Full text provided by www.sciencedirect.com

Get the most out of your metagenome: computational analysis of environmental sequence data

Jeroen Raes, Konrad Ulrich Foerstner and Peer Bork

New advances in sequencing technologies bring random shotgun sequencing of ecosystems within reach of smaller labs, but the complexity of metagenomics data can be overwhelming. Recently, many novel computational tools have been developed to unravel ecosystem properties starting from fragmented sequences. In addition, the so-called 'comparative metagenomics' approaches have allowed the discovery of specific genomic and community adaptations to environmental factors. However, many of the parameters extracted from these data to describe the environment at hand (e.g. genomic features, functional complement, phylogenetic composition) are interdependent and influenced by technical aspects of sample preparation and data treatment, leading to various pitfalls during analysis. To avoid this and complement existing initiatives in data standards, we propose a minimal standard for metagenomics data analysis ('MINIMESS') to be able to take full advantage of the power of comparative metagenomics in understanding microbial life on earth.

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

Corresponding author: Bork, Peer (bork@embl.de)

Current Opinion in Microbiology 2007, **10**:490–498

This review comes from a themed issue on
Genomics
Edited by Claire M. Fraser-Liggett and Jean Weissenbach

Available online 23rd October 2007

1369-5274/$ – see front matter

DOI 10.1016/j.mib.2007.09.001

Introduction

Since the first publications of large-scale environmental shotgun sequencing projects [1–3], we witness exponentially increasing efforts to investigate the genetic basis of environmental diversity using this technique [4]. The combined 'metagenome' of a community complements traditional (16S) phylotyping approaches and genome sequencing of culturable ecosystem members [5]. Given recent advances in sequencing technologies, this approach promises to uncover the identity as well as functionality of the 'unculturable majority' of microbial species on earth, and might lay a firm basis for our understanding of ecosystem functioning. Today, about four times as many genes have been generated in five years of metagenomic sequencing than in over a decade of complete genome sequencing [4], and with dramatically dropping sequencing costs, metagenome projects will be initiated almost everywhere on earth. However, because of the great complexity of generated data, their analysis, an essential step in each project, is far from easy and requires accessible and user-friendly tools that are mostly not available yet. Consequently, we witness an emerging field within computational biology that aims not only at the development of those tools but also at an understanding of the ecosystem imprinted in the sequence data. Recently, several computational methods have been developed and applied to analyze the functional and phylogenetic composition of individual samples (environments) and to derive various properties of the inhabiting microbial communities. In addition, the comparison of results between different (sub)environments ('comparative metagenomics' [6••]) allows to draw more general conclusions about the relationship between metagenome properties and the habitat they were derived from. Here, following the workflow of a typical metagenomics data analysis, we review the state-of-the-art in methodologies that address each step and describe the discoveries they have led to. In addition, we point to possible interdependencies of different metagenome properties and various pitfalls in the analysis of metagenomics data. Finally, we suggest a minimal set of computational analyses to be performed to be able to properly describe and compare an environment.

A typical metagenomics data analysis workflow

Starting from raw reads, assembly is usually the first step to increase fragment length and gain insights in population structure. After that, gene calling is performed, as most (but not all) functional analyses are done at the protein level. Next, higher level metagenome descriptors are derived: basic properties (e.g. sequence composition), species composition, functional composition and population properties. These descriptors provide first insights in the communities but become even more powerful when compared with environments using comparative metagenomics approaches.

From reads to proteins: assembly and gene calling

In bacterial genome sequencing projects, the protocols to go from raw reads to complete and high-quality proteomes have become well established [7]. Metagenomes, however, are sending bioinformaticians back to the drawing board. Already the assembly process can pose a great challenge. One of the main reasons for this is the phylogenetic complexity of samples: while it is usually possible to assemble most of the genomes from environments containing a small number of (dominant) species (e.g. [2,8•]), samples with large species richness, such as soil [6••], can hardly be assembled given a sequencing depth of up to 100 Mb per (sub)sample. A recent study based on simulated metagenomes confirmed this trend [9•]. On top of this complexity come the added complications regarding high frequency of polymorphisms and genome variations that have been reported even up to the subspecies level [10–12]. Also, the presence of viruses and/or inserted phages might hamper the assembly by increasing the chance of chimeric contigs [13]. Novel short-read sequencing technologies are imposing further complications. Although strategies are being developed to assemble these novel datatypes [14,15], so far, no specific metagenome assembly software has been published or can be simply downloaded. Two strategies to alleviate the assembly problem were identified, namely (i) the use of reference sequences [11••] and (ii) the pre-binning of reads into phylogenetic groups based on sequence composition [2,8•,16••]. Although both can provide improvements, for the former approach, the number of reference genomes still is insufficient for complex metagenome assembly, while for the latter, the binning process only seems to be satisfyingly work in very simple communities (see below). Another alternative are 'greedy' assembly approaches in which multiple metagenomic samples are combined into one superassembly [11••,17••], though with the associated risk of cross-assembling different strains and species. In any case, while currently assemblers such as phrap [18], Arachne [19], JAZZ, Forge and Celera Assembler [20] are being adapted and used to assemble metagenomes, this research area needs further active developments to increase assembly quality for complex metagenomes, sometimes generated using a combination of different sequence technologies.

Owing to the generally limited assembly of these data, also gene prediction methods have to be adapted to deal with (i) massive amounts of fragmented genes on short sequences, (ii) the phylogenetic diversity in samples that hampers the usage of species-specific training sets and (iii) lower end quality of sequences leading to within-frame stop codons and frameshifts. Recent developments use heuristically estimated codon models based on the GC content of the small fragments to overcome this [21]. Alternatively, extrinsic strategies to find coding regions based on (a) their similarity to other regions in a reference database (e.g. known (meta)genomes and/or the sequence set under investigation) and (b) their synonymous versus non-synonymous substitution rate (indicating evolutionary constraints) can be applied [22,23]. To avoid gene prediction problems altogether, some studies (e.g. [24•,25]) base their whole downstream analysis on blastx annotation of reads, limiting themselves to the 'known fraction' of their dataset. The development of metagenome gene prediction software is still in its infancy, and a rigorous evaluation of new and existing methods is needed. Recently, two 'classical' gene predictors (Fgenesb and Critica/Glimmer) were evaluated for this purpose, but unfortunately not compared with the abovementioned heuristic/extrinsic approaches [9•]. Fgenesb performed markedly better (especially on non-assembled sequences), but about 20% of genes were missed and another 10% wrongly predicted, leaving ample room for improvement, judged by the current performance of these methods on full genomes. The considerable differences between these two tools are somewhat worrying given the wide range of methodologies that were used to analyze the metagenomes published so far, suggesting artificial differences in the resulting gene sets which hamper comparative analysis ([4]; see below).

Towards community understanding: delineation of metagenome descriptors

To go from 'a bag of genes' towards a proper understanding of an ecosystem, its inhabitants and its functioning, a range of techniques are being developed to derive parameters that are helpful in this process. These can be subdivided into the following categories: (i) basic descriptors; (ii) phylogenetic composition; (iii) functional composition and (iv) population properties, though they are interdependent (see below). When combined, they can give a first glimpse into biodiversity and ecosystem functioning. Here, we will describe some of the first published methods towards this goal.

While more technical descriptors such as average read length, contig size or assembly rate (as, for example reported by references [3,16••]) make the nature of the dataset more transparent, intrinsic, **basic metagenome descriptors**, such as sequence composition reflect already some environmental constraints. The descriptors range from the GC content [26] over codon usage [21] to oligonucleotide composition [27,28]. These measurements are currently being used for phylogenetic read classification and/or gene prediction (for the latter see reference [21]). Another basic parameter that can be derived from metagenomic shotgun reads is the effective genome size (EGS; a measure of average genome size that takes associated plasmids, inserted elements and phages into account [29]). It can be used to normalize for genome-size effects in comparative metagenomics analyses (see below), and, when combined with assembly information, allows estimation of species richness [3,17••]. It further can provide guidance on coverage issues, for example how much more sequencing is needed to capture most of unique sequences in a sample [3] or to complete the most dominant genome [6••].

To understand the contribution of the different inhabitants to the community, the deduction of the **species**

composition from metagenome data is of crucial importance, but far from trivial. Two distinct concepts with different aims should be discerned: (i) the classification of each read/contig to species (or at least some level of phylogenetic grouping), to possibly link functions of genes encoded by the reads to the community members exerting them and (ii) the quantitative determination of general species composition of the environment at hand. Several approaches for these two concepts have been recently proposed.

Assigning contigs and genes to taxonomic groups is, for complex samples, currently mostly done by 'best-BLAST-hit' mapping (e.g. [17••,24•,30•,31,32•,33]) because of the low computational efforts and the usage of the full spectrum of known genes as reference. On the down side, it is generally not very accurate, cannot deal with horizontal gene transfer (HGT) and does not allow mapping reads to internal nodes of the tree of life (leading to misleading mappings if the best-hit is from a phylogenetic group that is underrepresented in sequence databases) [9•,34•]. Though methods are being developed to deal with these problems, they still only allowed the correct assignment of 25% of reads of a missing species in simulations [35].

Alternatively, sequence composition based 'binning' approaches might be less influenced by biases in sequence databases. Various techniques based on oligonucleotide (2–8 mer) frequency signatures are being applied and developed ([2,9•,16••,27,28,36]; see McHardy and Rigoutsos in this issue). In a recent analysis on simulated metagenomes, the phylopythia binning tool outperforms a 'best-BLAST hit' and a basic oligonucleotide frequency method [9•] but was unfortunately not compared with any of the other recently developed methodologies. In addition, only results on larger contigs (>8 kb or >10 reads) were presented, while the big challenge in this field lies in assigning <1 kb-sized reads that dominate unassembled, complex samples. Given that currently only 60% of larger contigs can be correctly assigned using the best approach [9•], this problem is still far from being solved.

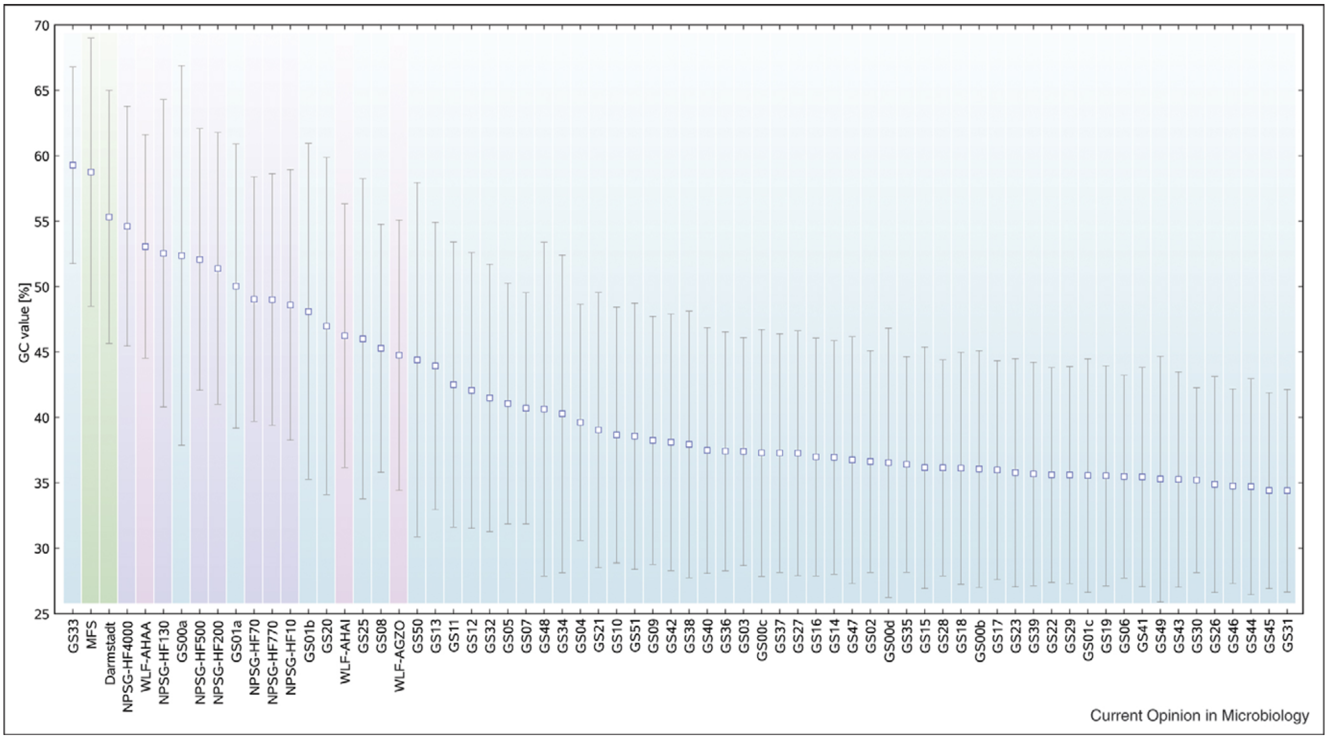
To estimate the species composition quantitatively ('who is there and how many of them are present?'), approaches based on single-copy or equal-copy marker genes, whose counts linearly scale with the number of individuals present (and not e.g. with genome size), are used. After early applications of this principle (using one marker gene at the time; e.g. [3]), a first large-scale phylogenetic approach was developed to map marker gene containing reads to (internal and external) nodes of a reference tree [34•]. This approach should also be more quantitative than classic 16S rRNA PCR based approaches, as it does not suffer from amplification bias or from quantification problems due to varying 16S copy numbers [34•]. However, as 16S methods have the advantage of being able to map to a much larger sampling of genera, both techniques should be regarded as complementary [34•].

To elucidate the **functional composition** of environments, first the predicted ORFs need functional annotation. Strategies analogous to genome sequencing projects have generally been used so far (see reference [4] for an overview). The most common approach is BLAST-based annotation by comparing ORFs against higher order databases such as NCBI Clusters of Orthologous Groups (COGs), TIGR funccats, STRING extended COGs, SEED, KEGG, and so on ([4] and refs therein). However, these techniques only allowed to functionally annotate ~25–50% of proteins per published metagenome [4], which might be due to the limited blast sensitivity for highly fragmented genes ([9•,37]; see below). In genome annotation, two additional methodologies are being employed to improve on this: profile-based homology searches [38] and gene context approaches [39]. The former are easy to implement by searching for protein modules using domain databases (e.g. [40–42]) while the latter offer more potential but need adaptation to the data. In particular, gene neighborhood analysis, a powerful function prediction concept for prokaryotic genomes that does not require homology for the ORFs to be annotated, can be used for function prediction in shotgun sequence data [4,23,43,44]. Recently, Harrington *et al.* published an improved methodology yielding high sensitivity on short contigs (including unassembled reads), allowing function prediction (using a combination of blast-based and profile-based homology and neighborhood approaches) for ~50–80% of proteins in four metagenomics samples [44]. While homology-based approaches will be useful to trace new functionally distinct (sub)families within known superfamilies, neighborhood-based approaches are particularly useful to discover and annotate completely novel proteins associated to known processes, especially in the light of biomining for novel industrially relevant enzymes and catalysts (see Figure 1).

Recently, the first methods have been developed to derive **population properties** from metagenomics data. For example, Johnson and Slatkin [10] estimated growth rates and mutation rates on the basis of site-frequency mutation spectra derived from the raw reads (while incorporating Phred quality scores) and von Mering *et al.* [34•] measured relative evolutionary rate on the basis of phylogenetic mapping of metagenomic marker genes. Rusch *et al.* [11••] and Gill *et al.* [45••] used an elegant technique dubbed 'fragment recruitment' to investigate the amount of genome rearrangements and gain first insights in population structure by aligning mated sequence reads to reference genomes.

While, in complex samples, all the descriptors discussed above (and probably more) are needed to move towards a better understanding of ecosystem functioning, in less

Figure 1



Current Opinion in Microbiology

GC content of the metagenomic soil and sea samples. The highest GC content is observed for both the soil samples while the ocean surface water samples have the lowest. The only exception is from a Global Ocean Sampling (GOS), sample taken in a hypersaline lagoon with an exceptionally high GC content. GOS samples taken from coastal (GS13) or fresh water, from a mangrove (GS32), embayment (GS5) or reef (GS25) also show a higher value—possibly because of mixing with soil. Contaminations (GS00a/Sargasso1 [55]) or a higher fraction of eukaryotes in the sample due to filter differences (GS01a, GS01 [3]) apparently also increase the GC content. Sample abbreviations: GSX, Global Ocean Sampling [11••]; MFS, Minnesota farm soil [6••]; Darmstadt, Darmstadt soil [33]; NPSG, North pacific subtropical gyre [24•]; WLF, Whale fall [6••].

complex ecosystems dominated by a few species, a combination of sequence composition binning and assembly seems already sufficient to (almost) completely sequence the community members. This, in turn, allows the assignment of some metabolic activities to individual ecosystem members. This allowed the reconstruction of metagenomic metabolic pathways and indicates cooperation between species/phylogenetic groups within one environment (e.g. [1,2,8•]) or assign specific roles for distinct species/phylogenetic groups in relation with their host (e.g. [16••]). As it has only been possible to start exploring 'who does what' in these simple ecosystems, a great challenge lies in the improvement of tools to address these issues in increasingly complex environments to understand the interrelationships of organisms living in soil, the ocean or in the human body.

Metagenomes in context: comparative analysis

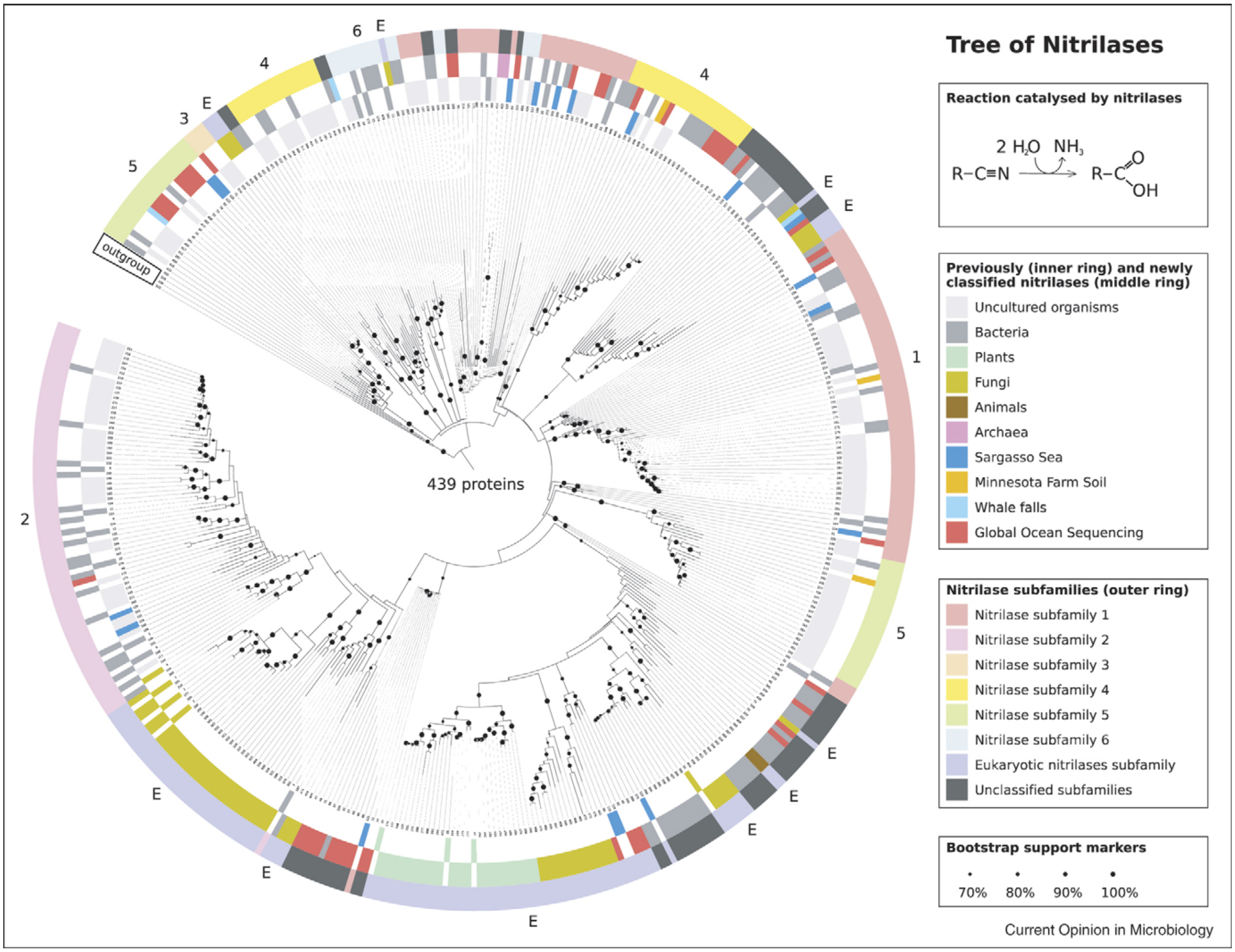
Although the analysis of individual metagenomes has greatly increased our understanding about microbial communities, genome sequence analysis has shown that great additive power comes from comparative approaches [46] as they provide context to the individual samples.

Comparison of different samples from *the same or similar* environments can reveal the influence of particular environmental factors on microbial communities. For example, in a gradual sampling of sea water from the surface to 4000 m depth, the increasing pressure and reduction of light was shown to influence the functional repertoire of organisms living at various depths [24•]. Similarly, comparisons of symbiont communities living in distinct murine intestines allowed linking disease (obesity) to the functional repertoire of the gut inhabitants (in this case the authors showed increased energy harvesting capacities [32•]).

Conversely, the comparison of *diverse* habitats' metagenomes allows the discovery of general trends that link metagenome and community properties with phenotypic features of environments. For instance, one of the most **basic properties**, the GC content, was shown to differ significantly between environments, ranging from high values in Minnesota farm soil to very low ones in surface sea water [26]. Although the original study was based on limited data, this observation is confirmed by the recent Global Ocean Survey (GOS) data [11••], where all open sea surface water sequences show low

GC content, while those from samples taken close to the land (e.g. lagoons and beaches) where water can mix with soil, show markedly higher values (see Figure 2). Likewise, a correlation between microbial genome size and environmental complexity was shown [29] along with the differences of evolutionary rates between environments [34•]. All these outcomes derived from the same environmental datasets seem somehow related: the Sargasso Sea surface water samples harbor the fastest-evolving, smallest genomes with the lowest GC content, while the Minnesota farm soil genomes appear to be the largest, have the highest GC content and evolve the slowest. Although individual links between GC content, genome size and replication/evolutionary rate have been

Figure 2



Biomining metagenomics data case study: nitrilases. Natural habitats are likely to harbor many novel enzymes with biotechnological potential. To detect them, functional, PCR-based or hybridization-based screening methods [56] can be complemented by computational mining of metagenomics datasets. To illustrate a typical computational screen, we extend here the study by Podar *et al.* that detected 17 new nitrilases in the Sargasso Sea (SGS) dataset [57] on an updated environmental dataset collection and UniRef (see Supplementary data for details and references [3,58–61] for similar studies). The tree (drawn using iTOL [62]) contains the 27 bait sequences and hits in UniRef (341), GOS (47), SGS (18; one more than reference [57]), MFS (3), WLF (3), and none in Acid Mine Drainage (AMD). All previously described nitrilases of uncultured species [63] were detected. The colored rings label genes described by [57] (inner ring), nitrilases found in this study (middle ring) and the previously proposed classification into subfamilies (outer ring [57]). The results imply a linear scaling of the identified enzymes with dataset size (at least for ocean surface water) as the nitrilases-per-screened-protein ratio is similar in Sargasso Sea and the much larger Global Ocean Sampling data. The many newly added nitrilases seem to challenge the proposed classification of this family into six bacterial and one eukaryotic subfamily indicating a need for systematic updates. Although most of the proposed subfamilies could be extended by new members (including from archaea!), some became uncertain because of low bootstrap values or even fell apart. Furthermore, novel subfamilies with probably distinct substrate specificities became apparent. The eukaryotes now clearly comprise several distinct subfamilies (e.g. at least two distinct fungal groups exist, only one being related to the bacterial subfamily 2), and several novel bacterial subfamilies can be assumed with confidence (black in our ring), some of which are deeply branching indicating a diverse substrate spectrum.

hypothesized before [47], the precise reasons for this observation remain unknown.

Not only is the breadth of the metagenomic **functional complement** linked to environmental factors (as measured from genome size), but also its composition. When one considers a habitat as one big intercommunicating 'soup' of organisms that carry their genes to maintain this interplay, the combined metagenome should reveal properties of the community and the environment as a whole. Indeed, this gene-centric approach has shown that the more similar the inferred functional composition of metagenomes is, the more related are the respective environmental phenotypes [6••,11••,32•]. Beyond general trends, comparative approaches can also pinpoint particular proteins, protein families and cellular processes that are likely to be responsible for the specific adaptations to particular environments, as could be shown in the first comparative study that used normalized overrepresentions and underrepresentions of such functional units [6••].

Comparative metagenomics can be also used to learn about differences in the **phylogenetic composition** of environments. It could be shown, for instance that the detectable taxonomic groups of microbes have distinct habitat preferences up to the subphylum level, which are remarkably stable in time [34•]. Likewise, a metagenomics study revealed a clear non-random distribution of phages in four ocean sampling sites, with a linear correlation between genetic and geographic distance [17••]. Finally, an analysis of the GOS data for aerobic anoxygenic photosynthesisers (AAnPs) showed great variation in diversity, abundance and composition of AAnP assemblages in different oceanic regions [48].

Interdependencies and pitfalls in comparative metagenomics

Despite the great potential of comparative metagenomics approaches, one should apply them cautiously. Various environment-specific biological factors (see above) and many (usually sample-specific) technical issues hamper the direct comparison of environments, as they influence each other and most results derived from these data.

As for the **basic descriptors**, differences in average genome size of samples (e.g. measurable by EGS [29]) will implicitly lead to differences in the relative functional composition of samples. For example, the sample with the smallest EGS should always have a significant overrepresentation of housekeeping genes as they are a constant fraction while other functional categories grow with genome size; no further biological conclusions should be drawn for differences here without proper normalization. Also, the observed differences in GC content [26] have an impact on homology searches, phylogenetic analyses and binning. However, the extent of such effects still needs to be determined.

The **phylogenetic complexity** of a sample strongly influences the (feasibility of) downstream analyses. For example, the lower the species complexity, the higher the coverage of each individual leading to better assembly and consequently better gene prediction. Longer contigs also improve the efficiency of neighborhood techniques for function prediction and increase the chance of correct phylogenetic assignment of fragments. Therefore, in these samples, the 'who does what' question will be easier to address.

Different **functional constraints** in environments can result in different evolutionary rates [34•] and thus can lead to skews in the gene and function detection rates (e.g. faster evolving genes are more difficult to capture by Blast and orthology assignment methods).

Limited sample coverage and phylogenetic diversity might hamper the direct comparison of **population genetic parameters** as robust estimates based on few data points are difficult, and abundant species might hide the real population structure in samples.

Besides the various biological factors, many **technical issues** related to sampling, sequencing and annotation influence downstream analyses. For instance, the frequent usage of filters or other selection methods for sampling directly influences the phylogenetic and functional composition of the sample. For example, Johnston *et al.* [49] described a surprising paucity of particular nitrogen-fixing genes in the first Sargasso Sea dataset [3], which was later criticized because of its failure to take into account that the main contributors to these genes (cyanobacteria) were probably not in the dataset because of the filtering [50•]. Likewise, in their comparison of the phylogenetic composition within several metagenomics datasets, von Mering *et al.* noticed a conspicuous lack of endospore-forming organisms, which could be linked with their ability to withstand DNA extraction protocols [34•].

Another effect comes from the sequencing technology and protocols. It is first reflected in the read length that depends on the technology (capillary (Sanger) sequencing versus 454 pyrosequencing) but also on parameter choices and other protocols (e.g. Sanger sequencing reads from the whale fall and soil datasets average at ~700 bp after quality clipping, while the sargasso sea ones are ~850 bp [29]). Together with coverage (partly also depending on the sequencing technology) this directly influences the amount of assembly. The resulting differences in contig length influence the success and quality of gene predictions, and the subsequent assignment of gene functions. Short reads as produced by the first generation of 454 GS20 sequencers are especially little informative as they often are insignificant in the BLAST statistics (e.g. in the mouse gut dataset ~95% of 454 reads were unassignable to known genes/COGs/KEGGs, while for Sanger reads this was ~20–30% [32•]).

In addition to the factors described above, the selected assembly, gene calling and annotation protocols themselves are yet another factor that complicates a direct comparison of samples, for example regarding functional composition. So far, a plethora of methodologies was used in the different projects [4], necessitating a uniform, standardized way of treating metagenomic data in order to be able to compare results from different projects (see Box 1). Only then and in conjunction with good coverage estimates, presence and absence as well as overrepresentation and underrepresentation of genes can be interpreted more confidently. (Given the estimated diversity, was coverage high enough to expect the absence of a gene by chance?) To measure functional and/or phylogenetic coverage, several techniques have been used, ranging from the analysis of single-copy, non-linked genes (mostly used when some full genomes can be almost assembled, e.g. sludge [8•]), via theoretical calculations based on the Lander–Waterman equation (e.g. [3,8•]) to rarefaction approaches (e.g. [6••]).

Box 1 Suggestion for a minimal metagenome sequence analysis standard (MINIMESS) to derive indicators for a dataset including annotation protocol, coverage estimates and community descriptors

The previously proposed Minimum Information about a Metagenomic Sequence (MIMS) standard covers detailed information about primary information such as sampling location and procedure, DNA isolation and sequencing and is an indispensable tool to interpret metagenomic datasets [51•,52]. While this is essential towards comparative metagenomics, the many interdependencies and pitfalls (see text) call for an additional, complimentary layer of reporting that provides a standardized description of the metagenome and its inferred community properties. A first prerequisite is the transparent and complete description of data treatment (e.g. assembly, gene calling and functional annotation protocol including the parameter settings). In addition, we propose the reporting of a basic set of metagenome descriptors, resulting from a standardized list of analyses to be performed on each published dataset. This set of descriptors provides an indispensable tool for the proper interpretation and post-analysis of the data and the comparison of metagenomes from independent samples.

(1) Basic sequence analysis: reporting of detailed assembly statistics (including contig composition), gene density, average gene length and fraction of predicted proteins with functional assignment.

(2) Species composition: quantitative description of species composition (marker gene approach, ideally complemented by 16S PCR based method) and species richness estimate.

(3) Functional composition: higher level functional content distribution (e.g. COG/KEGG/SEED, see main text).

(4) Species and gene coverage estimates.

(5) Linking of species and function: although phylomapping tools are just emerging (see dedicated section in this review), it would be favorable to provide a list of gene-species linkages, coming from phylogenetic assignment of reads/contigs based on homology, sequence composition and marker gene presence.

(6) Putative interfering *biological* factors: reporting of, for example GC content and average genome size (EGS).

(7) Putative interfering *technical* factors: reporting of read length and contig length distributions in relation to community complexity (see also coverage estimates).

Taken together, despite recent progress in method development to derive individual parameters for metagenomics samples, considerable effort has to go into the analysis of their interdependencies and the normalization of data from different production lines. Standards for some of the steps would be very helpful to make data comparable and thus add enormous value to them for little cost.

Minimal standards for annotation and analysis

The more sampling conditions for metagenomics datasets are reported, the more detailed can be the inferences of environmental constraints. Not only exact sample site location and sampling methodology should be mentioned but also broad measurements on the (also non-obvious) physical and chemical properties of the environment, as well as detailed descriptions of the habitat should be made. However, often this is a considerable effort beyond the scope of the individual project and sometimes it is not even known what would be needed to record. Hence, the development of a 'Minimum Information about a Metagenomic Sequence (MIMS)' standard has been proposed in the community to enforce some essential measurements when submitting data to public databases [51•,52]. Beyond the proposed sample information, we believe that a complementing set of standard analyses is also necessary for proper interpretation of metagenomic data (MINIMESS, see Box 1) and should enable normalization of the heterogeneous data for various comparative analyses (see previous paragraph). A first step would be a transparent analysis protocol with all the parameters of the various methods properly reported as they can have a considerable impact on the results.

Over time, metagenomics data generation and analyses protocols will diversify further and there is a need for an accepted infrastructure [53,54] that can cope not only with the heterogeneous data but also with agreements on how to annotate and compare them. It is early days in environmental genomics and important findings will require robust and accurate tools and approaches—what we have reviewed here is just the beginning of a new exciting field emerging in computational biology.

Acknowledgements

The authors would like to thank Christian von Mering and members of the Bork group for stimulating discussions. This work was supported by the EU 6th Framework Programme (Contract No: LSHG-CT-2004-503567).

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.mib.2007.09.001.

References and recommended reading

Papers of particular interest, published within the annual period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

1. Hallam SJ, Putnam N, Preston CM, Detter JC, Rokhsar D, Richardson PM, DeLong EF: **Reverse methanogenesis: testing the hypothesis with environmental genomics**. *Science* 2004, **305**:1457-1462.

2. Tyson GW, Chapman J, Hugenholtz P, Allen EE, Ram RJ, Richardson PM, Solovyev VV, Rubin EM, Rokhsar DS, Banfield JF: **Community structure and metabolism through reconstruction of microbial genomes from the environment**. *Nature* 2004, **428**:37-43.

3. Venter JC, Remington K, Heidelberg JF, Halpern AL, Rusch D, Eisen JA, Wu D, Paulsen I, Nelson KE, Nelson W *et al.*: **Environmental genome shotgun sequencing of the Sargasso Sea**. *Science* 2004, **304**:66-74.

4. Raes J, Harrington ED, Singh AH, Bork P: **Protein function space: viewing the limits or limited by our view?** *Curr Opin Struct Biol* 2007, **17**:362-369.

5. Handelsman J, Rondon MR, Brady SF, Clardy J, Goodman RM: **Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products**. *Chem Biol* 1998, **5**:R245-R249.

6. •• Tringe SG, von Mering C, Kobayashi A, Salamov AA, Chen K, Chang HW, Podar M, Short JM, Mathur EJ, Detter JC *et al.*: **Comparative metagenomics of microbial communities**. *Science* 2005, **308**:554-557.

The first comparative metagenomics study introducing the notion of a gene-centric view of environments.

7. Stothard P, Wishart DS: **Automated bacterial genome analysis and annotation**. *Curr Opin Microbiol* 2006, **9**:505-510.

8. • Garcia Martin H, Ivanova N, Kunin V, Warnecke F, Barry KW, McHardy AC, Yeates C, He S, Salamov AA, Szeto E *et al.*: **Metagenomic analysis of two enhanced biological phosphorus removal (EBPR) sludge communities**. *Nat Biotechnol* 2006, **24**:1263-1269.

An elegant linkage of functional and phylogenetic information in a simple system to reconstruct sludge metabolic pathways.

9. • Mavromatis K, Ivanova N, Barry K, Shapiro H, Goltsman E, McHardy AC, Rigoutsos I, Salamov A, Korzeniewski F, Land M *et al.*: **Use of simulated data sets to evaluate the fidelity of metagenomic processing methods**. *Nat Methods* 2007, **4**:495-500.

An important pioneering effort to compare metagenomic data analysis tools, which should be expanded to include the recent wave developments.

10. Johnson PL, Slatkin M: **Inference of population genetic parameters in metagenomics: a clean look at messy data**. *Genome Res* 2006, **16**:1320-1327.

11. •• Rusch DB, Halpern AL, Sutton G, Heidelberg KB, Williamson S, Yooseph S, Wu D, Eisen JA, Hoffman JM, Remington K *et al.*: **The Sorcerer II Global Ocean Sampling Expedition: Northwest Atlantic through Eastern Tropical Pacific**. *PLoS Biol* 2007, **5**:e77.

An impressive amount of data that will keep researchers busy for several years to come

12. Tettelin H, Masignani V, Cieslewicz MJ, Donati C, Medini D, Ward NL, Angiuoli SV, Crabtree J, Jones AL, Durkin AS *et al.*: **Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: implications for the microbial "pan-genome"**. *Proc Natl Acad Sci USA* 2005, **102**:13950-13955.

13. Salzberg SL, Yorke JA: **Beware of mis-assembled genomes**. *Bioinformatics* 2005, **21**:4320-4321.

14. Sundquist A, Ronaghi M, Tang H, Pevzner P, Batzoglou S: **Whole-genome sequencing and assembly with high-throughput, short-read technologies**. *PLoS ONE* 2007, **2**:e484.

15. Warren RL, Sutton GG, Jones SJ, Holt RA: **Assembling millions of short DNA sequences using SSAKE**. *Bioinformatics* 2007, **23**:500-501.

16. •• Woyke T, Teeling H, Ivanova NN, Huntemann M, Richter M, Gloeckner FO, Boffelli D, Anderson IJ, Barry KW, Shapiro HJ *et al.*: **Symbiosis insights through metagenomic analysis of a microbial consortium**. *Nature* 2006, **443**:950-955.

See annotation to reference [45••].

17. •• Angly FE, Felts B, Breitbart M, Salamon P, Edwards RA, Carlson C, Chan AM, Haynes M, Kelley S, Liu H *et al.*: **The marine viromes of four oceanic regions**. *PLoS Biol* 2006, **4**:e368.

See annotation to reference [30•].

18. Green P: Phrap (www.phrap.org).

19. Jaffe DB, Butler J, Gnerre S, Mauceli E, Lindblad-Toh K, Mesirov JP, Zody MC, Lander ES: **Whole-genome sequence assembly for mammalian genomes: Arachne 2**. *Genome Res* 2003, **13**:91-96.

20. Celera assembler (http://sourceforge.net/projects/wgs-assembler/).

21. Noguchi H, Park J, Takagi T: **MetaGene: prokaryotic gene finding from environmental genome shotgun sequences**. *Nucleic Acids Res* 2006, **34**:5623-5630.

22. Krause L, Diaz NN, Bartels D, Edwards RA, Puhler A, Rohwer F, Meyer F, Stoye J: **Finding novel genes in bacterial communities isolated from the environment**. *Bioinformatics* 2006, **22**:e281-e289.

23. Yooseph S, Sutton G, Rusch DB, Halpern AL, Williamson SJ, Remington K, Eisen JA, Heidelberg KB, Manning G, Li W *et al.*: **The Sorcerer II Global Ocean Sampling Expedition: expanding the universe of protein families**. *PLoS Biol* 2007, **5**:e16.

24. • DeLong EF, Preston CM, Mincer T, Rich V, Hallam SJ, Frigaard NU, Martinez A, Sullivan MB, Edwards R, Brito BR *et al.*: **Community genomics among stratified microbial assemblages in the ocean's interior**. *Science* 2006, **311**:496-503.

Well-designed comparative metagenomics study along an oceanic depth gradient

25. Edwards RA, Rodriguez-Brito B, Wegley L, Haynes M, Breitbart M, Peterson DM, Saar MO, Alexander S, Alexander EC Jr, Rohwer F: **Using pyrosequencing to shed light on deep mine microbial ecology**. *BMC Genomics* 2006, **7**:57.

26. Foerstner KU, von Mering C, Hooper SD, Bork P: **Environments shape the nucleotide composition of genomes**. *EMBO Rep* 2005, **6**:1208-1213.

27. McHardy AC, Martin HG, Tsirigos A, Hugenholtz P, Rigoutsos I: **Accurate phylogenetic classification of variable-length DNA fragments**. *Nat Methods* 2007, **4**:63-72.

28. Teeling H, Waldmann J, Lombardot T, Bauer M, Glockner FO: **TETRA: a web-service and a stand-alone program for the analysis and comparison of tetranucleotide usage patterns in DNA sequences**. *BMC Bioinformatics* 2004, **5**:163.

29. Raes J, Korbel JO, Lercher MJ, von Mering C, Bork P: **Prediction of effective genome size in metagenomic samples**. *Genome Biol* 2007, **8**:R10.

30. • Culley AI, Lang AS, Suttle CA: **Metagenomic analysis of coastal RNA virus communities**. *Science* 2006, **312**:1795-1798.

Two studies that chart the vastly understudied world of virus diversity.

31. Poinar HN, Schwarz C, Qi J, Shapiro B, Macphee RD, Buigues B, Tikhonov A, Huson DH, Tomsho LP, Auch A *et al.*: **Metagenomics**

to paleogenomics: large-scale sequencing of mammoth DNA. *Science* 2006, **311**:392-394.

32. • Turnbaugh PJ, Ley RE, Mahowald MA, Magrini V, Mardis ER, Gordon JI: **An obesity-associated gut microbiome with increased capacity for energy harvest**. *Nature* 2006, **444**:1027-1031.

See annotation to reference [45••].

33. Treusch AH, Kletzin A, Raddatz G, Ochsenreiter T, Quaiser A, Meurer G, Schuster SC, Schleper C: **Characterization of large-insert DNA libraries from soil for environmental genomic studies of Archaea**. *Environ Microbiol* 2004, **6**:970-980.

34. • von Mering C, Hugenholtz P, Raes J, Tringe SG, Doerks T, Jensen LJ, Ward N, Bork P: **Quantitative phylogenetic assessment of microbial communities in diverse environments**. *Science* 2007, **315**:1126-1130.

Study providing evidence for evolutionary rate difference between environments and a clear habitat preference of micro-organisms, which seems to be remarkably stable in time.

35. Huson DH, Auch AF, Qi J, Schuster SC: **MEGAN analysis of metagenomic data**. *Genome Res* 2007, **17**:377-386.

36. Abe T, Sugawara H, Kinouchi M, Kanaya S, Ikemura T: **Novel phylogenetic studies of genomic sequence fragments derived from uncultured microbe mixtures in environmental and clinical samples**. *DNA Res* 2005, **12**:281-290.

37. Tress ML, Cozzetto D, Tramontano A, Valencia A: **An analysis of the Sargasso Sea resource and the consequences for database composition**. *BMC Bioinformatics* 2006, **7**:213.

38. Koonin EV, Tatusov RL, Galperin MY: **Beyond complete genomes: from sequence to structure and function**. *Curr Opin Struct Biol* 1998, **8**:355-363.

39. Huynen MA, Snel B, von Mering C, Bork P: **Function prediction and protein networks**. *Curr Opin Cell Biol* 2003, **15**:191-198.

40. Bateman A, Coin L, Durbin R, Finn RD, Hollich V, Griffiths-Jones S, Khanna A, Marshall M, Moxon S, Sonnhammer EL *et al.*: **The Pfam protein families database**. *Nucleic Acids Res* 2004, **32**:D138-D141.

41. Letunic I, Copley RR, Pils B, Pinkert S, Schultz J, Bork P: **SMART 5: domains in the context of genomes and networks**. *Nucleic Acids Res* 2006, **34**:D257-D260.

42. Mulder NJ, Apweiler R, Attwood TK, Bairoch A, Bateman A, Binns D, Bork P, Buillard V, Cerutti L, Copley R *et al.*: **New developments in the InterPro database**. *Nucleic Acids Res* 2007, **35**:D224-D228.

43. Overbeek R, Fonstein M, D'Souza M, Pusch GD, Maltsev N: **The use of gene clusters to infer functional coupling**. *Proc Natl Acad Sci USA* 1999, **96**:2896-2901.

44. Harrington ED, Singh AH, Doerks T, Letunic I, Von Mering C, Jensen LJ, Raes J, Bork P: **Quantitative assesment of protein function prediction from metagenomics shotgun sequences**. *Proc Natl Acad Sci* 2007, **104**:13913-13918.

45. •• Gill SR, Pop M, Deboy RT, Eckburg PB, Turnbaugh PJ, Samuel BS, Gordon JI, Relman DA, Fraser-Liggett CM, Nelson KE: **Metagenomic analysis of the human distal gut microbiome**. *Science* 2006, **312**:1355-1359.

Three studies allowing the first large-scale metagenomic insights into the intriguing world of symbiotic microbial communities.

46. Fraser CM, Eisen J, Fleischmann RD, Ketchum KA, Peterson S: **Comparative genomics and understanding of microbial biology**. *Emerg Infect Dis* 2000, **6**:505-512.

47. Bentley SD, Parkhill J: **Comparative genomic structure of prokaryotes**. *Annu Rev Genet* 2004, **38**:771-792.

48. Yutin N, Suzuki MT, Teeling H, Weber M, Venter JC, Rusch DB, Beja O: **Assessing diversity and biogeography of aerobic anoxygenic phototrophic bacteria in surface waters of the Atlantic and Pacific Oceans using the Global Ocean Sampling expedition metagenomes**. *Environ Microbiol* 2007, **9**:1464-1475.

49. Johnston AW, Li Y, Ogilvie L: **Metagenomic marine nitrogen fixation—feast or famine?** *Trends Microbiol* 2005, **13**:416-420.

50. • Remington KA, Heidelberg K, Venter JC: **Taking metagenomic studies in context**. *Trends Microbiol* 2005, **13**:404.

Comment that highlights the importance of taking technical factors (such as filtering) into account

51. • Field D, *et al.*: **Towards a richer description of our complete collection of genomes and metagenomes: the "Minimum Information about a Genome Sequence" (MIGS) specification.** *Nat Biotechnol* (In community review).

Crucial proposal to provide a minimal amount of information on sampling, and so on. when submitting/publishing (meta)genome sequences.

52. Field D, Morrison N, Selengut J, Sterk P: **Meeting report: eGenomics: cataloguing our complete genome collection II**. *Omics* 2006, **10**:100-104.

53. Markowitz VM, Ivanova N, Palaniappan K, Szeto E, Korzeniewski F, Lykidis A, Anderson I, Mavromatis K, Kunin V, Garcia Martin H *et al.*: **An experimental metagenome data management and analysis system**. *Bioinformatics* 2006, **22**:e359-e367.

54. Seshadri R, Kravitz SA, Smarr L, Gilna P, Frazier M: **CAMERA: a community resource for metagenomics**. *PLoS Biol* 2007, **5**:e75.

55. Mahenthiralingam E, Baldwin A, Drevinek P, Vanlaere E, Vandamme P, Lipuma JJ, Dowson CG: **Multilocus sequence typing breathes life into a microbial metagenome**. *PLoS ONE* 2006, **1**:e17.

56. Handelsman J: **Metagenomics: application of genomics to uncultured microorganisms**. *Microbiol Mol Biol Rev* 2004, **68**:669-685.

57. Podar M, Eads JR, Richardson TH: **Evolution of a microbial nitrilase gene family: a comparative and environmental genomics study**. *BMC Evol Biol* 2005, **5**:42.

58. Kannan N, Taylor SS, Zhai Y, Venter JC, Manning G: **Structural and functional diversity of the microbial kinome**. *PLoS Biol* 2007, **5**:e17.

59. Zhu Y, Pulukkunat DK, Li Y: **Deciphering RNA structural diversity and systematic phylogeny from microbial metagenomes**. *Nucleic Acids Res* 2007, **35**:2283-2294.

60. van Loo B, Kingma J, Arand M, Wubbolts MG, Janssen DB: **Diversity and biocatalytic potential of epoxide hydrolases identified by genome analysis**. *Appl Environ Microbiol* 2006, **72**:2905-2917.

61. Rhee JK, Ahn DG, Kim YG, Oh JW: **New thermophilic and thermostable esterase with sequence similarity to the hormone-sensitive lipase family, cloned from a metagenomic library**. *Appl Environ Microbiol* 2005, **71**:817-825.

62. Letunic I, Bork P: **Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation**. *Bioinformatics* 2007, **23**:127-128.

63. Robertson DE, Chaplin JA, DeSantis G, Podar M, Madden M, Chi E, Richardson T, Milan A, Miller M, Weiner DP *et al.*: **Exploring nitrilase sequence space for enantioselective catalysis**. *Appl Environ Microbiol* 2004, **70**:2429-2436.