Molecular eco-systems biology: towards an understanding of community function

Jeroen Raes and Peer Bork

Published in final edited form as:
Nature Reviews Microbiology. 2008 Sep ; 6(9): 693-699 | doi: 10.1038/nrmicro1935
Nature Publishing Group (U.K.) ►

Final Draft

Molecular eco-systems biology: towards an understanding of community function

Jeroen Raes[1] *and Peer Bork*[1]

[1] European Molecular Biology Laboratory, Meyerhofstrasse 1, D-69117 Heidelberg, Germany

ABSTRACT | Systems-biology approaches, which are driven by genome sequencing and high-throughput functional genomics data, are revolutionizing single-cell-organism biology. With the advent of various high-throughput techniques that aim to characterize complete microbial ecosystems (metagenomics, meta-transcriptomics and meta-metabolomics), we propose that the time is ripe to consider molecular systems biology at the ecosystem level (eco-systems biology). Here, we discuss the necessary data types that are required to unite molecular microbiology and ecology to develop an understanding of community function and discuss the potential shortcomings of these approaches.

Over the past decade, the advent of robotics has enabled a paradigm shift in molecular biology: a change of emphasis from reductionistic approaches and 'single-protein' studies to global investigations of increasingly more complex systems of molecules and their interrelationships. These 'systems approaches' are used to investigate processes as a whole and enable models to be built to predict the behaviour of a system in response to various external cues, disturbances or modifications of its composition [1]. After ground-breaking work on the properties of small networks that consisted of a few genes, the wiring of complete cells and microbial organisms is now being investigated and modelled [2,3]. However, as free-living organisms constantly interact with each other and the environment, systems biologists are already looking towards the next big challenge — unravelling the complexity of complete ecosystems.

A microbial ecosystem can be defined as a system that consists of all the microorganisms that live in a certain area or niche and that function together in the context of the other biotic (plants and animals) and abiotic (temperature, chemical composition and structure of the surroundings) factors of that niche. Communities range from being simple (for example, one- or two-species-dominated bioreactors and biofilms that are growing on ore-mine effluents or medical implants) to complex (for example, symbiotic human gut flora, plant rhizospheres, soil communities and ocean dwelling or even airborne microorganisms, such as those present in clouds). The complexity of the interactions in ecosystems depends on the number of species and the population structure, variation in food and energy supply and the geography of the habitat [4]. Eco-systems biology seeks to understand, as a whole, the immensely complex set of molecular processes and interactions that contribute to ecosystem functioning — the total sum of ecosystem-level processes, such as matter, nutrient and energy cycling [5]. This understanding should ultimately lead to predictive modelling of ecosystems, allowing the *in silico* investigation of ecosystem properties. Important issues that could be addressed by an ecosystems approach include estimating the relative importance of ecosystem members in ecosystem functioning and productivity, the effect of nutrient availability on species composition or the resilience of the ecosystem to disturbances.

To be successful, however, any systems-biology study requires data on three important aspects of the system: the 'parts list'; the connectivity between the parts; and the placement of connectivity in the context of time and space [6,7]. Fig.1 shows the current status of these three data levels at various system scales. In single-organism systems biology, the parts list is generally established; almost 700 complete bacterial and archaeal genomes are available and some functional knowledge is available for approximately 70–80% of the encoded genes [8,9]. For several model organisms, large-scale efforts have determined the connectivity among the parts (the physical and genetic interactions between genes) [2]. This, together with an ever increasing amount of temporal, spatial and structural data, means that model microorganism systems biology is ready to enter the third phase and progress towards its final goal — the modelling and manipulation of complete organisms. The recent advent of several new large-scale technologies in microbial ecology, which have allowed high-throughput monitoring of genes (metagenomics), transcript and protein levels (meta-transcriptomics and meta-proteomics) and metabolites (meta-metabolomics) (Fig.1), are paving the way to an expansion of systems biology to the ecosystem level and are promising insights into these systems parts lists, connectivity and their temporal and spatial context at previously unforeseen scales. Here, we review these developments and assess how close we are to modelling complete microbial ecosystems.

Metagenomes provide the parts list

For several decades, ribosomal RNA studies have charted the species-level parts lists of environments [10]. However, unless the microorganisms that are identified can be cultured, their functional roles remain largely unknown. Functional assays of samples (for example, using BIOLOG plates to measure ecosystem substrate-usage phenotypes) can provide insights into some of the processes that occur in communities, but do not provide information on which community members are involved. Techniques such as RNA-based stable-isotope probing [11,12], fluorescence in situ hybridization (FISH)–microautoradiography [13], isotope arrays [14] and FISH–secondary-ion mass spectrometry [15] and its variants [16] allow substrate usage and specific processes to be linked to species, but are limited to particular substrates, are subject to cross-feeding, are not applicable in all environments and do not generally provide molecular details on the genes that are involved [17]. Environmental DNA cloning and screening enable specific ecosystem functions to be linked to genes [18], but such linking to bacterial or archaeal species is rare, and when successful, necessitates the co-cloning of a phylogenetic marker [19]. Novel techniques that are based on single-

cell isolation and simultaneous PCR of a phylogenetic marker with a functional gene of interest show promise [20,21], but have not yet been scaled up to high-throughput simultaneous analysis of a large number of genes or functions and still have sensitivity issues [22].

Environmental shotgun sequencing [23-27] has recently provided ecosystems biology with a possible global 'one-does-all' method. The random sampling of sequence data from the combined community members (the metagenome) provided a first unbiased and large-scale glimpse into the total molecular parts list of communities, and allowed the researchers (in theory) to simultaneously investigate genes, their functions and the individuals that exert them [23]. This promise has led researchers from all over the world to initiate metagenomic sequencing projects — more than 100 projects have been completed or are currently underway [28]. In addition, novel sequencing technologies with increasingly longer read lengths and the rapidly falling cost of sequencing will only expedite this process.

Metagenomic sequencing has so far added more than 10 billion bp to sequence databases [9,28]. The larger projects usually sequence approximately 50–100 Mb per environment, which should provide a firm foundation to start investigating the functioning of the underlying communities. However, this process is far from easy. Deriving ecosystem functioning from metagenomes requires careful sampling and DNA-extraction designs, followed by a considerable amount of far-from-trivial sequence-data processing (assembly and gene prediction on short reads), including the prior determination of a set of metagenome descriptors that describe the basic functional and phylogenetic composition of a sample [29] (Box 1; Fig.2). Unfortunately, these descriptors are also interlinked and are influenced by various biological and technical factors, and therefore yield a rich spectrum of pitfalls (for example, observed phylogenetic composition is dependent on sampling strategy and observed functional composition is dependent on sequence coverage and read length [29] (Fig.2)). In addition, the phylogenetic assignment of sequence reads, which is of paramount importance to the linking of molecular functions to species, remains a serious challenge in complex samples [23,29]. However, for most metagenomic samples, up to 75% of genes can be functionally characterized using targeted computational methodologies that combine homology and gene neighbourhood [8,9], and in simple communities, genes can be assigned to species (because complete genomes can be assembled), which means a parts list — the proteins, their function and their host organism — can be established. Given that more and more bioinformatics tools are being developed to analyse metagenomes and the standardization of data and analysis is being discussed (indispensable for comparison of independent studies [29]), it is likely that in the near future, metagenomic sequencing will provide a workable parts list for a large number of different ecosystems.

Part lists to ecosystem properties

If a parts list has been generated using sufficient sequence coverage and in a reasonably unbiased way (Box 1), several basic ecosystem properties can be derived that should help to characterize the microbial community in the sample. Here, we delineate some standard properties that are used in (microbial) ecology and propose possible metrics that are easily obtainable from the raw sequences.

Community structure: species richness, evenness and diversity

Calculations can be made using rarefaction approaches that are based on 16S sampling in conjunction with metagenomic sequencing. Alternatively, if an average genome size is known or predicted [30], these metrics can be predicted from assembly statistics [31].

Functional potential or breadth of the community: COG richness

Calculations can be made by rarefaction of COG (clusters of orthologous group) counts on randomly sampled reads from the environment.

Global functional complementarity between community members: COG richness per genome equivalent

Can be calculated using COG richness and effective genome size [30], and should be a measure that correlates with the amount of within-community functional overlap.

Adding connectivity to parts lists

The measurements discussed above, which are based on the parts lists, allow a first glimpse into ecosystem functioning and structure. However, for a full systems approach, the more detailed wiring between the parts list needs to be deduced. Assuming we have a reasonable molecular parts list for an ecosystem, can we investigate the connectivity between its members? In cellular systems, connectivity refers to protein–protein interactions and modifications (such as phosphorylation), substrate and end-product transfer and regulatory interactions. In ecosystems, this concept encompasses an even wider range of interactions at various levels. These include ecological interactions between the carriers of function (organisms), such as competition, predation and structural interactions (such as mat formation). Many of these processes also have a molecular basis; for example, through direct cell–cell attachment [32] or though communication using various signalling molecules that bind specific receptors and therefore activate signalling pathways and instigate various forms of behaviour [33, 34]. Also included is metabolic cooperation, in which the interaction is based on a sometimes mutual exchange of metabolites, such as the biogeochemical cycling of elements, nutrients and electrons or coordinated breakdown of complex polymers by multiple organisms. At the molecular level, this refers to the presence of complementary pathways in different organisms and the active or passive transport of metabolites in and out of the cell. As many of these processes are also linked to the abundance of organisms (for example, quorum sensing [33]) the nature and presence of these molecular interactions are highly dependent on population structure, which might vary over time (discussed in the next section). In addition, in many environments, the physical and

geographical heterogeneity of the local habitat can determine the interactions that are possible [35].

Given this complexity of cellular interactions, which are analogous to those of multicellular organisms [36], the reconstruction of ecosystem-wide molecular networks will be far from trivial. However, as more relevant data have become available, some aspects of cooperative molecular networks can already be derived.

Data sources to probe connectivity

Metabolic cooperation

Metabolic cooperation has historically been studied using co-culture experiments, in which synergistic relationships between different strains are observed. This synergy has been found to occur by the transfer of intermediate metabolites (for example, the degradation of glucose through acetate to methane by *Acetobacterium woodii* and *Methanosarcina barkeri*), the transfer of electron carriers (for example, hydrogen and formate) or the removal of limiting by-products (for example, methanol and oxygen; reviewed in Ref. 37). However, to understand and model cooperation in complex, natural communities, the co-culture approach needs to be replaced by a more high-throughput and systematic approach that will allow chemical and microscopic monitoring of the various players in an ecosystem under a range of perturbations.

Alternatively, genome-content analysis of ecosystem members could be used to infer metabolic cooperation. When the complexity of the ecosystem is low, metagenomic sequencing can yield enough coverage to complete the genomes of the most dominant members of the community. Having complete genomes enables the examination of patterns of metabolic complementarity between organisms and the proposal of hypotheses of cooperation between community members. For example, the reconstruction of several members of an acid-mine drainage sample allowed the authors to propose that one member, *Leptospirillum* sp. group III, could be a cornerstone species that carries out the fixation of nitrogen for the other community members that lack these pathways [26]. In an analysis of four gut symbionts in a gutless worm, evidence was found for syntrophic cycling of sulphate and sulphide (or other intermediate sulphur compounds) between gammaproteobacterial and deltaproteobacterial symbionts and for the existence of additional hydrogen syntrophy [38]. Finally, metabolic network reconstruction from the genome analysis of two endosymbionts (*Candidatus Baumannia cicadellinicola* and *Candidatus Sulcia muelleri*) in a sharpshooter revealed not only their role in providing nutrients, vitamins and cofactors to the host, but also their extensive metabolic cooperation — for example, in amino acid biosynthesis and complementing genome reduction between the two bacteria [39,40]. Using computational techniques, such as flux-balance analysis, simple *in silico* models of metabolic cooperation can be built that are based on genomic data. These models can be predictive of growth and metabolic fluxes and allow insights into synergistic reactions [41].

When ecosystem complexity is high, however, current coverage of metagenomic sequencing is inadequate. For example, to obtain eight times the coverage of only the most dominant member in the Minnesota soil metagenome, approximately 2–5 Gbp would need to be sequenced (in the pilot study, only 120 Mb were generated [25]). Such a sequencing effort is not impossible (for example, 6.3 Gbp were sequenced in the Global Ocean Sampling survey [42]), but has not been achieved as of yet for a single sample. With the rapidly dropping sequencing costs, metagenomic sequencing could remain a hypothesis-generating tool for ecosystem-wide network reconstruction. However, current developments in cultivation methodologies [43] and DNA amplification from single cells [44], combined with high-throughput cell sorting, will probably push single-genome sequencing to environmental scales and will allow us to avoid some of the weaknesses of metagenomics data [45]. Single-genome sequencing is therefore likely to be the input data source of choice for this type of analysis in the near future [46-49].

Many examples of the collaboration and cooperation of microorganisms, including the formation of complex consortia or biofilms and cell–cell communication, seem to occur at micrometre-range distances in open systems [50]. Therefore, analogous to protein complexes in cellular systems biology, the observation of physical cell–cell interactions between organisms provides strong indications of functional interactions in the ecosystem network. This is especially true when evidence for interactions, for example, from FISH microscopy, is combined with chemical measurements of metabolized compounds [51]. Although microscopy data is scattered and no high-throughput approaches to detect cell–cell interactions (an 'ecosystems yeast two-hybrid' at the cellular level) have yet been described, ongoing advances in high-throughput and three-dimensional microscopy, combined with automated image-analysis techniques should allow data to be gathered on a larger scale [52]. Until then, indirect measures might provide a solution. For example, investigating taxon occurrence patterns could provide signs of metabolic cooperation. Indeed, studies of both macroorganisms and microorganisms have indicated clear non-random distribution patterns [53,54]. However, as other factors, such as competition, niche (species-composition cycles and biogeography can be predicted from habitat parameters [55] or organismal physiological traits [56]) and sampling, might also contribute to the patterns observed, further studies will need to show what information can be extracted from such data.

Cell–cell signalling — communication and quorum sensing

Evidence is accumulating that microorganisms do not live as isolated individuals, but as populations of cells that are continuously producing, sensing and responding to chemical signals, which allows them to communicate and cooperate. The best studied of these processes is quorum sensing, a process in which bacteria can 'measure' the cell density of their population to initiate processes such as bioluminescence, biofilm formation, sporulation and virulence [33,34]. Inter-species communication is less understood, although the discovery of more examples of this phenomenon has strengthened the general notion that these processes are more ubiquitous than previously thought [33,57]. These observations herald the exciting prospect of reconstructing the various inter-species small-molecule-based signalling cascades that drive social behaviour in environments. Although the data are currently too fragmented to be used in a global systems

approach, the modelling of specific processes could constitute a proof-of-principle case study. To include this aspect of microbial interactions in global ecosystems biology, an integrated effort is needed to detect both the production of, and the response to, the plethora of small molecules that are produced by these organisms. Environmental metabolomics approaches, combined with metagenomic, meta-transcriptomic [58] and meta-proteomic [59,60] data, should eventually allow the reconstruction of ecosystem-wide combined protein small-molecule networks, similar to those that have been achieved for single organisms (see Further information for a link to the STITCH chemical–protein interactions resource [61]). This approach could ultimately result in the molecular modelling of community multicellular behaviour types other than quorum sensing, such as dispersal, nutrient acquisition and biofilm formation [62].

Spatial and temporal variation

Previous studies have detected variation in species composition in various habitats, both spatially (reviewed, for example, in Refs [35,63-65]) and temporally (for example, Refs [66,67]). Spatio–temporal variation has been linked to variation in environmental conditions [35,68], even to the point at which environmental parameters can be predictive of species composition [55]. Similar spatio–temporal variation has been observed from a functional point of view [69]. Comparative metagenomics approaches [24,25,29] recently charted the molecular basis of spatial functional variation of environments from the kilometre [25,42,70] to centimetre [38] and even millimetre scale [71] (Fig.3), and with time-series metagenomics studies underway [28], studying temporal (and spatio–temporal [72]) aspects should become possible at the molecular level. The recent development of phylochips, metagenome-based microarrays and high-throughput sequencing-based monitoring will further expedite the amount of dynamic data that is available (for example, Refs [73-77]; reviewed in Refs [78,79]). By providing a quick and cheap read-out of variation in species content and molecular function in environments, these techniques will allow the simultaneous discovery of new genes and species that are involved in specific processes (for example, from ecosystem perturbation experiments) or linked to environmental conditions (for example, from seasonal time series). Therefore, these advances lay the foundations to investigate the dynamic nature of molecular ecosystem networks in time and space.

Conclusions

Many datasets that will facilitate ecosystems biology are now being gathered. Metagenomics studies are collating the parts lists from which some general ecosystem properties, as well as first insights into metabolic cooperation, can be extracted. Other technologies that will gather additional, complementary data types, such as the environmental counterpart of high-throughput functional genomics (a cornerstone of cellular systems biology), are still in their infancy. However, technologies such as large-scale automated monitoring of chemicals and meta-metabolomics are developing rapidly. The interpretation and integration of these data will be challenging and will necessitate the development of novel computational approaches [29], but these challenges will be overcome. Thus, the reconstruction of larger ecological networks at the molecular level will become feasible. Integration of these molecular networks into the vast body of macro-ecological theory should lead to a more thorough understanding of the wiring of the main biological systems on the Earth.

Databases

Candidatus Baumannia cicadellinicola:
http://www.ncbi.nlm.nih.gov/sites/entrez?db=genomeprj&cmd=search&term=Baumannia%20cicadellinicola

Candidatus Sulcia muelleri:
http://www.ncbi.nlm.nih.gov/sites/entrez?db=genomeprj&cmd=Retrieve&dopt=Overview&list_uids=19805

Methanosarcina barkeri:
http://www.ncbi.nlm.nih.gov/sites/entrez?db=genomeprj&cmd=search&term=Methanosarcina%20barkeri

Further information

Jeroen Raes's homepage:
http://www.embl.de/~raes/

Peer Bork's homepage:
http://www.bork.embl.de/j/

STITCH chemical–protein interactions:
http://stitch.embl.de/

Acknowledgements

The authors thank L. Jensen, A. Singh and other members of the Bork group for valuable comments. The author's research is funded by the FP7 programme (grant number HEALTH-F4-2007-201,052).

Corresponding Author

Jeroen Raes and Peer Bork are at the European Molecular Biology Laboratory, Meyerhofstrasse 1, D-69117 Heidelberg, Germany. Email: raes@embl.de Email: bork@embl.de

References

1. Bork, P. Is there biological research beyond Systems Biology? A comparative analysis of terms. Mol. Syst. Biol. 1, 2005.0012 (2005).
2. Joyce, A. R. & Palsson, B. O. The model organism as a system: integrating 'omics' data sets. Nature Rev. Mol. Cell Biol. 7, 198–210 (2006).
3. Kitano, H. Systems biology: a brief overview. Science 295, 1662–1664 (2002).
4. McMahon, K. D., Martin, H. G. & Hugenholtz, P. Integrating ecology into biotechnology. Curr. Opin. Biotechnol. 18, 287–292 (2007).
5. Azam, F. & Worden, A. Z. Oceanography. Microbes, molecules, and marine ecosystems. Science 303, 1622–1624 (2004).
6. Bork, P. & Serrano, L. Towards cellular systems in 4D. Cell 121, 507–509 (2005).

7. Palsson, B. Two-dimensional annotation of genomes. Nature Biotechnol. 22, 1218–1219 (2004).
8. Harrington, E. D. *et al.* Quantitative assessment of protein function prediction from metagenomics shotgun sequences. Proc. Natl Acad. Sci. USA 104, 13913–13918 (2007).
9. Raes, J., Harrington, E. D., Singh, A. H. & Bork, P. Protein function space: viewing the limits or limited by our view? Curr. Opin. Struct. Biol. 17, 362–369 (2007).
10. Pace, N. R. A molecular view of microbial diversity and the biosphere. Science 276, 734–740 (1997).
11. Lu, Y. & Conrad, R. *In situ* stable isotope probing of methanogenic archaea in the rice rhizosphere. Science 309, 1088–1090 (2005).
12. Whiteley, A. S., Manefield, M. & Lueders, T. Unlocking the 'microbial black box' using RNA-based stable isotope probing technologies. Curr. Opin. Biotechnol. 17, 67–71 (2006).
13. O'Donnell, A. G., Young, I. M., Rushton, S. P., Shirley, M. D. & Crawford, J. W. Visualization, modelling and prediction in soil microbiology. Nature Rev. Microbiol. 5, 689–699 (2007).
14. Adamczyk, J. *et al.* The isotope array, a new tool that employs substrate-mediated labeling of rRNA for determination of microbial community structure and function. Appl. Environ. Microbiol. 69, 6875–6887 (2003).
15. Orphan, V. J., House, C. H., Hinrichs, K. U., McKeegan, K. D. & DeLong, E. F. Methane-consuming archaea revealed by directly coupled isotopic and phylogenetic analysis. Science 293 484–487 (2001).
16. Kuypers, M. M. & Jorgensen, B. B. The future of single-cell environmental microbiology. Environ. Microbiol. 9, 6–7 (2007).
17. Neufeld, J. D. & Murrell, J. C. Witnessing the last supper of uncultivated microbial cells with Raman–FISH. ISME J. 1, 269–270 (2007).
18. Handelsman, J. Metagenomics: application of genomics to uncultured microorganisms. Microbiol. Mol. Biol. Rev. 68, 669–685 (2004).
19. Beja, O. *et al.* Bacterial rhodopsin: evidence for a new type of phototrophy in the sea. Science 289, 1902–1906 (2000).
20. Ottesen, E. A., Hong, J. W., Quake, S. R. & Leadbetter, J. R. Microfluidic digital PCR enables multigene analysis of individual environmental bacteria. Science 314, 1464–1467 (2006).
21. Stepanauskas, R. & Sieracki, M. E. Matching phylogeny and metabolism in the uncultured marine bacteria, one cell at a time. Proc. Natl Acad. Sci. USA 104, 9052–9057 (2007).
22. Lau, S. C. & Liu, W. T. Recent advances in molecular techniques for the detection of phylogenetic markers and functional genes in microbial communities. FEMS Microbiol. Lett. 275, 183–190 (2007).
23. Eisen, J. A. Environmental shotgun sequencing: its potential and challenges for studying the hidden world of microbes. PLoS Biol. 5, e82 (2007).
24. Tringe, S. G. & Rubin, E. M. Metagenomics: DNA sequencing of environmental samples. Nature Rev. Genet. 6, 805–814 (2005).
25. Tringe, S. G. *et al.* Comparative metagenomics of microbial communities. Science 308, 554–557 (2005).
26. Tyson, G. W. *et al.* Community structure and metabolism through reconstruction of microbial genomes from the environment. Nature 428, 37–43 (2004).
27. Venter, J. C. *et al.* Environmental genome shotgun sequencing of the Sargasso Sea. Science 304, 66–74 (2004).
28. Liolios, K., Tavernarakis, N., Hugenholtz, P. & Kyrpides, N. C. The Genomes On Line Database (GOLD) v.2: a monitor of genome projects worldwide. Nucleic Acids Res. 34, D332–D334 (2006).
29. Raes, J., Foerstner, K. U. & Bork, P. Get the most out of your metagenome: computational analysis of environmental sequence data. Curr. Opin. Microbiol. 10, 490–498 (2007).
30. Raes, J., Korbel, J. O., Lercher, M. J., von Mering, C. & Bork, P. Prediction of effective genome size in metagenomic samples. Genome Biol. 8, R10 (2007).
31. Angly, F. *et al.* PHACCS, an online tool for estimating the structure and diversity of uncultured viral communities using metagenomic information. BMC Bioinformatics 6, 41 (2005).
32. O'Toole, G., Kaplan, H. B. & Kolter, R. Biofilm formation as microbial development. Annu. Rev. Microbiol. 54, 49–79 (2000).
33. Bassler, B. L. & Losick, R. Bacterially speaking. Cell 125, 237–246 (2006).
34. Keller, L. & Surette, M. G. Communication in bacteria: an ecological and evolutionary perspective. Nature Rev. Microbiol. 4, 249–258 (2006).
35. Martiny, J. B. *et al.* Microbial biogeography: putting microorganisms on the map. Nature Rev. Microbiol. 4, 102–112 (2006).
36. Shapiro, J. A. Thinking about bacterial populations as multicellular organisms. Annu. Rev. Microbiol. 52, 81–104 (1998).
37. Schink, B. Synergistic interactions in the microbial world. Antonie Van Leeuwenhoek 81, 257–261 (2002).
38. Woyke, T. *et al.* Symbiosis insights through metagenomic analysis of a microbial consortium. Nature 443, 950–955 (2006).
39. McCutcheon, J. P. & Moran, N. A. Parallel genomic evolution and metabolic interdependence in an ancient symbiosis. Proc. Natl Acad. Sci. USA 104, 19392–19397 (2007).
40. Wu, D. *et al.* Metabolic complementarity and genomics of the dual bacterial symbiosis of sharpshooters. PLoS Biol. 4, e188 (2006).
41. Stolyar, S. *et al.* Metabolic modeling of a mutualistic microbial community. Mol. Syst. Biol. 3, 92 (2007).
42. Rusch, D. B. *et al.* The Sorcerer II Global Ocean Sampling expedition: northwest Atlantic through eastern tropical Pacific. PLoS Biol. 5, e77 (2007).
43. Giovannoni, S. & Stingl, U. The importance of culturing bacterioplankton in the 'omics' age. Nature Rev. Microbiol. 5, 820–826 (2007).
44. Zhang, K. *et al.* Sequencing genomes from single cells by polymerase cloning. Nature Biotechnol. 24, 680–686 (2006).
45. Steward, G. F. & Rappe, M. S. What's the 'meta' with metagenomics? ISME J. 1, 100–102 (2007).
46. Hutchison, C. A. & Venter, J. C. Single-cell genomics. Nature Biotechnol. 24, 657–658 (2006).
47. Dethlefsen, L. & Relman, D. A. The importance of individuals and scale: moving towards single cell microbiology. Environ. Microbiol. 9, 8–10 (2007).
48. Marcy, Y. *et al.* Dissecting biological "dark matter" with single-cell genetic analysis of rare and uncultivated TM7 microbes from the human mouth. Proc. Natl Acad. Sci. USA 104, 11889–11894 (2007).

49. Ochman, H. Single-cell genomics. Environ. Microbiol. 9, 7 (2007).

50. Battin, T. J. *et al.* Microbial landscapes: new paths to biofilm research. Nature Rev. Microbiol. 5, 76–81 (2007).

51. Boetius, A. *et al.* A marine microbial consortium apparently mediating anaerobic oxidation of methane. Nature 407, 623–626 (2000).

52. Pepperkok, R. & Ellenberg, J. High-throughput fluorescence microscopy for systems biology. Nature Rev. Mol. Cell Biol. 7, 690–696 (2006).

53. Gotelli, N. J. & McCabe, D. J. Species co-occurrence: a meta-analysis of JM Diamond's assembly rules model. Ecology 83, 2091–2096 (2002).

54. Horner-Devine, M. C. *et al.* A comparison of taxon co-occurrence patterns for macro- and microorganisms. Ecology 88, 1345–1353 (2007).

55. Fuhrman, J. A. *et al.* Annually reoccurring bacterial communities are predictable from ocean conditions. Proc. Natl Acad. Sci. USA 103, 13104–13109 (2006).

56. Follows, M. J., Dutkiewicz, S., Grant, S. & Chisholm, S. W. Emergent biogeography of microbial communities in a model ocean. Science 315, 1843–1846 (2007).

57. Wingreen, N. S. & Levin, S. A. Cooperation among microorganisms. PLoS Biol. 4, e299 (2006).

58. Bailly, J. *et al.* Soil eukaryotic functional diversity, a metatranscriptomic approach. ISME J. 1, 632–642 (2007).

59. Lo, I. *et al.* Strain-resolved community proteomics reveals recombining genomes of acidophilic bacteria. Nature 446, 537–541 (2007).

60. Ram, R. J. Community proteomics of a natural microbial biofilm. Science 308, 1915–1920 (2005).

61. Kuhn, M., Campillos, M., von Mering, C., Jensen, L. J. & Bork P. STITCH: interaction networks of chemicals and proteins. Nucleic Acids Res. 36, D684–D688 (2008).

62. West, S. A., Griffin, A. S., Gardner, A. & Diggle, S. P. Social evolution theory for microorganisms. Nature Rev. Microbiol. 4, 597–607 (2006).

63. Bell, T. Larger islands house more bacterial taxa. Science 308,1884 (2005).

64. Green, J. L. Spatial scaling of microbial eukaryote diversity. Nature 432, 747–750 (2004).

65. Horner-Devine, M. C., Lage, M., Hughes, J. B. & Bohannan, B. J. A taxa-area relationship for bacteria. Nature 432, 750–753 (2004).

66. Palmer, C., Bik, E. M., Digiulio, D. B., Relman, D. A. & Brown, P. O. Development of the human infant intestinal microbiota. PLoS Biol. 5, e177 (2007).

67. Thompson, J. R. Genotypic diversity within a natural coastal bacterioplankton population. Science 307, 1311–1313 (2005).

68. Lozupone, C. A. & Knight, R. Global patterns in bacterial diversity. Proc. Natl Acad. Sci. USA 104, 11436–11440 (2007).

69. Torsvik, V. & Ovreas, L. Microbial diversity and function in soil: from genes to ecosystems. Curr. Opin. Microbiol. 5, 240–245 (2002).

70. DeLong, E. F. *et al.* Community genomics among stratified microbial assemblages in the ocean's interior. Science 311, 496–503 (2006).

71. Kunin, V. *et al.* Millimeter-scale genetic gradients and community-level molecular convergence in a hypersaline microbial mat. Mol. Syst. Biol. 4, 198 (2008).

72. Seymour, J. R., Mitchell, J. G., Pearson, L. & Waters, R. L. Heterogeneity in bacterioplankton abundance from 4.5 millimetre resolution sampling. Aquat. Microb. Ecol. 22, 143–153 (2000).

73. He, Z. *et al.* GeoChip: a comprehensive microarray for investigating biogeochemical, ecological and environmental processes. ISME J. 1, 67–77 (2007).

74. Moisander, P. H. Application of a *nifH* oligonucleotide microarray for profiling diversity of N_2-fixing microorganisms in marine microbial mats. Environ. Microbiol. 8, 1721–1735 (2006).

75. Palmer, C. Rapid quantitative profiling of complex microbial populations. Nucleic Acids Res. 34, e5 (2006).

76. Polz, M. F., Bertilsson, S., Acinas, S. G. & Hunt, D. A(r)Ray of hope in analysis of the function and diversity of microbial communities. Biol. Bull. 204, 196–199 (2003).

77. Rich, V. I., Konstantinidis, K. & Delong, E. F. Design and testing of 'genome-proxy' microarrays to profile marine microbial communities. Environ. Microbiol. 10, 506–521 (2008).

78. Gentry, T. J., Wickham, G. S., Schadt, C. W., He, Z. & Zhou, J. Microarray applications in microbial ecology research. Microb. Ecol. 52, 159–175 (2006).

79. Wold, B. & Myers, R. M. Sequence census methods for functional genomics. Nature Methods 5, 19–21 (2008).

80. Committee on metagenomics: challenges and functional applications. The New Science of Metagenomics: Revealing the Secrets of our Microbial Planet (The National Academies, Washington DC, 2007).

81. Field, D. *et al.* The minimum information about a genome sequence (MIGS) specification. Nature Biotechnol. 26, 541–547 (2008).

82. Markowitz, V. M. *et al.* An experimental metagenome data management and analysis system. Bioinformatics 22, e359–e367 (2006).

83. Seshadri, R., Kravitz, S. A., Smarr, L., Gilna, P. & Frazier, M. CAMERA: a community resource for metagenomics. PLoS Biol. 5, e75 (2007).

84. Kanehisa, M. From genomics to chemical genomics: new developments in KEGG. Nucleic Acids Res. 34, D354–D357 (2006).

85. Letunic, I., Yamada, T., Kanehisa, M. & Bork, P. iPath: interactive exploration of biochemical pathways and networks. Trends Biochem. Sci. 33, 101–103 (2008).

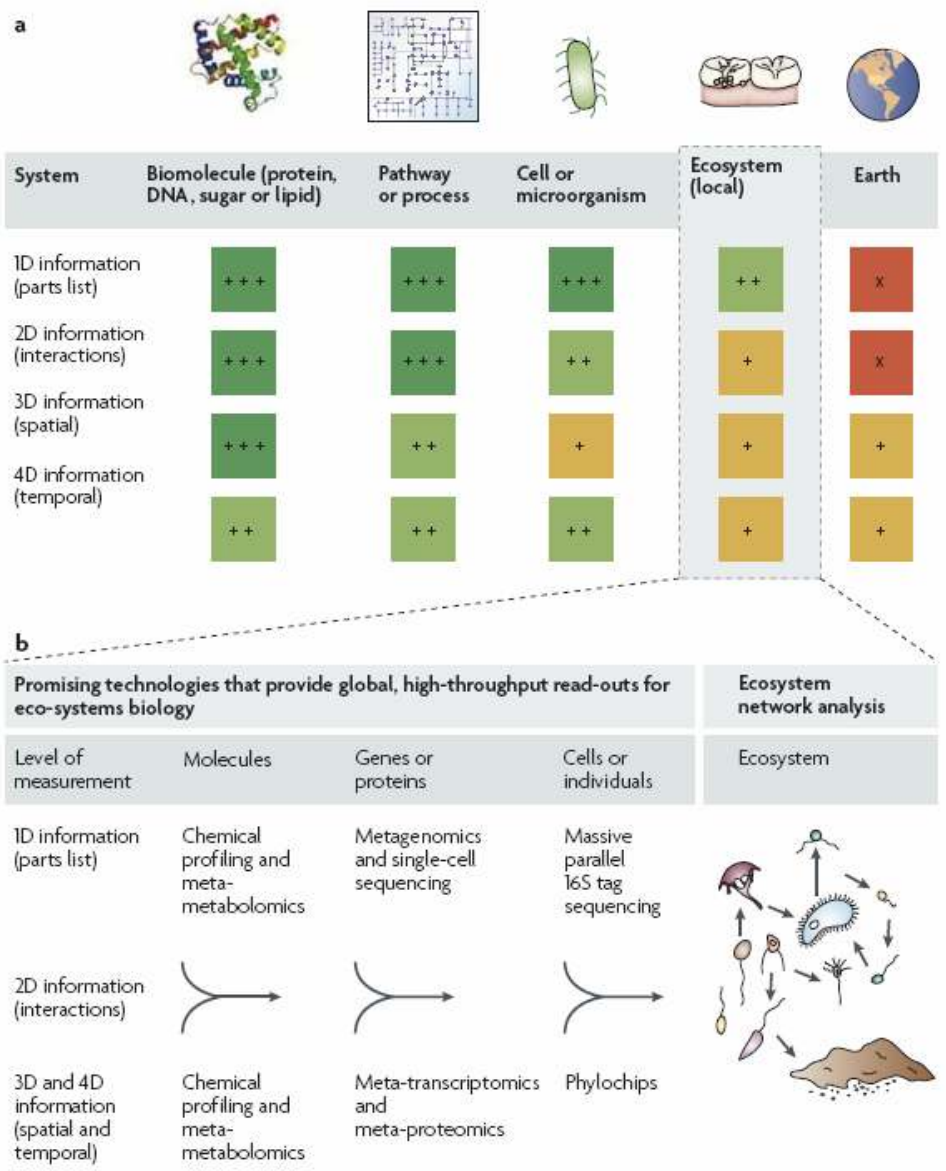

Figure 1. Systems biology: from proteins to environments. a | Different spatial scales at which systems biology can be performed (based on the 'dimension' definitions in Ref. [6]). The columns show data availability for each scale and the rows indicate the aspect of the system that is targeted by the data (+++, ample data available and good knowledge of the system aspect; ++, a number of high-throughput data sets available and fair knowledge of the system aspect, but more data are still needed to build comprehensive models; +, a few scattered non-high-throughput data sets available and model building is restricted to case studies; x, almost no data available). **b** | At the ecosystem scale, read-outs are available at different levels: molecules (ranging from trace elements to small signalling compounds to metabolism intermediates), genes or proteins, and cells or individuals. Here, we show some of the more promising high-throughput approaches to the generation of data that would facilitate eco-systems biology. No high-throughput tools are currently available that can map interactions, and this information will need to be inferred from other data sources (see the main text).

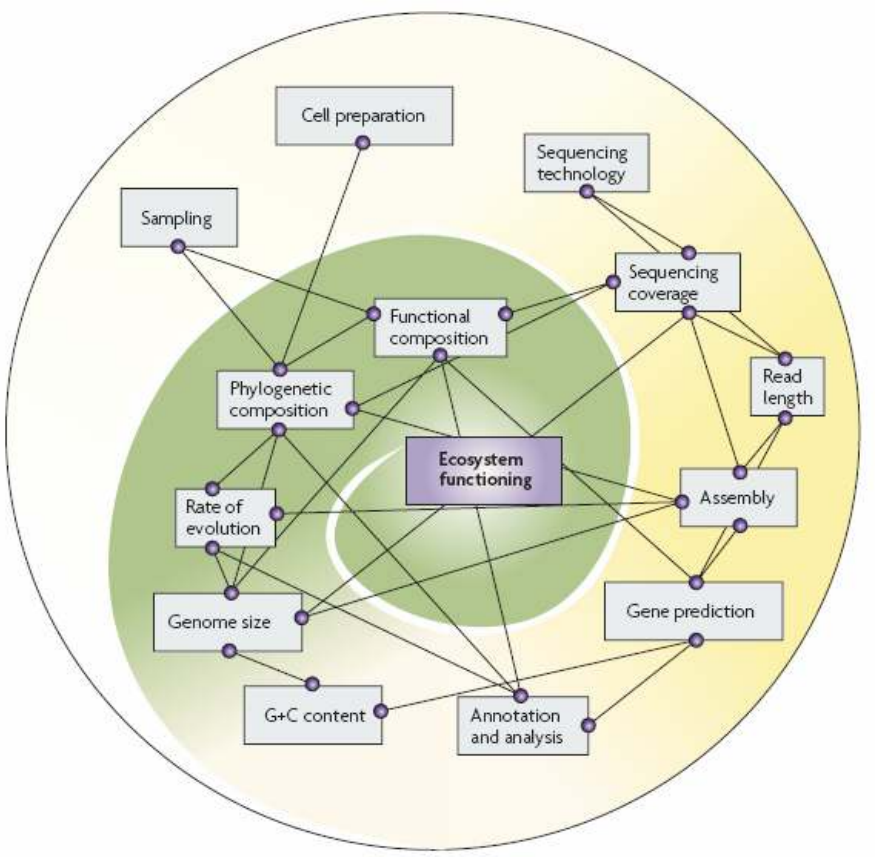

Figure 2. From metagenomes to ecosystem functioning: influencing factors and hidden dependencies. An overview of the factors that are required to analyse metagenomes and understand the molecular basis of ecosystem functioning (the total sum of ecosystem-level processes, such as matter, nutrient and energy cycling). Lines between factors indicate interdependencies (for example, perceived ecosystem functional composition depends on the functional annotation of genes and sampling protocol influences the observed phylogenetic composition; reviewed in Ref. 29). All these factors must be assessed when analysing a particular metagenome and it should be noted that all the factors are interrelated, which is important to our understanding of ecosystem functioning.

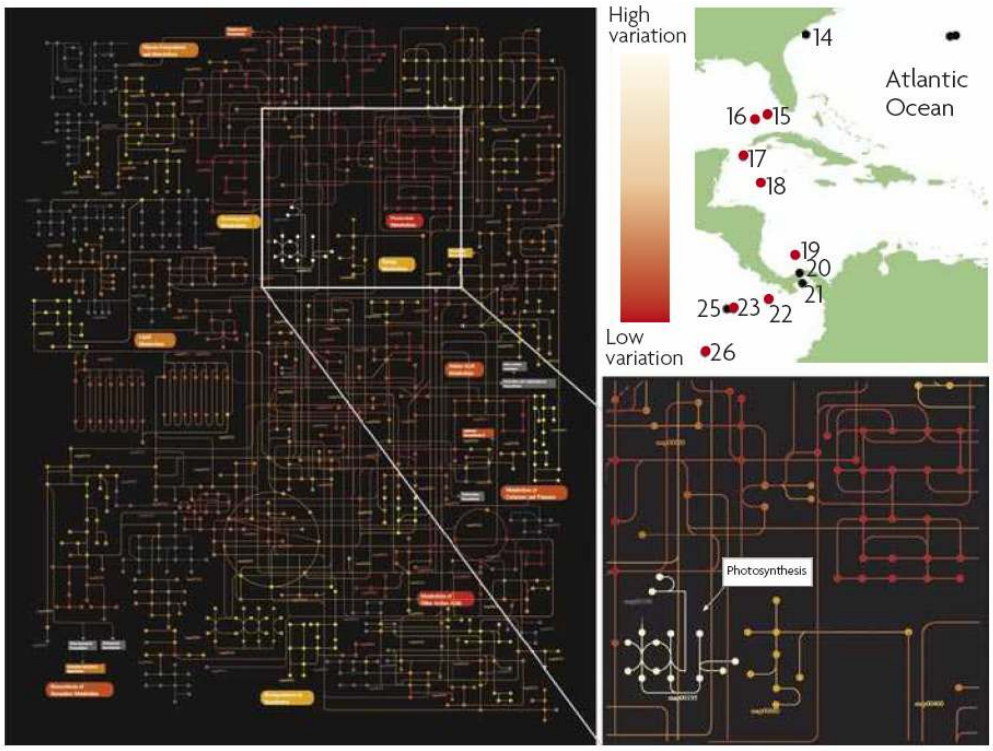

Figure 3. Visualizing complex environmental patterns. Novel visualization techniques will be needed to describe complex data and patterns. The example shown here is a summary of the metabolic variation along a longitudinal transect of ocean surface water samples (data from Ref. [42]; the samples used (red) were selected for similarity in habitat type). Colour intensity shows the contribution to the overall variance for different KEGG [84] maps that are involved in central metabolism (for example, red indicates maps with low contribution and therefore low variability over sites, whereas yellow indicates maps with high contribution) and grey indicates no significant KEGG mapping for these samples. The inset shows the large contribution of photosynthesis to the overall functional variation among samples.

Box 1. Generating representative metagenomics data.

To maximize the information content of environmental sequencing projects and allow effective post-analysis comparisons, the guidelines provided here could prove helpful. It should be noted, however, that these ideal-case-scenario guidelines might be subject to project (for example, financial and logistical) constraints.

Detailed sampling-methodology description and meta-data recording
To correctly interpret and compare metagenomics projects, an exact description of how the sample was taken is paramount (for example, filtering, enrichment procedures and DNA extraction). As much additional data about the sample as possible should also be recorded. This could range from the exact geographical location (for example, longitude, latitude, depth, height, time or date) to biochemical habitat measurements (for example, pH, levels of oxygen, phosphate or nitrate, and salinity) to patient information (for example, gender, age and disease or nutritional state) [29,80]. The 'Minimum Information about a Metagenome Sequence' specification should allow this information to be captured [81].

Sufficient coverage
A pilot study of the environment using rarefaction approaches should allow an estimation of its phylogenetic and functional complexity [29]. This could then be used to estimate the amount of sequencing that is required for the dataset to be representative.

Variability assessment
Ideally, multiple samples should be taken at the same site, at different time points or under different conditions to allow the biological variation at the site to be determined. Experimental variability should also be investigated [80].

Transparent and complete description of data treatment
Full details of computational data treatment should be provided for reproducibility, to assess the presence of data-treatment artefacts in functional conclusions and enable comparative metagenomics (for example, on assembly, gene calling and functional annotation) [29].

Reporting of a minimal set of metagenomic analyses and descriptors
To allow proper interpretation, post-analysis and comparison of independent samples and projects, a standardized set of minimal metagenome analyses and descriptors was proposed (MINIMESS [29]). Providing these data together with the raw data should allow those researchers who do not have access to bioinformatics resources to make optimal use of results.

Public availability
As for any sequencing project, data should be released to the general public. In addition to depositing assembled contigs in GenBank, the European Molecular Biology Laboratory (EMBL) and the DNA Data Bank of Japan (DDBJ), the raw reads should also be made available through the National Center for Biotechnology Information (NCBI) and European Bioinformatics Institute (EBI) trace archive. Other resources, such as CAMERA and IMG/M [82,83], allow further meta-data and analyses to be linked to deposited sequences.