ECOLOGY LETTERS

Ecology Letters, (2010) 13: 776–791 doi: 10.1111/j.1461-0248.2010.01464.x

REVIEW AND SYNTHESIS

Integration of molecular functions at the ecosystemic level: breakthroughs and future goals of environmental genomics and post-genomics

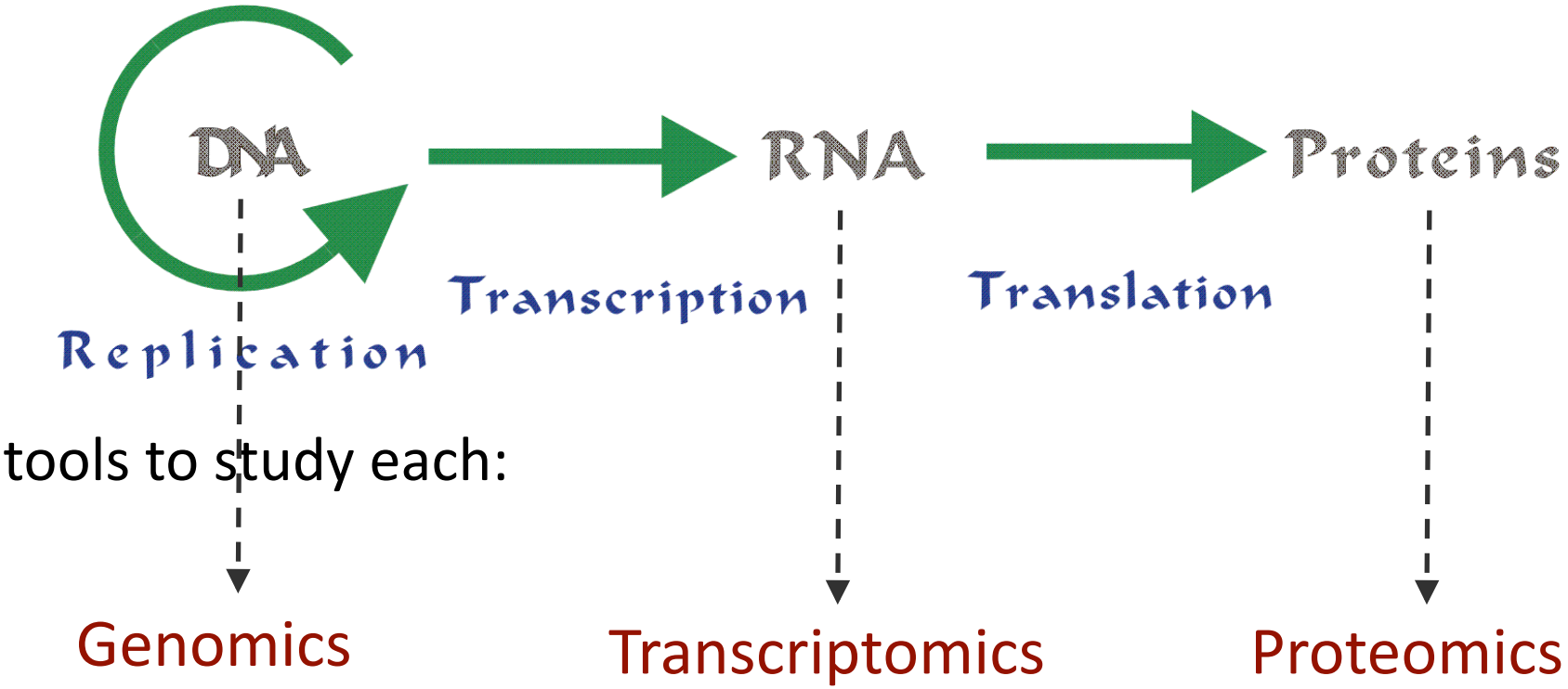
Philippe Vandenkoornhuyse,* Alexis Dufresne, Achim Quaiser, Gwenola Gouesbet, Françoise Binet, André-Jean Francez, Stéphane Mahé, Myriam Bormans, Yvan Lagadeuc and Ivan Couée

= a unifying framework

Abstract Environmental genomics and genome-wide expression approaches deal with large-scale sequence-based information obtained from environmental samples, at organismal, population or community levels. To date, **environmental genomics, transcriptomics and proteomics** are arguably the most powerful approaches to discover completely novel ecological functions and to link organismal capabilities, organism–environment interactions, functional diversity, ecosystem processes, evolution and Earth history. Thus, environmental genomics is **not merely a toolbox** of new technologies but also **a source of novel ecological concepts and hypotheses**. By **removing previous dichotomies** between ecophysiology, population ecology, community ecology and ecosystem functioning, environmental genomics enables the integration of sequence-based information into higher ecological and evolutionary levels. However, environmental genomics, along with transcriptomics and proteomics, **must involve pluridisciplinary research**, such as new developments in bioinformatics, in order to integrate high- throughput molecular biology techniques into ecology. In this review, the validity of environmental genomics and post-genomics for studying ecosystem functioning is discussed in terms of major advances and expectations, as well as in terms of potential hurdles and limitations. Novel avenues for improving the use of these approaches to test theory-driven ecological hypotheses are also explored.

what are "**environmental genomics, transcriptomics and proteomics**"?

let's go back to
the "central dogma" of molecular biology



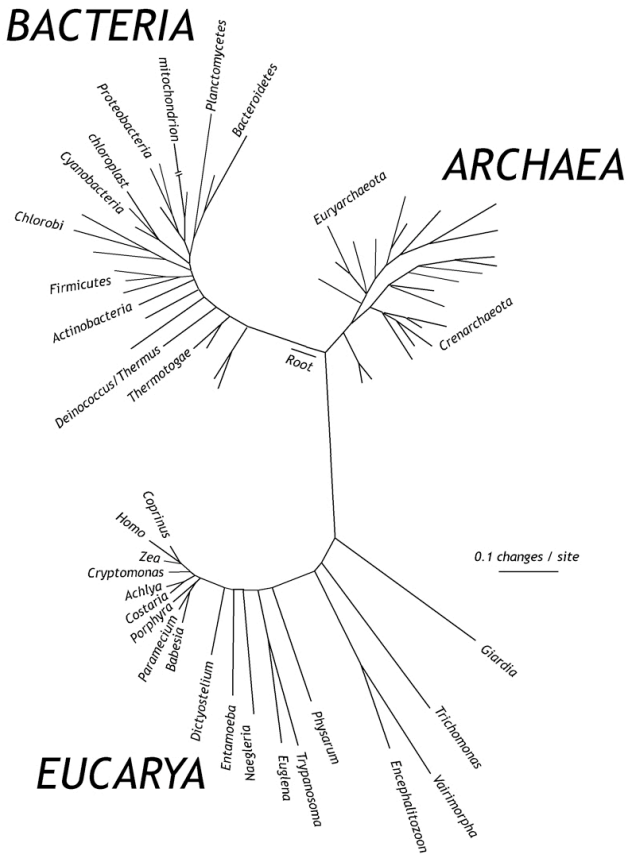
more on why these different levels of information do not have a 1:1:1 correspondence (the first is easiest to see: not all genes are expressed all the time) in a future week's discussion

So what are "**environmental**" –omics studies?

p. 777

- mixed mostly prokaryotic microbial communities *= the majority of research so far*
- mixed prokaryotic and eukaryotic microorganismal communities
- small-size euks *(q to you: why not large euks?)*
- mutli-species networks of higher euks, e.g. root mats, mixed-species insect swarms
- higher euk. tissues containing symbionts
- non-cultivable species

this side: smaller cells, smaller genomes, and high "coding density" = the density of genes versus non-coding DNA in the genome



http://www.landcareresearch.co.nz/research/soil/Ant_soils/Images/soil_microbes_tol_800.jpg

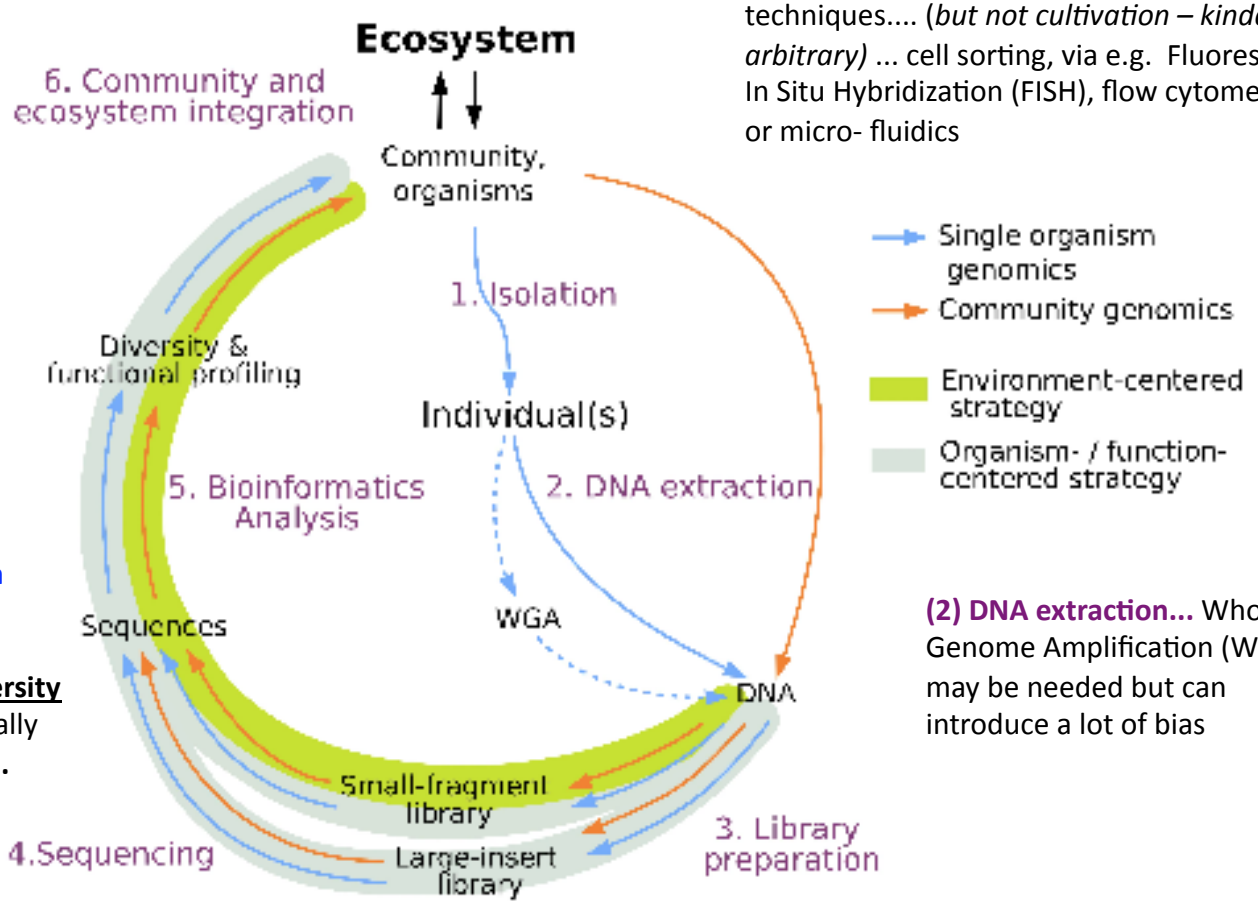
this side: larger cells, much larger genomes and low "coding density"

Box S1 : Strategies in environmental genomics

- communities as a whole
- or single (micro)organisms isolated

(6) Community and ecosystem integration : ... requires... (meta)transcriptomics, (meta)proteomics, physiological and biochemical characterisation, experimental analysis, microscale and large-scale environmental data measurements, mathematical modelling.

(1) Isolation : using a variety of techniques.... (*but not cultivation – kinda arbitrary)* ... cell sorting, via e.g. Fluorescent In Situ Hybridization (FISH), flow cytometry or micro- fluidics



(5) Bioinformatics :

- assemble sequences into larger fragments of genomes, or even near-complete genomes ... may represent composite genomes **aka "population genomes"**
- predict genes and **identify their potential functions to determine diversity and functional potentialities** of naturally occurring organisms and communities.
- **plus "Diversity"**

(2) DNA extraction... Whole-Genome Amplification (WGA) may be needed but can introduce a lot of bias

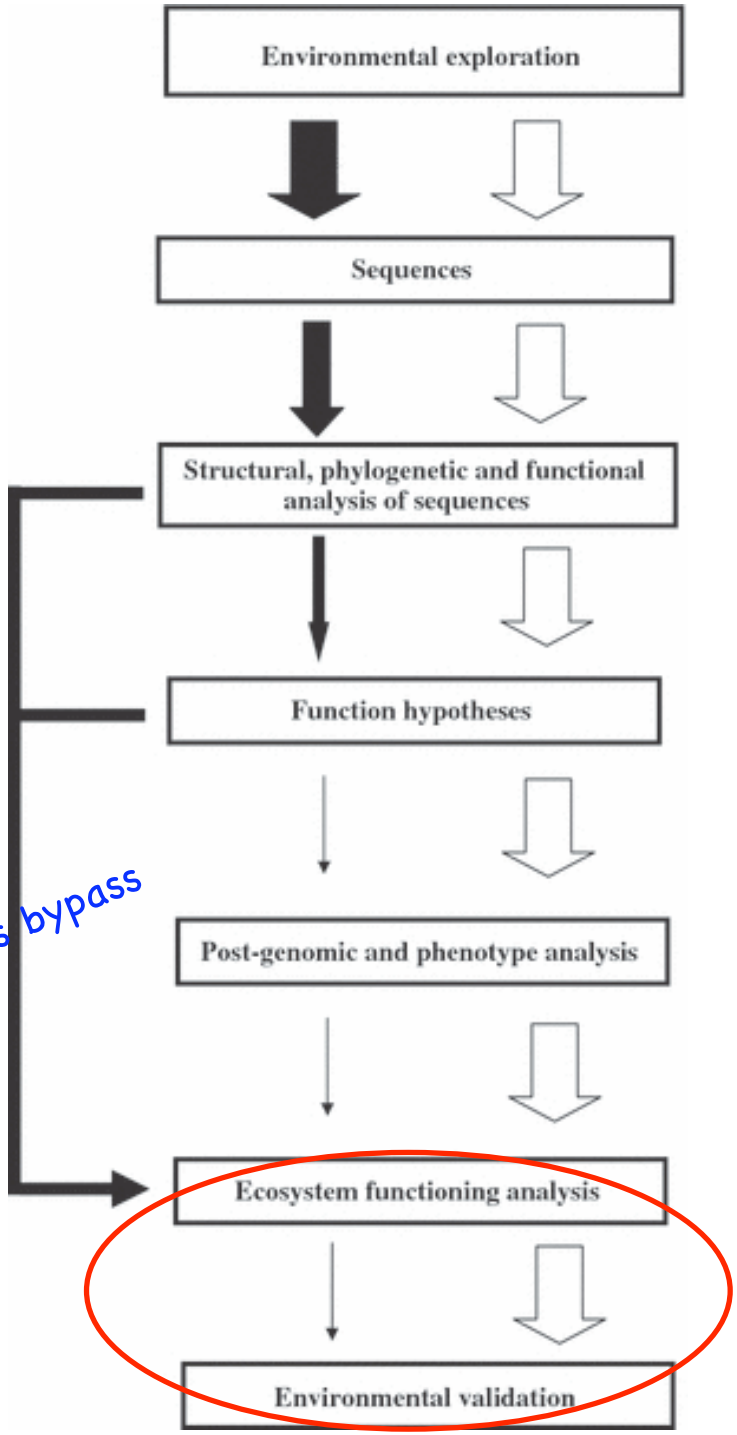
(3), (4) Library preparation and sequencing :

i. **"Large-insert clone libraries":** Large DNA fragments cloned into vectors such as bacterial artificial chromosomes and fosmids. *Screening* of large-insert libraries allow to select clones bearing functionally or phylogenetically relevant genes before sequencing (function- / organism-centered strategy)

ii. **Small-fragment libraries** (now generally without cloning): shotgun sequencing, genomic DNA is fragmented into nsmall size fragments which are then randomly sequenced (environment-centered and function- / organism-centered strategy). Used for high-throughput sequencing methods e.g. pyrosequencing (454), Illumina etc...

what we do vs. what we'd like to do in linking sequences to ecosystems...

Figure 1. Real-life and ideal fluxes of analysis and information in environmental genomics. Current throughputs of analysis and information-processing are given as black arrows, whereas the ideal throughputs to be achieved are shown as white arrows. Arrow thickness reflects the efficiency of the analyses.



more on Step 6: community and ecosystem integration – how it's done now...

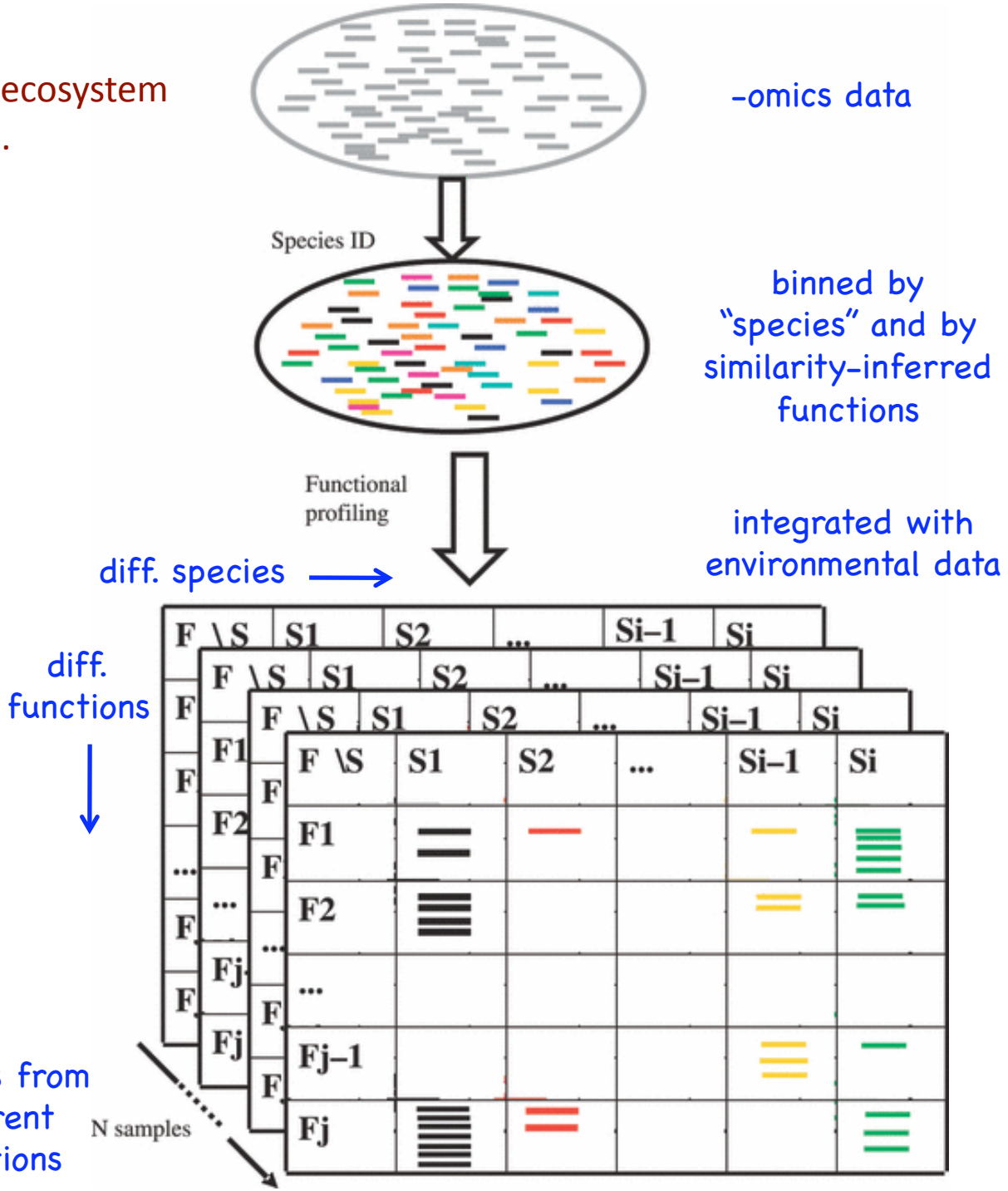


Figure 3. Spatio-temporal three-dimensional organisation of sequence-derived datasets. The set of environmental genomic, *cDNA*, or protein sequences (grey bars) is ascribed to a set of i Species (S), thus resulting in species-labelled sequences (colour bars). The aim of functional analysis and profiling is to ascribe species-labelled sequences to a set of j functional categories (F), thus resulting in a 'potential function × species' understanding of the ecosystem. The third dimension of the matrix corresponds to spatio-temporally replicated samples, such as samples subjected to various environmental constraints, or samples at different points in time. This kind of dataset can be analysed not only to understand the mechanisms induced by a forcing variable, but also to select and parameterize the components that have to be included in a model.

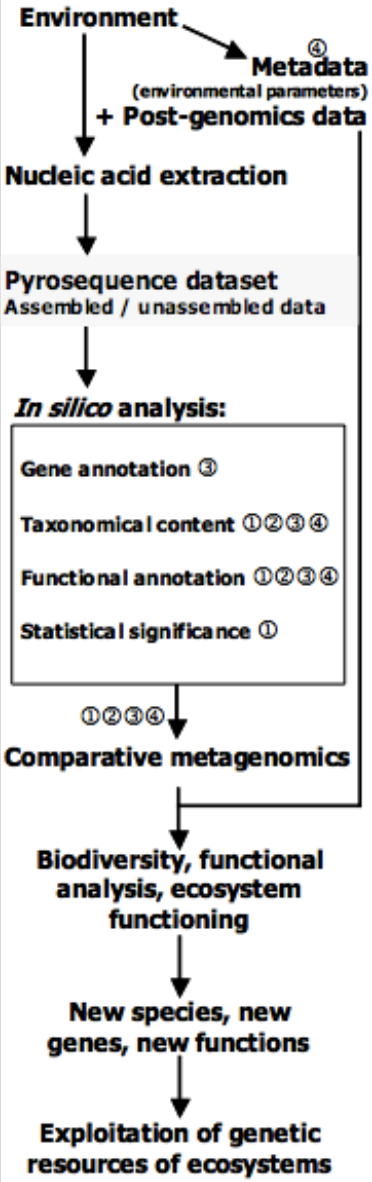
Paper covers many aspects of Environmental Genomics:

- **As unifier throughout ecology and biology**
- **Molecular data allowing integration of diversity and functions**
- **Major breakthroughs *(proteorhodopsin, ammonia-oxidizing Crenarchaea, AMD, GOS)***
- **Current limitations of functional integration**
  - **Sampling and sequencing *(biases, depth)***
  - **Gene ID and functional characterisation *(biases, error prop, unknowns)***
  - **Difficulty of "true" functional assignment beyond characterization challenges**
  - **Plasticity of gene expression**
  - **Linking to environmental phenotypes**
- **Improvements from an ecological point of view**
  - **Using ecological and evolutionary info to inform functional identification**
  - **Bioinformatics and statistics**
  - **Future iterative improvements of functional annotations and genome assembly**
  - **Reanalysis of growing datasets with new Qs, new approaches**
  - **Mathematical modelling for itegration/ ecological validation**
- **New frontiers *(species concept, interdisciplinary studies, population genomics, manipulative experiments, more spatiotemporal sampling)***

Stage of analysis	ADVANTAGES	LIMITATIONS
Sampling	No culture- or growth-related bias	Spatio-temporal heterogeneity
	Direct environmental sampling; large multi-species sampling; large multi-tissue sampling	Cost of representative or exhaustive sampling
	Analysis of complex experimental designs involving populations and communities	Careful **ecological assessment of environmental sampling** and of experimental designs
	Possible **long-term storage** of DNA, RNA, or protein samples	Availability of **reliable protocols for the extractions** of nucleic acids and proteins
Sequencing	High-throughput technologies for DNA, RNA and proteins	Possibilities of **sequencing bias**; poor sequencing of less-represented genomes
	Decreasing cost of sequencing and mass spectrometry	Cost of sequencing for large sample collections, in relation to the exhaustiveness of sampling
	Long-term public databases	Exponential increase of the amount of sequence data; cost and maintenance of database infrastructure
Information processing and functional analysis of organisms, communities and ecosystems	Biodiversity and phylogenetic analysis	Taxonomic **bias in databases**
	Functional profiling of naturally occurring organisms and communities	Assembly of short genomic fragments giving a partial view of organismal functional capacities
	Link function and diversity and answer the question 'who is doing what?	Functional bias in database; computational demand for bioinformatics analyses; poor quality of annotations and **amplification of annotation errors**
	Discovery **of novel ecologically relevant functions**	Functional inferences from genomics data in the absence of transcriptomic and/or proteomic data; biased conclusions on the basis of apparent absence of function
	Identifying links between diversity, functional changes and environmental variables	**Experimental bottleneck of functional characterization of new genes**
	Evolvability of genomics data analysis through improvement of annotations	Computational cost of re-annotating sequences
	Re-analysis of genomics data in the light of novel environmental data	**Comprehensive environment variable surveys**; environment variable databases; environment-dedicated bioinformatics tools; exponential increase of environmental data; **increased complexity of the comparison between environmental data and genomics data**
	Comparison of present-day ecosystem functioning with earth history and paleo-ecosystem functioning Combination of synchronic and diachronic analysis	
	Identifying links between diversity, functional changes and environmental variables	**Confusing the reality of ecosystem functioning with the reconstructed image from environmental genomics**

Box 2 Selection of bioinformatics tools for analysing environmental genomics data

The advent of massive sequencing methods is yielding increasing amounts of genomic data and metadata. Sophisticated statistical analysis using dedicated databases is necessary to interpret these data. Various algorithms and websites have been developed in order to bring the power of these statistical analyses within the reach of scientists that are not bioinformatics specialists. Such analytical tools will enhance understanding of ecosystems if they are incorporated into robust ecological and statistical frameworks.



Many bioinformatics tools have recently been published. Most of these publications describe algorithms that enhance the robustness of statistical analysis at all levels of characterisation (species richness, functional richness, frequency of gene families, taxonomic classification, *de novo* genome assembly). MEGAN V3 (Metagenome Analysis Software) (Mitra et al. 2009), IMG/M (Integrated Microbial Genomes) (Markowitz et al. 2008), MG-RAST (Meta Genome Rapid Annotation using Subsystem Technology) based on the SEED framework (Meyer et al. 2008) and CAMERA (Community Cyberinfrastructure for Advanced Marine Microbial Ecology Research and Analysis) (Seshadri et al. 2007) are the main bioinformatics resources available online.

① MEGAN V3.6, a freely available software, estimates the taxonomical content by analysing BLAST (Altschul et al. 1997) comparisons of a set of reads against one or more reference databases, such as NCBI (National Center for Biotechnology Information) or genome-specific databases. The comparative analysis of the functional content of metagenome datasets is performed by tools using Clusters of Orthologous Groups of Proteins or Gene Ontology.

② IMG/M consists of 65 microbial community samples that are part of 20 metagenome studies and 5062 publicly available single microbial genomes data. The efficiency of genome analysis by IMG/M stems from this comparative context. This data management system provides tools for analysing the functional capability of metagenome data generated by shotgun sequencing.

③ MG-RAST is a fully-automated server dedicated to the annotation of metagenomes. It provides sequence annotation, phylogenetic classification and metabolic reconstructions. It accepts the upload of files directly in the format delivered by 454 pyrosequencing systems. The pipeline includes 394 completed public metagenomes from the curated SEED database and comparison tools.

④ CAMERA includes environmental metagenomic and genomic sequence data (43 datasets), associated environmental parameters (metadata), and precomputed search results, in order to carry out cross-analysis of environmental samples through BLAST search.

MEGAN V3: http://www-ab.informatik.uni-tuebingen.de/software/megan
IMG/M: http://img.jgi.doe.gov/cgi-bin/m/main.cgi
MG-RAST: http://metagenomics.nmpdr.org/
CAMERA: http://camera.calit2.net

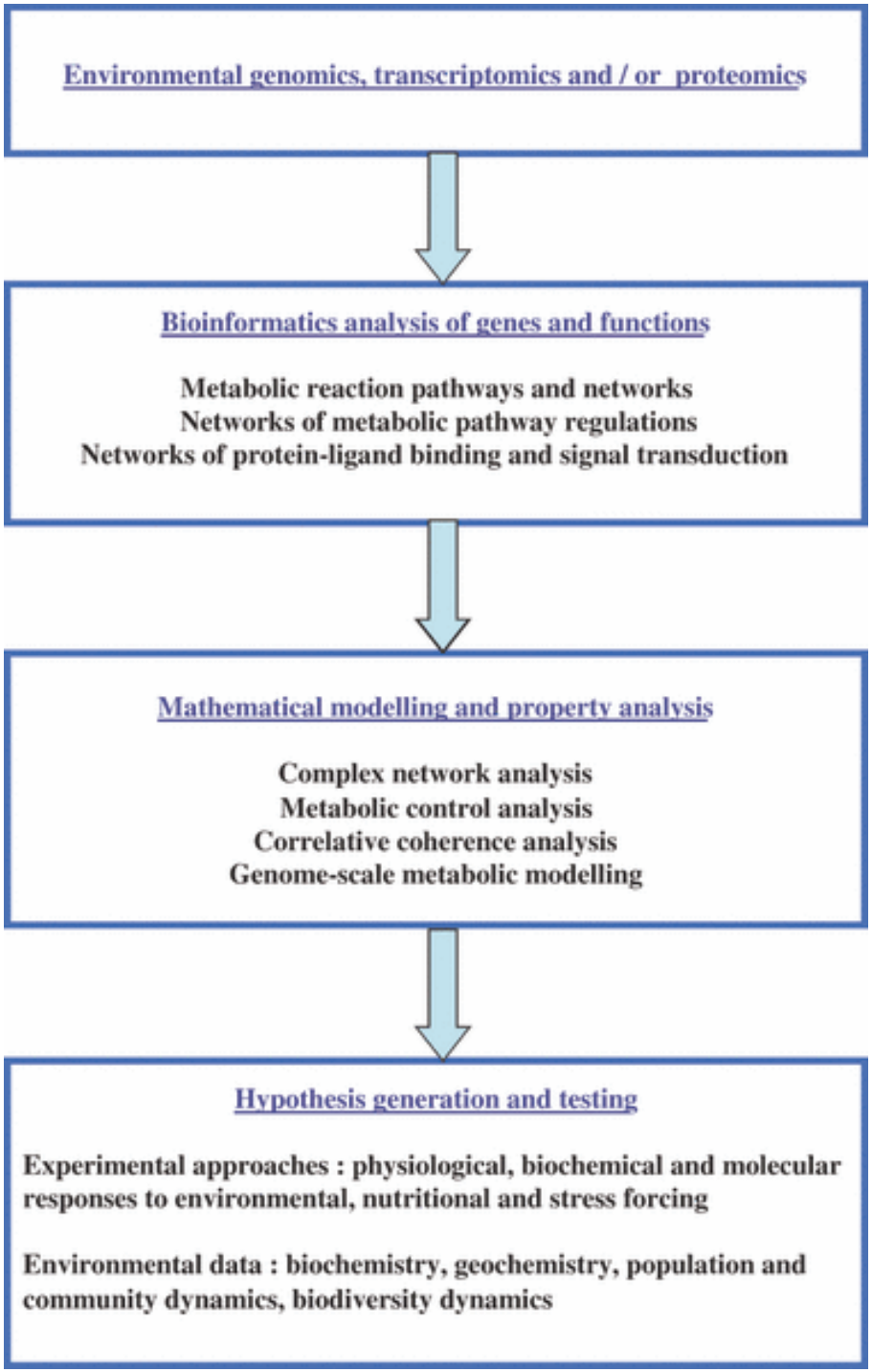


Figure 2. Mathematical modelling in environmental genomics analysis. Reconstructed networks from environmental genomics data (Box S2) can be analysed by various methods of mathematical modelling (Getz 2003; Feist *et al.* 2008; Westerhoff & Palsson 2008; Fuhrman 2009), that can assess and quantify their dynamic properties and generate hypotheses on community and ecosystem functioning. Hypothesis testing can then be carried out by experimental and environmental verification approaches, with the subsequent possibility of iterations between the different steps of the process. The main steps in this flowchart are derived from the description of the systems biology paradigm by Palsson (2006).