are more and more complex, while the gain from complex models might be negligible, given the error rate in the multiple sequence alignments. Similarly, the reconstruction of genome evolution history may be highly influenced by gene annotation and homolog clustering methods.

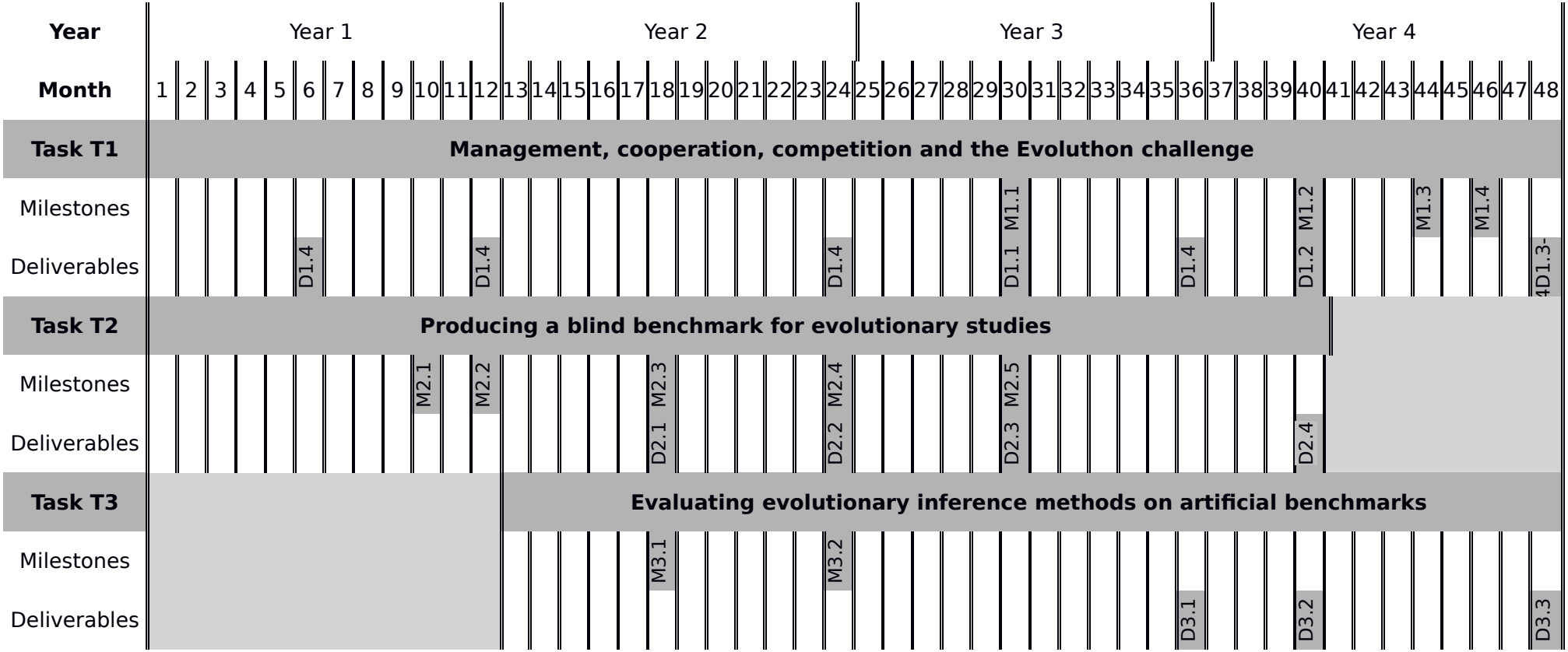
These tests will help us to improve the methods traditionally devised in the cocoon team, detecting convergent evolution, gene interactions, or multi-scale evolution. Indeed we have an expertize in comparing different levels of evolution, host/parasite, gene/species, transposable element/genome, genotype/phenotype. The benchmarking will set up a quality measure of these methods.

We will also diversify the scope of methodologies which can be tested with Aevol, by focusing on approaches that are central in computational genomics today. We assess the reliability of methods for reconstructing the demography of ancestral populations from current genomes, which is a very debated topic and has never been tested on a blind benchmark.. We will test several selection detection methods at different temporal scales as the dn/ds ratio or McDonald-Kreitman test. We will test methods attempting to detect sites of proteins that correlate with convergent phenotypic changes. Such methods have been used widely recently, and are at the center of the ANR project Convergenomix awarded in 2015. Based on the result of the assessments of all those popular methods on our simulated data, we will revisit past works that used such methodologies in the scientific literature to put a new light on the reliability of the current knowledge in deep or recent evolution.

Eventually, we will propose improvements of some of the methods and pipelines based on these tests. This is the most prospective part, because by definition of the whole project we don’t know precisely in advance where are the influential parameters. However, our experience of artificial evolution and knowledge of the standard methods used in phylogenomics gives us little doubt that at least a few steps will be good candidates for improvement, which would be best addressed by novel probabilistic approaches.

DELIVERABLE D3.1. An article on the limits of the current phylogenetic methods, identifying key bottlenecks in phylogenomic pipelines. Submission date: M36
DELIVERABLE D3.2. Proposed improvements to current methods Submission date: M40
DELIVERABLE D3.3. Test report for improved methods Submission date: M48
MILESTONE M3.1. Test of a phylogenetic pipeline on Aevol data Expected date: M18
MILESTONE M3.2. Test of a demographic historical study on Aevol data Expected date: M24

Gantt chart



See previous sections for a detailed description of deliverables and milestones

Partner 1: INRIA

Staff expenses One Ph-D student is requested, who will achieve the goals of task 2, that is, implementing mecanistic processes responsible for the patterns that will be tested. This implies transforming Aevol into a benchmarking tool, and will require computational expertise, knowledge in biology, if possible not much knowledge in phylogeny and molecular evolution.

Instruments and material costs For the achievement of the project we will need one desktop (2800 euros) and one laptop (1400 euros) for this partner. The laptop will serve for the Ph-D student and the desktop will be used to generate the benchmarks.

Outsourcing 8 000 euros for 4 publications in biology journals for the project (like Genome Research, Molecular Biology and Evolution, PLoS Computational Biology, Genome Biology, all with author processing charges around 2000 euros)

General and administrative costs & other operating expenses The amount resquested here contains:

- 8 000 euros for the organization of an international meeting “Evoluthon”, presenting the results of the contest. This covers half of the planned expenses and we will find the other half with other partners (SMBE, Inria, CNRS)

- 10 200 euros will cover the travel and stay expenses for members of the team (4 members and 4 years) to promote the initiative and the results, in visits to partner teams or presentations to conferences (ECAL, RECOMB, ISMB).

Partner 2: CNRS

Staff expenses We request hiring one Post-doctoral intership of 24 months. Her/his task will be to implement task 3, that is, using the benchmarks produced in task 2 on standard methods, on new methods involving multi-scale cophylogenies, and to give a feedback on task 2 on new processes to implement.

Instruments and material costs The 9800 euros requested here contain 8400 euros of participation to the functioning of a computing cluster at the LBBE, on which we will perform the benchmarking evaluation, central to the project. It will consist in running diverse inference methods, some of them requiring substantive amount of computing power. Then 1400 euros will be dedicated to a laptop for the post-doctoral researcher.

General and administrative costs & other operating expenses We plan two participations to international conferences (4000 euros) for the hired post-doctoral researcher (conferences SMBE, Evolution...), as well as one mission for one of the two participant (2000 euros). Then we request 1800 euros for the participation to missions due to collaborations and national workshops during the 4 years of the project.

		Partner 1 ***Inria Beagle***	**Partner 2** ***CNRS LBBE***
Staff expenses (1Ph-D, 1 post-doc)		122 400	105 840
Instruments and material costs (including the scientific consumables)		4 200	9 800
Building and ground costs			
Outsourcing / subcontracting		8 000	
General and administrative costs & other operating expenses	Travel costs (including the organization of the Evoluthon workshop)	18 200	7 800
	Administrative management & structure costs**	12 224	9 875
Sub-total		165 024	133 315
Requested			**298 339**

III. Impact and benefits of the project

We are proposing a research project which is definitely contributing to the 2019 ANR work programme, B11 “Research in support of major cross-disciplinary challenges”, and in particular Theme 1: Mathematics, computer science, automation and signal processing to meet the challenges of biology and health. Our project is best described by one of the requirements of this Theme (we emphasize keywords):

> “the development of concepts and **new methods** using mathematical, **computing** and biostatistical tools for **the simulation of complex biological systems**, digital simulation, **high-performance computing** and the associated optimisation and immersive simulation (virtual and augmented) to integrate and represent multimodal, **multi-scale data**”

Indeed, the simulations produce multi-scale data by essence. We have population, organism, gene and nucleotide levels in the simulator, which is the guarantee that it can afterwards be used to test either singlescale or multiscale methods. Moreover, we will in particular test the methods on multiscale co-evolution methods as developed by the Cocoon team.

We plan several kinds of outcome from this project. The broadness of the methodologies that are encompassed in the project will lead to a broad impact of the results. We expect to have a better view on the reliability of a wide range of methods, used on a daily basis in diverse fields as human history, medicine, biology of conservation. We expect to improve the methods, but we also expect a feedback on our way to use ad-hoc or generalist simulations. Those won’t disappear after our project because Aevol cannot bear all aspects of testing methods by simulations. However we can expect that even the standard of ad-hoc simulations is improved. For example, thanks to the proof of principle of the use of Aevol to benchmark comparative genomics methods (Biller et al, 2016a), it can be expected that future genome evolution simulators including genome rearrangements also include intergene sequences, or at least intergene sizes, given their importance in the evolutionary distance estimations.

Evoluthon is also likely to impact computational biology and artificial life. Indeed, Aevol_BM, the version of Aevol dedicated to the production of benchmarks, will include some biological properties that, to the best of our knowledge, have never been tested in silico. Hence, it will enable implementing many evolutionary scenarii (e.g. bottlenecks, mutational biases, environmental changes...) and study their consequences on the genomic structure with a degree of realism that has never been reached so far.

We expect a good international visibility thanks to the organization of the Evoluthon contest, as well as high impact publications. We will also negotiate a presence at some well known international conferences in Evolution or Computational Biology, like SMBE or RECOMB, as a prize for the contest winner. Dissemination of the software in the artificial life community will be achieved through the organization of satellite tutorials during the yearly International Conference on Synthesis and Simulation of Living Systems (Alife). We plan to release all publication in open access mode and all software in open source license, in order to maximize our visibility, and the transparency of our approach. Finally we will necessarily leave some aspects of molecular evolution without a good benchmarking.

For example we don’t see easy ways of including, within the four next years and with our working means, phylogeography, host-pathogen co-evolution, symbiosis, microbiota, evolution of sex, biased gene conversion, protein folding, which are major issues in Evolution today. We hope that the publicity we will give to this initiative will lead to the spread of the approach, in order to generalize the reliability improvements of what we know about our history.

IV. References related to the project

Adami (2006) Digital genetics: unravelling the genetic basis of evolution, *Nat. Rev. Genet.*, 7(2), 109-118

Altenhoff, *et al* (2016) Standardised Benchmarking in the Quest for Orthologs, *Nature Methods*, 13, 425

Arenas, Posada (2014) Simulation of genome-wide evolution under heterogeneous substitution models and complex multispecies coalescent histories *Mol biol evol*, 31, 1295-1301

Barrick et al (2009) Genome evolution and adaptation in a long-term experiment with Escherichia coli, *Nature* 461 (7268), 1243

Batut, Parsons, Fischer, **Beslon**, Knibbe (2013) In silico experimental evolution: a tool to test evolutionary scenarios, *BMC bioinfo* 14 (15), S11

Beslon *et al.* (2013), An alife game to teach evolution of antibiotic resistance. *Proc of European Conference on Artificial Life, Tormina, July 2017*, pp. 43-50

Beslon, Schneider (2017) ISEE-Resistance: Using in silico experimental evolution to sensitize providers on antibiotic resistance. *Antimicrobial Resistance and Infection Control*. 6(52):2

Biller, Knibbe, **Beslon, Tannier** (2016a). Comparative genomics on artificial life, *Computability in Europe*, *LNCS*, 35-44

Biller, Guéguen, Knibbe, **Tannier** (2016b). Breaking good: accounting for fragility of genomic regions in rearrangement distance estimation, *Genome biol evol* 8, 1427-1439

Boussau, **Daubin** (2010). Genomes as documents of evolutionary history. Trends Ecol Evol. Apr;25(4):224-32.

Bradnam *et al* (2013) Assemblathon 2: evaluating de novo methods of genome assembly in three vertebrate species, *GigaScience* **2**:10

Carvajal-Rodríguez (2008) Simulation of genomes: a review *Current genomics* 9 (3), 155-159

Carvajal-Rodríguez (2010) Simulation of genes and genomes forward in time *Current genomics* 11 (1), 58-61

Chan, *et al* (2018) A Likelihood-Free Inference Framework for Population Genetic Data using Exchangeable Neural Networks. Bioarxiv doi: https://doi.org/10.1101/267211

Cooper GM, Shendure J. Needles in stacks of needles: finding disease-causal variants in a wealth of genomic data. *Nat Rev Genet*. 2011 Aug 18;12(9):628-40. doi: 10.1038/nrg3046.

Crombach, Hogeweg (2008) Evolution of evolvability in gene regulatory networks. *PLoS comp biol* 4(7):e1000112.

Dalquen, *et al* (2011) ALF—a simulation framework for genome evolution, *Mol biol Evol*, 29 (4) 1115–1123

Davin, Tricou, **Tannier**, de Vienne, Szollosi (2018) Zombi: simulating genome evolution with transfers and extinct species. Biorxiv.

Dessimoz, et al (2012) Toward community standards in the quest for orthologs, *Bioinformatics*, 28 6

Duchemin, **Daubin, Tannier** (2015) Reconstruction of an ancestral Yersinia pestis genome and comparison with an ancient sequence, *BMC genomics*, 16, S9

Earl *et al* (2014) Alignathon: a competitive assessment of whole-genome alignment methods, *Genome Research*, 24: 2077-2089

Edgar, et al (2018), http://www.drive5.com/evolver/

Eriksson (2004) Statistical and combinatorial aspects of comparative genomics, *Scandinavian journal of statistics*

Felsenstein, (2003) *Inferring phylogenies*, Sinauer Associates

Fertin, Labarre, Rusu, **Tannier**, Vialette, *Combinatorics of genome rearrangements*, MIT press

Gatesy, Springer (2013). Concatenation versus coalescence versus “concatalescence”, *PNAS* March 26.

Groussin, Brochier, Gouy, **Boussau**, **Daubin** (2016). Gene Acquisitions from Bacteria at the Origins of Major Archaeal Clades Are Vastly Overestimated. *Mol Biol Evol*. 33(2):305-10.

Groussin, et al (2018) Species tree aware methods produce biochemically more realistic resurrected proteins, *to appear*

Hindré, Knibbe, **Beslon,** Schneider (2012) New insights into bacterial adaptation through in vivo and in silico experimental evolution *Nature Reviews Microbiology*, 10, 352

Kashtan, Alon (2005) Spontaneous evolution of modularity and network motifs. *PNAS*. 102(39):13773-13778;

Knibbe, Coulon, Mazet, Fayard, **Beslon** (2007) A long-term evolutionary pressure on the amount of noncoding DNA, *Mol. Biol. Evol.* 10:2344--2353

Knibbe, Parsons (2014) What happened to my genes? Insights on gene family dynamics from digital genetics experiments. *Proc of 14th Intl. Conf. on the Synthesis and Simulation of Living Systems*, pp. 33-40.

Kuo CH, Janzen FJ (2003), Bottlesim: a bottleneck simulation program for long-lived species with overlapping generations, *Molecular Ecology Notes*, Oct. 9

Lehman, …, **Beslon**, et al (2018) The Surprising Creativity of Digital Evolution: A Collection of Anecdotes from the Evolutionary Computation and Artificial Life Research Communities, arXiv:1803.03453

Lenski, Ofria, Pennock, Adami (2003) The evolutionary origin of complex features, *Nature 423 139–144*

Liard, **Rouzaud-Cornabas**, **Beslon** (2017). A 4-base model for the Aevol in-silico experimental evolution platform. *In European Conference on Artificial Life*. pp. 265-266.

Liu et al. (2010). A maximum pseudo-likelihood approach for estimating species trees under the coalescent model, *BMC Evolutionary Biology*, Oct 2010.

Liu, et al (2015) Genome-wide identification and evolutionary analysis of positively selected miRNA genes in domesticated rice, *Molecular Genetics and Genomics*, 290, 593-602

Luksza, Lässig (2014) Predictive fitness model for influenza, *Nature*, 507, 57-61

Mallo, *et al* (2015) Simphy: Phylogenomic simulation of gene, locus, and species trees, *Syst biol*, 65, 334-344

Mirarab, Bayzid, Warnow (2014). Evaluating summary methods for multi-locus species tree estimation in the presence of incomplete lineage sorting. *Syst Biol.* 65(3).

Moore, Höhna, May, Rannala, Huelsenbeck (2016). Critically evaluating the theory and performance of Bayesian analysis of macroevolutionary mixtures. *PNAS* 113 (34) 9569-9574.

Nelson-Sathi et al. (2015). Origins of major archaeal clades correspond to gene acquisitions from bacteria. *Nature*. 517(7532).

Pugh, Ken (2011). *Lean-Agile Acceptance Test-Driven Development: Better Software Through Collaboration*. Addison-Wesley.

Rabosky (2014). Automatic detection of key innovations, rate shifts, and diversity-dependence on phylogenetic trees. *PLoS One*. 9(2):e89543.

Rabosky, Mitchell, Chang (2017). Is BAMM Flawed? Theoretical and Practical Concerns in the Analysis of Multi-Rate Diversification Models. *Syst Biol*, 66(4).

Randall, *et al* (2016) An experimental phylogeny to benchmark ancestral sequence reconstruction, *Nature communications*, 7

Romiguier, Ranwez Douzery, Galtier, Genomic evidence for large, long-lived ancestors to placental mammals , *Molecular biology and evolution* 30 (1), 5-13

Rosenzweig, Pease, Besansky, Hahn, Powerful methods for detecting introgressed regions from population genomic data, *Molecular ecology*, 2016

Scaduto, *et al* (2010) Source identification in two criminal cases using phylogenetic analysis of HIV-1 DNA sequences, *PNAS*, 107, 21242-21247

Sjöstrand, *et al* (2013) GenPhyloData: realistic simulation of gene family evolution, *BMC Bioinf*, 14, 209

Slatkin *et al* (2016) Ancient DNA and human history, *PNAS*, 113, 6380-6387

Song , Liu, Edwards, Wu (2012). Resolving conflict in eutherian mammal phylogeny using phylogenomics and the multispecies coalescent model. *PNAS* 109(37):14942–14947

Strope, Abel, Scott, Moriyama (2009) Biological Sequence Simulation for Testing Complex Evolutionary Hypotheses: indel-Seq-Gen Version 2.0 *Molecular Biology and Evolution*, 26, 11

Szöllősi, **Boussau,** Abby, **Tannier, Daubin** (2012). Phylogenetic modeling of lateral gene transfer reconstructs the pattern and relative timing of speciations *PNAS*, 109, 17513-17518

Szöllősi, **Tannier**, Lartillot, **Daubin** (2013), Lateral Gene Transfer from the Dead, *Syst Biol*, 62, 3, 386–397

Thrall, *et al* (2011) Evolution in agriculture: the application of evolutionary approaches to the management of biotic interactions in agro-ecosystems, *Evol Appl*. 4(2): 200–215

Wilke, et al (2001) Evolution of digital organisms at high mutation rates leads to survival of the flattest. *Nature* 412(6844):331.

Identité / Personal details

Nom et prénom / Name and first name:	Tannier Eric
Identifiant chercheur / Researcher ID :	0000-0002-3681-7536
Page web personnelle / Personal Webpage :	http://lbbe.univ-lyon1.fr/-Tannier-Eric-.html

Position actuelle / Current position[1]

Organisme(s) public(s) français / French public organisation(s)

Code RNSR / RNSR code	Organisme / Organisation	Laboratoire / Laboratory	Code unite / Unit code	Code postal / Postcode	Ville / Town
201120997E	Inria	Beagle		69603	Villeurbanne

Organisme(s) privé(s) français / French private organisation(s)

Siret	Etablissement / Organisation	Direction service / Department unit	Code postal / Postal code	Ville /Town

Organisme(s) étranger(s) / Foreign organisation

Etablissement / organisation	Laboratoire / Laboratory	Ville / Town	Pays / Country

Titre / Function

Chargé de Recherches

Autres activités / Other activities

Activités de direction, encadrement, enseignement, activité d'évaluation dans des commissions ou en tant qu'expert scientifique / Executive board, supervision of student, teaching, memberships in panels or individual scientific reviewing activities

- Teaching in Computational Biology, Computer Science, History, Research Ethics, Epistemology, for Master and Doctoral Students and for a general public
- Member of the administration council of Inria
- Member of the « environmental and social responsibilities » committee at Inria
- Member of the Open Science Committee of the Ministry of Research
- Member of the Ethics Platform of Université de Lyon
- Evaluator for the ANR and FNRQ
- Editor for Discrete Mathematics and Theoretical Computer Science, and for Peer Community in Evolutionary Biolgy

Positions antérieures / Previous positions

Début / Start date	Fin / End date	Ville / Town	Etablissement / Organisation	Fonction / Function
2003	2004	Lyon	Inria	Post-doc
2003	2003	Rome	CNR	Post-doc
2002	2002	Bonn	Discrete Mathematics Institute	Post-doc
1999	2002	Grenoble	Université Fourrier	Ph-D

Formation supérieure / Education[2]

Habilitation to supervise researches in 2011
Ph-D in 2002 in Discrete Mathematics
Master in 1999 in Contemporaneous History
Master in 1997 in Combinatorics
Engineer in 1997 in Computer Science and Applied Mathematics

1 *Compléter la ou les sections appropriées / Fill the appropriate field(s)*
2 *Les jeunes chercheurs ayant soutenu leur thèse de doctorat depuis moins de 10 ans doivent préciser le nom de leur directeur de thèse. Les non-titulaires d'un PhD indiquent la date de leur dernier diplôme académique. / Young researchers who obtain their PhD up to 10 years ago must provide the name of their PhD supervisor. Researchers without a PhD must indicate the date of their last academic degree.*

Productions scientifiques / Scientific productions

Projets de recherche, prix, distinctions, bourses, etc. / Grants, prizes, awards, fellowships, etc.

- Executive Committee (Principal Investigator of a work package) of the ANCESTROME project, funded 2.2 million euros by the « projet investissement d'avenir » bioinformatics, 2012-2017
- Principal Investigator of a Technological Action of Inria, funded 150keuros, 2016-2018
- Principal Investigator of a Eco-NET project, funded 40 keuros (Europe) for an international project with Hugary and Bosnia, 2008-2010

5 publications majeures / 5 most relevant publications

1	**Davin AA, Tannier E, Williams TA, Boussau B, Daubin V, Szollosi GJ (2018)** Gene transfers can date the tree of life, *Nature ecology & evolution*, **vol. 2** pp.904-909.
2	**Biller P, Guéguen L, Knibbe C, Tannier E (2016)** Breaking Good: Accounting for Fragility of Genomic Regions in Rearrangement Distance Estimation, *Genome Biology and Evolution*, **vol. 8** pp.1427-39.
3	**Abby S S, Tannier E, Gouy M, Daubin V (2012)** Lateral gene transfer as a support for the tree of life, *Proceedings of the National Academy of Sciences of The United States of America*, **vol. 109** pp.4962-4967.
4	**Miklos I, Tannier E (2012)** Approximating the number of Double Cut-and-Join scenarios, *Theoretical Computer Science*, **vol. 439** pp.30-40.
5	**Fertin G, Labarre A, Rusu I, Tannier E, Vialette S (2009)** *Combinatorics of Genome Rearrangements*, MIT press, 2009

Publications des 3 dernières années / Publications in the past 3 years[3]

- D Hasic, E Tannier
Gene tree reconciliation including transfers with replacement is hard and FPT
Accepted by Journal of Combinatorial Optimization
https://link.springer.com/article/10.1007/s10878-019-00396-z
arXiv preprint arXiv:1709.04459

- **D Hasic, E Tannier**
Gene tree species tree reconciliation with gene conversion
Accepted by Journal of Mathematical Biology
https://link.springer.com/article/10.1007/s00285-019-01331-w
arXiv preprint arXiv:1703.08950

- Anselmetti Y, Duchemin W, Tannier E, Chauve C, Bérard S (2018)
Phylogenetic signal from rearrangements in 18 Anopheles species by joint scaffolding extant and ancestral genomes, *BMC Genomics*, **vol. 19** pp.96-96.

- **Anselmetti Y, Luhmann N, Bérard S, Tannier E, Chauve C (2018)**
Comparative Methods for Reconstructing Ancient Genome Organization
in: *Methods in Molecular Biology*, , pp.343-362.

- **Davin AA, Tannier E, Williams TA, Boussau B, Daubin V, Szollosi GJ (2018)**
Gene transfers can date the tree of life, *Nature ecology & evolution*, **vol. 2** pp.904-909.

- **Duchemin W, Gence G, Arigon Chifolleau AM, Arvestad L, Bansal M S, Berry V, Boussau B, Chevenet F, Comte N, Davin AA, Dessimoz C, Dylus D, Hasic D, Mallo D, Planel R, Posada D, Scornavacca C, Szollosi G, Zhang L, Tannier E, Daubin V (2018)**
RecPhyloXML - a format for reconciled gene trees, *Bioinformatics*, **vol.** pp.1-7.

- **Dumas JG, Roch JA, Tannier E, Varrette S (2018)**
Théorie des codes, , 2018 .

- **Chauve C, Rafiey A, Davin AA, Scornavacca C, Veber P, Boussau B, Szollosi G, Daubin V, Tannier E (2017)**
MaxTiC: Fast ranking of a phylogenetic tree by Maximum Time Consistency with lateral gene transfers, *Peer Community In Evolutionary Biology*, **vol.** pp.1-18.

---

3 *Les pré-prints sont acceptés. Si disponible, indiquer en fin de référence le lien open access pour en améliorer l'accessibilité. / Preprints are allowed. If available, indicate the open-access link of each reference to improve its accessibility*.

- **Duchemin W, Anselmetti Y, Patterson M, Ponty Y, Berard S, Chauve C, Scornavacca C, Daubin V, Tannier E (2017)**
DeCoSTAR: Reconstructing the Ancestral Organization of Genes or Genomes Using Reconciled Phylogenies, *Genome Biology and Evolution*, **vol. 9** pp.1312-1319.
- **Fertin G, Jean G, Tannier E (2017)**
Algorithms for computing the double cut and join distance on both gene order and intergenic sizes, *Algorithms for Molecular Biology*, **vol. 12** pp.16-16.
- **Jacox E, Weller M, Tannier E, Scornavacca C (2017)**
Resolution and reconciliation of non-binary gene trees with transfers duplications and losses, *Bioinformatics*, **vol. 33** pp.980-987.
- **Biller P, Guéguen L, Knibbe C, Tannier E (2016)**
Breaking Good: Accounting for Fragility of Genomic Regions in Rearrangement Distance Estimation, *Genome Biology and Evolution*, **vol. 8** pp.1427-39.
- **Biller P, Knibbe C, Beslon G, Tannier E (2016)**
Comparative Genomics on Artifical Life, *Computability in Europe,* pp.1-10.
- **Bulteau L, Fertin G, Tannier E (2016)**
Genome rearrangements with indels in intergenes restrict the scenario space, *BMC Bioinformatics*, **vol. 17** pp.225-231.
- **Fertin G, Jean G, Tannier E (2016)**
Genome Rearrangements on Both Gene Order and Intergenic Regions, *WABI,* , **vol. 9838** pp.162-173.
- **Groussin M, Daubin V, Gouy M, Tannier E (2016)**
Ancestral Reconstruction: Theory and Practice
in: *The Encyclopedia of Evolutionary Biology*, , pp.70-77.
- **Lays C, Tannier E, Henry T (2016)**
Francisella IglG protein and the DUF4280 proteins: PAAR-like proteins in non-canonical Type VI secretion systems?, *Microbial Cell*, **vol. 3** pp.445-447.
- **Noutahi E, Semeria M, Lafond M, Seguin J, Boussau B, Guéguen L, El-Mabrouk N, Tannier E (2016)**
Efficient Gene Tree Correction Guided by Genome Evolution, *PLoS One*, **vol. 11** pp.e0159559-e0159559.

Valorisation

brevet, licence, création d'entreprise, développement de logiciel, base de données, prototype, etc. / patent, creation of a start-up, software development, database, prototype, etc.

- Software Treerecs, user friendly phylogenetic reconciliation, project leader https://project.inria.fr/treerecs/
- Software Decostar, evolution of gene interactions and reconstruction of ancestral genomes, project leader http://pbil.univ-lyon1.fr/software/DeCoSTAR/
- Format RecphyloXML, unified format for phylogenetic reconciliation, project co-leader http://pbil.univ-lyon1.fr/software/DeCoSTAR/

Model of CV - Appendix

*Submit **in one document PDF** the CVs of the scientific coordinator (the French coordinator and the foreign coordinator (country referent) for PRCI) and the principal investigators of the other partners. Any other information included in the appendix will not be taken into account by the scientific evaluation panels. Applicants are strongly advised to draft their CV in English as evaluations may be carried out by non-French-speakers.*

Identité / Personal details

Nom et prénom / Name and first name:	Daubin Vincent
Identifiant chercheur / Researcher ID :	https://orcid.org/0000-0001-8269-9430
Page web personnelle / Personal Webpage :	http://lbbe.univ-lyon1.fr/-Daubin-Vincent-.html

Position actuelle / Current position

Organisme(s) public(s) français / French public organisation(s)

Code RNSR / RNSR code	Organisme / Organisation	Laboratoire / Laboratory	Code unite / Unit code	Code postal / Postcode	Ville / Town
199411998X	CNRS/ Université Lyon 1	LBBE	UMR5558	69622	Villeurbanne

Organisme(s) privé(s) français / French private organisation(s)

Siret	Etablissement / Organisation	Direction service / Department unit	Code postal / Postal code	Ville /Town

Organisme(s) étranger(s) / Foreign organisation

Etablissement / organisation	Laboratoire / Laboratory	Ville / Town	Pays / Country

Titre / Function

2013 – : CNRS Directeur de Recherche, 2ème classe, Laboratoire BBE Lyon, France.

Autres activités / Other activities

Activités de direction, encadrement, enseignement, activité d'évaluation dans des commissions ou en tant qu'expert scientifique / Executive board, supervision of student, teaching, memberships in panels or individual scientific reviewing activities

Responsable du département « co-évolution multi-échelle » du LBBE à partir du 1er janvier 2021; Expert scientifique à l'ERC, membre du panel de l'école doctorale E2M2, enseignement en phylogénie moléculaire (M2)…

Positions antérieures / Previous positions

Début / Start date	Fin / End date	Ville / Town	Etablissement / Organisation	Fonction / Function
2004	2013	Lyon	CNRS	chargé de recherche 1ere classe
2002	2004	Tucson, USA	University or Arizona	Post-Doc

Formation supérieure / Education

2002: thèse en évolution moléculaire, Lyon France

Productions scientifiques / Scientific productions

Projets de recherche, prix, distinctions, bourses, etc. / Grants, prizes, awards, fellowships, etc.

5 publications majeures / 5 most relevant publications

1	Davin AA, Tannier E, Williams TA, Boussau B, Daubin V, Szollosi GJ (2018), Gene transfers can date the tree of life, Nature ecology & evolution, vol. 2 pp.904-909.
2	Lassalle F, Perian S, Bataillon T, Nesme X, Duret L, Daubin V (2015), GC-Content Evolution in Bacterial Genomes: The Biased Gene Conversion Hypothesis Expands, PLoS Genetics, vol. 11 pp.e1004941-e1004941.
3	Szollosi G J, Tannier E, Daubin V, Boussau B (2015), The Inference of Gene Trees with Species Trees, Systematic biology, vol. 64 pp.e42-e62.
4	Batut B, Knibbe C, Marais G, Daubin V (2014), Reductive genome evolution at both ends of the bacterial population size spectrum, Nature reviews Microbiology, vol. 12 pp.841-50.
5	Szollosi GJ, Boussau B, Abby SS, Tannier E, Daubin V (2012), Phylogenetic modeling of lateral gene transfer reconstructs the pattern and relative timing of speciations, Proceedings of the National Academy of Sciences of The United States of America, vol. 109 pp.17513-17518.

Publications des 3 dernières années / Publications in the past 3 years

Davin AA, Tannier E, Williams TA, Boussau B, Daubin V, Szollosi GJ (2018)
Gene transfers can date the tree of life, Nature ecology & evolution, vol. 2 pp.904-909.

Duchemin W, Gence G, Arigon Chifolleau AM, Arvestad L, Bansal M S, Berry V, Boussau B, Chevenet F, Comte N, Davin AA, Dessimoz C, Dylus D, Hasic D, Mallo D, Planel R, Posada D, Scornavacca C, Szollosi G, Zhang L, Tannier E, Daubin V (2018)
RecPhyloXML - a format for reconciled gene trees, Bioinformatics, vol. pp.1-7.

Chauve C, Rafiey A, Davin AA, Scornavacca C, Veber P, Boussau B, Szollosi G, Daubin V, Tannier E (2017)
MaxTiC: Fast ranking of a phylogenetic tree by Maximum Time Consistency with lateral gene transfers, Peer Community In Evolutionary Biology, vol. pp.1-18.

Duchemin W, Anselmetti Y, Patterson M, Ponty Y, Berard S, Chauve C, Scornavacca C, Daubin V, Tannier E (2017)
DeCoSTAR: Reconstructing the Ancestral Organization of Genes or Genomes Using Reconciled Phylogenies, Genome Biology and Evolution, vol. 9 pp.1312-1319.

Venner S, Miele V, Terzian C, Biemont C, Daubin V, Feschotte C, Pontier D (2017)
Ecological networks to unravel the routes to horizontal transposon transfers, PLoS Biology, vol. 15 pp.e2001536-e2001536.

Groussin M, Boussau B, Szollosi G, Eme L, Gouy M, Brochier-Armanet C, Daubin V (2016)
Gene Acquisitions from Bacteria at the Origins of Major Archaeal Clades Are Vastly Overestimated, Molecular Biology and Evolution, vol. 33 pp.305-10.

Duchemin W, Daubin V, Tannier E (2015)
Reconstruction of an ancestral Yersinia pestis genome and comparison with an ancient sequence, BMC Genomics, vol. 16 Suppl 10 pp.S9-S9.

Lassalle F, Perian S, Bataillon T, Nesme X, Duret L, Daubin V (2015)
GC-Content Evolution in Bacterial Genomes: The Biased Gene Conversion Hypothesis Expands, PLoS Genetics, vol. 11 pp.e1004941-e1004941.

Szollosi GJ, Arellano Davin A, Tannier E, Daubin V, Boussau B (2015)
Genome-scale phylogenetic analysis finds extensive gene transfer among fungi, Philosophical Transactions of The Royal Society B-Biological Sciences, vol. 370 pp.201403335-201403335.

Szollosi G J, Tannier E, Daubin V, Boussau B (2015)
The Inference of Gene Trees with Species Trees, Systematic biology, vol. 64 pp.e42-e62.

Valorisation
brevet, licence, création d'entreprise, développement de logiciel, base de données, prototype, etc. / patent, creation of a start-up, software development, database, prototype, etc.