European Cooperation in Science and Technology
- COST -

Secretariat

Brussels, 9 June 2011

COST 4123/11

MEMORANDUM OF UNDERSTANDING

Subject :	Memorandum of Understanding for the implementation of a European Concerted Research Action designated as COST Action ES1103**:** Microbial ecology & the earth system: collaborating for insight and success with the new generation of sequencing tools.

Delegations will find attached the Memorandum of Understanding for COST Action ES1103 as approved by the COST Committee of Senior Officials (CSO) at its 182nd meeting on 17 May 2011.

MEMORANDUM OF UNDERSTANDING

For the implementation of a European Concerted Research Action designated as

COST Action ES1103

MICROBIAL ECOLOGY & THE EARTH SYSTEM: COLLABORATING FOR INSIGHT AND SUCCESS WITH THE NEW GENERATION OF SEQUENCING TOOLS.

The Parties to this Memorandum of Understanding, declaring their common intention to participate in the concerted Action referred to above and described in the technical Annex to the Memorandum, have reached the following understanding:

1. The Action will be carried out in accordance with the provisions of document COST 4154/11 Rules and Procedures for Implementing COST Actions, or in any new document amending or replacing it, the contents of which the Parties are fully aware of.

2. The objective of the Action is to lay the foundation for the systematic exploration of microbial diversity in the European Union. The Action will secure or increase the quality of each individual study, ensuring that data is gathered, analysed and curated to agreed standards.

3. The economic dimension of the activities carried out under the Action has been estimated, on the basis of information available during the planning of the Action, at EUR 56 million in 2011 prices.

4. The Memorandum of Understanding will take effect on being accepted by at least five Parties.

5. The Memorandum of Understanding will remain in force for a period of 4years, calculated from the date of the first meeting of the Management Committee, unless the duration of the Action is modified according to the provisions of Chapter IV of the document referred to in Point 1 above.

_____

A. ABSTRACT AND KEYWORDS

The microbial world is a vast frontier of intrinsic scientific importance and profound practical importance. The exploration of this frontier has been revolutionised by the introduction of molecular techniques. However, recent advances have only served to emphasise the enormity of the task before us. The improvements in sequencing technology have enormous implications for those at this frontier. Nevertheless description of this huge resource and the discovery of the rule governing its occurrence transcend the ability of not simply any one research group, but of any one nation. The purpose of this Action is to coordinate research groups across Europe to meet this challenge in the belief that if we agree upon common protocols and procedures we will share and pool knowledge to create a whole which is far greater than the sum of the parts. The Action will not only seek to document this frontier but to analyse it, to seek patterns, generate hypotheses and to test theories and thus deepen our knowledge. Perhaps most importantly of all, the Action will be preparing and enabling the next generation of researchers to use the next generation of technologies to ensure that Europe can lead the world in the exploration of this frontier.

Keywords: Microbial diversity, Pyrosequencing, massively parallel sequencing Pyronoise, bioinformatics, metagenomics, mathematical modelling

B. BACKGROUND

B.1 General background

The microbial world and the microbes therein, are both of enormous practical importance and intrinsically interesting. The seas, lakes, soils, sediment, animal and vegetable microbiome's and engineered biological systems represent a scientific frontier of astronomical scale.

This proposal is about the rational exploration of the microbial world. A task so challenging that it transcends the ability of not simply any one research group, but of any one nation. The members of the Action believe that this task can be achieved if scientists from across Europe are able to collaborate and coordinate their studies. Microbial diversity is sometimes described as a frontier, or an unknown land. Each study represents a scouting party, exploring this land. Though valid in their own right they can, on their own tell us little of the underlying "map". COST support will provide a medium and a mechanism for all these "scouting parties" to share protocols and pool observations so that together we can see the bigger picture and become more than the sum of the parts. Gaps that need filling will be found as will commonalities and patterns, thus future "explorers" will go forth more intelligently and effectively. For this synergistic pooling of resources to happen it is necessary to construct a forum in which researchers from across Europe can use to agree and coordinate protocols, to pool their data and to share the insights anticipated.

COST offers a systematic way in which to do this. There is no obvious transnational body for the coordination of microbial ecology research. The many other national (e.g. NERC, NRF) and international (ESF, ESA and FP7) funding schemes are focused on the individual projects that constitute, at best a handful of "scouting parties". Each study though undoubtedly excellent will necessarily be focussed and finite in scope and finite in impact. Moreover, such studies are rarely able to support a suite of bio-informaticians, mathematicians or theorists.

COST is particularly relevant to this challenge: this new generation of sequencing technology will define microbial ecology for the next generation or scientists. So the emphasis on early stage researchers, the next generation, is particularly important. Moreover, the technology is changing fast, so the flexibility offered by COST is essential as sequencing technologies and the associated bioinformatics tools are likely to be superseded within the lifetime of the project. New partners, approaches and thus tactics can be incorporated in the COST programme as and when necessary to ensure that the community can exploit novel developments in the field and still meet our broad strategic goals.

The benefits of COST will be, if not infinite, scalable and flexible. As more and more people collaborate in agreed protocols, the pool of shareable data will become larger and larger, the sum will indeed be greater than the parts. A COST project will be able to foster greater collaboration between the more classical community of microbial ecologists and the theoreticians and mathematicians from whom they have much to learn (and *vice versa*). There are great hopes for the science that will come from the interdisciplinary culture the Action seeks to foster. However, the most important benefit will be the inspiration and education of the next generation of scientists who will have the knowledge, the know-how and the confidence to complete what the task the generations before them have just begun.

This Action is quintessentially about people. By enabling the community it will maximise the effectiveness of a number of other initiatives which are about the recording of data in databases such as SILVA, Greengenes and CAMERA. By improving the quality of the data gathering process the Action will enhance the effectiveness of these investments.

B.2 Current state of knowledge

Ever since Antonie Philips van Leeuwenhoek first observed bacteria, microbiologists and there scientific forbears have been fitfully coming to terms with the extent and diversity of the microbial world. The term fitfully is appropriate, for periods of complacency have always been interrupted by revolutionary and surprising revelations about the extent and nature of the microbial world. The field is in the midst of one such period of revolution and revelation right now.

Most recently, the introduction and application of molecular microbial ecology has transformed our understanding of all such systems. The ability to infer the nature, physiology and even activity of microbes by examining the sequences they contain has breathed new life into this old field. The methods have been applied to environments varying from the floors of the oceans to the peaks of mountains and the atmosphere above. It is not possible to recount in this brief document all the insights and benefits we have accrued, they include the discovery of many new phyla, whole new physiologies the junking of text book assumptions about engineered systems and faltering, but unprecedented insights into how such systems might function. After more than two decades of

research by 1000s of researchers it is gradually dawning on the community that we are only scratching at the surface. An initial period of euphoria is gradually being supplanted by a deepening insight into the magnitude of the task before us and the heartening belief that the best has yet to come!

For though there has been an explosion of knowledge we have a very partial picture of the microbial world. Microbial systems are difficult to observe and scientists have only a primitive theoretical framework within which to operate. Two interlinked responses are emerging to this challenge. The simplest and most obvious is to simply gather more data. Theoretically based calls for bigger sample sizes have been followed up by experimental work that has indeed shown that larger sample sizes reveal greater diversity (the rare biosphere). However, there has also been a growing realisation that pure technology will never be enough and a body of quantitative theory to guide our work is needed.

The question of sample size, and to some extent resources, has been addressed, though not wholly resolved, by the advent of a new generation of sequencing technologies. Pyrosequencing and the like are heralding a step change in our understanding and could have a transformative effect on the field. The community will go from sample sizes of hundreds to sample sizes of 10s to 100s of thousands of samples with a proportionally spectacular drop in costs.

Any initial hubris associated with the success of this new technology has evaporated as scientists have come to realise that the large amount of data generated has become the bottleneck in our analyses, and further the interpretation of the data is complicated by sequencing errors. Several new challenges have arisen with the introduction of these methods. The valid interpretation of the data requires relatively sophisticated bio-informatics tools. Mathematical modelling and bioinformatics are becoming so important it is vital we improve communication between the modelling and bioinformatics and experimental sides of the community. Furthermore, even when these barriers can be overcome, the samples sizes are still not sufficient to document the majority of the taxa in many communities. More subtly, even with perfect sequencing and perfect sample sizes the data per se will yield only modest insight. To stretch our earlier analogy, a list of names is not a map.

Thus theoretical and mathematical developments are essential. They can be applied in two ways: Firstly, to analyse and interpret the raw data so that the diversity observed is correctly described and the extent of the unobserved diversity and thus the magnitude of the task ahead. Secondly, to use that data to test and explore theories of microbial community assembly or theories pertaining to other aspects of microbial ecology and to seek patterns, theoretically explicable or otherwise in the, data.

B.3 Reasons for the Action

At the moment, tens perhaps hundreds of research groups across Europe are gearing up to exploit this new technology and to spend millions of Euros analysing myriad microbial environments. The purpose of COST is to help coordinate and inform these initiatives to ensure that the European Scientific Community maximises the benefits of this investment by increasing the quality of each individual study and, more importantly ensuring that data is gathered, analysed and curated to agreed standards so that data can be pooled and compared to generated new and collective theoretical insights.

The Action will do this by ensuring that data is generated analysed and curated in such a way that all these data can be shared with confidence. This in turn will facilitate the comparison of datasets, avoid unwarranted duplication and allow a firm foundation for future work to capture patterns and promote theoretical developments.

COST is particularly relevant to this challenge: this new generation of technology will define microbial ecology for a generation. The emphasis on networking for the young investigators is particularly important. Moreover, the technology is changing fast, so the flexibility offered by COST is essential as sequencing technologies and the associated bioinformatic tools are likely to be superseded within the lifetime of the project. New partners, approaches and thus tactics can be incorporated in the COST programme as and when necessary to ensure that novel developments in the field can be exploited and that the broad strategic goals are still met.

At present, the primary motivation is scientific. However, we have in the past advocated a systematic survey of the microbial world along the lines of a Geological Survey. For we believe that a systematic survey of microbial diversity will have both scientific and economic benefits. This Action will highlight the benefits of a systematic approach to this most valuable of resources, and demonstrating that such a microbial survey might be cost effective and exercise.

The alternative to this Action is, potentially at least, a tragic waste of resource. Each study and each country finding its own way, adopting different standards of analysis and yielding mountains of data that cannot be compared or pooled.

B.4 Complementarity with other research programmes

There has never been a COST Action on the evaluation of microbial diversity from a phylogenetic or genomic perspective before. However, this field is so important and the technological push and scientific pull are both so strong that it can be reasonably expected many analogous related projects to be initiated in the course of the project. Thus there are ESF EUROCORES programmes such as Eurodiversity and EuroEEFG (evolutionary and evolved functional genomes). However none of these programmes address the fundamental methodological issues of this Action. Moreover, the fundamental methodological issues the Action wishes to pursue are generic and our activities will be hopefully synergistic with these and other programmes. The Action is particularly keen to complement the work of two multi-national initiatives: the Census of Marine Life and Terragenome, which are seeking to comprehensively document the marine and soil environments respectively and which both must address the fundamental methodological issues this Action will address. However, almost more important will be the myriad small projects that will be at once enabled by the exponentially dropping price of sequencing and blocked by the methodological issues that CISME will address. This COST Action will not be merely complementary to, but highly synergistic with, the known and plausible future activities in this area.

C. OBJECTIVES AND BENEFITS

C.1 Main/primary objectives

The main objective of the Action is to lay the foundation for the systematic exploration of microbial diversity in the European Union by creating catalytic and synergistic interactions between the myriad individual national and transnational studies of microbial diversity using next generations sequencing and the skills of advanced theoreticians and other numerate scientists across Europe. Participation in the Action will secure or increase the quality of each individual study, ensuring that data is gathered, analysed and curated to agreed standards so that the results can be pooled and compared to generate new and collective theoretical insights and deeper and wider map of this frontier in Science.

C.2 Secondary objectives

1. To agree and disseminate protocols for the use of pyrosequencing data for diversity estimates including techniques for designing sampling particular environments, distinguishing samples, primer choice, minimum standards for metadata and the curation of datasets.
2. To disseminate protocols for the generation and analysis of high quality diversity estimates. Primarily, but not exclusively, using the 16S gene the use of Bayesian protocols for the removal of sequencing noise will be explored and compared with other approaches for determining diversity. Various protocols for the rational estimation of total diversity using Bayesian estimates of species abundance patterns and other non-parametric methods will be considered. Authoritative recommendations on, and protocols for, determining sequencing depth will be generated. Finally the impact of sequencing error on diversity and sample size estimates will be determined.
3. To disseminate protocols for the generation and analysis of high quality meta-community, proteomics and RNA expression data. Using tools and concepts developed in WG2 the Action will develop protocols for the informed collection analysis and collation of all the genetic material in a given environment and the assessment of the proportion of the material recovered in a sample of a given size.
4. The systematic sharing and comparison of datasets within and between specific environments to both calibrate contemporary models in diversity and to detect and document consistent patterns within and between environments and microbial groups.

5. The systematic sharing and comparison of genomic datasets within and between specific environments
6. Towards a European Microbiological survey: an evaluation of the merits of European or national microbiological surveys analogous to the national geological surveys.

C.3 How will the objectives be achieved?

Protocols will in the first instance be discussed and agreed by workshops lead and coordinated by the appropriate theme leader (see below). These first workshops should be held early in the programme. However, given the likely pace of change in the field and the prospect of new and exciting technologies, it is likely that the protocol oriented workshops may have to be repeated again later in the process. Further liaison will also be required to anticipate that some laboratories will be branching out from the descriptive studies of the DNA coding for the 16S molecule to more challenging studies of metagenomic, proteomics and activities based studies from RNA analyses. Dissemination will be challenging. For though some approaches may be simply documented and placed on a web-site, others especially the more mathematical ones will benefit from demonstrations. National representatives will become local (i.e. national) centres of best practice and demonstrate such techniques to their counterparts. Small inter-laboratory exchanges (Short Term Scientific Missions); these could be didactic (though workshops are more efficient) or more likely to facilitate informal comparisons of methods within and between laboratories. The laboratory and computational resources required to develop these protocols should be held by the individual laboratories and research groups; The Action will bring the members of these lab's and groups together.

The systematic sharing of data should bring together "data generators" and theoreticians. Even when data generators have established theoretical ideas, they can often find that others will have new and ingenious angles on a particular subject.

Two forms of sharing are envisaged; between those with studying the same environment and those studying differing environments. Plainly it makes sense to conduct the former first. The sharing can be based around environment and theoretical theme and both of which can be proposed by members of the COST network and then chosen by the Management Committee. Successful workshops will require an appropriate mix of theoreticians and colleagues with data they feel able to share and appropriate facilitation. Importantly, such workshops should result in a publication capturing the insights obtained.

However, as the ambitions of the Action stretch beyond any individual environment or theory we will hold a "finale" to bring together as many participants as possible and to evaluate the progress the Action has made towards a systematic mapping of microbial diversity.

C.4 Benefits of the Action

The benefits of this Action are twofold. Individual researchers and research group will produce work of higher quality more quickly, and their results will be more widely accepted. Thus the exploration of a wide variety of microbial domains will be enhanced and accelerated and deeper and more authoritative insights into how such systems work will be gained.

Collectively the scientific community as a whole will gain tremendous benefits. The collective production of data of an agreed and assured quality will allow researchers to share data across sites revealing patterns and geographies the no single researcher or research group could possibly hope to attain. For example, Europe stretches from the Arctic area down to Africa. There are several environmental gradients, the study of which is known to be insightful tremendous potential to explore biogeographical patterns within the network. It is anticipated that the presence or absence of consistent patterns within and between environments will give unprecedented insights into the mechanisms governing the formation and change of such communities. These insights and possibilities can be formalised in new or existing models that can be calibrated and tested using this new resource. This in turn would catalyse the rational and ultimately predictive exploration of the microbial world with tremendous intellectual and economic benefits.

This project is correspondingly highly geared. Making the very best use of a very large investment in the sequencing of microbial systems that Europe is on the cusp of making. The corollary is, that if this investment is not made “the whole will be less than the sum of the parts” as individual datasets will be of highly variable, and often unknown, quality. The quality of individual endeavours will be reduced and prospects for collective and synergistic insights weakened and perhaps eliminated.

C.5 Target groups/end users

In the first instance the target groups will be microbial ecologists and associated theoreticians and bio-informaticians. Most microbial ecologists are now keenly aware of the potential of next generation sequencing. However, the Action will draw in numerate scientists from a mathematical, physics or engineering background who could have a vital role to play in the analysis of this data. However, many microbial ecosystems are of profound practical importance and it is believed that the standardisation and dissemination of protocols for the analysis and understanding of the next generation of sequencing data will fast track the extension of this technology to areas such as agriculture, environmental protection and human and veterinary medicine.

D. SCIENTIFIC PROGRAMME

D.1 Scientific focus

This Action aims to provide a structured, but not excessively detailed work plan flexible enough to permit the adjustment, and also inclusion, at the implementation stage, of perspectives and activities and innovations not foreseen or realised during the preparation of the proposal. The Action will therefore seek to engage with other researchers, and countries, beyond those who have participated in the proposal to date.

The scientific programme has been developed to achieve the objectives set out in Section C. There are broadly speaking 4 foci to this proposal. The basic analysis of the data, the search for patterns therein and the testing of the hypotheses formulated on that basis and finally the prospect for the systematic exploration of the microbial world in Europe and elsewhere. These foci map on to the 5 main tasks which in turn will correspond to the Working Group structure.

Task 1. To agree and disseminate protocols for the generation of the data

Furthermore the new generation of sequencing technology is generating new techniques for getting the most samples from a single pyrosequencing run (barcoding) and new challenges for storing the data. For instance when will it be cheaper to resequence the DNA than to store the results on a computer! Last and by no means least, it would be extremely valuable to consider and agree when where and how samples should be taken the minimal metadata (pH, temp etc.) requirements on a generic (i.e. for all samples) and an environment specific basis (e.g. soil, seas, and lakes) and the numbers and kinds of samples that are appropriate.

Task 2. To generate protocols for the analysis of the data on diversity

The excitement generated by the use of the next generation of sequencing has also been matched by a considerable amount of controversy. Broadly speaking two kinds of error are under discussion: sequencing error and erroneous estimates of diversity. The two forms of error are linked. The sequencing error is largely caused by the tendency of the new generation of massively parallel sequencing to miscall homopolymers (sequences of DNA with the same base), though there are undoubtedly problems with chimera as well (when the DNA from two separate organisms give one sequence). The effect of this error is to increase the apparent diversity of the sample and lead to inflated numbers of unique sequences. It also skews the underlying distribution of the sequences which is used to estimate the true diversity. There are a number of approaches to resolving the issue of sequencing error. The best, but perhaps more difficult approach is to reanalyse the data to remove the sequencing error. The simpler, but arguably less satisfactory approach is to analyse the data in such a way that the effects of sequence error are minimised. The relative merits of these approaches need to be discussed and evaluated and the effect on the quality and clarity of our picture of microbial diversity considered. The Action is particularly concerned about the understanding of the long term impact of use the simpler approach. Will it condemn us and future generations to a degraded picture of diversity that will prevent us from seeing the fundamental patterns in the data.

In short long term harm could be caused to the knowledge base if, for short term gain, it is neglected to analyse the data correctly. The costs and consequences of the contrasting approaches should be evaluated. Even with very large sample sizes, the observed diversity typically falls well below the true diversity of the system under study. A range of mathematical tools exist to permit the extrapolation from a sample to the diversity of the true community. Again there are differing perspectives on how best to achieve this and differing techniques. A discussion of and recommendations for the determination of diversity is required. This debate must be informed by the debate on the sequencing problem, because the strategy for coping with sequencing error will inevitably impinge on the strategy for estimating diversity.

Task 3. To extend the application to metagenomics and RNA

Though task 2 is ambitious and sophisticated, for many researchers, its successful accomplishment will merely set the scene for deeper studies. In particular, an examination of all the genetic material, its expression as RNA and as protein has the potential to render further insights. Indeed it seems inevitable that as the costs of sequencing drop and drop more and more groups will extend their work to this logical next phase. The task of this WG is to ensure that this next step is taken efficiently and rationally. For example the current modus operandi of ad-hoc sequencing of as much as material as possible will be replaced with protocols for rationally sequencing to a depth sufficient to capture a given fraction of the genomic material. Furthermore, this work will be dogged by the same sequencing errors as the previous WG but now with different consequences. For one of the most powerful reasons to garner and compare meta-genomic sequences is to gain insights into evolution: the master force behind all microbial ecology. High levels of sequencing error will diminish the ability to gain such insights and so the Action should strive to promote the highest standards in respect to the analysis of raw sequencing data. Indeed comparisons between separate studies may be impossible if different procedures are adopted by different laboratories.

Assuming data of suitable quality can be gathered it will be possible to study evolution and biogeography at the same time. In addition to noise the analysis of whole genome data brings new challenges. For example, the individual segments of DNA distributed throughout the genomes of all the members of community need to be assembled into longer stretches of DNA. This problem has been well studied for single genomes but for meta-genomes no good algorithm exists. This can provide a wealth of information on the community structure and function. Ecological insights may also be gained by determining when certain genetic material is expressed and by examining RNA or protein (proteomics). Though proteomics is not going through the same revolution that genomics is, it will offer an important and complementary perspective.

Task 4. To see patterns and to test hypotheses

Whilst it is obviously beneficial for individual research groups to generate data whilst working to the highest and agreed standards, the greatest added value will come from comparing the data sets. Such comparisons can be used in an at least 3 simple and powerful ways: (i) To Speculate; by simply looking at the data observing novel patterns that might relate to geography or the environment. The Action will seek new patterns or testing for new ones; an approach that might generate hypotheses. (ii) Test theoretical concepts; The Action will evaluate established theoretical concepts that might relate to time, distance, physiology or environmental variable. (iii) To test novel hypotheses; it is inevitable that new hypotheses will arise in the lifetime of the project. These may be wholly or partly tested by the datasets emerging across Europe in the coming years. The nature of the comparisons will ultimately depend on the extent of the collaborations. The greater the intellectual and geographical spread of our Action the more scope the Action will have. Those the participants in the call have interests that spread from the very highest arctic to Africa and encompass soils, sediments freshwater and engineered systems.

Task 5. Towards a European Microbiological Survey

Most, perhaps all, European nations have a geological survey, in recognition of the fundamental importance of geology to the economy, ecology and infrastructure of a modern society. However, it is not unreasonable, to suggest that Europe's microbiological resources could have as much impact in the 21st century as our geological resources had in the 19th and 20th century. This impact may be direct and economic as the organisms and the genes therein are harnessed to generate new fuels or other products, or indirect as microbes mediate climate change or dispersed pollutants. However, at present the exploration of the microbial world is an essentially *ad hoc* process. Even avowed surveys or censuses are not wholly rationally statistically informed undertakings. By the end of the Action, the community will have a much clearer ideas of both the magnitude of the task of exploring the microbial world and the costs and benefits of doing so. Participants in the Action would be in an excellent position to consider the prospects of undertaking such study on a European scale. It would be premature at this juncture to state what conclusions those conclusions might be. However, they might vary from a bespoke transnational body to a loose confederation of national research, or simply conclude that such a survey is not required. However, the Action would allow the participants to not only to determine the value of such a survey but to be able to identify the most cost effective and accessible manner in which it might be undertaken. This will generate a Roadmap for the exploration of European microbial diversity at a national and transnational level.

D.2 Scientific work plan – methods and means

This COST Action will comprise five Working Groups (WG) that will address research the corresponding tasks 1-5. Throughout the WG will meet every six months (see Section F). These meetings will form the main mechanism for synthesising and consolidating of existing knowledge and exploration of new ideas, collaborations and research directions relevant to each WG. The WG meeting topics and programme are sufficiently flexible to allow collaboration with other COST Actions or European networks and groups to minimise risk of duplication and maximise scientific benefits. It is anticipated that the WG 1 and 2 will have more prominence initially and WG 2 and WG3 in the middle of the Action and WG 5 and WG4 will be more important towards the end.

However, flexibility is of paramount importance in such a dynamic field. WG meetings will be a mixture of open meetings and highly focused, thematic, workshops that are limited to approximately 30 people, to ensure meetings are highly focused and participative. However, open calls will be issued to ensure a wide variety of researchers, from a diverse range of institutions, can attend and input into the events. A full conference held in year 4 will disseminate research widely and comprise the main forum for the discussion of, and recommendations with respect to, any putative national or transnational microbiological survey.

WG1: Data Generation Protocols

This Working Groups will consider the “nitty gritty” of generating the data, including best practice in bar-coding and curation, the most appropriate regions of 16S genes and the consequences of using differing regions or other methodological difference such as differing DNA extraction methods. They will propose minimal metadata requirements for various environments.

The protocols should as far as possible represent a rational consensus reflecting wherever possible known best practice ascertained through a workshop supported by polling those who cannot attend and a critical consideration of the literature. Sort term scientific missions (STSM) will be required in-certain instances to clarify the relative merits of different methodologies. The protocols may well have to be revisited in the course of the Action.

Deliverables: (i) A review recommendations and consensus about the data generation protocols that will encompass the choice or sequences, barcoding, curation, sampling strategy and minimal meta-data. The consequences of modest and severe departures from the consensus will be noted and knowledge gaps identified; (ii) a community that can operationalise these recommendations; (iii) a community that can revise the consensus in the light of ongoing technical developments.

WG2: The analysis of raw data for diversity estimates

This Working Group will consider how the raw data is converted into “species” or more correctly operational taxonomic units (OTU) and patterns in OTUs into diversity estimates. In particular they will review and critique the strategies for dealing with sequencing error and where possible through workshops and STSMs will obviate any technical or knowledge barriers to the best approaches. In making such recommendations they should pay particular attention to the effects on subsequent attempts to determine the diversity or sampling effort using the best available and foreseeable technology. They should, consider how diversity estimate can be made consider the long term advantages and disadvantages of each one and issues of resources where appropriate. Groups with differing expertises will teach their methodologies in a workshop environment to spread best practice.

Deliverables: (i) A review recommendations and consensus about the generation of OTU data including an evaluation of the long term and short term consequences of the differing strategies for dealing with sequencing error. (ii) A review recommendation and consensus about how to estimate diversity and sample sizes from OTU data and (iii) A trans-European cohort able implement the best practice identified and to spread best practice to their compatriots and collaborators.

WG3: The analysis of raw data for metagenomics, RNA expression and beyond

This WG will consider the analysis of metagenomic data gathered directly from the DNA or from RNA expressed in the environment. Building on the work in WG2, WG3 will be able to make concrete recommendations about how diversity estimates can be used to make estimates about samples size and coverage in metagenomic studies. The challenge of de-noising must also be confronted by this community and strategies for so doing can be considered and protocols generated on this basis. For example the merits of investing in computing power to make the best use of the data gathered can be set against the alternative strategy of simply sequencing more.

The impact of sequencing error on inferences about evolution should be considered. For example assembly algorithms will be compared using synthetic *in-silico* data sets and novel ones will be potentially developed. More prosaically, the challenge of the sheer volume and complexity of the data must be brought to the fore as must strategies for identifying the functions of the very large amounts of sequence.

Deliverables: (i) A review recommendation and consensus about the analysis of metagenomic data including an evaluation of the long term and short term consequences of the differing strategies for dealing with noisy sequence data, assembly algorithms, and gene function prediction; (ii) a review recommendation and consensus about how to determine the appropriate depth of sequencing; (iii) a trans-European cohort able to implement the best practice identified and to spread best practice to their compatriots and collaborators.

WG4: Data comparison pooling and exploration

This WG will coordinate those with data, ideas and theories to seek patterns in the data. STSM and generalist workshops will be commissioned, in response to requests from the community, by the management group in liaison with the WG leader. It is anticipated favouring STSMs proposed by and involving Early Stage Researchers (ESR). These workshops and missions will seek patterns, test models or examine hypotheses. Though plainly speculative, this is potentially an exciting and highly productive exercise with each workshop generating insights and publications.

Deliverables: (i) Manuscripts describing the exploration of patterns, the testing of theories or the testing of novel hypotheses; (ii) a community, in which ESR play a prominent role, committed to the forgoing in the future.

WG5: Towards a systematic survey

This WG will reflect upon the Action as a whole and if and how the aspirations of the Action forward can and should be taken formally or informally. This WG will be made up of leaders and members of the other WGs and will address the hypothesis that a systematic survey of microbial diversity is warranted and feasible. Drawing upon the outputs and insights of all the WGs it will be able to make authoritative recommendations in this regard.

Deliverables: A Roadmap for the exploration of European microbial diversity.

E. ORGANISATION

E.1 Coordination and organisation

The organisation of the COST Action will follow the usual procedures in regard to management through the formation of a Management Committee (MC) and, for this Action, five Working Groups (WGs).

The COST Action will be led by the MC. MC meetings will take place every six months, usually linked with the scientific meeting of a Working Group (WG) or with a workshop. Meanwhile, restricted MC meetings (Chair and Vice-chair(s), WG Co-ordinators) will take place every three months, by video- or tele-conference, to insure an efficient co-ordination of the activities, to review critical points and to maintain a clear focus on the objectives and milestones. The tasks of the Management Committee will be the following:

- Appointment of Action Chair, Vice-Chair(s), WG co-ordinators and an STSM Evaluation Committee.
- Planning of Management Committee meetings and of Scientific Meetings and Workshops.
- Assessment and report of the progress made by the different Working Groups and STSMs to meet their respective objectives and milestones, in the framework of the focus and direction of the Action.
- WG co-ordinators are expected to prepare short update reports every three months. If WGs are failing to meet their milestones or objectives in a timely fashion the MC will convene at the earliest opportunity, by telephone/video conference if necessary, to discuss mitigation actions.
- Promotion of co-operation and of data exchange between the Working Groups.
- Promotion and approval of STSMs, according to the recommendations of the evaluation committee.
- Establishment of a conference secretariat for the final Action conference which should include a local organising committee and a scientific committee.
- Preparation of Annual Reports.

- Establishment and updating of an Action specific website for internal communication, advertising the Action conference and dissemination of STSM findings, WG reports and other results.
- Organisation of contacts and common workshops with the appropriate ongoing COST Actions and other relevant technology or scientific platforms (including those identified in Section B.4), to address problems of common interest.
- Seek to deliver the Action outputs with the most efficient use of COST resources.

A simple collaboration and confidentiality agreement will be set up for each STSM or other data sharing activity.

E.2 Working Groups

Five Working Groups will be established as outlined in section D. Each Working Group will be led by a Chair and Deputy Chair. Excellent male and female ESR with a track record in the relevant fields who are willing to be nominated for these positions have been identified. The WG leaders will be responsible for:

- Co-ordination of the activities within the WG to ensure the timely attainment of the objectives and milestones.
- Planning the appropriate scientific meetings.
- Ensuring the production of final reports and encouraging the production or journal papers
- Promoting the set-up of joint research, including STSMs.
- Reporting (every six months) on the WG progress to the Action Chair and Management Committee (MC).
- Participation in the plenary and restricted MC meetings.

It is envisaged that the Working Group meetings will be organised and held individually (i.e. without other WG). This will enhance the exchange of information and ideas, to stimulate the synergy between scientists, institutes and countries to maximise effort and productivity.

E.3 Liaison and interaction with other research programmes

The Action would be delighted to host join workshops with appropriate ongoing COST Actions. However, the current suite of COST schemes does not yet have an appropriate Action. The Action will liaise with various research programmes (most notably the International Census of Marine Microorganisms and Terragenome) either directly through shared goal oriented workshops or indirectly by having fundamental participants from the organizations as experts involved in this Action. Collaborations will be sought proactively throughout the Action's programme.

E.4 Gender balance and involvement of early-stage researchers

This COST Action will respect an appropriate gender balance in all its activities and the Management Committee will place this as a standard item on all its MC agendas. The Action will also be committed to considerably involve early-stage researchers. This item will also be placed as a standard item on all MC agendas.

A significant number of the experts identified in the Action are women including the co-author of this proposal, many are internationally recognised researchers. These leading figures will be encouraged to mentor young female scientists and encourage others to participate in the Action. Likewise there are a number of Early-Stage Researchers involved in this Action. Early stage researchers will be strongly encouraged to take on a co-ordination role of WGs, chair discussions in workshops and sessions at the Action conference, take a lead author position in preparing scientific journal outputs and be proactive in steering the Action agenda to support their training and development. More established researcher will be encouraged to support the ESR to help them develop "Savoir Faire" and confidence. In addition to proactively seeking to promote the gender and early-stage researcher balance, Short-Term Scientific Missions (STSMs), within and also between the different Working Groups, will help to co-ordinate the research through the establishment of adequate co-operation between the participating institutes. Exchange and mobility

of scientists will not only strengthen the co-operation within and between the Working Groups, but also favour the training of young scientists, in the framework of STSMs. Funds for STSMs will be awarded competitively by an STSM evaluation committee. This committee will be appointed by the Management Committee and contain one co-ordinator and one representative of each Working Group. The STSM evaluation committee will award funding according to the relevance to the COST Action objectives and the quality of the candidate. The majority of STSM is expected to lead, or contribute, to a journal publication with multinational authors.

F. TIMETABLE

The duration of the Action is 4 years. WGs will meet every six months, with an initial workshop involving all WGs mapping out the objectives and activities for the next four year. Initially most of the WG meetings will pertain to WG1 and WG2, alternating. Problems identified in the meetings will be addressed by STSM, the results of which will feed back to the next relevant WG. . Where appropriate these workshops will be held with other COST Actions or ESF Networking Programmes to allow for the cross-fertilisation of outputs and ideas.

Plenary Management Committee meetings will take place every six months, linked with the scientific meeting of a WG where possible. Meanwhile, restricted meetings (Action Chair, Vice-chair(s) and WG co-ordinators) will take place every three months, by video- or tele-conference (to reduce travel costs and environmental impact) where appropriate. These meetings ensure an efficient co-ordination of the activities and maintain a clear focus on the objectives and milestones.

The preliminary timetable for the Action is:

	Year 1				Year 2			
	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4
MC Meetings	*		*		*		*	
WG Meetings	All		WG1			WG1		WG1
			WG2		WG3			WG2
								WG3
STSMs			--------	---------	--------	---------	----------	
Scientific activity	|--------	----------	--------	--------	---------	--------	---------	--
Protocols report						*		

	Year 3				Year 4			
	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4
MC Meetings	*		*		*		*	
WG Meetings	WG4		W4		WG4			WG4
						WG5		WG5
STSMs	-------		--------	--------	--------	-------	--------	|
Scientific activity	--------	----------	--------	--------	---------	------	|	
Conference							*	
Final report							*	

G. ECONOMIC DIMENSION

The following COST countries have actively participated in the preparation of the Action or otherwise indicated their interest: AT, BE, DE, DK, EE, ES, FI, FR, IT, NL, NO, RO, SE, UK. On the basis of national estimates, the economic dimension of the activities to be carried out under the Action has been estimated at 56 Million € for the total duration of the Action. This estimate is valid under the assumption that all the countries mentioned above but no other countries will participate in the Action. Any departure from this will change the total cost accordingly.

H. DISSEMINATION PLAN

H.1 Who?

The primary target for dissemination is other researchers. It seems inevitable that the examined methods will, in the lifetime of the project become a routine part of microbial ecology. Although established laboratories are obviously important in this endeavour, it is highly likely that the rapid decrease in cost will bring in many new players. Thus participant in the Action are alive to the possibility of the creation of a whole new generation of microbial ecologists transformed by technology. The Action must reach out to all potential newcomers and it is expected that our many national representatives will cascade knowledge down with their country and region.

It is also important that the Action does not confine activities to biologists. The Action is particularly anxious to reach out to and indeed stimulate the generation of a new cadre of bioinformaticians, modellers and theoreticians who will be able to take advantage of the new data rich era.

Practitioners (engineers, agriculturalists, medical and veterinary microbiologists) may also benefit from the work. Though this is not primarily about problem solving the protocols and concepts generated by the Action will be of use to those that solve problems involving microbes.

Finally the Action hopes to target policy makers. In particular the Action would like each and every nation in Europe to recognise the intrinsic and economic value of the microbial resource within their borders. If it is concluded that they should more actively explore the microbial diversity of each nation the Action will have to not only make this case but to cost it and layout the mechanisms by which such a survey might be undertaken nationally or transnationally.

H.2 What?

The Action will use a number of complementary strategies to reach out to fellow researchers. The most enduring and authoritative mode of dissemination is the peer reviewed paper or opinion piece. This will be an excellent way to publish methodological improvements and comparisons and the anticipated outcomes of WG3 (patterns, tests of theories or hypotheses). However, where participants in the Action have agreed and recommended protocols they should make them available widely and quickly. These can be published on the website along with the rationale. Where necessary pre-publication material will be password protected to ensure the Action comply with rules concerning prior-disclosure. However, it is likely that some of the most important protocols associated with WG3 will involve novel and computationally intensive computer programmes which will not be easy for most members of the community to implement. The Action will therefore hold master-classes during our WG meetings to disseminate these methods, the associated software and the teaching material available to others to cascade best practice across Europe.

The Action will, in addition to the above, reach out to practitioners by encouraging member of the Action to publicise our findings and activities in the practitioner based meetings forums.

Policy makers can be reached by direct lobbying of the relevant national government organisations) and opinion pieces in prominent International Multidisciplinary journals supported by press releases and traditional and non-traditional (the blogosphere) journalism.

All dissemination formal and informal will be recorded on the COST Action website. The website will be managed by an ESR associated with the Action co-ordinator or a WG Chair. Each WG will have an ESR responsible for their WG's presence on the website. The website will provide information for the wider community relating to conferences and reports. The website can also be a mechanism for engaging with international researchers interested in this area.

H.3 How?

Knowledge, data and methods coming out of the COST Action activities will be integrated and presented at appropriate National and International Conferences (e.g. The International Society for Microbial Ecology) may include a special session at the high profile European Geosciences Union assembly. Recommendations will be reported to Journal Editorial Boards and professional bodies

The findings will be written up as a series of high profile journal publications. Five journal publications, led by WG Chairs, will include:

- Review of the state of the art in data generation (Year 2)
- Review of the state of the art in Noise Removal, OTU definition and chimera detection (Year 2),
- Review with recommendations and associated software for diversity estimation and sample size estimation for diversity studies and meta-genomics studies (Year 2),
- Review of the methods and perspectives in analysing validly collected data to look for patterns, test theories and hypotheses (Year 3).
- Review of the prospects for the rational exploration of the microbial world (Year 4).

These review papers will help consolidate the state of the art in these areas and provide a valuable resource for academics and professionals. Additional papers will result from STSMs. It is anticipated that a number of papers with multi-national authorship will come out of WG 3. These papers will be published in a wide range of journals to maximise readership.

Journal papers will be supplemented by a stakeholder report, to ensure methods and outputs are widely accessible to practitioners.