OTU - Observing the Unknown

I. Context, positioning and objectives of the project

Objectives of the project and research hypotheses

“*To a first approximation all species are extinct*” [1], and those that are not are mostly unknown. At least 5 billion species have existed in the past[1], and between 100 million[2–5] and a trillion[6] are suspected to live today. Of all those, we only know a tiny fraction: between 2 and 3 million species [7], extinct and extant. Describing species’ biodiversity and understanding the evolutionary processes that shaped it during the last 4 billion years while focusing only on a minor component of it is thus highly biased. It is like trying to understand the present document by reading only three characters (this is the actual highest estimate of the proportion of known species suspected to have ever lived). As an illustration, we recently demonstrated that overlooking the role of hidden lineages in studies of gene flow lead to systematic errors, resulting sometimes in conclusions opposite to those published[8–10]. New and original approaches are needed to specifically define and characterise invisible biodiversity, classically referred to as **ghost diversity**, to better apprehend their abundance and their distribution in time and space. The discovery of unknown microbial diversity may reveal novel pathogens with the potential to drive disease emergence and disrupt socio-economic stability.

For a long time, much of this ghost diversity was thought to be, and indeed remain,inaccessible. First, because many extinct species have left no traces in the fossil record because they lived in periods or places where fossilization was not favorable, or because simply they would not fossilize (like most bacteria and archaea). Second, because many extant unknown species, especially microorganisms, cannot be cultured in the lab, which prevents their proper description and characterization.

In recent years however, important advances changed our vision of ghost biodiversity and how it could be approached. First, genomic traces left by ghost lineages in extant genomes (through horizontal gene transfers or introgressions) could be used to reveal hidden lineages[11–14]. Second, metagenomes are now routinely obtained from various environments, enabling the direct exploration and quantification of unknown extant species from available data[15]. Third, coalescent-based genomics models can now infer past demographics and population structures from modern genomes[16,17], two signatures that could be useful to explore past speciation events involving ghost lineages.

This project will explore ghost diversity through these three complementary approaches, each corresponding to a work package detailed below. Together, we believe that studying ghost biodiversity as a research object in its own right, using such an interdisciplinary strategy, is both original and timely.

Methodology

WP1: Ghost lineages leave traces through gene transfers

This year, we provided the first proof of concept that identifying and quantifying horizontal gene transfers in groups of organisms where they are common allows ghost lineages to be placed along branches of the studied phylogenetic tree[14]. This opens new exciting lines of research to characterize and quantify ghost lineages in the Tree of Life, even for species without fossil records.

Objectives and methods of WP1: Work is needed to transform the proof of concept into a usable inference tool for detecting ghost lineages from empirical data. Our preliminary results show that graph neural networks trained on simulated data could quantify (and not just qualify) ghost lineages along the branches of a phylogeny. But limitations remain. To unlock them, we propose (i) to improve the realism of simulations on which models are trained to improve the predictions. This implies the amelioration of simulation and inference software to get rid computational bottlenecks ; (ii) to develop new methods for gene transfer detection (*i.e. reconciliation* methods) that directly integrate ghost species into the model of inference ; (iii) to demonstrate the benefit of having an estimate of the number of ghost lineages along each branch of a phylogenetic tree in macroevolutionary studies, for instance for birth–death diversification models[18]. These models for the inference of past macroevolutionary dynamics face well-known challenges: extinction rates are difficult to estimate reliably[19], and speciation and extinction rates can be unidentifiable because several combinations yield the same net diversification rate[20]. These issues of unidentifiability can be resolved and models improved by incorporating fossils or robust prior knowledge of past diversity[21]. We anticipate that estimates of ghost lineages will be highly valuable to achieve this goal, especially for clades where fossils are unavailable.

WP2: Ghost lineages in metagenomics data

Thanks to high throughput sequencing of all DNA fragments sampled in an environment, environmental metagenomics revolutionised the exploration of unknown biodiversity[15,22,23]. Indeed, unknown species can be detected either using sequence similarity of the sequenced "reads" or assembling them into large fragments of genomes (MAGs for Metagenome-Assembled Genomes). Doing so revealed that only a small fraction of sequence reads (~35%) are assigned to previously known species[24]. Some attempts to quantify and identify what we call here ghost diversity from metagenomic data have been done, but applied only to a small fraction of available metagenomes[24,25] due to computational limitations. The largest study assembling MAGs[26] (over 50,000) analyzed "only" 10,000 metagenomes—about 0.2% of the 5 million currently available in the public database SRA. Yet, this work expanded the known phylogenetic diversity of bacteria and archaea by 44%! This highlights both the tremendous amount of unknown diversity hidden in environmental metagenomic data, and the need for new efficient methods to estimate ghost biodiversity from the huge amount of available data.

<u>Objectives and methods of WP2</u>: we propose to study the 2 million biosamples of environmental metagenomes publicly available in SRA. For each sample, taxonomic assignation of each read (obtained with STAT[27]) and curated geographical information (coordinates, biome, etc.) are already available, thanks to the work performed on a related project (Virome@tlas, dedicated to the exploration of the virosphere). Preliminary results on a small subset of samples show that the proportion of reads left unassigned (*i.e.* they resemble nothing known in the reference databases) roughly correlates with fine-grained estimates of the proportion of "unknown species" (using SingleM[24]). This encourages us to go further and (i) perform simulations on *in silico* "mock" community (of known composition and abundancy) to evaluate the impact to the fraction of unassigned sequences of parameters such as the completeness of the database, the complexity of the community and technical artifacts related to library preparation and sequencing ; (ii) lead some comparative analyses on a subset of biological and ecological data to extend the estimate of unknown diversity with other approaches (MAG reconstruction, Metphlan[25]), before extrapolating to the whole dataset ; (iii) explore the geographic context and distribution of ghost (and non-ghost) diversity across ecosystems and re-evaluate large-scale geographical patterns of diversity (*e.g.*, latitudinal diversity gradient[28]).

WP3: Shadows of ghost lineages in past speciation events

Over the past 15 years, advances in evolutionary genomics have demonstrated the remarkable potential of genome analyses to uncover hidden patterns. Good illustrations are coalescent-based genomics models (*e.g.*, PSMC[16], MSMC[17], and related approaches[e.g. 29]) that can reconstruct past demographic history, *i.e.* the history of effective population sizes (*Ne*), and/or past population structures, sometimes from a single individual[16]. It appears as a promising direction to explore the potential for such approaches to detect **past speciation events**, thought to be associated with both *Ne* and structure variations over time. Identifying these ghost speciation events (for which only a single descendant species is known) would provide an indirect access to ghost lineages, without requiring any event of gene flow in secondary contact as in WP1.

<u>Objectives and methods of WP3</u>: We propose to explore signatures (or *scars*) of speciation events by studying extant genomes, inspired by demographic inference methods. This will be done with two parallel approaches: (i) forward-in-time simulations with the tool SLiM[30] to determine the conditions (parameters associated with speciation) under which a speciation event leaves a detectable signal in extant genomes; (ii) empirical analysis of biological data to see if variations in *Ne* and population structure as obtained with coalescent-based methods[16,17] coincide with speciation events visible in the phylogeny. If such coincidences are confirmed, this could open the way to new methods for detecting genomic *scars* of past speciation events. <u>Note</u>: This work package is the most exploratory in this project. Unlike the previous ones, no

preliminary results were obtained yet, and the approach proposed is risky, notably for validating detected *scars* of speciation. We are aware of the difficulties associated with this workpackage, but we are also convinced that this is how research works: risky ideas lead to important breakthroughs.

Visualizing ghost lineages in the tree of life

In addition to the three work-packages presented above, the project will extend functionalities of the popular visualisation tool Lifemap[31] – developed and maintained by the coordinator of this project – in order to integrate visualisation of estimates of ghost lineages abundance along the branches of the Tree of Life. This is an important feature for communicating to a broad and general audience (Lifemap has ~15 000 visitors/month) the questions addressed in this project and the importance of ghost diversity.
In conclusion, this ambitious project proposes to define new ways to quantify and characterize ghost diversity, and to illustrate the use of such knowledge for a better understanding of the macroevolutionary history and geographical distribution of past and present biodiversity.

Added-value in terms of scientific contribution and knowledge production

This project represents a conceptual and methodological shift in how biodiversity is studied, by focusing explicitly on the unknown fraction of it. While ghost lineages have long been seen both as inaccessible and as a source of noise and errors in analyses, we propose to make it a research subject on its own. The project aims at developing original methods to detect, quantify, and characterize these hidden lineages using complementary sources of information: genomic signatures of horizontal gene transfers (WP1), large-scale metagenomic data (WP2), and traces of past speciations inferred from modern genomes (WP3). These approaches combine theoretical, computational, and empirical developments to reveal portions of the Tree of Life previously unreachable. The project's added value lies in its potential to redefine biodiversity by extending it beyond the observable species, providing a more accurate framework to reconstruct macroevolutionary dynamics.

Ability of the project to address the research issues covered by Theme A.02: Living Earth

By specifically focusing on a fraction of biodiversity that has long been overlooked and neglected – ghost biodiversity – it will improve fundamental knowledge on biodiversity and its evolutionary dynamics. The project is truly interdisciplinary, combining various approaches such as evolutionary biology, genomics, metagenomics, phylogenetics, and modelling to characterize hidden diversity across all environments.

II. Partnership

This project covers all aspects of genomics, from phylogenomics to metagenomics through evolutionary genomics, and relies on various skills ranging from deep learning to empirical data analysis through modeling. Members of this project cover all necessary aspects. The coordinator Damien M. de Vienne (CR CNRS, LBBE, WP1 (coord), WP2, WP3) is an evolutionary biologist and phylogeneticist. Recently, he coordinated research on ghost lineages, their negative impact for the study of gene flow[8–10,32], and their possible identification[14]. These results stimulated new interest in ghost lineages, especially in population genetics[33–35]. He recently opened to metagenomics by joining the Virome@tlas project and by coordinating the project Terra-Incognita-Seq (two master thesis) dedicated to the exploration of unknown species in metagenomics data (with V Navratil and A Haudry). He is the author of Lifemap, a popular tool for the exploration of the tree of life[31]. Laurent Jacob (DR CNRS, LCQB, WP1) is an expert in machine learning for genomics[36,37], with an emphasis on molecular evolution[38,39]. He was involved in our previous endeavor for neural inference of ghost lineages. Bastien Boussau (DR CNRS, LBBE, WP1) is a specialist in phylogenetics and phylogenomics, recently developing deep learning approaches in these fields. With L Jacob and DM de Vienne, he was involved in the preliminary studies for the detection of ghosts with neural networks. Julien Clavel (CR CNRS, LEHNA, WP1) is a specialist in phylogenetic modelling and comparative methods. He is one of the developers and maintainers of a popular package dedicated to diversification analyses on phylogenies[40]. Annabelle Haudry (MC UCBL, LBBE, WP2 (coord), WP3) is an expert in population and comparative genomics. As an associate professor, she recently co-created the first environmental genomics master class in France and developed new metagenomics projects involving students. Vincent Navratil (IR

UCBL, PRABI, WP2) is specialized in large-scale sequence analysis. He is the head of the bioinformatics platform PRABI and the coordinator of Virome@tlas. Carina Mugal (CR CNRS, LBBE - WP3) brings expertise in population genomics and mathematical population genetics, essential for detecting genomic signatures of past speciation events and accessing ghost lineages. Laure Segurel's (CR CNRS, LBBE - WP3) expertise in evolutionary genomics and human population genetics will support the identification and interpretation of demographic signals related to ghost speciation events. Sylvain Mousset (MC UCBL, LBBE - WP3) contributes extensive experience in evolutionary modeling and computational genomics, crucial for simulating and analyzing patterns left by past speciation events in extant genomes.

III. Bibliography

1. Raup, D. M. *Extinction: Bad Genes or Bad Luck?* (W.W. Norton, New York, 1991).
2. May, R. M. Tropical arthropod species, more or less? *Science* **329**, 41–42 (2010).
3. Mora, C. *et al.* How Many Species Are There on Earth and in the Ocean? *PLoS Biol.* **9**, e1001127 (2011).
4. Heywood, V. H. *et al. Global Biodiversity Assessment*. vol. 1140 (Cambridge university press Cambridge, 1995).
5. Louca, S. Simulating trees with millions of species. *Bioinformatics* **36**, 2907–2908 (2020).
6. Locey, K. J. *et al.* Scaling laws predict global microbial diversity. *Proc. Natl. Acad. Sci.* **113**, 5970–5975 (2016).
7. Schoch, C. L. *et al.* NCBI Taxonomy: a comprehensive update on curation, resources and tools. *Database* **2020**, baaa062 (2020).
8. Tricou, T. *et al.* Ghost lineages can invalidate or even reverse findings regarding gene flow. *PLOS Biol.* **20**, 1–16 (2022).
9. Tricou, T. *et al.* Comment on "On the impact of incomplete taxon sampling on the relative timing of gene transfer events". *PLOS Biol.* (2024).
10. Tricou, T. *et al.* Ghost lineages highly influence the interpretation of introgression tests. *Syst. Biol.* **71**, 1147–1158 (2022).
11. Andam, C. P. *et al.* Biased gene transfer in microbial evolution. *Nat. Rev. Microbiol.* **9**, 543–555 (2011).
12. Andam, C. P. *et al.* Ancient origin of the divergent forms of leucyl-tRNA synthetases in the Halobacteriales. *BMC Evol. Biol.* **12**, 85 (2012).
13. Szöllősi, G. J. *et al.* Lateral Gene Transfer from the Dead. *Syst. Biol.* **62**, 386–397 (2013).
14. Tricou, T. *et al.* Gene flow can reveal ghost lineages. *Evol. J. Linn. Soc.* kzaf014 (2025) doi:10.1093/evolinnean/kzaf014.
15. Parks, D. H. *et al.* Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. *Nat. Microbiol.* **2**, 1533–1542 (2017).
16. Li, H. *et al.* Inference of human population history from individual whole-genome sequences. *Nature* **475**, 493–496 (2011).
17. Schiffels, S. *et al.* Inferring human population size and separation history from multiple genome sequences. *Nat. Genet.* **46**, 919–925 (2014).
18. Morlon, H. *et al.* Phylogenetic Insights into Diversification. *Annu. Rev. Ecol. Evol. Syst.* **55**, 1–21 (2024).
19. Rabosky, D. L. Extinction rates should not be estimated from molecular phylogenies. *Evolution* **64**, 1816–1824 (2010).
20. Louca, S. *et al.* Extant timetrees are consistent with a myriad of diversification histories. *Nature* **580**, 502–505 (2020).
21. Wright, A. M. *et al.* Integrating Fossil Observations Into Phylogenetics Using the Fossilized Birth–Death Model. *Annu. Rev. Ecol. Evol. Syst.* **53**, 251–273 (2022).
22. Bernard, G. *et al.* Microbial Dark Matter Investigations: How Microbial Studies Transform Biological Knowledge and Empirically Sketch a Logic of Scientific Discovery. *Genome Biol. Evol.* **10**, 707–715 (2018).
23. Escudeiro, P. *et al.* Functional characterization of prokaryotic dark matter: the road so far and what lies ahead. *Curr. Res. Microb. Sci.* **3**, 100159 (2022).
24. Woodcroft, B. J. *et al.* Comprehensive taxonomic identification of microbial species in metagenomic data using SingleM and Sandpiper. *Nat. Biotechnol.* 1–6 (2025).
25. Blanco-Míguez, A. *et al.* Extending and improving metagenomic taxonomic profiling with uncharacterized species using MetaPhlAn 4. *Nat. Biotechnol.* 1–12 (2023).
26. Nayfach, S. *et al.* A genomic catalog of Earth's microbiomes. *Nat. Biotechnol.* **39**, 499–509 (2021).
27. Katz, K. S. *et al.* STAT: a fast, scalable, MinHash-based k-mer tool to assess Sequence Read Archive next-generation sequence submissions. *Genome Biol.* **22**, 270 (2021).
28. Cao, K. *et al.* Species packing and the latitudinal gradient in beta-diversity. *Proc. R. Soc. B Biol. Sci.* **288**, 20203045 (2021).
29. Wang, K. *et al.* Tracking human population structure through time from whole genome sequences. *PLoS Genet.* **16**, e1008552 (2020).
30. Haller, B. C. *et al.* SLiM 4: Multispecies Eco-Evolutionary Modeling. *Am. Nat.* **201**, E127–E139 (2023).
31. de Vienne, D. M. Lifemap: Exploring the Entire Tree of Life. *PLOS Biol.* **14**, e2001624 (2016).
32. Tannier, E. *et al.* HGTs are not SPRs: In the Presence of Ghost Lineages, Series of Horizontal Gene Transfers do not Result in Series of Subtree Pruning and Regrafting. *Mol. Biol. Evol.* **42**, msaf128 (2025).
33. Pang, X.-X. *et al.* Detection of Ghost Introgression Requires Exploiting Topological and Branch Length Information. *Syst. Biol.* **73**, 207–222 (2024).
34. Forsythe, E. S. *et al.* Detecting cryptic ghost lineage introgression in four-taxon genomic datasets. 2025.04.28.651118 Preprint at https://doi.org/10.1101/2025.04.28.651118 (2025).
35. Tolman, E. R. *et al.* GhostParser: A highly scalable phylogenomic approach for the identification of ghost introgression. 2025.08.21.671585 Preprint at https://doi.org/10.1101/2025.08.21.671585 (2025).
36. Chen, D. *et al.* Biological sequence modeling with convolutional kernel networks. *Bioinformatics* **35**, 3294–3302 (2019).
37. Villié, A. *et al.* Neural Networks beyond explainability: Selective inference for sequence motifs. *Transactions on Machine Learning Research*.

38. Nesterenko, L. *et al.* Phyloformer: Fast, Accurate, and Versatile Phylogenetic Reconstruction with Deep Neural Networks. *Mol. Biol. Evol.* **42**, msaf051 (2025).
39. Leroy, A. *et al.* Graph Neural Networks for Likelihood-Free Inference in Diversification Models. Preprint at https://www.biorxiv.org/content/10.1101/2025.08.14.670341 (2025).
40. Morlon, H. *et al.* RPANDA: an R package for macroevolutionary analyses on phylogenetic trees. *Methods Ecol. Evol.* **7**, 589–597 (2016).