# Metagenomics: Microbial diversity through a scratched lens

Ben Temperton[1]* and Stephen J. Giovannoni[1]

[1]Department of Microbiology, Oregon State University, Corvallis, OR 97331

*Corresponding Author

Mailing address: Department of Microbiology, Oregon State University, 220 Nash Hall, Corvallis, OR, 97331. Phone: (541) 737-3502. Email: btemperton@gmail.com

## Abstract

Since nucleic acids were first extracted directly from the environment and sequenced, metagenomics has grown to one of the most data-rich and pervasive techniques for understanding the taxonomic and functional diversity of microbial communities. In the last decade, cheaper sequencing has democratized the application of metagenomics and generated billions of reads, revealing staggering microbial diversity and functional complexity. However, cheaper sequencing has come at the cost of reduced sequence length, resulting in poor gene annotation and overestimates of bacterial richness and abundance. Recent improvements in sequencing technology are beginning to provide reads of sufficient length for accurate annotation and assembly of whole operons and beyond, that will once again enable experimental testing of gene function and re-capture the early successes of metagenomic investigations.

## Diversity in sharp focus

The revelation of the ‘Great Plate-Count Anomaly’ by Staley and Konopka in 1985 [1] highlighted that contemporary understanding of microbial metabolism was highly skewed towards a small fraction of readily culturable bacteria. To address this issue, direct extraction and cloning of environmental DNA began to unravel novel phylogenetic [2,3] and functional [4] diversity. In 1998, the term ‘metagenomics’ was coined [5] and this exciting new field hinted at the scale of genetic variability within natural microbial populations [6] and associated phage [7]. In 2000, Béjà et al. [8] cloned 130-kb environmental DNA fragments from seawater into BAC libraries and found bacteriorhodopsin, a mechanism for ATP generation from light. This type of photochemistry previously had been know to occur only in hypersaline ponds; its discovery in the oceans is perhaps one of the most heralded successes of metagenomics [9]. Four years later, two landmark studies demonstrated the power of metagenomics to explore environmental microbiology: Venter et al. exposed the magnitude of microbial diversity in the surface water of Sargasso Sea, identifying 148 novel phylotypes and 1.2 million novel genes in a single study [10]; Tyson et al. demonstrated that when diversity was low, metagenomics could be used to reconstruct genomes of uncultured bacteria to reveal complete metabolic pathways, providing insight into their nutritional requirements and biogeochemical functions [11]. Encouraged by the success of their previous exploration of marine microbial diversity, Venter expanded their previous study on a global scale. In 2007, the first results of the Global Ocean Survey (GOS) were published [12], revealing ~390 new species and ~6 million predicted protein sequences in ~4000 protein clusters of which 42.6% had no known homology [13]. The unprecedented volumes of data generated by this project catalyzed the advancement of the bioinformatic techniques required to assemble, analyze and contextualize novel genes, phyla and pathways. Subsequent and ongoing voyages ensure that the GOS project remains to this day the largest metagenomic survey undertaken.

## An era of open-access science

Early adopters of metagenomics had the foresight to understand that the datasets from a single metagenomic investigation were too large for comprehensive analysis by a single research group and opened up datasets for public access, often prior to initial publication. This in turn fuelled the development of third-party tools for data management and analysis (e.g. MG-RAST [14], CAMERA [15], IMG-ER [16]) as well as standards for collection of metadata to assist in downstream analysis [17]. At the current time (February 2012), MG-RAST currently holds 111 publically available metagenomic projects, comprising 7,444 datasets, 459 million sequences and 8.4 × $10^{10}$ base pairs. Public availability of metagenomic datasets has enabled a broad range of bioinformatic investigations into novel clades, genetic diversity and microbial pan-genomes [18-22]. Fragment recruitment of metagenomic fragments to full genome sequences from related isolates has highlighted the prevalence of 'hypervariable regions' (HVRs) across multiple strains and species, where the genomic content of the sequenced isolate is not representative of the population as a whole [12,20,23] (Grote et al., *in submission*). Comparative metagenomic studies have shed light on community adaptation to a local environment, particularly in nutrient cycling (e.g. [24-28]), with new algorithms to deal with the statistical challenges of large numbers of observations with minimal or no replication [29-31] (Beszteri et al., *in submission.)*. Whereas the first culture-independent investigations into microbial diversity were performed on 38 16S rRNA clones [2], amplification of DNA with barcoded primers [32] to allow sequencing of multiple samples in a single run, coupled with massively reduced sequencing costs enabled investigators to identify tens of thousands of unique sequences [33] in a single study, providing an unprecedented insight into microbial diversity. Large-scale projects such as the Earth Microbiome Project (http://www.earthmicrobiome.org/) and *Tara* Oceans Expedition [34] continue to explore environmental microbial diversity with rich metadata to better model and understand microbial ecology at the systems level. Similar programs to map the human microbiome are also ongoing [35,36], making metagenomics one of the most data-rich, fastest-growing and exhaustively reviewed scientific fields [9].

## Data rich, information-poor: The cost of metagenomic democratization

Until 2006, metagenomics had required the cloning of environmental DNA into vectors and their subsequent Sanger sequencing and fragment assembly. With costs of Sanger sequencing approaching ~$500 per Mb [37], large scale metagenomics projects were limited to those with significant financial resources. Use of replicated experimental samples was extremely limited, preventing statistically rigorous analysis of biological variation and correlation of taxonomic abundance with nutrient metadata [38]. The pairing of metagenomics with an emergent pyrosequencing technology [39] marked the beginning of a sequencing revolution. In rapid succession 454 (Roche), Illumina, SOLiD (Life Technologies) and Pacific Biosciences drove the cost of sequencing down to < $0.10 per Mb (Fig. 1) (albeit with no similar decrease in the cost of data storage and computation [40]), democratizing metagenomic analysis, with a concomitant explosion of studies across a wide range of environments to investigate the diversity and functional capacity of all domains of life (e.g. *Eukarya*: [41,42]; *Archaea*: [4,43]; *Bacteria* [21,33,36,44,45]) and their associated viral communities [46-49], even from as bizarre a locale as the windshield splatter from a single road trip [50]. However, second generation sequencers generated reads ranging from ~35-500 bp depending on the technology used  – far shorter than the 130 kb fragment used to identify the 747 bp gene encoding proteorhodopsin by Béjà et al., or even the length of Sanger sequences used in the GOS study (~1000 bp). Furthermore, each advance in sequencing technology introduced its own biases, which, coupled with non-standardized metadata collection, made it difficult to compare the results of one investigation to previous datasets. Despite the reduction in sequencing costs, use of replicated samples was still poor [38]. In 16S rRNA diversity studies, the reduced sequence length of pyrosequencing prevented the use of full-length sequences; relying instead on shorter hypervariable regions (V1 through V8) sequenced both separately and in combination. Species richness of samples was significantly biased by primer choice, with V6 and V1+V2 overestimating richness while V3, V7 and V7+V8 underestimated richness [51]. Issues of over-estimation of the ‘rare’ biosphere from sequencing error [52-54] and PCR chimeras [55] required correction by post-sequencing analysis with tools such as AmpliconNoise [56]. Bias introduced during sample preparation, such as variation in DNA extraction efficiencies of different taxa [57] and poor representation of low G+C taxa [58,59] required fine-tuning of experimental design and greater use of controls.

In shotgun metagenomic studies, annotation of genes is improved by inclusion of identifying features such as promotors, riboswitches, co-operonic genes and signature protein domains. The probability of capturing such features on the same fragment is proportional to the length of the fragment. Separation of genes from their distinguishing features is significant in Sanger sequences (~1000 bp) compared to BAC sequences (~130 kbp). With the even shorter lengths of pyrosequencing, the issue is exacerbated. Furthermore, increased sequencing errors of ~1-3% (depending on sequencing technology) frequently introduced frameshift mutations into reads [60]. Consequently, the number of identifiable homologs on shorter reads was 20-30% lower than for Sanger reads from a bacterial metagenome and 70% lower for viral metagenomic samples [61]. As early as 2008, Wommack et al. concluded that despite the reduced cost per basepair of pyrosequencing, the cost per unit of information was comparable with Sanger sequencing [61]. Whilst the continued reduction of sequencing costs has tilted the cost-benefit balance firmly in favor of next-generation sequencing, the issues of annotating short fragments remain [48,62]. It is no coincidence that the ubiquitous contextualization of the ecological significance of novel, experimentally derived microbial function overwhelmingly choose the GOS and longer read datasets for their analyses [63-66]. Furthermore, the rate of discovery of putative protein sequence has dwarfed the rate at which protein structure and function can be characterized, and has made manual curation of functional genes unfeasible (Fig. 1). Of the 20.6 million protein sequences in the current (2012_03) release of the UniProt database, only 2.8% have had their existence confirmed either at the protein or the transcript level (http://www.ebi.ac.uk/uniprot/TrEMBLstats/). Instead, automated annotation via homology transfer from similar sequences with known function was favored. However, homology and sequence similarity are not synonymous. As the number of automated annotations of increasingly diverse putative proteins continues to increase exponentially, transfer of homology has resulted in ‘homology creep’ to non-homologous sequences [67]. Automated annotation can be drastically improved via incorporation of a measure of evolutionary distance (phylogenomics) but requires near-full length protein sequences [68]. Only very recently has sequencing technology improved sufficiently to approach such lengths (~700-1100 bp) [69], whilst maintaining sufficient coverage to allow shotgun metagenomics to achieve its full potential.

Despite dramatic increases in sequencing efficiency (Fig. 1), inadequate coverage, which translates into “under sampling”, remains a major issue in metagenomics. Under sampling compromises some of the favorite experimental design strategies of ecologists, who often seek to understand how functional aspects of microbial communities vary along clines, such as gradients in latitude, temperature, or productivity. Under sampling can be ameliorated by binning data at courser resolution, for example, by COG instead of by species. However, such a shift in strategy elicits a significant cost - the loss of genome context, which is often crucial for interpreting function.

## Recapturing lost information: assembly and SAGs

To overcome the effects of short reads on accurate annotation, assembly of fragments into longer contigs has been attempted. Whilst numerous assembly programs have successfully reconstructed genomes from clonal organisms [70], successful assembly of contigs from metagenomic data is more limited. Problems with metagenomic assembly arise from poor community coverage [71], significant genomic variance in natural populations [12,20] and a high risk of chimeric sequence generation [62,72,73]. Furthermore, repetitive reads present a dichotomy for metagenome assembly [74]. During the assembly of a clonal genome, repetitive reads are problematic during construction of de Brujin graphs and are therefore removed from analysis [75,76]. Conversely, in a metagenome, repetitive reads are likely to come from higher coverage of dominant organisms and should therefore be assembled together. To tackle this issue, new ‘digital normalization’ algorithms to remove redundant reads and improve assembly are now emerging [77]. The likelihood of meaningful assembly is directly related to the complexity of the sampled community and the genetic variability of its constituents. In relatively simple systems, a combination of deep sequencing with short reads combined with longer reads for fragment recruitment and robust chimera checking are yielding some successes in assembling both microbial [11,62,78-80] and viral [81] genomes. Even with long reads, assemblies of metagenomes from complex communities (>400 taxa) are problematic. However, the longer reads have an inherent advantage of capturing full gene information for more accurate annotation [62]. The emergence of new metagenome assembly tools such as MetaORFA [82], Genovo [83], MetaVelvet [84], Meta-IDBA [85] and SEAStAR [86] may improve the utility of short sequences in metagenomes from complex communities.

To circumvent assembly problems most metagenomic studies rely heavily on complete or draft genomes to identify fragmentary sequences and interpret natural variation in an organismal context. Many online tools, e.g. MG-RAST, use this strategy, and have limited power to resolve metagenomic data originating from uncharted sectors of microbial diversity.  Advances in culturing technology have made more genomes from relevant organisms available, but the uncultured part of microbial diversity remains substantial.  To bridge this gap, researchers are increasingly turning to single cell genome amplification. Individual cells are isolated either by micromanipulation or by fluorescent-activated cell sorting (FACS) flow cytometry before lysis and multiple displacement amplification (MDA) of DNA to concentrations suitable for sequencing [87]. DNA can then be deeply sequenced with short-read, high throughput sequencing and readily assembled into large contigs, often assisted by scaffolding using longer Sanger reads [88,89]. Single-cell genomics is a significant advance that is providing draft genomes from organisms, many of them important, that have so far evaded cultivation. This technique was recently used to sequence the genome of SAR86, an important, highly abundant, but as-yet uncultured marine aerobic chemoheterotroph [90]. With ever-decreasing sequencing costs and rapid FACS cytometry cell isolation, it is not difficult to imagine that high-throughput metagenomics of important community representatives, and/or populations within a community from single-cell amplified genomes is imminent. Such an approach will avoid issues of chimeric assemblies (other than contaminants) and will enable functional annotation with intact synteny. Amplification-free, single-molecule DNA sequencing technologies such as those implemented in the MinION™ and GridION™ (Oxford Nanopore) and PacBio *RS* (Pacific Biosciences) will further reduce the cost, simplify assembly and improve the accuracy of single-cell metagenomics, perhaps even removing the need for cell isolation entirely. Accurate assembly of genomes from single cells and/or reads long enough to contain complete operons will have two major advantages. Firstly, complete genes will enable fine-scale intra and inter-specific phylogenomic analyses, improving our understanding of community structure and how bacterial diversity is derived and maintained via periodic selection [91] and viral predation [23,92,93]. Secondly, accurate annotation of genes will dramatically improve the capacity of systems biologists to model community connectivity and thus the effects of perturbations on microbial biogeochemical processes [94].

## The continued importance of culturing

Accurate annotation may provide insight into the metabolic potential of an organism and its community. However, it is often difficult to predict the ecological significance of an annotated gene without first considering its function in the broad context of metabolism. In some cases, this can be achieved by reconstructing metabolic pathways and testing predictions by experimentation with axenic cultures. The demonstration that SAR11 bacteria are methylovores was an example of this strategy [64].  In other cases, particularly those in which a gene is functioning in non-canonical pathways that are not represented in KEGG of other databases, the only choice may be exploratory experiments with cells in culture. SAR11 proteorhodopsin provides an apt example. Whilst the abundance and biochemistry of proteorhodopsin as an ATP-generating proton pump had previously been described from metagenomic data, exposure to light did not significantly improve the growth of SAR11 in axenic cultures, as would have been predicted from the annotated function. A decade after proteorhodopsin was first reported in metagenomic datasets, Steindler et al. showed that it provides an important source of ATP under conditions of carbon starvation, with cells grown in the dark forced to consume endogenous reserves of carbon for survival [95]. None of this surprises the average genome scientist, who by now, accustomed to the principles of systems biology, understands that selection is acting to shape the output from genes functioning in a concerted way.

## The ‘post-*Beagle’* era of metagenomics

There is little doubt that metagenomics has revolutionized our perspective of microbial taxonomic and functional diversity, and the scale of the generation of testable hypotheses from patterns observed through metagenomic studies has led to favorable comparisons with Darwin’s *Beagle* voyage [96]. However, like *The Origin of Species*, the power of observational science lies not in the data collection, but in the analysis and experimentation. After his four-year voyage on the *Beagle,* Darwin’s *magnum opus* resulted from two decades of experimental evidence to test his hypotheses on common ancestry, convergent evolution and descent with modification [97]. Similarly, the discovery of proteorhodopsin by Béjà et al. in marine metagenomic datasets was more potent for its confirmation through *in vitro* cloning and purification from complete gene sequences on 130-kb fragments [8]. Although early metagenomic investigations with pyrosequencing provided more data, the increased error rates in sequencing, assembly and annotation would have made the success story of bacteriorhodopsin less likely. Darwin’s lack of training as an ornithologist resulted in erroneous classification of some Galápagos finches as blackbirds and required careful curation by John Gould before the true extent of adaptive selection was apparent [97,98]. Analogously, genomic fragments annotated via transfer of homology from similar sequences with known function suffer from a similar issue. Greater efforts in novel protein curation are required for accurate predictions of taxonomic and functional diversity to better elucidate their roles in biogeochemical cycling. It is worth remembering, however, that metagenomics is still in its infancy and that the difficulties of short read lengths and fragmented genes are likely to be transient. Improvements to sequencing biochemistry and new methodologies are now increasing read lengths to a critical point where near-complete genes can be captured on a single read whilst maintaining depth of coverage, significantly improving annotation even when assembly into full genomes is difficult. Lessons have been learned for replicated experimental design to allow for robust statistical analysis and standards for metadata collection will improve comparisons between datasets [38,99]. Broad, shallow, replicated sequencing across large numbers of samples, followed by targeted deep sequencing and single-cell genomics will allow investigators to define a hypothesis; identify samples most likely to provide insight and then identify metabolic potential within single genomes and the community as a whole [38]. Confirmation of predicted biochemistry in axenic and community cultures of important taxa [100] will improve our ability to more accurately predict, contextualize and, importantly, test novel function and its role within bacterial populations and their wider communities, and will continue to drive the metagenomic revolution in microbial ecology.

**Acknowledgements**

We would like to thank Cameron Thrash and Bánk Beszteri for comments and critical review of this manuscript. We would also like to thank Predrag Radivojac for provision of data used in Figure 1.

1. Staley JT, Konopka A: **Measurement of in situ activities of nonphotosynthetic microorganisms in aquatic and terrestrial habitats.** *Annu. Rev. Microbiol.* 1985, **39**:321–346.

2. Schmidt TM, DeLong EF, Pace NR: **Analysis of a marine picoplankton community by 16S rRNA gene cloning and sequencing.** *Journal of Bacteriology* 1991, **173**:4371–4378.

3. Vergin KL, Urbach E, Stein JL, DeLong EF, Lanoil BD, Giovannoni SJ: **Screening of a fosmid library of marine environmental genomic DNA fragments reveals four clones related to members of the order Planctomycetales.** *Applied and Environmental Microbiology* 1998, **64**:3075–3078.

4. Stein JL, Marsh TL, Wu KY, Shizuya H, DeLong EF: **Characterization of uncultivated prokaryotes: isolation and analysis of a 40-kilobase-**

**pair genome fragment from a planktonic marine archaeon.** *Journal of Bacteriology* 1996, **178**:591–599.

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**. *Chemistry & biology* 1998, **5**:R245–R249.

6. Schleper C, DeLong EF, Preston CM, Feldman RA, Wu KY, Swanson RV: **Genomic analysis reveals chromosomal variation in natural populations of the uncultured psychrophilic archaeon Cenarchaeum symbiosum.** *Journal of Bacteriology* 1998, **180**:5003–5009.

7. Breitbart M, Salamon P, Andresen B, Mahaffy JM, Segall AM, Mead D, Azam F, Rohwer F: **Genomic analysis of uncultured marine viral communities.** *Proc. Natl. Acad. Sci. U.S.A.* 2002, **99**:14250–14255.

**8. Béjà O, Aravind L, Koonin EV, Suzuki MT, Hadd A, Nguyen, L.P., Jovanovich SB, Gates CM, Feldman RA, Spudich JL: **Bacterial rhodopsin: evidence for a new type of phototrophy in the sea**. *Science* 2000, **289**:1902–1906.

This study marked the discovery of bacteriorhodopsin in marine bacterioplankton, and is often heralded as one of the great successes for metagenomic investigations. This is in part due to the comprehensive analysis of gene function and kinetics via heterologous gene expression in *E. coli*, made possible by the long read lengths generated with BAC clones.

**9. Gilbert JA, Dupont CL: **Microbial metagenomics: beyond the genome.** *Annu. Rev. Marine. Sci.* 2011, **3**:347–371.

An excellent and comprehensive review of metagenomics that describes the scale and breadth of metagenomic investigations, and the resulting computational challenges.

*10. 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.

This study was the pilot study for the Global Ocean Survey and marks a shift in the scale of metagenomic investigations, identifying over 1.2 million new genes. In its concluding remarks, the paper correctly predicts a significant reduction in sequencing costs and the possibility of targeting rare, uncultured taxa within a community.

11. 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.

**12. 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.

A follow-up paper to the 2004 pilot study by Venter et al., this investigation extended the analysis across the world's oceans and generated (and continues to expand) the largest shotgun metagenomic dataset currently available. As a consequence, this paper spawned numerous bioinformatic analyses of the data both by the original investigating group and others.

13. 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.

14. Meyer F, Paarmann D, D'Souza M, Olson R, Glass E, Kubal M, Paczian T, Rodriguez A, Stevens R, Wilke A, et al.: **The metagenomics RAST server – a public resource for the automatic phylogenetic and functional analysis of metagenomes**. *BMC Bioinformatics* 2008, **9**:386.

15. Sun S, Chen J, Li W, Altintas I, Lin A, Peltier S, Stocks K, Allen EE, Ellisman M, Grethe J, et al.: **Community cyberinfrastructure for Advanced Microbial Ecology Research and Analysis: the CAMERA resource.** *Nucleic Acids Res.* 2011, **39**:D546–51.

16. Markowitz VM, Chen IMA, Chu K, Szeto E, Palaniappan K, Grechkin Y, Ratner A, Jacob B, Pati A, Huntemann M, et al.: **IMG/M: the integrated metagenome data management and comparative analysis system**. *Nucleic Acids Res.* 2011, **40**:D123–D129.

17. Garrity GM, Field D, Kyrpides N, Hirschman L, Sansone S-A, Angiuoli S, Cole JR, Glöckner FO, Kolker E, Kowalchuk G, et al.: **Toward a Standards-Compliant Genomic and Metagenomic Publication Record**. *OMICS: A Journal of Integrative Biology* 2008, **12**:157–160.

18. Medini D, Donati C, Tettelin H, Masignani V, Rappuoli R: **The microbial pan-genome**. *Current Opinion in Genetics & Development* 2005, **15**:589–594.

19. Kettler GC, Martiny AC, Huang K, Zucker J, Coleman ML, Rodrigue S, Chen F, Lapidus A, Ferriera S, Johnson J, et al.: **Patterns and**

**Implications of Gene Gain and Loss in the Evolution of Prochlorococcus**. *PLoS Genet* 2007, **3**:e231.

20. Wilhelm LJ, Tripp HJ, Givan SA, Smith DP, Giovannoni SJ: **Natural variation in SAR11 marine bacterioplankton genomes inferred from metagenomic data.** *Biol. Direct* 2007, **2**:27.

21. Dinsdale EA, Edwards RA, Hall D, Angly F, Breitbart M, Brulc JM, Furlan M, Desnues C, Haynes M, Li L, et al.: **Functional metagenomic profiling of nine biomes**. *Nature* 2008, **452**:629–632.

22. Rusch DB, Martiny AC, Dupont CL, Halpern AL, Venter JC: **Characterization of Prochlorococcus clades from iron-depleted oceanic regions.** *Proc. Natl. Acad. Sci. U.S.A.* 2010, **107**:16184–16189.

23. Rodriguez-Valera F, Martin-Cuadrado A-B, Rodriguez-Brito B, Pašić L, Thingstad TF, Rohwer F, Mira A: **Explaining microbial population genomics through phage predation**. *Nat Rev Micro* 2009, **7**:828–836.

24. Zhang Y, Gladyshev VN: **Trends in Selenium Utilization in Marine Microbial World Revealed through the Analysis of the Global Ocean Sampling (GOS) Project**. *PLoS Genet* 2008, **4**:e1000095.

25. Sebastian M, Ammerman JW: **The alkaline phosphatase PhoX is more widely distributed in marine bacteria than the classical PhoA**. *ISME J* 2009, **3**:563–572.

26. Luo H, Benner R, Long RA, Hu J: **Subcellular localization of marine bacterial alkaline phosphatases**. *Proc. Natl. Acad. Sci. U.S.A.* 2009, **106**:21219–21223.

27. Temperton B, Gilbert JA, Quinn JP, McGrath JW: **Novel analysis of oceanic surface water metagenomes suggests importance of polyphosphate metabolism in oligotrophic environments.** *PLoS ONE* 2011, **6**:e16499.

28. Larsen PE, Collart FR, Field D, Meyer F, Keegan KP, Henry CS, McGrath J, Quinn J, Gilbert JA: **Predicted Relative Metabolomic Turnover (PRMT): determining metabolic turnover from a coastal marine metagenomic dataset**. *Microbial Informatics and Experimentation* 2011, **1**:4.

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

30. Kristiansson E, Hugenholtz P, Dalevi D: **ShotgunFunctionalizeR: an R-package for functional comparison of metagenomes**. *Bioinformatics*

2009, **25**:2737–2738.

31. Beszteri B, Temperton B, Frickenhaus S, Giovannoni SJ: **Average genome size: a potential source of bias in comparative metagenomics.** *ISME J* 2010, **4**:1075–1077.

**32. Hamady M, Walker JJ, Harris JK, Gold NJ, Knight R: **Error-correcting barcoded primers for pyrosequencing hundreds of samples in multiplex**. *Nature Methods* 2008, **5**:235–237.

In this study, Hamady et al. created DNA barcodes that could be attached to a primer to allow different samples to be sequenced in the same sequencing run. Barcode 'tags' were robust to sequencing error, enabling bioinformatic separation of samples post-sequencing. This method has now become the standard method for amplicon metagenomics to maximize the number of samples that can be processed in a single run.

33. Gilbert JA, Steele JA, Caporaso JG, Steinbrück L, Reeder J, Temperton B, Huse S, McHardy AC, Knight R, Joint I, et al.: **Defining seasonal marine microbial community dynamics.** *ISME J* 2012, **6**:298–308.

34. Karsenti E, Acinas SG, Bork P, Bowler C, De Vargas C, Raes J, Sullivan M, Arendt D, Benzoni F, Claverie J-M, et al.: **A holistic approach to marine eco-systems biology.** *PLoS Biol.* 2011, **9**:e1001177.

35. Turnbaugh PJ, Ley RE, Hamady M, Fraser-Liggett CM, Knight R, Gordon JI: **The Human Microbiome Project**. *Nature* 2007, **449**:804–810.

36. Qin J, Li R, Raes J, Arumugam M, Burgdorf KS, Manichanh C, Nielsen T, Pons N, Levenez F, Yamada T, et al.: **A human gut microbial gene catalogue established by metagenomic sequencing.** *Nature* 2010, **464**:59–65.

37. Kircher M, Kelso J: **High-throughput DNA sequencing - concepts and limitations**. *Bioessays* 2010, **32**:524–536.

*38. Knight R, Jansson J, Field D, Fierer N, Desai N, Fuhrman JA, Hugenholtz P, van der Lelie D, Meyer F, Stevens R, et al.: **Unlocking the potential of metagenomics through replicated experimental design.** *Nature Biotechnology* 2012, **30**:513–520.

This paper highlights the need for replication in metagenomic experimental design for statistically robust analysis and suggests initial broad, replicated and shallow sequencing across many samples followed by targeted, deeper sequencing of samples most likely to provide the most information. It is an excellent resource for consideration by those embarking on their first metagenomic investigation.

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

40. Sboner A, Mu XJ, Greenbaum D, Auerbach RK, Gerstein MB: **The real cost of sequencing: higher than you think!** *Genome Biol.* 2011, **12**:125.

41. Craft JA, Gilbert JA, Temperton B, Dempsey KE, Ashelford K, Tiwari B, Hutchinson TH, Chipman JK: **Pyrosequencing of Mytilus galloprovincialis cDNAs: tissue-specific expression patterns.** *PLoS ONE* 2010, **5**:e8875.

42. Cuvelier ML, Allen AE, Monier A, McCrow JP, Messié M, Tringe SG, Woyke T, Welsh RM, Ishoey T, Lee J-H, et al.: **Targeted metagenomics and ecology of globally important uncultured eukaryotic phytoplankton.** *Proc. Natl. Acad. Sci. U.S.A.* 2010, **107**:14679–14684.

43. Meyerdierks A, Kube M, Kostadinov I, Teeling H, Glöckner FO, Reinhardt R, Amann R: **Metagenome and mRNA expression analyses of anaerobic methanotrophic archaea of the ANME-1 group.** *Environmental Microbiology* 2010, **12**:422–439.

44. Frias-Lopez J, Shi Y, Tyson GW, Coleman ML, Schuster SC, Chisholm SW, DeLong EF: **Microbial community gene expression in ocean surface waters**. *Proc. Natl. Acad. Sci. U.S.A.* 2008, **105**:3805.

45. Chistoserdova L: **Methylotrophy in a Lake: from Metagenomics to Single-Organism Physiology**. *Applied and Environmental Microbiology* 2011, **77**:4705–4711.

46. 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.

47. Bench SR, Hanson TE, Williamson KE, Ghosh D, Radosovich M, Wang K, Wommack KE: **Metagenomic characterization of Chesapeake Bay virioplankton.** *Applied and Environmental Microbiology* 2007, **73**:7629–7641.

48. Victoria JG, Kapoor A, Li L, Blinkova O, Slikas B, Wang C, Naeem A, Zaidi S, Delwart E: **Metagenomic Analyses of Viruses in Stool Samples from Children with Acute Flaccid Paralysis**. *Journal of Virology* 2009, **83**:4642–4651.

49. Kristensen DM, Mushegian AR, Dolja VV, Koonin EV: **New dimensions of the virus world discovered through metagenomics**. *Trends in*

*Microbiology* 2010, **18**:11–19.

50. Kosakovsky Pond S, Wadhawan S, Chiaromonte F, Ananda G, Chung WY, Taylor J, Nekrutenko A, The Galaxy Team: **Windshield splatter analysis with the Galaxy metagenomic pipeline**. *Genome Research* 2009, **19**:2144–2153.

51. Youssef N, Sheik CS, Krumholz LR, Najar FZ, Roe BA, Elshahed MS: **Comparison of Species Richness Estimates Obtained Using Nearly Complete Fragments and Simulated Pyrosequencing-Generated Fragments in 16S rRNA Gene-Based Environmental Surveys**. *Applied and Environmental Microbiology* 2009, **75**:5227–5236.

*52. Reeder J, Knight R: **The “rare biosphere”: a reality check**. *Nature Methods* 2009, **6**:636–637.

An excellent review describing how sequencing error has resulted in a significant over-estimation of the number of 'rare' taxa in metagenomic datasets.

*53. Gomez-Alvarez V, Teal TK, Schmidt TM: **Systematic artifacts in metagenomes from complex microbial communities**. *ISME J* 2009, **3**:1314–1317.

This investigation describes the discovery that metagenomic datasets from pyrosequencing contain significant numbers of replicated reads that are artifacts from the emulsion polymerase chain reaction used during sequencing.

*54. Kunin V, Engelbrektson A, Ochman H, Hugenholtz P: **Wrinkles in the rare biosphere: pyrosequencing errors can lead to artificial inflation of diversity estimates.** *Environmental Microbiology* 2010, **12**:118–123.

This paper experimentally confirmed the over-estimation of diversity by 16S rRNA amplicon metagenomics (reviewed in [52]). Kunin et al. constructed an amplicon library using DNA from *E. coli* MG1655 and found that diversity was over-estimated by two orders of magnitude as a result of sequencing error. To alleviate this, the paper recommends countermeasures of stringent end-trimming of reads and operational taxonomic unit clustering at 97% sequence identity - practices that are now standard approaches in metagenomic analyses.

55. Haas BJ, Gevers D, Earl AM, Feldgarden M, Ward DV, Giannoukos G, Ciulla D, Tabbaa D, Highlander SK, Sodergren E, et al.: **Chimeric 16S rRNA sequence formation and detection in Sanger and 454-pyrosequenced PCR amplicons**. *Genome Research* 2011, **21**:494–504.

56. Quince C, Lanzén A, Davenport RJ, Turnbaugh PJ: **Removing Noise**

**From Pyrosequenced Amplicons**. *BMC Bioinformatics* 2011, **12**:38.

57. Feinstein LM, Sul WJ, Blackwood CB: **Assessment of Bias Associated with Incomplete Extraction of Microbial DNA from Soil**. *Applied and Environmental Microbiology* 2009, **75**:5428–5433.

58. Temperton B, Field D, Oliver A, Tiwari B, Mühling M, Joint I, Gilbert JA: **Bias in assessments of marine microbial biodiversity in fosmid libraries as evaluated by pyrosequencing.** *ISME J* 2009, **3**:792–796.

*59. Danhorn T, Young CR, DeLong EF: **Comparison of large-insert, small-insert and pyrosequencing libraries for metagenomic analysis**. 2012, doi:10.1038/ismej.2012.35.

In this investigation, Danhorn et al. confirmed the hypothesis of [58] that the under-representation of taxa in fosmid libraries was likely a factor of G+C content, highlighting that different sequencing approaches are likely to harbor different biases which must be accounted for.

60. Rho M, Tang H, Ye Y: **FragGeneScan: predicting genes in short and error-prone reads**. *Nucleic Acids Res.* 2010, **38**:e191–e191.

**61. Wommack KE, Bhavsar J, Ravel J: **Metagenomics: Read Length Matters**. *Applied and Environmental Microbiology* 2008, **74**:1453–1463.

At a time when the greatly increased coverage and cheaper costs of pyrosequencing were emerging, this paper sounded an important cautionary note with regards to the amount of information contained in the shorter reads, particularly with regard to functional annotation of shotgun metagenomic reads. Recent analyses in [62] confirmed that even with the improvements in sequencing the issue still remains.

62. Mende DR, Waller AS, Sunagawa S, Järvelin AI, Chan MM, Arumugam M, Raes J, Bork P: **Assessment of Metagenomic Assembly Using Simulated Next Generation Sequencing Data**. *PLoS ONE* 2012, **7**:e31386.

63. Schwalbach MS, Tripp HJ, Steindler L, Smith DP, Giovannoni SJ: **The presence of the glycolysis operon in SAR11 genomes is positively correlated with ocean productivity**. *Environmental Microbiology* 2010, **12**:490–500.

64. Sun J, Steindler L, Thrash JC, Halsey KH, Smith DP, Carter AE, Landry ZC, Giovannoni SJ: **One Carbon Metabolism in SAR11 Pelagic Marine Bacteria**. *PLoS ONE* 2011, **6**:e23973.

65. Chen Y, Patel NA, Crombie A, Scrivens JH, Murrell JC: **Bacterial flavin-containing monooxygenase is trimethylamine monooxygenase**.

*Proc. Natl. Acad. Sci. U.S.A.* 2011, **108**:17791–17796.

66. Feingersch R, Philosof A, Mejuch T, Glaser F, Alalouf O, Shoham Y, agrave OBEJ: **Potential for phosphite and phosphonate utilization by Prochlorococcus**. 2011, **6**:827–834.

**67. Friedberg I: **Automated protein function prediction--the genomic challenge**. *Briefings in Bioinformatics* 2006, **7**:225–242.

An excellent review describing the issues of functional annotation of metagenomic shotgun sequencing reads.

68. Eisen JA: **Phylogenomics: improving functional predictions for uncharacterized genes by evolutionary analysis**. *Genome Research* 1998, **8**:163–167.

69. Glenn TC: **Field guide to next-generation DNA sequencers**. *Molecular Ecology Resources* 2011, **11**:759–769.

70. Earl D, Bradnam K, St John J, Darling A, Lin D, Fass J, Yu HOK, Buffalo V, Zerbino DR, Diekhans M, et al.: **Assemblathon 1: A competitive assessment of de novo short read assembly methods**. *Genome Research* 2011, **21**:2224–2241.

71. Wooley JC, Godzik A, Friedberg I: **A primer on metagenomics.** *PLoS Comput. Biol.* 2010, **6**:e1000667.

72. Pignatelli M, Moya A: **Evaluating the Fidelity of De Novo Short Read Metagenomic Assembly Using Simulated Data**. *PLoS ONE* 2011, **6**:e19984.

73. Charuvaka A, Rangwala H: **Evaluation of short read metagenomic assembly**. *BMC Genomics* 2011, **12**:S8.

74. Desai N, Antonopoulos D, Gilbert JA, Glass EM, Meyer F: **From genomics to metagenomics.** *Current Opinion in Biotechnology* 2012, **23**:72–76.

75. Zerbino DR, Birney E: **Velvet: Algorithms for de novo short read assembly using de Bruijn graphs**. *Genome Research* 2008, **18**:821–829.

76. Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJM, Birol I: **ABySS: A parallel assembler for short read sequence data**. *Genome Research* 2009, **19**:1117–1123.

77. Brown C, Howe A, Zhang Q, Pyrkosz A: **A single pass approach to reducing sampling variation, removing errors, and scaling *de novo***

**assembly of shotgun sequences**. *arXiv* 2012, [no volume].

78. Denef VJ, Mueller RS, Banfield JF: **AMD biofilms: using model communities to study microbial evolution and ecological complexity in nature**. *ISME J* 2010, **4**:599–610.

79. Mueller RS, Denef VJ, Kalnejais LH, Suttle KB, Thomas BC, Wilmes P, Smith RL, Nordstrom DK, McCleskey RB, Shah MB, et al.: **Ecological distribution and population physiology defined by proteomics in a natural microbial community**. *Molecular Systems Biology* 2010, **6**:1–12.

80. Luo C, Tsementzi D, Kyrpides NC, Konstantinidis KT: **Individual genome assembly from complex community short-read metagenomic datasets**. 2011, **6**:898–901.

81. Garcia-Heredia I, Martin-Cuadrado A-B, Mojica FJM, Santos F, Mira A, Antón J, Rodriguez-Valera F: **Reconstructing Viral Genomes from the Environment Using Fosmid Clones: The Case of Haloviruses**. *PLoS ONE* 2012, **7**:e33802.

82. Ye Y, Tang H: **An ORFome assembly approach to metagenomics sequences analysis.** *J Bioinform Comput Biol* 2009, **7**:455–471.

83. Laserson J, Jojic V, Koller D: **Genovo: De NovoAssembly for Metagenomes**. *Journal of Computational Biology* 2011, **18**:429–443.

84. Namiki T, Hachiya T, Tanaka H, Sakakibara Y: **MetaVelvet: an extension of Velvet assembler to de novo metagenome assembly from short sequence reads**. *Proceedings of the 2nd ACM Conference on Bioinformatics, Computational Biology and Biomedicine* 2011, [no volume].

85. Peng Y, Leung HCM, Yiu SM, Chin FYL: **Meta-IDBA: a de Novo assembler for metagenomic data**. *Bioinformatics* 2011, **27**:i94–i101.

86. Iverson V, Morris RM, Frazar CD, Berthiaume CT, Morales RL, Armbrust EV: **Untangling Genomes from Metagenomes: Revealing an Uncultured Class of Marine Euryarchaeota**. *Science* 2012, **335**:587–590.

87. Lasken RS: **Single-cell genomic sequencing using Multiple Displacement Amplification**. *Current Opinion in Microbiology* 2007, **10**:510–516.

88. Ishoey T, Woyke T, Stepanauskas R, Novotny M, Lasken RS: **Genomic sequencing of single microbial cells from environmental samples**. *Current Opinion in Microbiology* 2008, **11**:198–204.

89. Chitsaz H, Yee-Greenbaum JL, Tesler G, Lombardo M-J, Dupont CL, Badger JH, Novotny M, Rusch DB, Fraser LJ, Gormley NA, et al.: **Efficient de novo assembly of single-cell bacterial genomes from short-read data sets**. *Nature Biotechnology* 2011, **29**:915–921.

**90. Dupont CL, Rusch DB, Yooseph S, Lombardo M-J, Alexander Richter R, Valas R, Novotny M, Yee-Greenbaum J, Selengut JD, Haft DH, et al.: **Genomic insights to SAR86, an abundant and uncultivated marine bacterial lineage**. *ISME J* 2011, doi:10.1038/ismej.2011.189.

Although the relative abundance of SAR86 in marine metagenomic datasets had highlighted it as an important member of the bacterioplankton, it was resistant to culturing attempts and thus a complete genome was not available. Dupont et al. used a combination of metagenomics and SAGs to re-construct the genome of SAR86 from environmental samples, demonstrating the importance of this technique for understanding important non-cultured organisms.

91. Cohan FM, Perry EB: **A Systematics for Discovering the Fundamental Units of Bacterial Diversity**. *Current Biology* 2007, **17**:R373–R386.

92. Thingstad TF: **Elements of a theory for the mechanisms controlling abundance, diversity, and biogeochemical role of lytic bacterial viruses in aquatic systems**. *Limnology and Oceanography* 2000, **45**:1320–1328.

93. Wilmes P, Simmons SL, Denef VJ, Banfield JF: **The dynamic genetic repertoire of microbial communities**. *FEMS Microbiology Reviews* 2009, **33**:109–132.

94. Giovannoni SJ, Vergin KL: **Seasonality in Ocean Microbial Communities**. *Science* 2012, **335**:671–676.

95. Steindler L, Schwalbach MS, Smith DP, Chan F, Giovannoni SJ: **Energy Starved Candidatus Pelagibacter Ubique Substitutes Light-Mediated ATP Production for Endogenous Carbon Respiration**. *PLoS ONE* 2011, **6**:e19725.

96. Gilbert JA, O'Dor R, King N, Vogel TM: **The importance of metagenomic surveys to microbial ecology: or why Darwin would have been a metagenomic scientist**. *Microbial Informatics and Experimentation* 2011, **1**:5.

97. Jones S: *Darwin's Island*. Little, Brown Book Group; 2009.

98. Gould J: *Birds. Part 3 of The zoology of the voyage of H.M.S. Beagle [Internet]*. Smith Elder & Co; 1841.

99. Field D, Garrity G, Gray T, Morrison N, Selengut J, Sterk P, Tatusova T, Thomson N, Allen MJ, Angiuoli SV, et al.: **The minimum information about a genome sequence (MIGS) specification**. *Nature Biotechnology* 2008, **26**:541–547.

100. Giovannoni S, Stingl U: **The importance of culturing bacterioplankton in the “omics” age**. *Nat Rev Micro* 2007, **5**:820–826.

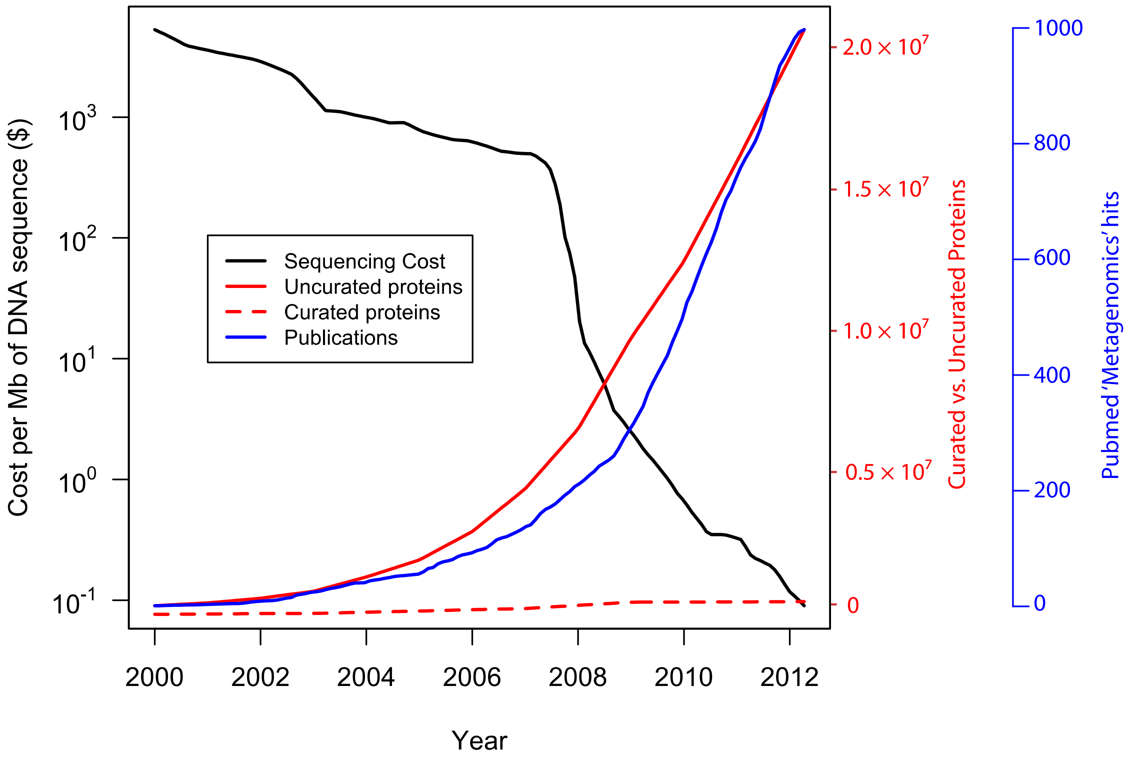

**Figure 1** - Cost of DNA sequencing and its impact on genomics and metagenomics. Y-axis 1 (black): The cost per Mb of DNA sequencing on a log scale (data from http://www.genome.gov/sequencingcosts/). Y-axis 2 (red): The total number of sequences in the UniProt (http://www.uniprot.org/) database for automatically annotated (solid, TrEMBL database) and manually annotated (dashed, SwissProt database) proteins (data courtesy of Predrag Radivojac*).* Y-axis 3 (blue): The total number of metagenomics publications in PubMed (http://www.ncbi.nlm.nih.gov/pubmed/). The search term "metagenomics"[MeSH Terms] OR "metagenomics" was used to retrieve publication records in XML-format and binned by month and year according to the 'DateCreated' element.