Available online at www.sciencedirect.com

SciVerse ScienceDirect

Current Opinion in Biotechnology

From genomics to metagenomics

Narayan Desai[1], Dion Antonopoulos[2,3], Jack A Gilbert[2,4], Elizabeth M Glass[1] and Folker Meyer[1,2]

Next-generation sequencing has changed metagenomics. However, sequencing DNA is no longer the bottleneck, rather, the bottleneck is computational analysis and also interpretation. Computational cost is the obvious issue, as is tool limitations, considering most of the tools we routinely use have been built for clonal genomics or are being adapted to microbial communities. The current trend in metagenomics analysis is toward reducing computational costs through improved algorithms and through analysis strategies. Data sharing and interoperability between tools are critical, since computation for metagenomic datasets is very high.

Addresses

[1] Mathematics and Computer Science Division, Argonne National Laboratory, 9700 South Cass Avenue, Argonne, IL 60439, USA
[2] Institute for Genomics and Systems Biology, Argonne National Laboratory, 9700 South Cass Avenue, Argonne, IL 60439, USA
[3] Biosciences Division, Argonne National Laboratory, 9700 South Cass Avenue, Argonne, IL 60439, USA
[4] Department of Ecology and Evolution, University of Chicago, 5640 South Ellis Avenue, Chicago, IL 60637, USA

Corresponding author: Meyer, Folker (folker@anl.gov)

Current Opinion in Biotechnology 2012, **23**:72–76

This review comes from a themed issue on
Analytical biotechnology
Edited by Wei E. Huang and Jizhong Zhou

Available online 5th January 2012

0958-1669/$ – see front matter

DOI 10.1016/j.copbio.2011.12.017

Metagenomics: a brief history and current trends

Metagenomics has changed in the past 13 years from sequencing of cloned DNA fragments using Sanger technology [1–4] to direct sequencing of DNA without heterologous cloning [5,6]. Studies like the Global Ocean Survey (GOS) led to the creation of novel tools like fragment recruitment [2]. The vast majority of the tools, however, were developed in the context of 'clonal genomics' projects. Currently, technical aspects limit metagenomics. Consider, for example, the inability of current tools to reliably assemble longer contigs from multistrain, multispecies mixtures. Therefore, metagenomics cannot be viewed simply as an extension of genomics; instead, it requires a different paradigm of bioinformatics tool design. Just as human genome analysis is moving from one to many sequenced representative genomes, bioinformaticians dealing with microbial community sequence data need to consider metagenomes as complex mixtures of (pan)genomes [7].

The cost of DNA sequencing has plummeted [8] in recent years, making this technology accessible to more researchers. This has resulted in the generation of large numbers of massive datasets. The primary factor driving metagenomic research in next-generation sequencing technologies is financial. Short-read sequencing platforms provide an inexpensive means to deeply sample a microbial community. However, these sequencing technologies are less than ideal since they were driven by the human genome resequencing effort, which does not necessarily require long reads. Hence, metagenomics researchers have found themselves analyzing large, short-read datasets using tools built for long reads and, most important, for clonal datasets.

Data analysis is the key limiting factor in metagenomic studies. DNA sequence data production is no longer the bottleneck in microbial studies; instead, increased data volumes are posing significant challenges to the existing analysis tools and indeed to the community providing analysis systems. Even simple computing of sequence similarity results is becoming a limiting factor in metagenome analysis. This growth in dataset size, in conjunction with computational complexity of analysis, has left the metagenomics community in an unsustainable position, in terms of both financial cost and feasibility of analysis itself.

From genomics to metagenomics

Many metagenomics projects have been using a version of the standard genomics analysis workflow: sequence as deeply as possible, assemble reads into consensus 'contigs,' and annotate these contigs [9•,10,11••]; however, this is a costly strategy when using any of the current sequencing platforms. It is also important that each analysis strategy match the scientific goals of the study. For example, assembling into contigs, deleting singleton reads, and subsequently performing an analysis of the microbial communities will result in significant biases.

Challenges of metagenomic data

Metagenomic sequencing data from a complex community often represents only a minute fraction of the community. For example, to produce a dataset representing onefold coverage of a microbial community from a gram of soil would require more than 6000 HiSeq2000 runs at a cost of $267 million. And this is no guarantee, since unknown community composition and relative abundances limit our ability to calculate the coverage required to 'sequence to extinction' *robustly*.[1] Metagenomics currently lacks the tools to determine whether sufficient coverage is available for the type of analysis planned or whether one can interpret data of a certain depth for a community of a given complexity the way that is being planned. Therefore, the standard low coverage in metagenomic studies generates a dataset that reflects a random subsampling of the genomic content of the individual community members. If money was no object and complete coverage of all individuals was achieved, then metagenomics would require tools similar to those used for genomic analysis. Since this is currently infeasible, however, new tools must be developed.

The role of contigs and data fidelity

While consensus sequences are useful for *de novo* sequencing or resequencing in clonal genomics, they are of limited use for metagenomics. Using a consensus sequence built from many reads shields us from individual read errors but also obscures community structure throughout the taxonomic tree by collapsing strain-level, species-level, genus-level and family-level variation that could help with pangenome comparisons.

Clonal genomics uses consensus sequences to increase quality by integrating over multiple observations (reads); however, since metagenomic sequence data represents a random subsample of genetic information from many potential genomes, each single read potentially provides new data. This situation not only presents a challenge to sequencing providers; it also complicates making any statements about variation. For instance, what allows us to distinguish variation from a read error in a complex microbial community at low sequence coverage? We need ways of determining the quality of a sequencing run and the fidelity of any derived statements. Artifacts from the sequencing process need to be determined for the analysis of each dataset as a fundamental metric.

Researchers interested in microbial community composition and microbial community function look at contigs as a means to an end rather than an end product. We assume here that a contig from a metagenomic assembly is 'virtual' unless proven real by *in vitro* methods such as PCR amplification or restriction analysis. A contig that is potentially a hybrid of many strains and even species, while informative, has only limited information content once one is interested in the composition of the strains or species within that contig. However, using recruiting of short reads to a contig is a computationally inexpensive method for displaying species or strain differences.

In the days of Sanger-based clonal genomics, the read coverage of the target genome proved useful as a means of understanding the expected quality of the assembled consensus sequence. The number of gaps in the (un)finished genome sequence was used as another indicator of quality. Metagenomics lacks a corresponding instrument.

We believe that the scientific question motivating the sequencing needs to govern the depth (and type?) of sequencing applied, rather than the simple model 'more is better' and monetary considerations.

One of the fundamental tenets that bioinformaticians should hold true is that the tools they develop must be useful to and usable by the intended audience. In microbial metagenomics, one of the key groups is microbial ecologists. It is therefore essential that the dialog between these groups be open and capable of enabling continuous feedback; new tool developments quickly change the playing field and create new requirements. One good example is the migration from providing a catalog of functions and taxonomic composition to creating an extrapolation of these catalogs to additional metrics of community capability. Predictive elative Metabolic Turnover, PRMT [12••], converts metagenomic sequence data into relative metrics for the consumption or production of specific metabolites. Bioinformaticists should focus on exploring new ways to interpret metagenomic sequence data, with continuous reference to the needs of the community.

The information required by researchers can be generalized to who is out there and what are they doing.

Who is out there? or 'Binning' to taxonomic units

Binning tools were initially developed to allow placing of metagenomic contigs into taxonomic units [13•]. These tools typically require contigs of at least 1000 bp (the longer the better) to allow training the models used for classification [14]. It is not clear to us that current assembly algorithms produce contigs that are well suited to binning, since the algorithms optimize a mathematical function, constructing the shortest common superstring [15] without preserving species or strain identities in the process. Binning of long Sanger reads with the assumption that the data were from a clonal source made sense. Binning a contig created from short reads without considering all the implications of studying a complex mix of species and/or strains is fundamentally flawed and likely to reflect artifacts of the assembly process. In many ways it is accurate to say that genomicists pretend a metagenomes consists of a series of clonal contigs, rather than accepting the new challenge of analyzing a complex mixture of genomes.

[1] Even with perfect 16s amplicon based characterization of microbial communities, strain variants cannot be distinguished based on single-gene amplicons. One OTU as detected by a primer targeting a ribosomal gene is likely to consist of many different strains.

While this assumption might hold true for very simple systems such as acid mine drainage, initial unpublished data convince us that most currently used assembly algorithms [16••,17••,18,19] do not account for the presence of strain-level or species-level variation in a metagenomic mixture. One of the main assumptions deeply embedded into modern assembly tools is that in order to avoid being confused by repetitive elements, the assembly tools need to detect coverage and eliminate regions of higher coverage (read repeats) from the initial rounds of the assembly and add them back in later. However, doing so in a metagenome where defining the 'average abundance' is an impossible task, accepting only a small coverage band for assembly is introducing a serious bias into the data. It is arguable that assembly therefore should not be used to create contigs that can subsequently be binned to taxonomic units. We feel, however, that using contigs as artificial constructs for the purpose of binning reads is acceptable as long as it is understood that the contigs probably do not exist in nature.

What are they doing? or Functional assignment and finding genes

Using BLASTX [20] for defining functions of short DNA reads is tempting but computationally not tractable [21]. Alternative codes such as BLAT [22] or RAPSearch [23] are computationally friendlier but involve some loss of sensitivity. BLAT and RAPSearch also require an additional gene prediction stage such as FragGeneScan, MetaGeneMark, Metagene, or MetaGeneAnnotator [24••,25–27]. (BLASTX performs the prediction of potential coding genes implicitly but is limited to genes with representatives in the known protein databases.) Methods based on other classification techniques abound [28,29] but have not gained a lot of traction because the community has, so far, not sufficiently validated the results.

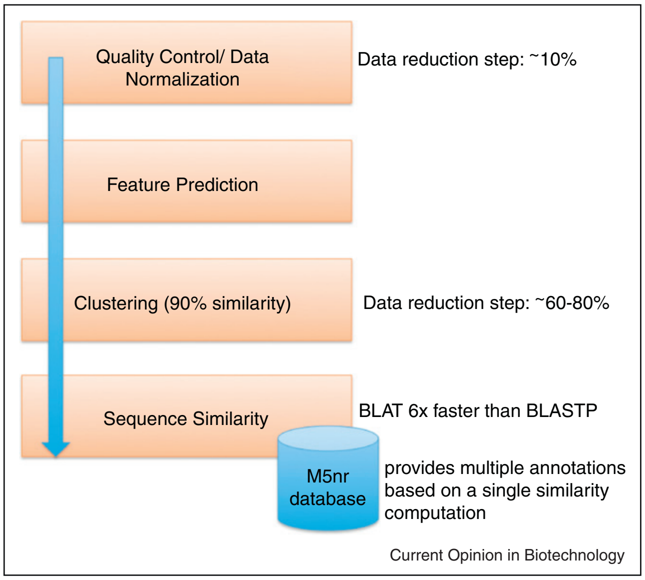
The computational cost of metagenome sequence analysis is so high that only a subset of the traditional tool set from clonal genomics is typically applied. While finding short, noncoding RNAs is clearly interesting to many and might lead to numerous new insights into the biology of the biome studied [30], the computational cost is high. The cost for running BLASTX analysis for large datasets on Amazon's EC2 cloud [21] is several times the cost of running the sequencing instrument, with sequencing cost dropping much faster than computing cost. Running analyses that are significantly more expensive than BLASTX, such as CRISPR [31•] or RFAM [32], is not current practice, with the trend going toward reducing the cost of computation.

Sharing data and analysis results

The cost-imbalance associated with bioinformatics analysis vs. data generation means that ideally the community should analyze a dataset only once and rely on a community archive to distribute the raw data and the analysis results. This culture of trusting third-party analysis results is currently not well established in metagenomics. Traditional data products (e.g. GenBank files) are not well suited to the requirements of metagenomics and fail, for example, to represent community properties. For metagenomics to meet the new challenges of the 'data bonanza' and to flourish, the sharing of data and analysis results needs to be established so the community of analysis providers (MG-RAST, IMG/M, or CAMERA [33•,34•,35]) do not spend their time and resources reanalyzing the same data. In the past this has been made difficult by the fact that the community sequence archives (SRA) [36] were not built to primarily support the current metagenomics use-cases. Important progress in this area has been seen with MG-RAST [33•], which has taken over the role of the de facto community archive and makes both raw data and computational results available for download and further analysis.

Sharing of data and results is clearly part of the solution as our ability to produce data grows faster than our computers [13•,37]. We can do this only if we create reliable data products and include more than mere sequence data. Derived data and provenance information needs to be exchangeable between groups and analysis providers in standard formats. In this context the Genome Standards Consortium [38••] has been a pioneer in standards development and in the context of their M5 project [39] started to work on exchangeable data formats. This will enable researchers to download data and analysis results.

Figure 1


Quality Control/ Data Normalization
Data reduction step: ~10%
Feature Prediction
Clustering (90% similarity)
Data reduction step: ~60-80%
Sequence Similarity
BLAT 6x faster than BLASTP
M5nr database
provides multiple annotations based on a single similarity computation
Current Opinion in Biotechnology

The MG-RAST system has made performance improvements to deal with the onslaught of metagenomics sequence data. This figure outlines, at a high level, the data reduction steps as well as changes in tools being used to speed computation. Overall, MG-RAST has seen a 40–60× improvement in speed throughout.

In addition it will enable the development of a series of downstream analysis and visualization tools.

Future directions

Metagenomic analytics is a rapidly changing field, with major revisions to both tools and analysis pipelines occurring frequently. We expect the major trends of data growth and algorithmic complexity to hold for at least a few years. For the foreseeable future, analysis strategy will be guided by tradeoffs among scalability, sensitivity and performance. These decisions must be made carefully, with specific goals, in order to determine the proper approach. MG-RAST has attempted to balance performance improvements with sensitivity of analysis (Figure 1). Studies such as the acid mine drainage study were in the past possible only on very simple microbial communities. With easier access to sequencing, we believe that by improving the tools available to the community we will see many such studies in the coming years, continuing to improve our understanding of the microbial world surrounding us.

Acknowledgements

This work was supported by the U.S. Department of Energy, under Contract DE-AC02-06CH11357.

References and recommended reading

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

- • of special interest
- •• of outstanding interest

1. 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.

The acid mine drainage community studied here using Sanger sequencing has only a handful of OTUs. In the coming years we expect to see similar studies for more complex communities if all the technical hurdles can be overcome.

2. 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.

3. Lorenz P, Schleper C: **Metagenome? a challenging source of enzyme discovery**. *J Mol Catal B: Enzymatic* 2002, **19**–**20**:13-19.

4. 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.

5. Gilbert JA, Field D, Huang Y, Edwards R, Li W, Gilna P, Joint I: **Detection of large numbers of novel sequences in the metatranscriptomes of complex marine microbial communities**. *PLoS ONE* 2008, **3**:e3042.

6. Gilbert JA, Field D, Swift P, Thomas S, Cummings D, Temperton B, Weynberg K, Huse S, Hughes M, Joint I, Somerfield PJ, Mühling M: **The taxonomic and functional diversity of microbes at a temperate coastal site: a 'multi-omic' study of seasonal and diel temporal variation**. *PLoS One.* 2010, **5**:e15545 doi: 10.1371/journal.pone.0015545.

7. Medini D, Donati C, Tettelin H, Masignani V, Rappuoli R: **The microbial pan-genome**. *Curr Opin Genet Dev* 2005, **15**:589-594.

8. Pushkarev D, Neff NF, Quake SR: **Single-molecule sequencing of an individual human genome**. *Nat Biotechnol* 2009, **27**:847-850.

9. • 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.

The first of a series of marine datasets produced by JCVI. This study is viewed by many as the start of environmental shotgun metagenomics work for more complex communities.

10. 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.

11. •• 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.

While this study was done by a large international consortium, it will serve as a blueprint for many studies by smaller groups in the near future.

12. •• Larsen PE, Collart F, 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**. *Microb Inform Exp* 2011, **1**:4.

A novel way of using metagenomic data to predict microbial metabolism.

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

The authors argue that accurate phylogenetic classification can only be achieved with sequences of 1000 bp or more.

14. Kunin V, Copeland A, Lapidus A, Mavromatis K, Hugenholtz P: **A bioinformatician's guide to metagenomics**. *Microbiol Mol Biol Rev* 2008, **72**:557-578.

15. Dan Gusfield: *Algorithms on Strings, Trees and Sequences: Computer Science and Computational Biology*. edn 1. Cambridge University Press; 1997.

16. •• Zerbino DR, Birney E: **Velvet: algorithms for de novo short read assembly using de Bruijn graphs**. *Genome Res* 2008, **18**:821-829.

Velvet is a very fast and easy to use assembly algorithm that is well suited to first-pass testing.

17. •• Chevreux B, Pfisterer T, Drescher B, Driesel AJ, Müller WE, Wetter T, Suhai S: **Using the miraEST assembler for reliable and automated mRNA transcript assembly and SNP detection in sequenced ESTs**. *Genome Res* 2004, **14**:1147-1159.

Mira is possibly the best assembly tool for microbial sequences we currently have.

18. Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJ, Birol I: **AByS.S: a parallel assembler for short read sequence data**. *Genome Res* 2009, **19**:1117-1123.

19. Maccallum I, Przybylski D, Gnerre S, Burton J, Shlyakhter I, Gnirke A, Malek J, McKernan K, Ranade S, Shea TP *et al.*: **ALLPATHS 2 small genomes assembled accurately and with high continuity from short paired reads**. *Genome Biol* 2009, **10**:R103.

20. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs**. *Nucleic Acids Res* 1997, **25**:3389-3402.

21. Wilkening J, Wilke A, Desai N, Meyer F: **Using clouds for metagenomics: a case study**. *IEEE International Conference on Cluster Computing (Cluster 2009)*; *New Orleans, LA, August 31–September 04: 2009*.

22. Kent WJ: **BLAT – the BLAST-like alignment tool**. *Genome Res* 2002, **12**:656-664.

23. Ye Y, Choi JH, Tang H: **RAPSearch: a fast protein similarity search tool for short reads**. *BMC Bioinform* 2011, **12**:159.

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

This article describes a gene prediction methodology for metagenomic data that includes support for noisy data. One of the first pieces of software that acknowledges the existence of errors in metagenomic DNA data.

25. Zhu W, Lomsadze A, Borodovsky M: **Ab initio gene identification in metagenomic sequences**. *Nucleic Acids Res* 2010, **38**:e132.

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

27. Noguchi H, Taniguchi T, Itoh T: **MetaGeneAnnotator: detecting species-specific patterns of ribosomal binding site for precise gene prediction in anonymous prokaryotic and phage genomes**. *DNA Res* 2008, **15**:387-396.

28. Krause L, Diaz NN, Goesmann A, Kelley S, Nattkemper TW, Rohwer F, Edwards RA, Stoye J: **Phylogenetic classification of short environmental DNA fragments**. *Nucleic Acids Res* 2008, **36**:2230-2239.

29. Lingner T, Asshauer KP, Schreiber F, Meinicke P: **CoMet – a web server for comparative functional profiling of metagenomes**. *Nucleic Acids Res* 2011, **39**:W518-W523.

30. Shi Y, Tyson GW, DeLong EF: **Metatranscriptomics reveals unique microbial small RNAs in the ocean's water column**. *Nature* 2009, **459**:266-269.

31. • Grissa I, Vergnaud G, Pourcel C: **The CRISPRdb database and tools to display CRISPRs and to generate dictionaries of spacers and repeats**. *BMC Bioinform* 2007, **8**:172.

CRISPRs first described in the acid mine drainage paper and discovered through environmental shotgun metagenomics now have become mainstream in microbial bioinformatics.

32. Griffiths-Jones S, Bateman A, Marshall M, Khanna A, Eddy SR: **Rfam: an RNA family database**. *Nucleic Acids Res* 2003, **31**:439-441.

33. • Meyer F, Paarmann D, D'Souza M, Olson R, Glass EM, 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 Bioinform [electronic resource]* 2008, **9**:386.

The most widely used analysis workflow and data repository for environmental metagenomics, mostly focused on automated analysis for microbial ecology.

34. • Markowitz VM, Ivanova NN, Szeto E, Palaniappan K, Chu K, Dalevi D, Chen IM, Grechkin Y, Dubchak I, Anderson I *et al.*: **IMG/M: a data management and analysis system for metagenomes**. *Nucleic Acids Res* 2008, **36**:D534-D538.

A system focused on detailed analysis of contigs obtained through metagenomics, mostly focused on the genomics.

35. Sun S, Chen J, Li W, Altintas I, Lin A, Peltier S, Stocks K, Allen EE, Ellisman M, Grethe J, Wooley J: **Community cyberinfrastructure for Advanced Microbial Ecology Research and Analysis: the CAMERA resource**. *Nucleic Acids Res* 2011, **39**:D546-D551.

36. Wheeler DL, Barrett T, Benson DA, Bryant SH, Canese K, Chetvernin V, Church DM, Dicuccio M, Edgar R, Federhen S *et al.*: **Database resources of the National Center for Biotechnology Information**. *Nucleic Acids Res* 2008, **36**:D13-D21.

37. Stein L: **The case for cloud computing in genome informatics**. *Genome Biol* 2010, **11**:207 doi: 10.1186/gb-2010-11-5-207.

38. •• Field D, Amaral-Zettler L, Cochrane G, Cole JR, Dawyndt P, Garrity GM, Gilbert J, Glöckner FO, Hirschman L, Karsch-Mizrachi I *et al.*: **The Genomic Standards Consortium**. *PLoS Biol* 2011, **9**:e1001088.

The GSC strives to provide data describing data (metadata) that allows meta-analysis across many datasets.

39. Gilbert JA, Meyer F, Knight R, Field D, Kyrpides N, Yilmaz P, Wooley J: **Meeting report: GSC M5 roundtable at the 13th International Society for Microbial Ecology Meeting in Seattle, WA, August 22–27, 2010**. *Stand Genomic Sci* 2010, **3**:235-239.