Biosurveillance Next-Gen des changements dans la structure et fonctionnement des écosystèmes

Next Generation Biomonitoring of change in ecosystems structure and function

TABLE OF CONTENTS

Project summary

In a just published Opinion paper in *Trends in Ecology & Evolution*, we advocate that a next-generation, global-scale, ecological approach to biomonitoring will emerge in the coming decade, which can detect ecosystem change accurately, cheaply and generically. Next-generation sequencing (NGS) of DNA sampled from the Earth's environments, would provide data for the relative abundance of operational taxonomic units or ecological functions. Machine-learning methods would then be used to reconstruct the ecological networks of interactions implicit in the raw NGS data in order to detect and predict ecosystem change.

In this Next Generation Biomonitoring (NGB) project, we will examine whether NGS samples from five distinct ecosystems undergoing global change can be used to reconstruct hypothetical networks of interaction using machine learning. We will then compare these reconstructed networks with the current state of knowledge for these systems to test whether NGS and machine learning approaches can be used to reconstruct valid ecological networks. These tests will include examining the NGS networks for specific, established interactions through to detailed comparisons against already-known ecological networks, built using classic network construction approaches. The five systems we will work on represent a cross-section of the organisational scales, drivers of change and data quality we would expect that a NGB approach could be applied to. From microbial interaction networks to macro-biome networks of interacting invertebrates, and across drivers of change such as invasion, disease, conservation, management and climate, the project will determine whether ecosystem change can be detected using an NGB approach. We will troubleshoot many of the technical, methodological and ecological problems facing the development of an NGB approach, such as the variable quality of NGS databases, taxa biases, identification errors, zero-rich data and asymmetric abundance distributions, and develop statistical approaches for detecting change and determining the size and power of biomonitoring programs.

Ultimately, we envision the development of autonomous samplers that would sample nucleic acids and upload NGS sequence data to the cloud for network reconstruction, using methods that we will develop in the project. Large numbers of these samplers, in a global array, would allow sensitive automated biomonitoring of the Earth's major ecosystems at high spatial and temporal resolution, revolutionising our understanding of ecosystem change.

Table 1. Summary of permanent researchers involved in the project

Partner	Surname	Name	Position	Months	Role & responsibilities in the project
1. INRA UMR 1347 Agroécologie	BOHAN	David	DR	12	Leader of WP6, Quantitative ecologist
	PETIT-MICHAUT	Sandrine	DR	8	Expert in agro-ecology
	RICCI	Benoît	CR	8	Expert in modelling agricultural systems
2. INRA UMR1202 BioGeCo	VACHER	Corinne	CR	10	Leader of WP1, Microbial and network ecologist
	DELZON	Sylvain	DR	6	Expert in plant tolerance to drought
	BURLETT	Régis	IE	3	Expert in plant physiology measurements
	SEGURA	Raphaël	TR	3	Field sampling and ecophysiology measurements
	LALANNE	Céline	AI	3	Molecular biologist
	SALIN	Franck	IR	3	Sequencing facility manager (PGTB)
	GUICHOUX	Erwan	IR	3	Coordinator of the NGS activities of PTGB
	CHANCEREL	Emilie	IE	6	Molecular biologist and bioinformatician
	BOURY	Christophe	AI	3	Molecular biologist at PGTB
3. INRA UMR 1349 IGEPP	CANARD	Elsa	CR	12	Expert in ecological network structure
	PLANAGENEST	Manuel	PR	4	Expert in agro-ecology and molecular testing of trophic interactions
4. Université de Lille 1 UMR 8198 EEP	MASSOL	François	CR	6	Leader of WP4, theoretical ecologist specialized in evolutionary and spatial ecology
	DUPUTIE	Anne	MC	6	Expert in shifts in plant ranges and climate change
	GALLINA	Sophie	IR	6	Expert in Bioinformatics
	HAUTEKEETE	Nina	MC	6	Expert in plant ecology and plant-pollinator networks
	PIQUOT	Yves	MC	6	Expert in plant ecology and plant-pollinator networks
	POUX	Céline	MC	6	Expert in metabarcoding and bioinformatics
5. CNRS CEFE UMR 5175	DAVID	Patrice	DR	12	Leader of WP3, evolutionary community ecologist
	JARNE	Philippe	DR	10	Evolutionary community ecologist
	SCHATZ	Bertrand	DR	10	Plant-insect interactions and pollination ecology
6. AgroParisTech / INRA UMR 518 MIA	ROBIN	Stéphane	DR	8	Leader of WP2, statistical network inference and analysis
	AUBERT	Julie	IR	9.6	Statistical network inference and analysis
	CHIQUET	Julien	CR	4.8	Statistical network inference and analysis
	DONNET	Sophie	CR	4.8	Statistical network inference and analysis
	BARBILLON	Pierre	MC	4.8	Statistical network inference and analysis
	OUADAH	Sarah	MC	4.8	Statistical network inference and analysis
	DELATTRE	Maud	MC	4.8	Statistical network inference and analysis
	LEBARBIER	Emilie	MC	4.8	Statistical network inference and analysis
7. CIRAD UMR C53 PVBMT	RAVIGNÉ	Virginie	CR	6	Leader of WP5, evolutionary ecology of plant pests
	BECKER	Nathalie	CR	12	Expert in host microbe interactions
	CHIROLEU	Frédéric	CR	3	Expert in population dynamics and statistics
	DUYCK	Pierre-François	CR	2	Expert in fruit fly ecology
	FACON	Benoît	CR	12	Expert in fruit fly ecology
	GLENAC	Serge	TR	3	Expert in fruit fly trapping and rearing
	PAYET	Jim	TR	3	Expert in fruit fly trapping and rearing
	ROUMAGNAC	Philippe	CR	2	Expert in rice virome
	VERNIERE	Christian	CR	2	Expert in rice bacteriome
8. Imperial College, London, UK	TAMADDONI-NEZHAD	Alireza			Expert in logic-based machine learning

Changes that have been made in the full proposal

A rice system of foliar bacteria and fungi that confer resistance to a fungal phyto-pathogen has been included.

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

Global change affects network structure and biodiversity-derived ecosystem services. The diversity of species and their interactions supports many of the ecosystem services on which humanity relies[1,2]. Global change threatens the provision of these services by affecting species ranges and altering niches. Global change-derived drivers of change include new pathogens, diseases and invasive pest species, and mitigation such as management and conservation. Networks have become the standard ecological method for studying systems of interacting species and their functions[3]. ***To detect, monitor and understand change in ecosystem functioning, we need to develop methods that integrate network ecology into biomonitoring*[4].**

The potential of next-generation sequencing (NGS) for learning networks of interactions is enormous[5,6]. Creating and measuring change in classically constructed ecological networks is so costly in time and money it precludes their use in biomonitoring. Recently, however, networks have been reconstructed directly from ecological sample data using machine-learning approaches[2,7], and validated both against the literature and by direct testing for feeding interactions[8]. This demonstrated that machine learning can rapidly and robustly recover ecological interaction information from raw data[2,9]. Here we extend this idea to learning from a generic form of ecological data. The diversity of barcode sequences[10,11] from next-generation sequencing (NGS) of environmental DNA (eDNA from species or operational taxonomic units, OTUs) is here used as data. We believe that this approach will soon become predominant in deciphering interactions because: (1) nucleic acids are ubiquitous, being shared by all forms of life; (2) NGS platforms generate millions of DNA sequences for a few hundred dollars; (3) they can characterize organisms in complex environmental samples (e.g. soil, water, plant tissues, faeces, pellet, gut content, etc.); (4) many sequences can be identified at the species- or genus-level by using reference databases; and, (5) the interactions between these species could be identified based on species abundance patterns[12] and knowledge on their functional traits[2]. ***This enormous potential of NGS data for resolving interaction networks can only be fully exploited with learning approaches.***

Machine-learning algorithms require specific development. To date, machine learning of ecological networks has been done using two very different approaches to learning: statistical and logic-based machine learning. Logic-based approaches learn network links using a logical pattern based approach. The same sentence in two languages can be translated because both have similar logical patterns of nouns, verbs and adjectives and interaction links between species or OTUs can be learnt by searching for logical patterns in ecological data. Inductive Logic Programming (ILP) has automatically generated quantitative food webs for species and functional groups from classic ecological abundance data[2,7]. Subsequent literature and molecular testing of species gut contents revealed that these networks were valid[7]. Statistical approaches are rather familiar, making use of the variance and co-variance structure of data-sets. Bayesian, statistical approaches, implemented in the *R* package Saturnin[13-15], have provided a "proof-of-concept" that learning from NGS data works. The network of interactions was reconstructed between an invasive fungal pathogen and other microbial species in the foliar system using NGS data[13] (Figure 1). ***However, it is necessary to demonstrate the validity of these different learning approaches with further NGS datasets and combine their best features to improve the learning of networks from NGS data.***

Validation of NGS-reconstructed networks in the NGB project. The process of NGS analysis, network reconstruction and validation will be done in five study ecosystems that reflect the possible diversity of NGB biomonitoring. These systems come either with expectations for particular OTU (species) interactions and/or classic ecological networks, which will allow us to validate the NGS reconstructed networks to be validated will be done in a series of sequential steps[2,7]. We will ask: i) do the identified NGS networks and perform as expected from network theory[16]; ii) are the networks valid when compared to text-mining results from the literature[17]; and, iii) do the NGS networks perform similarly to their classic counterparts? ***We will test whether NGS networks are scientifically valid and ecologically relevant.***

Application to biomonitoring of ecosystems in a changing world. Future NGS-based biomonitoring could be done at various scales, from a single station monitoring a particular site through to a great many stations monitoring effects at continental scales. In this project, we will begin the development of this future biomonitoring by testing whether next-generation biomonitoring (NGB) approaches to biomonitoring could detect ecosystem structural and functional change. This research will make use of the replication in the networks reconstructed from the NGS samples of the study systems to ask the question 'can we detect change in structure given the natural variation among NGS-based ecological networks?' Our approach will be rather similar to a power analysis, asking how many NGS samples are necessary to detect a change greater than the natural variation using network structural statistics[18]. Together, the power analysis and interpretation of ecological effects will allow us to propose NGS-based biomonitoring protocols for detecting prescribed levels of change in ecosystems. ***NGB networks from NGS data will give rapid, sensitive and statistically robust detection of ecosystem change.***

I.1. Objectives and scientific hypotheses

Scientific objective - A Revolution in Biomonitoring.

Our vision is to develop and test a generic NGB approach that will detect ecosystem-wide change more rapidly, sensitively and cheaply than current biomonitoring. Using a unique combination of NGS and Machine Learning, NGB will reconstruct species interaction networks to identify change in ecosystem properties, revolutionising both our understanding of ecosystems and our ability to predict and mitigate global change[4].

Scientific hypotheses

The NGB project has a clear set of scientific hypotheses and research goals that support the scientific objective. These *project* hypotheses and goals will be tested and done in all five systems in parallel. We will test whether:

- H_1 - *machine learning approaches can be used to reconstruct hypothetical networks of ecological interaction from NGS data, in five distinct ecosystems affected by global drivers of change;*
- H_2 - *the NGS networks reconstructed for each ecosystem are similar to the already-known ecological networks for the system, constructed using classic ecological approaches;*
  - $H_{2,i}$ - *the OTUs and hypothesised links identified in the network reconstruction perform as we expect from network theory*[16]*;*
  - $H_{2,ii}$ - *the reconstructed networks are valid, when compared to text-mining results from the literature*[17]*;*
  - $H_{2,iii}$ - *the NGS networks perform similarly to their classic counterparts;*
- H_3 - *we can detect change in NGS networks due to global drivers;*
  - $H_{3,i}$ - *we can characterise natural network variation;*
  - $H_{3,ii}$ - *we can define change in network structure attributable to global drivers.*

We will also test system-specific hypotheses that relate network structure to global drivers of change and change to ecological function (Table 3).

I.2. Originality and relevance in relation to the state of the art

The NGB approach builds on scientific and technological elements that already exist and are either being used practically or in test situations: i) NGS-based approaches are being used to biomonitor some ecosystem functions subject to change; ii) networks have been learnt from classic ecological data, and more recently from NGS data, pioneered by members of this consortium; and, iii) there have been calls to integrate network approaches into biomonitoring to improve decision-making in ecosystem management. What this consortium does that is *entirely original* is to, for the first time, put forward a convincing argument and proof-of-concept that these elements could be combined to build a NGB approach that would provide a standardised and sensitive biomonitoring, potentially of all the Earth's ecosystems, at high resolution and in real-time[4]. We argue that by reconstructing highly replicated networks of ecological interactions, such biomonitoring would provide the global standard of ecosystem information and revolutionise our ability to measure, understand and predict how all ecosystems respond to environmental change.

Contribution of the consortium to the State of the art

NGB consortium members have contributed to a series of methodological and theoretical developments that are a proof-of-concept for the NGB approach.

Development 1: Can machine learning be used to accelerate network (re)construction from ecological data?

Machine learning uses the variation in ecological data-sets, habitually employed by ecologists to test hypotheses for past reproduction, migration or predation. We have recently reconstructed a replicated, agro-ecological food web from a large herbicide treatment data-set[19,20], using logic-based machine learning[21]. The trophic links in the invertebrate species abundance data-set were transformed to a logarithmic treatment-ratio across the herbicide levels. Correlated changes in the ratios of pairs of species in the data-set, driven by herbicide, were hypothesised to be due to trophic interactions. However, correlations in data arise for many reasons, including chance, and do not alone imply a trophic interaction. To avoid spurious interactions, the learning was guided using 'background information' that serves as a model for a trophic interaction. It posited that a trophic interaction is one in which: i) predator x co-occurs at the same sample

points as prey y; ii) x has appropriate mouthparts to consume y; and, iii) calling on a basic hypothesis of trophic ecology, x should have a larger body size than y - big things eat small things[21,22]. Validation was done through an analysis of the literature[7,21]. The frequency of learnt links correlated well with the frequency at which the link was found in the literature[7,23]. In essence, machine learning can reconstruct a network that might have been hypothesised from expert knowledge and the literature.

Development 2: Can we learn new science using network reconstruction? Machine learning should not simply reproduce what we already know, but also learn new science. In the above-mentioned network, illogical links involving spiders as prey were hypothesised. These were tested by examining the gut contents of 'spider-predator' species using spider-specific, DNA primers. Several species of spiders were present in the predator guts[8], demonstrating the potential of machine learning to discover new ecological science.

Development 3: Could we reconstruct networks from NGS data? The great beauty of NGS methods is that the nucleic acids they work with are common to all life forms and ubiquitous. In principle, NGS can be applied to the identification of OTUs and functions in samples from any environment with minimal change in protocol. These benefits have driven the huge interest in eDNA as a source of data[24-28]. Our argument is that, if coupled, machine learning and NGS data could lay the foundation of a generic and rapid network-based biomonitoring system, which would require relatively little refinement to fit the environmental context in which it is deployed. We have recently attempted learning the microbial network on the leaves of oak trees (*Quercus robur*) to identify potential antagonists of the causal agent of powdery-mildew, *Erysiphe alphitoides* (Figure 1). It was proposed that most of the co-occurrence links should be explained by environmental requirements shared between OTUs[13]. Using environmental correlation as background information, statistical inference was used to reconstruct a network of 26 OTU nodes. These hypothesised links between the resident microbial OTUs and *E. alphitoides*, some of which indicate mechanisms that facilitate or suppress invasion and disease, are driving new research in this system.

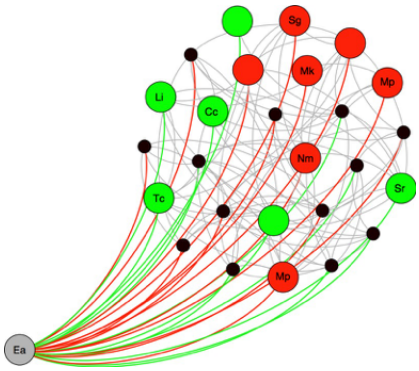

Figure 1. Microbial network of the oak tree (*Quercus robur* L.) susceptible to the foliar fungal pathogen, *Erysiphe alphitoides* (Ea). Each node represents a microbial OTU and red and green links indicate hypothesised co-exclusions and co-associations, respectively.

These consortium contributions demonstrate that machine learning: i) can reconstruct valid networks from ecological abundance data; ii) learns new ecological science leading to new research; and, iii) could be extended to NGS data, building ecological networks from ubiquitous eDNA. NGB approaches could, therefore, become a generic method for the rapid construction and sensitive detection of change in ecological networks. However, Weiss et al.[29] recently cautioned that machine-learning methods may vary widely in sensitivity and precision. We thus need to demonstrate that the merging of NGS data and learning can reconstruct networks for other systems (H_1), that these networks are valid (H_2) and can reliably detect change due to global drivers of change (H_3). We also need to determine which, of logical and statistical machine-learning approaches, meet the requirements of NGB for high-quality learning and speed.

Achieving the project scientific objective - Building an NGB approach using test ecosystems.

The NGB consortium partners research five ecosystems that we will use to test H_1 - H_3 (Figure 2). These ecosystems have been chosen because they reflect the diversity and complexity of ecosystem change biomonitoring and each is well-enough characterised that formal testing of an NGS reconstructed network is possible. Specifically, each ecosystem comes with, at least, system-specific hypotheses linking structure and function to global drivers and expectations for key interactions between known OTUs. In some systems, such

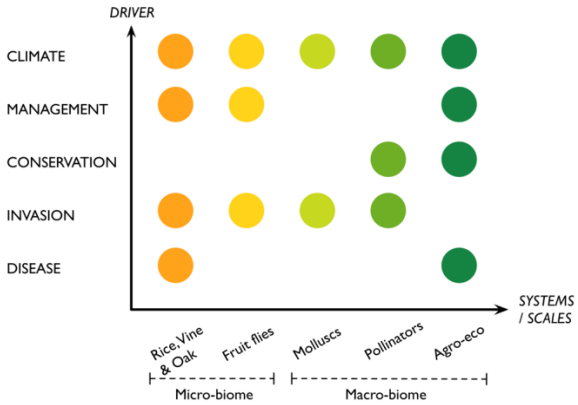

Figure 2. Summary diagram of scales and drivers of the five ecosystems being tested in the NGB project

as those for the pollinators (see Figure 6, Section II.3) and agro-ecological ecosystems, classic networks are available for testing, and networks for the fruit-fly system will become available during the project. The five systems exist at two scales of biological organisation; the micro-biome of interacting bacterial and fungal species that are habitually identified by NGS OTUs and a macro-biome scale of networks of interacting invertebrates identified using taxonomy.

I.3. Risk management

The risks to this project are predominantly technological and developmental challenges that are part of the aims of the NGB project (see also[4]). Here we detail these challenges, and our risk strategy.

NGS sequencing: The process of NGS sequencing, going from eDNA samples to the bioinformatic analysis of sequence data, can prove to be difficult in new systems. We avoid many of these problems by choosing systems where much of this development work has already been done (Table 3). We have, in addition, two strategies for managing these potential sequencing problems. First, we have deliberately constructed the work packages (WP) around the idea of a workflow that progresses from sampling to sequencing, network reconstruction and validation, and ultimately to detection of change and protocol development (see Figure 4). All ecosystems will follow this workflow in parallel. Thus, should any problems arise, in any system, these can be discussed and solutions found across the whole project because all the other systems are simultaneously going through the same set of processes. Second, we mutualise the NGS meta-barcoding step, reducing costs and the potential for errors in sequencing. The extracted eDNA samples from all ecosystems will be sequenced centrally by the Plateforme Génome Transcriptome de Bordeaux (PGTB) at Partner 2. These strategies mean that there is strength and support in depth in the project – redundancy – that will lead to great resilience to risk. We would also note that this will promote communication, collaboration and publication across the consortium, as all WP leaders have to actively involve themselves in all the ecosystems.

NGS identification errors: NGS data is susceptible to primer biases, where the primers used preferentially amplify some OTUs at the expense of others. This bias can be reduced using appropriate molecular approaches[30,31,32]. Furthermore, issues with sequencing errors, noise and statistical artefacts of the data (e.g. zero-rich data, asymmetric distributions) have all been studied at length and appropriate bioinformatics approaches exist to deal with these[33].

NGS OTU Databases: A major challenge is the quality of OTU databases. While certain gene databases are reasonably well populated and robust, coverage of all genes and taxa is still incomplete and databases contain incorrectly assigned sequences or omit entire taxonomic groups[34]. A significant, global effort is required to provide robust, well-curated and maintained sequences databases for use in NGB; fortunately, this is already underway for many taxonomic groups (e.g. International Barcode of Life[35]). NGB could accelerate this process by providing the "big picture" impetus for these shared sequence databases that would transform NGS. The unknown OTUs identified during the project will be uploaded to all appropriate online databases.

Learning: Weiss et al.[29] recently showed that reconstruction of microbial networks from NGS data varied widely in sensitivity and precision. A challenge for the NGB project is to develop statistical and logic-based learning methodologies that can reconstruct networks which satisfy H_1-H_3. We believe that our past results (section 1.2) demonstrate that the learning approaches we have developed to date work well, in proof of concept. We are confident that further development during the project will improve the quality of the reconstruction and the genericity of the methods.

Validation and the statistics of network variation: There is, in NGS data-sets, variation necessary to reconstruct networks[9,13,36]. What is unknown is the level of between-replicate variation that exists in the NGS data we will sample for the five ecosystems; some ecosystems are likely to be much more variable across replicate networks than others. Finding this out is one aspect of our goals to reconstruct and test NGS networks, and does not of itself represent a risk. It does affect, however, our ability to test, at the system-specific level, whether NGS networks are similar to classically constructed networks, and to link NGS network variation with both functional change and anthropogenic drivers of change. For these systems, which might satisfy only some of our project hypotheses, H_1-H_3, the aim will be to understand why NGS reconstructed networks might differ from our classic ecological networks and expectations.

II. Project organisation and means implemented

II.1. Scientific coordinator

The consortium is led by Dr David A. Bohan, who represents the UMR Agroécologie at INRA-Dijon. He has been a Principal Investigator for 12 years, at both Rothamsted Research in the UK and at INRA. Dave is a quantitative ecologist with an interest in ecosystem functions/services, replicated networks and learning methodologies, and is currently the editor of Advances in Ecological Research. His interest in ecosystem change began as a consortium member of the FarmScale Evaluations (FSE) of the impacts of GM herbicide-tolerant crops on farmland biodiversity. This trial remains the largest of its kind ever conducted, and played an important role in changing the direction of agriculture in the UK and Europe away from GMO cropping. Dave has since investigated how farmland management affects ecosystem structure and function, using simulation models, network approaches and learning.

II.2. Consortium and Complementarity

The consortium comprises key units from INRA, AgroParisTech, CNRS, CIRAD and the Universities, which bring expertise in the five ecosystems that are used as test cases (Table 3) and in Machine Learning (Figure 3). This consortium has worked together intensively, being involved in common grants (see Table 2), producing papers[e.g. 4,9,13,17,23,42,44,51,56,70] and Thematic Issues of Advances in Ecological Research on Invasion Ecology[L1,L2], Ecosystem Services[L3,L4] and Agricultural Systems[L5]. As experts in the test case ecosystems, Partners 1-5 and 7 bring established hypotheses for the interaction structure and functions of their ecosystems (from specific interaction links through to already constructed networks), practical knowledge for how to sample the ecosystem and, in some cases, existing NGS data for the ecosystem. Partners 6 and 8 bring expertise in statistical and logic-based machine learning, respectively, which are the two main types of machine learning that might be used to reconstruct NGS networks. Partners 6 and 8 have already worked on problems of learning ecological networks of interaction. In particular P8 and P1 have worked together on large-scale ecological networks that have been tested and validated. P6 and P2 have worked together to produce ecological interaction networks from NGS data. These two research lines provide, in proof of concept, learning to build ecological networks from NGS data in each of the five ecosystems.

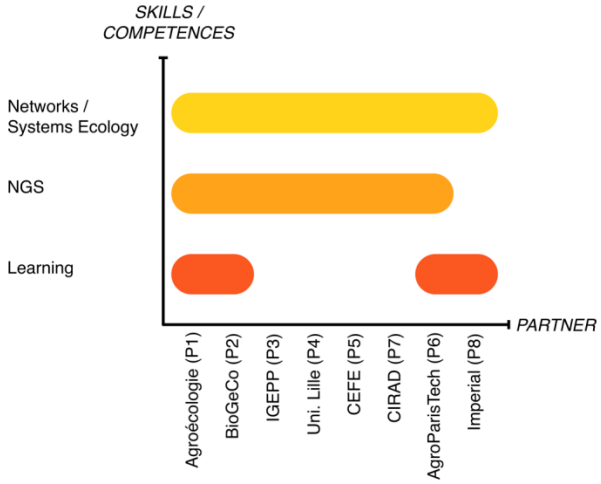

Figure 3. Schematic of the combinations of skills brought to the project by each partner.

Knowledge of NGS methodologies is widespread through the consortium. Each partner will be responsible for doing their own NGS sample processing. Due to the costs of NGS sequencing, however, we will mutualize the NGS sequencing for all ecosystems, at PGTB (Partner 2).

Partner 1 - INRA UMR 1347 Agroécologie: The INRA UMR Agroécologie is the lead partner of this consortium of researchers. The UMR Agroécologie, created in 2012, brings together ecologists, agronomists and geneticists around issues related to the design and evaluation of agricultural systems that meet agricultural productivity and environmental compliance criteria. The project members of this UMR are part of the Pôle GESTAD "Sustainable management of weeds". This Pôle aims to elucidate the links between agricultural systems, the functioning of communities and agro-ecosystem services delivered by these communities in order to design and evaluate sustainable production systems. Dr David A. Bohan will act as the project Coordinator. His role is to ensure that the project runs to schedule and produces high quality research.

Partner 2 - INRA/Université de Bordeaux UMR 1202 BioGeCo and associated sequencing service (PGTB): The UMR BIOGECO (Biodiversity, Genes & Communities) brings together plant geneticists, physiologists and

pathologists, entomologists and community ecologists with the aim of understanding the eco-evolutionary processes underlying biodiversity, across all levels of organization (from genes to communities) and to promote sustainable management. BIOGECO investigates the biodiversity-functioning relationships in various ecosystems and is internationally acknowledged for its expertise on forest biodiversity. BIOGECO's molecular biologists work predominantly at the PGTB, which is part of a federation of seven technology platforms at the Functional Genomic Center of Bordeaux. The PGTB specializes in next-generation sequencing solutions adapted to many types of research projects including: the development of molecular markers using ddRAD-Seq; targeted sequencing; metagenomics and metabarcoding; and, whole transcriptome and sequencing of small genomes. The PGTB team (seven scientists) also have skills in degraded DNA analysis and SNP genotyping using MassArray technology. C. Vacher is the leader of the "Functional Ecology & Genomics" team. She investigates the functional role of the plant microbiota by combining environmental genomics, plant physiology and network ecology. Her role in the project is to ensure that WP1 produces high-quality NGS data for network reconstruction. She will act as a link between the sequencing service at BIOGECO and the project partners.

Partner 3 - INRA UMR 1349 IGEPP: The INRA UMR IGEPP brings together ecologists, agronomists, geneticists and modellers working in the field of agricultural sciences on plant protection, genetics and environment. Its aims are focused on the development of new plant protection and cropping systems that are sustainable, respectful of the environment, and have low inputs. A key competency of IGEPP is in understanding agro-ecosystem ecological complexity which unites plant genetics and pathogens/pests ecology across scales of biological organisation from the gene to the agroecosystem. Elsa Canard is a permanent INRA researcher at IGEPP, with skills in community ecology and a particular focus on the functioning of ecological interaction networks. Elsa will coordinate the work on the agro-ecosystem. Manuel Plantegenest is a professor at Rennes Agrocampus Ouest working at IGEPP, with expertise in community and landscape ecology of pests and their natural enemies.

Partner 4 - Université de Lille 1 UMR 8198 EEP:
The Evolution, Ecology, Palaeontology (EEP) unit at the University of Lille is a new laboratory (founded in 2015) comprising ecologists, geneticists, physiologists and palaeontologists. The main research axes of EEP are: (i) genomic and (ii) evolutionary ecology aspects of mating systems, especially in plants; (iii) adaptation of ecological communities to environmental changes; (iv) species interactions and comparative immunology; (v) macro-evolutionary patterns from fossil records; (vi) paleo- and macro-ecology of aquatic communities. François Massol is a permanent CNRS researcher at EEP, specialized in evolutionary and spatial ecology. His research follows two main axes: (i) the mechanisms affecting biodiversity in spatially structured systems and (ii) the evolutionary dynamics of traits and their ecological impacts, especially the evolution of dispersal and the coevolution of interaction networks. François will lead WP4 and coordinate the plant-pollinator system.

Partner 5 - CNRS CEFE UMR 5175: With more than 150 permanent researchers, the Centre for Evolutionary and Functional Ecology (CEFE) is the largest centre for ecological research in France and a very important one in Europe. The CEFE gathers researchers from different funding institutions (CNRS, INRA, IRD, EPHE), as well Montpellier SupAgro International Center for Higher Education in Agricultural Sciences and the two universities of Montpellier, and is very active in teaching ecology and evolutionary biology. The CEFE is also embedded within national initiatives of excellence such as the excellence laboratory (LABEX) "Mediterranean Center for Environment and Biodiversity" (CEMEB) and the national Human-Milieu observatory on the Mediterranean littoral. The main research themes developed at the CEFE include (i) human impacts and conservation biology, (ii) the evolution of life-history traits in a changing world, (iii) the role of biodiversity in functioning, and (iv) the impacts of global changes on ecosystem functioning. Patrice David is an evolutionary biologist, studying mating systems and community eco-evolutionary dynamics. He will lead WP3.

Partner 6 - AgroParisTech UMR 518 MIA: The UMR 518 MIA comprises two groups that develop statistics for life sciences. The 'Statistics and Genome' group is specialized in the analysis of genomic data, particularly from a network perspective. Collaborations already exist between P6, P2 and P4. The 'Modelling and Risks in Environmental Statistics (MORSE)' group is specialized in Ecology. Both groups have a considerable expertise in statistical network modelling. Members of both groups will participate in the project, usefully bringing their statistical expertise at biological scales from the genome to ecosystem. Stéphane Robin is a statistician

with great experience in the development of statistical methodologies for the life sciences. He has published over 60 papers in international peer-reviewed journals and has been PI on ten projects, including the 'Projet Investissement d'Avenir' on Algorithmics, Bioinformatics and Statistics for NGS (ABS4NGS) project. Stephane will lead WP2 and coordinate the statistically based network reconstruction.

Partner 7 - CIRAD UMR C53 PVBMT: Partner 7 is composed of researchers from two research units (BGPI in Montpellier and PVBMT in La Réunion) with a longstanding collaboration in the study of Mediterranean and tropical crop pests. These groups are specialized in the study of both phytophagous insects and viral and bacterial plant pathogens, using population genetics, molecular epidemiology and more recently metagenomics approaches. PVBMT is the leader laboratory on crop protection and biodiversity in La Réunion and South West Indian Ocean, and is, as such, involved in important European projects aiming to document and monitor regional biodiversity. Virginie Ravigné, who will lead this partner grouping and WP4, is a CIRAD researcher based at PVBMT. Virginie is a theoretical evolutionary biologist, and her research focuses on understanding the eco-evolutionary dynamics of plant pests by accounting for complex life-history strategies. She is the author of 42 peer-reviewed publications, supervisor of 11 MSc and 4 PhD students and work package leader in Flagship Fondation Agropolis programmes on biological invasions and disease emergence since 2010.

Parner 8 - Imperial College, London, Dept of Computing: The Department of Computing at Imperial College is one of the largest in the UK. It has particular expertise in Logic and Artificial Intelligence research, encompassing foundational studies in Logic and a variety of Artificial Intelligence disciplines. Work within the Logic and Artificial Intelligence Theme on machine learning research in bioinformatics has led to pioneering work on network analysis, modelling and alignment. Alireza Tamaddoni-Nezhad is Research Fellow at the Department of Computing and the Coordinator of Syngenta University Innovation Centre (UIC) at Imperial College London. His work lies in the areas of Machine Learning and Data Science, and in particular Symbolic and Logic-based Machine Learning (e.g. Inductive Logic Programming) and applications to Knowledge Discovery in Life Sciences. Alireza will contribute to the logic-based learning in WP2 and co-supervision of the PhD with Partner 1.

Table 2. Collaborative grants received by the consortium that contribute to the NGB project

Title of the call for proposals, source of funding	Project title	Coordinator	Starting/End date	Grant amount	Part.	Name	Person. Month
					Of the person involved in this proposal		
FACCE SURPLUS	PREAR	D.A. Bohan	04/16-03/19	716k€	1	D.A. Bohan	12
FACCE ERA-NET C-IPM	BioAWARE	D.A. Bohan	06/17-05/20	857k€	1	D.A. Bohan	12
ANR Agrobiosphere	PEERLESS	P. Franck	01/13-12/17	807k€	1,4	D.A. Bohan, M. Plantegenest	9
ANR Agrobiosphere	AgroBioSE	V. Bretagnolle	01/14-12/17	745k€	2	D.A. Bohan	8
CESAB/FRB/TOTAL	COREIDS	P. David, F. Massol	10/14-06/17	189k€	5, 4, 1	P. David, F. Massol, D.A. Bohan	1
INRA MEM	Learn-Biocontrol	C. Vacher	07/16 - 06/18	146k€	2, 1, 6, 8	C. Vacher, D.A. Bohan, S. Robin, A. Tamaddoni-Nezhad, C. Pauvert	
INRA MEM	Brassica-Div-Patho	C. Mougel	07/16 - 12/17	50k€	2, 6	C. Vacher, S. Robin, C. Pauvert	
Région Aquitaine	Athéné	C. Vacher	09/16 - 08/20	202k€	2	C. Vacher, S. Delzon, R. Burlett, R. Ségura	14, 5, 5, 5
LABEX CEBA	Drought	D. Bonal	09/16 - 08/19	268k€	2	C. Vacher, S. Delzon, R. Burlett, C. Lalane	3, 3, 2, 1
LABEX COTE	MicroMic	C. Vacher	03/17 - 09/19	181k€	2	C. Vacher, R. Burlett, C. Lalane, R. Ségura	10, 3, 1, 3
INRA SPE	MIRAAC	E. Canard	01/17 – 12/18	21k€	3	E. Canard, M. Plantegest	
ANR PRC	ARSENIC	F. Massol, N. Loeuille	11/14 - 11/18	499k€	4, 5	F. Massol, A. Duputié, N. Hautekèete, Y.Piquot, B. Schatz	24, 12, 6, 6,15

Region Hauts-de-France	AREOLAIRE	A. Duputié	10/15-11/18	150k€	4	A. Duputié, F. Massol, N. Hautekèete, Y.Piquot	18, 9, 8,8
CPER Hauts-de-France	CLIMIBIO	X. Vekemans, Y. Piquot	01/15-01/20	24k€	4	Y. Piquot, A. Duputié, S. Gallina, N. Hautekèete, F. Massol, C. Poux	3
ANR Bioadapt	AFFAIRS	P. David	10/12 – 03/17	350k€	5, 4	P. David, P. Massol, P. Jarne	30, 14, 21
ANR Société de l'information et de la communication	Hydrogene	P. Peterlongo	2015-19	400k€	6	S. Robin, J. Chiquet, S. Ouadah, J. Aubert	22
Proj. Inv. Avenir ANR-10-BTBR-01	Amaizing	A. Charcosset	2011-19	10000k€	6	S. Robin	8.4
Fondation Agropolis Flagship projects	E-SPACE	C. Neema	12/15-12/19	800k€	7	V. Ravigné, P. Roumagnac, C. Vernière	4, 6, 6
ErAfrica	FruitFly	PF. Duyck	03/15-03/18	182k€	7	PF. Duyck	6
Ecophyto	AttractMyFly	L. Costet	03/15-03/18	100k€	7	PF. Duyck	6
ERA-NET Biodiversa	EXOTIC	B. Facon	01/14-01/19	765k€	7	B. Facon	12
PHC Protea, France/South Africa	PATHOPAST	P. Roumagnac	01/17-01/18	20k€	7	P. Roumagnac	2
FP7 EU Marie-Curie Fellowship	GEOMETAGENOMICS	P. Roumagnac	07/14-07/17	300k€	7	P. Roumagnac	24

II.3. Methodology and Workpackages

The research in each of the five ecosystems will proceed in parallel, as a workflow, which is reflected in the project Work packages (Figure 4). The workflow in each ecosystem progresses in a sequence of logical steps, from NGS sampling of the ecosystem via network learning and validation to the detection of change in network structure and the development of protocols that specify how this NGB approach might be used practically as a biomonitoring protocol.

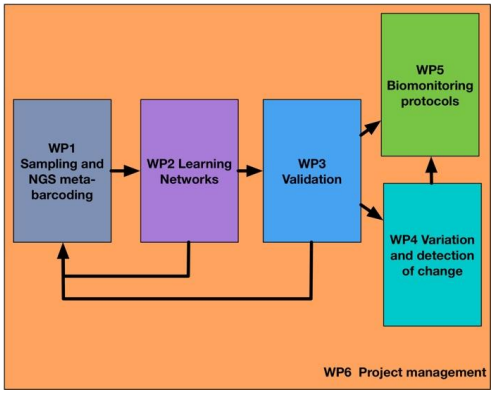

Figure 4. Schematic diagram of the project work packages.

WP1 Sampling and NGS meta-barcoding

Leader: Corinne Vacher *Start-End months*: 1-24 *Total person-months*: 161

Objective:

- To produce replicated NGS data for each system that can be used to learn network structure

Description of work: Environmental samples will be collected in the five ecosystems (Table 3) and DNA will be extracted by each Partner. DNA samples will be sent to the sequencing service (PGTB), at Partner 2. A molecular biologist will be employed by partner 2 to develop, where necessary, the molecular methods required to amplify the barcode regions and functional genes for all five biological systems. In interaction with all partners, the technician will then amplify the chosen regions. Most of the sequencing will be performed on a MiSeq platform. Bioinformatics analyses will be performed using a combination of existing pipelines (e.g. QIIME, Mothur, Usearch, Swarm, Frogs).

Approach: The current state of knowledge for each ecosystem is summarised in Table 3. The sampling design, molecular methods and bioinformatics analyses are all system-specific and will be done as a series of tasks as detailed in Figure 5. For systems 2, 3 and 5, where fully developed NGS approaches do not yet exist, a preliminary task (T1.1) will be to collect test samples to develop the NGS approach at PGTB (T1.3). It should be noted that this development will use, as a basis, existing NGS approaches from similar systems. For all systems, statistically valid sampling protocols will be produced (T1.2), based upon the outlines in Table 3. The sampling will then be conducted in each ecosystem and the DNA extracted (T1.4). The NGS sequencing will then be done at PGTB (T1.5), before being analysed using existing bioinformatics pipelines against OTU

databases, to produce OTU tables that can be passed to WP2 for the network reconstruction phase (T1.6). Bioinformatics and molecular biological methods will be employed to troubleshoot primer biases, sequencing errors, noise and statistical artefacts of the data (e.g. zero-rich data, asymmetric distributions).

Tasks:

- T1.1 To collect test samples for the development of molecular methods (month 0 to 6)
- T1.2 To elaborate sampling designs in collaboration with statisticians (month 0 to 6)
- T1.3 To develop molecular methods at PGTB (month 6 to 18)
- T1.4 To collect environmental DNA samples for learning ecological networks (months 6 to 18)
- T1.5 To sequence the DNA samples at PGTB (months 18 to 22)
- T1.6 Bioinformatics analyses of the sequence data (months 18 to 24)

Deliverables:

- D1.1 Metabarcoding protocols for microbial and macrobial communities (month 18)
- D1.2 OTU tables for learning ecological networks (month 24)

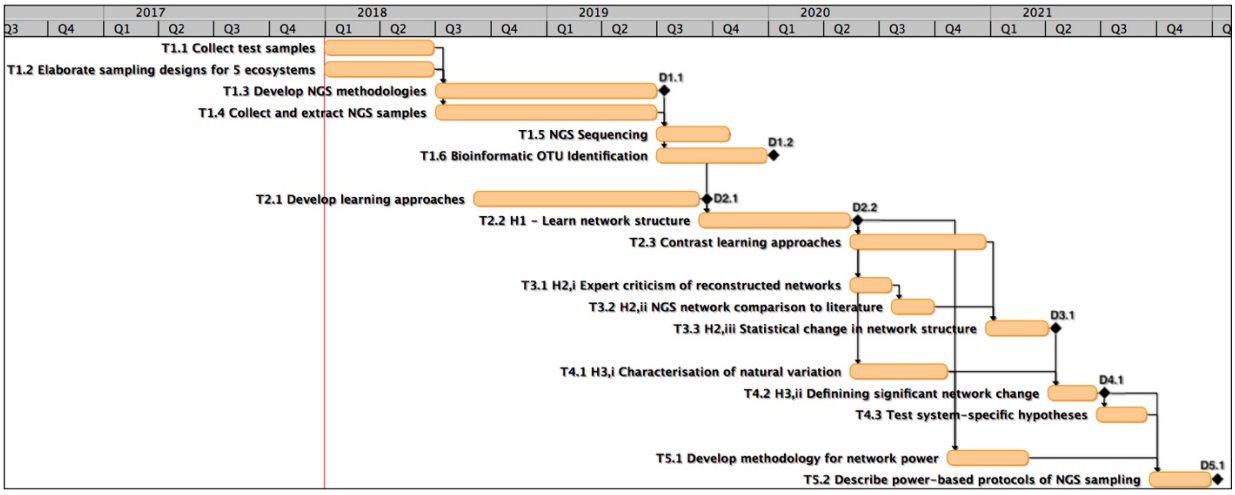

Figure 5. Gantt chart for the time relationships between tasks and deliverables.

WP2 Learning Networks

Leader: Stéphane Robin *Start-End months*: 8-36 *Total person-months*: 134

Objectives:

- To develop machine learning approaches for reconstructing networks from NGS data
- Test hypothesis H_1 "machine learning approaches can be used to reconstruct hypothetical networks of ecological interaction from NGS data"
- Compare and contrast the statistical and logical learning approaches

Description of work: The learning will be done predominantly by two Ph.D. students based at P6, for the statistical learning, and at P1 (co-supervision with P8), for the logic-based learning. The learning will commence using ecological data-sets already available to the project. Initially, the aim will be to develop the methodologies so that these can be applied to the OTU data tables obtained for each system in WP1. The learning will then be used to test whether H_1 is satisfied. Finally, the relative value and quality of the different methodologies, for use in an NGB approach, will be compared.

Approach: Statistical and logical inference techniques will be used to reconstruct the system networks, based on the methods previously used for ecological networks, but with appropriate development to NGS data. For T2.1, statistical network reconstruction methods exist, predominantly in a Gaussian setting that is not appropriate to NGS data. We will, therefore, consider more general approaches, initially examining tree-based methods[13,37,38] that have already demonstrated their efficiency. These approaches will need to be generalized to account for heterogeneous conditions across the 5 ecosystems. The tree-based methods will also be combined with recently developed latent Gaussian models[39]. For the logic-based learning, Inductive

Logic Programming (ILP), implemented in the Progrol 5.0 language[7], will be used. For known OTUs in each system, we will guide network reconstruction using background information for particular interaction types, such as the trophic background information used for learning an agro-ecological network[7,21]. We will also implement Meta-interpretative Learning (MIL[40,41]) in order to infer generic background information (interaction rules) directly from the NGS tables that will link all OTUs, both known and unknown. The learning procedures will be coded as procedures in *R* or as standalone software. These will then be used to learn networks (T2.2) and test H_1, in a collaboration between the two PhD students and the other partners. Finally, in T2.3, the relative performance of the logic-based and statistical learning will be benchmarked for speed and quality of reconstruction of networks from NGS data. An assessment of the satisfaction of the project partners for the networks produced will also be done in collaboration with WP3.

Tasks:

- T2.1 Develop statistical and logical machine learning approaches using existing NGS data (month 8 to 20)
- T2.2 Testing H_1 - Learn network structure (month 20 to 29)
- T2.3 Compare and contrast learning methodologies for NGB approaches (month 29 to 36)

Deliverables:

- D2.1 Validated learning methods for reconstructing networks from NGS data (month 20)
- D2.2 Hypothesised NGS ecological networks for each ecosystem (month 29)

WP3 Validation

Leader: Patrice David *Start-End months*: 29-39 *Total person-months*: 86

Objectives:

- Test hypothesis H_2 "the NGS networks reconstructed for each ecosystem are similar to the already-known ecological networks for the system"
  - $H_{2,i}$ - the OTUs and hypothesised links identified in the network reconstruction perform as we expect from network theory
  - $H_{2,ii}$ - the reconstructed networks are valid, when compared to text mining results from the literature
  - $H_{2,iii}$ - the NGS networks perform similarly to their classic counterparts

Description of work: The validation work package seeks to test and validate all the NGS reconstructed networks using a general methodology. The approach progresses in a series of step, testing sub-hypotheses of H_2. These steps are based upon an existing approaches used by P1 and P8 to validate an agro-ecological network reconstructed from ecological abundance data[7,21].

Approach: Once reconstructed, in WP2, the networks will be validated in order to test the overarching test hypothesis, H_2. This validation will follow the methodology developed by Partners 1 and 8[7,21]. We will ask: i) do the reconstructed networks behave as we would expect from network theory[16]; ii) are the networks valid when compared to text mining results from the literature[17]; and, iii) do the NGS networks perform similarly to the classically constructed ecological networks that already exist? In T3.1, to test $H_{2,i}$, we will calculate standard metrics of network structure[18] and test these values against published networks of similar type and from similar ecosystems. Criticism of the networks by consortium partners will also further reveal whether 'expert-scientists' are satisfied with the networks produced. Where appropriate, we will produce literature networks for each ecosystem using literature resources (i.e. Science Direct or Google Scholar)

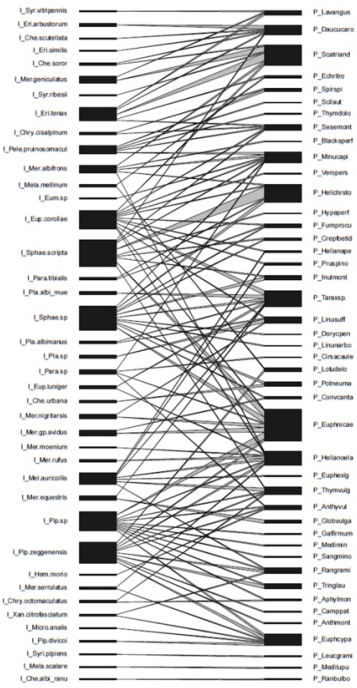

Figure 6. Quantitative interaction network between syrphid pollinators (left) and flowering plants (right).

as sources of data (T3.2). Standard methodologies exist and correlational bibliometric approaches will be used to compare the NGS and literature networks to test $H_{2,ii}$[7,42]. Finally, $H_{2,iii}$ will be tested in T3.3 by comparing NGS networks against existing expectations, under change, for links either between known pairs of OTUs in each system or, where they exist, whole networks (Figure 6). We would note that we expect this methodology to undergo some development during the project, but in principle it should allow testing of NGS network structure whether they contain known or unknown OTUs.

Tasks:

T3.1 $H_{2,i}$ - Do ecosystem experts understand the reconstructed networks? (month 29 to 31)
T3.2 $H_{2,ii}$ - Are NGS reconstructed networks comparable to literature networks? (month 31 to 33)
T3.3 $H_{2,iii}$ – Is change in reconstructed and classic network structure similar? (month 36 to 39)

Deliverables

D3.1 Validated NGS networks for all ecosystems (month 39)

WP4 Variation and detection of change

Leader: François Massol *Start-End months*: 29-45 *Total person-months*: 75

Objectives:

- To test hypothesis H_3 that we can "develop methods of characterising variation in network structure to determine natural variation and detecting changes consistent with global drivers of ecosystem change"
  - $H_{3,i}$ - *we can characterise natural network variation*
  - $H_{3,ii}$ - *we can define change in network structure attributable to global drivers*

Description of work: WP4 will work in close relation with WP5 to develop measures of network variation that can describe the natural change in network structure that occurs between replicates of the same ecosystem. It will define criteria for network change that, if superior to this natural background, represents a significant change in network structure. This will afford a test of H_3 and the system-specific hypotheses in Table 3.

Approach: The variability between replicate NGS networks will be used to describe and quantify the natural variation in network structure. We expect that across space and over time network structure will change due to different environmental conditions. Ecologically, such variation is extremely important as it describes the normal operating bounds of the ecosystem. In the presence of a driver of ecosystem change, we would expect the variation in structure to be greater than this natural background. To characterise network variation, T4.1 will use a combination of approaches from network science (engineering) and statistics. Engineering approaches to network variation include measure of network elasticity, typically applied to problems of telecommunications, which may have some utility for ecological networks. Statistical approaches will include evaluating variation in univariate network metrics (e.g. modularity; development of the methods in T3.1[18]) or Bayesian inference. $H_{3,ii}$ will be tested in T4.2 by establishing criteria for network change that is superior to the natural variation. Bayesian approaches will treat the dependency structure of the networks (e.g. temporal sampling structure in system 1 or the herbicide treatment in system 5) as part of the latent structure, in a Gaussian setting, and search for breakpoints in structure across the replicates. Finally, the methods and results of T4.2 will be used in T4.3 to examine the structural changes in NGS networks and test the system-specific hypotheses in Table 3.

Tasks:

T4.1 $H_{3,i}$ - Characterisation of natural variation in NGS network structure and function (month 29 to 34)
T4.2 $H_{3,ii}$ - Defining change in network structure attributable to global drivers (month 39 to 42)
T4.3 Test system-specific hypotheses (month 42 to 45)

Deliverables

D4.1 Definition of change in NGB network structure and function (month 42)

WP5 Biomonitoring protocols

Leader: Virginie Ravigné *Start-End months*: 34-48 *Total person-months*: 54

Objectives:

- To produce guidelines for building an NGB biomonitoring approach in a given ecosystem, that details NGS sampling and sequencing methodologies, network construction and validation and the characterization of variation and change

Description of work: WP5 collates information from all other WPs to deliver potential NGB protocols for any given ecosystem. It will detail the thought process and methodologies to consider in traversing the workflow that the NGB project follows. In addition, WP5 will take the findings from WP4 on natural variation to develop a methodology for calculating the power of any NGB sampling.

Approach: The aim of WP5 is to deliver an assessment of the power of any given NGB protocol, based upon the ecosystem of study. These protocols of NGB biomonitoring power will use knowledge gathered in the other WPs directly, but in particular from WP4. Univariate power analysis, to detect a given ecological effect and to estimate the appropriate number of samples, is well established in biomonitoring [49]. Power analysis for detecting change in network structure is much less well established, but will be necessary for NGB approaches to determine the statistics of the size of the array of samplers necessary to detect network change in real time. Using the natural variation in network structure[18], calibrated in WP4, WP5 will seek to establish rules for calculating NGB sampling protocols of appropriate power. Depending on the findings of WP4, with respect to measures of natural variation and change, our approach in T5.1 will initially use a combination of univariate network metrics, to measure network structure and variation (from WP3 and 4), and classic univariate power analyses. For T5.2, we will produce a series of protocols for using NGB. These will consider biomonitoring scenarios that state explicitly the set of expected taxa, in a particular ecosystem subject to a given driver of change, and that build on the experience gathered throughout the project.

Tasks:

T5.1 Develop methodology for calculating NGS sampling power (month 34 to 39)
T5.2 Describe power-based protocols for NGS sampling and NGB biomonitoring (45 to 48)

Deliverable:

D5.1 Consolidated protocols for conducting NGB biomonitoring (month 48)

WP6 Project management

Leader: David Bohan *Start-End months*: 1-48 *Total person-months*: 26.8

Objectives:

- To assure the smooth running of the project and the quality of the project outputs

Description of work: The NGB will be run by a management committee made up of all project partners, led by P1. This committee will meet every six months at one of the partner sites to assess progress against project deliverables and take necessary remedial action to assure data quality and the smooth running of the project. Scientific meetings will run every six months, with frequent, topic-led meetings as required.

II.4.Justification of resources

To achieve the work in the NGB project within a 790k€ budget has required considerable care. All field work and sampling involved in the project will piggy-back on the research ongoing within each partner group. These groups understand their systems well and their requests for additional manpower resources, such as technicians, reflect the extra sampling and DNA extraction work that will need to be done. Partners 1, 3, 4 and 6 make a request for a half PhD studentship each. The PhD students at partner 1 and 6 will conduct the supervised development of the logic-based and statistical machine learning, respectively. The PhD students at partners 3 and 4 will perform the sampling and analysis of the pollinator and agro-ecosystem, respectively. Half PhD studentships can be looked upon as potential risks to a project. However, we are confident that the other half of the studentship will readily be found by each partner and already have potential 'in principal'

interest from commercial collaborators should the project be funded. To save on the costs of NGS sequencing, we propose to mutualise the sequencing at the PGTB. Partner 2 will hire a post-doctoral researcher to perform the sampling and analysis of foliar microbial networks, in collaboration with the molecular biologist who will be hired at PGTB for the period of development and sequencing. These PhD and post-doctoral appointments will produce staff, with unique skill-sets, that we believe will be in high demand in the future. We have budgeted 25k€ for each ecosystem that we will sequence, and request funding for this. Additional molecular costs of DNA extraction were calculated at 2.50€ per sample, and are included in each partner budget. Cost of travel and subsistence are requested, for all partners, given the collaborative and global scale of the ecosystems in this project. Finally, partner 1 requests 10k€ for organising a project colloquium in Paris at the end of the project. As is noted in our dissemination plan, this colloquium will deliver project results to stakeholders, particularly those from science, policy and industry backgrounds. Other less costly budget items included by each partners include publication fees in open-access journals, Master student internships, and general operating costs (participation in international conferences, consumables, etc.).

III. Impact and benefits of the project

III.1. Expected Impact

Validated methods for rapidly reconstructing networks: The NGB project is designed to test whether a fusion of NGS data and machine-learning approaches can be used to reconstruct networks of ecological interactions more rapidly than is currently the case. The fundamental principle of machine learning is that it can recover the interactions that led to the structure in the data-set from observed variation and co-variation of variables. In using NGS data, sampled from five ecosystems across the range of system types that we might sample and monitor, the NGB project will demonstrate whether NGS data can be treated as a source of ecological network information. The project will also examine the quality and rapidity of network reconstruction provided by statistical and logic-based machine learning, and compare these two markedly different approaches. As with any model-fitting approach, the quality of the learning approach will be determined by its ability to reproduce expected interactions, ecological structures in the literature and networks already constructed. Our aim is to produce a methodology that could reconstruct networks rapidly, however, and so criterion such as the speed of learning will also determine the choice of the methodologies we will ultimately deliver. The expected short-term impact of this will be a set of validated methodologies that will allow ecologists and environmental scientists to reconstruct, both much more rapidly and at greater replication than current approaches, the ecological networks that structure and drive the functioning of ecosystems.

Methods for monitoring ecosystem change and learning new ecological science: Our current theory of ecosystem structure and function is largely based on a growing, but still quite limited, number of ecological networks, many of which have low numbers of replicates. These networks have permitted tests of fundamental theories, such as allometric diet breadth and body size or trophic cascades[21,22,43]. An increase in the rate of learning of replicated networks will therefore greatly advance our understanding of the generality of current theory. It will also drive new science. As shown by the agro-ecological example, learning methods can generate new scientific understanding, allowing the testing of identified links or nodes far more efficiently than current methods. The more networks that are learnt, the better and more general our theory for network structure and function will become. The expected short-term impact of these methods will be to revolutionise our understanding of ecosystem change.

Scientific roadmap for developing NGB approaches: The NGB work packages are structured around an idealised workflow for the learning and reconstruction of a network from NGS data. At its core, therefore, the project is oriented towards delivering methods and providing a roadmap that other scientists can use to create the NGB approach for their own system. The development of a practical NGB approach is not something that can be achieved by any one group of scientists, such as the NGB consortium, alone. While the project will have impact in delivering methods for sampling, reconstructing and validating networks from NGS data, and techniques for detecting change and predicting power, our impact will also be through advocating the importance of the NGB approach and detailing the research and development to be done. The NGB project will directly address problems of variable quality, sampling problems (taxa biases),

identification errors, zero-rich data and asymmetric abundance distributions in NGS samples, but these problems will require a concerted effort on behalf of all interested scientists to solve. In the case of the variable quality of the NGS databases, we expect that an NGB approach would have great medium-term impact in providing the "big picture" impetus for generating shared sequence databases that would transform NGS, but which, because of time and cost constraints, are still lacking.

New technological development and enterprise: In the recently published TREE opinion piece[4], we outline the technological developments that would be necessary to use an NGB approach in practice. We describe a global-scale sampling approach, using automated samplers, communication and in-cloud processing and storage. While we note that much of this technology already exists in a highly developed form - the smartphones in our pockets already contain processing power, storage and communications that permits interaction with, and storage of, data in the cloud - a considerable amount of technological research and development will be necessary to build an automated sampler and the infrastructure necessary to support a global array. We note in the paper that on the molecular biology side, there are already NGS sequencers that are the size of a USB key, but that sample processing equipment of the correct scale and sampling mechanisms to sample for eDNA in all biomes and environment need further development. The long-term impact of NGB would therefore be to develop a huge business opportunity, driving forward enterprise. Two opportunities, in particular, stand out. A global NGB array would become itself an industry requiring construction, servicing and resupply, potentially changing the business models of companies that supply the molecular reagents, for example. It would also lead to the development of new business supporting improved decision-making. We detail in the TREE paper an example of NGS sampling in West Africa for Ebola, and suggest that NGB could be used to monitor many diseases, improving epidemic detection and management (risk management) and potentially saving lives. We also describe hand-held samplers that might be used by farmers to detect much more sensitively invasive pests or diseases, leading to better pesticide management by the farmer himself and, by collating and distributing this data/information at larger scales, much improved environmental performance that would contribute to the green economy envisioned by the EU.

Better monitoring and prediction of ecosystem change: The objective of this project is to develop and test a generic NGB approach that will detect ecosystem-wide change more rapidly, sensitively and cheaply than current biomonitoring. As noted above, this will have great short-term impact on ecological science and in biomonitoring. The NGB approach will ultimately have most impact when used alongside existing methods of large-scale monitoring. The closest comparable example is remote sensing. Much as with remote sensing, NGB will use of large amounts of storage and processing in the cloud if we are to use it to reconstruct networks. One of the solutions to the costs of developing the NGB infrastructure would be to piggy-back on the existing structure used for large-scale remote sensing. The long-term impact of bringing these two approaches together will be to revolutionise environmental monitoring. For example, remote sensing of environmental change due to global drivers could routinely be coupled to changes in ecological function measured by NGB. Indeed, changes detected by NGB might serve as an early warning system for environmental change that, due to inertia and time lags in the ecosystem, may take time to appear in remote sensing data.

III.2. Relevance to the ANR 2017 Work Programme challenge

The NGB project addresses *Defi 1* "Gestion sobre des ressources et adaptation au changement climatique : vers une compréhension du changement global", *Orientation 3* "Évaluation et maîtrise du risque climatique et environnemental" and the Axe scientifique "Dynamique des écosystèmes en vue de leur gestion durable" of the ANR 2017 Work Programme challenge. Here we detail the relevance of the NGB project under key phrases taken from the programme.

Reinforce scientific collaboration at the national and international level: The NGB project uses a parallel workflow that fosters collaboration within the project. Specifically, this workflow will require all project partners to work with all other project partners to achieve the project goals and deliver shared publications. This promotes linkage, which we have detailed in Section I, to other National projects ongoing across the consortium. The consortium partners are also actively involved in past and ongoing International projects that will both feed into and use the results from the NGB project. There is, therefore, a ready-made framework of national and international collaborations into which the NGB project will fit. The NGB project

will be extremely interesting to a wide variety of researchers and organisations, at the national and international level. Through the recent TREE Opinion paper[4], which details the benefits of NGB and how it might be used, we have begun to interact with international researchers who wish to collaborate with this consortium and have had contact with IPBES and FutureEarth, via existing collaborators such as Guy Woodward (co-author of the TREE paper[4]). This collaboration and impact will continue to grow.

Economic impact and competitiveness: The development of an NGB approach and its use under real-world conditions will necessitate the development of an industry. The automated samplers will require technological development by equipment manufacturers working in molecular biology, robotics, computing and communications. The industry associated with in-cloud computing will need to be mobilised to develop the computational infrastructure, databases and learning algorithms necessary for reconstructing ecological networks from NGS data. An industry for servicing the automated samplers will also be needed, and indeed it is possible to imagine that this 'market' could become a significant source of revenue for reagent manufacturers. Finally, there are economic and job opportunities in the use of the data itself. Businesses may emerge to help farmers make decisions based upon the pests and diseases that are present in fields. This would minimise unwanted environmental impact, reduce risk and potentially increase on-farm productivity, thus contributing to the green economy. Disease monitoring could improve epidemic detection and management, potentially saving many lives and reducing some of the enormous economic loss. However, it should be stated that these potential economic impacts and competitiveness benefits are contingent on the success of the NGB project.

Fundamental knowledge of processes of ecosystem change: The NGB approach will revolutionise our understanding and knowledge of the processes of ecosystem change. Networks are increasingly being used to evaluate and understand ecosystem structure and function, but are difficult and expensive to construct. Consequently, relatively few networks have been constructed and these tend to have low numbers of replicates, limiting our ability to make general statements about change in ecosystems. A move to a combination of NGS, as a source of ecological data, and machine learning approaches for network reconstruction will vastly increase the rate of construction of ecological networks. This will allow us both to test the generality of ecological theories for network and ecosystem structure, such as allometric diet breadth and body size[21,22,43], and to develop new theories of ecosystem structure via the learning of background data. The NGB approach will also allow rapid, real-time detection of a change due to a global driver. The effects of this driver could be followed as they ripple-out across the network. Such new knowledge would allow the building of ever better understanding of the effects of drivers and will greatly advance our understanding of the link between ecosystem (network) structure and function.

Consequences of global change on ecosystems, services and society: Network approaches have tangible benefits for understanding and managing ecosystems and services that cannot be achieved using univariate biomonitoring indicators, for example. However, networks currently do not have a significant footprint in policy, in large part because of their costs of construction. Ultimately, the NGB approach could remove many of the limitations to the use of network approaches in the management of ecosystems by reconstructing networks quickly and cost-effectively. Recently, two TREE papers have discussed network approaches for the management of ecosystems, service and society[44,45], one of which was written by Dave Bohan[44] and several members of the NGB consortium. These papers make the case that the societal needs, the economics and the ecology of ecosystem management can be viewed as a set of social, economic and ecological networks. Moreover, as these networks can be analysed using the same methodologies, network ecosystem management gives a consolidated framework and much richer understanding of the interplay between societal needs, ecosystems and the services they deliver when subject to global change. NGB would supply the ecological networks for such a framework.

Understanding management of ecosystems and decision-making: The network frameworks developed by the QUINTESSENCE consortium[44] and Dee et al.[45] would allow the ecosystem and service consequences of social, economic and ecological change, driven by global change, to be examined. This will lead to improvement in our understanding of the interplay between these markedly different disciplines and how they contribute to ecosystem dynamics. Moreover, the framework would in turn allow the development of management goals and decision-making for ecosystems that meet social, economic and ecological (environmental) criteria, such as the goals of national and international policy. As noted above, NGB would supply the ecological networks for such a framework.

III.3. Dissemination and exploitation

Plan for dissemination and exploitation: The objective of the NGB project is to develop and test a generic next-generation biomonitoring (NGB) approach that will detect ecosystem-wide change more rapidly, sensitively and cheaply than current biomonitoring. Our strategy is to transmit the findings associated with this objective to all appropriate target groups at two distinct levels of dissemination. At the first level, we will seek to promote and publish on all the five ecosystems in joint papers, with the goal of showing that NGB is a generic methodology. We will also develop dissemination and exploitation approaches that present both statistical and logic-based machine learning approaches. The aim, in doing this, is to market the NGB approach as a generic concept that works in many systems. Second, and in parallel, the project will also publish and present to individual system audiences the findings in each ecosystem. This will involve publications in discipline-specific journals, with the goal of demonstrating to domain experts that the NGB approach meets domain-specific standards of quality. This combination of generic and specific levels of dissemination will maximise the exploitation potential of NGB and meets high scientific quality standards.

Target groups and messages: Our target groups for dissemination are domain-expert scientists, working in all the scientific aspects touched on by NGB and whom might use the project results and principles. This might also include companies, such as the water companies that are charged with assuring the 'good ecological status' of water under the European Water Framework Directive. There is also a higher order grouping of stakeholders with interests in NGB results. These include the policymakers, NGOs and institutions that legislate for, manage or use biomonitoring data. This group includes organisations such as National Governments (e.g. Ministère de l'Environnement, de l'Énergie et de la Mer and Ministère de l'agriculture, de l'agroalimentaire et de la forêt), IPBES and FutureEarth. Finally, there are the companies that might be interested in developing the technologies that we foresee being used in NGB, including molecular biology equipment suppliers, communications and computing hardware companies and software companies.

Communication activities and channels: The project will disseminate project results through a number of activities and channels. Primarily, our publicity will be through the production of scientific outputs and publications in high-impact journals. This consortium has an extremely strong track record of publishing in the highest impact journals. We expect this to continue, with generic level results being published in journals such as Current Biology, TREE and PNAS, and, where the results merit it, potentially in Nature and Science. Specific level results will be presented in domain journals such as Journal of Applied Ecology and Global Change Biology. We will also secure at least one Thematic Issue of Advances in Ecological Research during the project. Dissemination will also include specific demonstrations to key users of the approach, such as to water companies. In the final three months of the project, we will hold a Colloquium in Paris that will have the aim of presenting directly to user the benefits and opportunities of the method as well as future challenges. Presentations at conferences, such as those run by IPBES, will promote exploitation of the NGB project results amongst policymakers and NGOs. We are also actively involved in promoting the idea of global scale biomonitoring to the public, by doing interviews (Wired magazine) and presentations (potential TED talk) following on from the TREE paper. Finally, the project will maintain a website with appropriate areas dedicated to types of end-user; including scientists, biomonitoring companies, policymakers/NGOs and technology companies.

Management of results: The results of the NGB project will include the NGB protocols, NGS data, machine learning methods and software, change detection and power analytic methods and ecological networks. Statistical developments will be implemented and disseminated as freely available *R* packages (cran.r-project.org). The NGB protocols will be published on the project website and published in peer review journals, as part of our dissemination plan. To assure applicability and reuse of our data and methods (for example by FutureEarth), we will assure that our data will be compliant with GEOSS, Global Biodiversity Information Facility, CICES v4.3 and Copernicus standards. Adherence will ensure interoperability and re-use of existing data where appropriate. The NGS data generated by NGB will be included in the databases of the Barcode of Life Initiative, including the German barcode of life, International barcode of life and public repositories (GenBank, EMBL and DDBJ). The data will be made available within 6 months of the project end, once the major project publications are submitted.

Table 3. Ecosystem descriptions, proposed sampling designs and the current state of knowledge

System 1	***Response of plant-associated microbial networks to intensive farming and drought events (Partners 2 & 7)***
Objective and System hypothesis	To reconstruct foliar microbial networks of positive and negative associations between OTUs in rice, grapevine and Holm oak[12,46]. Foliar microbes act as a barrier effect to foliar pathogens and modulate foliar physiology[47-50] and their resilience is not understood[51]. We test the hypothesis that drought and intensive farming modify foliar networks and ablate the barrier effect.
Sampling design	Cross-sectional NGS data have been collected for rice and grapevine. A time-series sampling campaign[52] will collect 720 leaf samples (grapevine and oak x +/-drought stress x 60 time points x 3 replicates) to reconstruct 6 networks using the change-point detection learning method developed by Partner 6.
NGS protocols	Partners 2 and 7 have experience of meta-barcoding plant-associated bacterial, fungal and viral communities[13,53-56]. To assess the absolute abundance of OTUs, we will combine metabarcoding data and qPCR results[57], and link structure to function using a microbial function (nitrogenase gene) that may impact Holm oak performance under drought[58].
Preliminary results	Foliar microbial networks of English oak have been reconstructed using methods developed by Partner 6[9,13], which can be directly applied to grapevine powdery mildew (*Erysiphe necator*).
Network validation	Many plant-associated OTUs cannot be taxonomically assigned or cultivated. We will validate expected network links between known foliar pathogen and antagonist OTUs. Culturomics in rice will validate the inferred pathobiome of *Magnaporthe oryzae*.
System 2	***Response of host-parasite interaction networks to biological invasions (Partner 5)***
Objective and System hypothesis	To reconstruct mollusc species competition networks in lentic freshwaters of Grande-terre and Marie-Galante islands. Species invasion has led to marked changes in network structure, with local reductions and extinction of *Biomphalaria glabrata* that hosts the parasite *Schistosoma mansoni*[59], which we hypothesise has modified rates of schistosome transmission to humans and rats.
Sampling design	We will take freshwater samples from 100 sites across the islands, repeated twice a year (dry and wet seasons) over two years.
NGS protocols	NGS protocols will be developed from existing mollusc-specific COI primers[60]. Species-specific PCR will assess the presence of *S. mansoni*.
Preliminary results	Community of snails have been followed for more than 15 years. We have some understanding of why there is species turnover
Network validation	We will validate links between mollusc species as already understood from the long term sampling of communities.
System 3	***Response of host-gut microbiota interaction networks to land use change (Partner 7)***
Objective and System hypothesis	To learn tripartite trophic networks between host plants, fruit flies (*Tephritid* sp.) and gut bacteria in La Reunion island. Fruit fly gut bacteria help synthesize essential amino acids and minerals[61] and prevent the colonization of the gut by pathogenic bacteria[62]. We will test whether the gut microbiota of fruit flies determines host range and their invasibility.
Sampling design	A total of 1200 samples, corresponding to 2 land use types (jardin créole versus monoculture) x 3 sites x 4 dates x 50 fruit flies, will be collected for reconstructing 6 networks (3 per type of land use).
NGS protocols	Pyrosequencing 16S rRNA genes has been used to characterize the gut microbiota of six species of fruit flies[63].
Preliminary results	The gut microbiota differs among fruit fly genera and responds to abiotic environmental conditions[63].

Network validation	We will validate expected network links between known gut OTUs, fruit flies and host plants.
System 4	***Response of plant-pollinator interaction networks to climate change (Partner 4)***
Objective and System hypothesis	To learn bipartite plant-pollinator visitation networks. Land use and climate change are modifying plant-pollinator network structure and function[64]. Plant and pollinator species diversity increases from North to South in France. We will test whether this gradient is explained by changes in network connectance or modularity.
Sampling design	A total of 36 bee soup samples will be collected (one per month, for 6 months, at six sites). We will also collect pollen from each pollinator species (≤ 100 per site) for a total of ~2000 pollen samples.
NGS protocols	We will use dual-indexing PCR-based MiSeq on pollen samples[65] and mitogenome-based HiSeq identification of pollinators[66].
Preliminary results	The six sites have been sampled every month from March to October 2016. Syrphid-plant networks (Figure 6) are being, and soon bee networks will be, analysed through co-phenology as an explanatory variable of interaction probability.
Network validation	As "primary data" has been obtained without any use of NGS methods, the two obtained networks will be independent and can be compared to one another for validation of the NGS approach.
System 5	***Response of plant-herbivore interaction networks to herbicides (Partners 1 and 3)***
Objective and System hypothesis	To reconstruct bipartite invertebrate-plant trophic interaction networks, centred on the carabid beetles in agriculture. Change in farmland management (herbicide use), acting on weed plants, modifies network structure and function.
Sampling design	We will sample 20 fields, with two contrasted herbicide treatments, on two occasions (80 NGS networks). Pitfall trapping along two transects in each treatment will provide 1600 pooled regurgitate samples from the 10 most important species of carabid.
NGS protocols	NGS approaches will be developed, using pre-existing NGS protocols, based primarily on the short fragment of COI[67-69].
Preliminary results	Existing bipartite networks constructed from sample data[70,71], the literature[44] and logic-based machine learning[7,21].
Network validation	The existing networks, as independent data, can be used for comparison and validation.