Proposal for open discussion: Informatics challenges for next generation sequencing metagenomics experiments

Folker Meyer[1,2] and Nikos Kyrpides[3]

[1] Argonne National Laboratory, 9700 S.Cass Avenue, Argonne, IL, 60439, USA
[2]University of Chicago, 5801 South Maryland Avenue, Chicago, IL 60637, USA
[2]DOE Joint Genome Institute, 2800 Mitchell Drive, Walnut Creek, CA 945987, USA

Abstract. With DNA sequence data production no longer the bottleneck in microbial studies, a rapidly increasing number of researchers from diverse areas of interest can now use metagenomic tools to study their environment of interest. The large quantities of sequence data becoming available are posing significant challenges to the existing analysis tools and indeed to the community providing analysis portals.

1 Introduction

Direct sequencing of environmental DNA (aka "metagenomics") has been ongoing for several years [1-5]. These types of experiments were enabled by breakthroughs in DNA sequencing technology that lowered the cost for obtaining large quantities of DNA reads. Similar to the sequencing cost for the human genome costs for sequencing metagenomic DNA have been dropping dramatically since the early 2000s. Data analysis for complex microbial assemblages has proven to be one of the key component of any metagenomic experiment, leading to the development of a number of software packages and several portals offering analysis, data integration and visualization [6,7].With the advent of next generation sequencing [8,9] data analysis for metagenomic data sets became even more difficult. Existing tools are not efficiently working since reads got shorter and more abundant (see e.g.[10]) and computational requirements grew dramatically[11]. The length of reads went from an 700-900bp of Q20 reads with Sanger sequencing to 75-150bp for Illumina reads or about 450bp for 454 reads.

While only five years ago, data sets of several million base-pairs (MBp) were considered disruptive (take as an example the debate [2]). Data sets of this size can now be created with a single instrument run of e.g. a Roche 454 instrument (see Figure 1 for data set sizes). With sequencing no longer the bottleneck it used to be both in

financial terms and by the fact that few centers were capable of creating "large" data sets, the metagenome analysis ecosystem undergoing change.

2 Metagenome Data

Data set sizes grow rapidly (see Figure 1) and are outpacing the growth of computing equipment. As stated frequently by many authors, the growth trajectories of computing equipment and sequencing technology show dramatic differences, computing capabilities doubling every 18 month and sequencing roughly doubling every 5-6 months (for a recent discussion see:[12]).

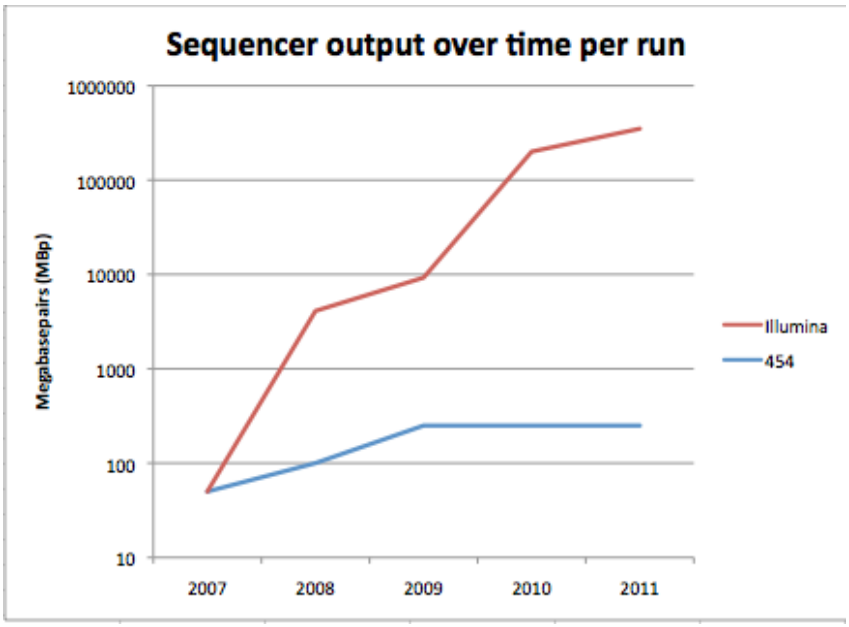

Figure 1: Data set sizes grow exponentially. over time for Illumina Solexa platform (red) and stay stable for the Roche 454 platform.

The number of data producers grows as well. The long discussed democratization of sequencing has finally arrived, allowing new individual institutes and universities to generate large scale sequencing data that just recently could be produced only from large sequencing facilities.

If 10 sequencing machines could be dedicated to global metagenomic sequencing, with the current state of the art technology of 200 gigabases (Gb) in around 10 days, we will be able to get 200 Gb of metagenomics sequences per day.

An influx prior to the advent of metagenomic data of that magnitude is likely to overwhelm the archives (SRA and Genbank and their international companions), which are struggling to keep up with a few big centers submitting large data quantities, it also represents demands on the analysis providers mentioned above that are beyond their capabilities.

Even to this day the current analysis portals do not provide an integration of the data from the Metahit project [10]. Published in early 2010, the MetaHit project produced 500 GBp of metagenomic data for gut microbial communities that will be an important resource for other researchers studying the human gut. However integrating even one single large experiment is proving to be a major challenge to the existing systems.

With the advent of the latest generation of sequencing instruments, even smaller centers have the ability to produce data sets of that size within two weeks. It is just the analysis bottleneck that prohibits widescale adoption of large shotgun metagenomics projects for many areas of research.

The argument made here is speculative in that we predict a certain number of sequencing instruments to be dedicated to running metagenomics experiments, however past submission history of our existing analysis portals MG-RAST and IMG/M can serve as evidence for the growing adoption of next generation sequencing (see Figure 2 below).

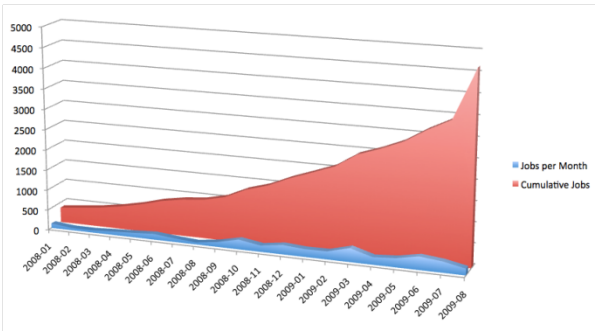

Figure 2: Number of data sets is growing fast (red) and the number of groups submitting is also rising (blue).

Analysis cost dominates the overall experimental costs. As shown by [11] the cost of running sequence analysis is significantly higher than the cost of sequencing.

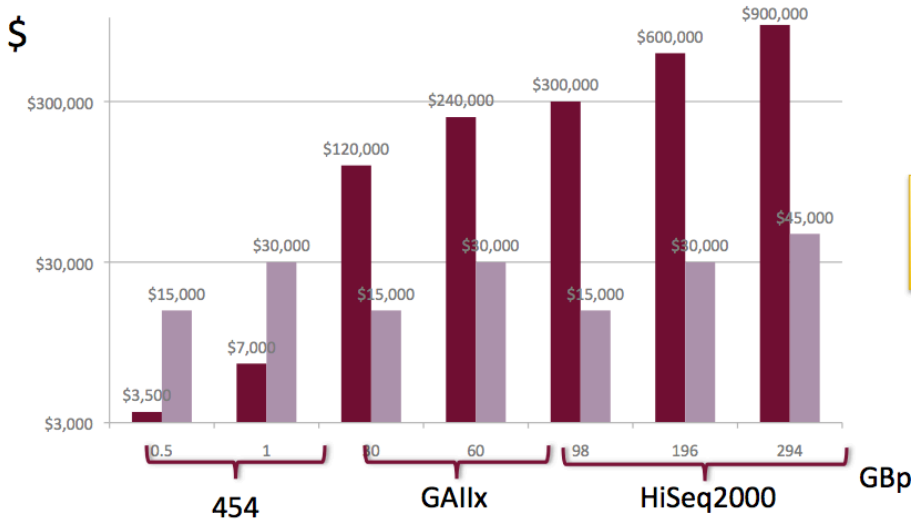

Figure 3: Computing cost dominate sequencing costs. While sequencing costs remain almost identical across platforms, the analysis costs vary with data set sizes. The cost of sequencing compared to the cost of running BLASTX analysis. Data from [11] using the Amazon EC2 cloud machine as a cost model.

Multiple analysis providers re-run the initial sequence analysis results using slightly different tools and parameters. Driven by historical factors, not by actual scientific need the various groups providing data portals for the metagenomics community ([11,12,13]) each run separate analysis pipelines, sharing significant parts of the value add process.

Given the cost of computing almost identical analysis, sharing of results would be very desirable at a time when significantly more data sets are being created. However due to the aforementioned implementation details, sharing the computational results is currently not possible.

In the current state of metagenomics, no single tool can provide all the answers to researchers, so submissions of data sets to multiple portals are the norm rather than the exception. This frequently leads to a multiple months wait time for researchers due to the need to re-compute the basic similarity analysis.

3 Metagenome Standards

Data standards are required to allow sharing of not only sequence sets but also computational results. If present these data standards would allow “instant” access to the metagenomic views and analysis tools provided by the other portals without incurring the extensive cost for re-computing the analysis.

However at the current state of development analysis provides lack the ability to even identify data sets that have been submitted to other portals before. The lack of experimental metadata, or better the universal adoption of metadata standards by the various communities producing metagenomes leads to more or less anonymous data sets. While efforts like GOLD [14] provided an invaluable service to the community using Sanger sequencing to produce complete microbial genomes in the past., the widespread adoption of metagenomic sequencing have led to a situation where only a subset of metagenomes is registered with GOLD.

Adoption of Metadata standards by the community is ongoing, but the existing standards proposed by the Genomics Standards Consortium [15,16] are only slowly being accepted. However with analysis providers updating their tools to enforce metadata standards compliance, the community of users will be guided towards metadata standards compliance.

The standards proposed by the GSC include minimal checklists that are required of about a dozen terms and the ability to create environmental packages that comprise many more parameters. With these packages, specific communities e.g. medical, soil or marine metagenomics can establish their specific metadata sets.

Machine readable metadata is absolutely required in a data ecosystem that contains several thousand data sets today and will contain several hundred thousand metagenomic data sets in the near future. The need for metadata goes beyond the description of sampling location and informatics analysis. While the recent discussion on the "rare biosphere" [17,18,19] has shown that informatics analysis plays a significant role and can in fact lead to significant false understanding of microbial diversity in a given sample, a similar discussion is already on the way regarding biome appropriate strategies for DNA isolation and handling [20,21]. Sampling strategies and the need for appropriate biological and technical replicates (in short statistically sound sampling) are likely next-in-line discussions that the community will have, now that the sequencing cost are no longer prohibiting the creation of replicates.

Report metagenomic data analysis is another area that will require significant community input. While a discussion about the pan-genome [22] has clearly shown that the existing data standards are inadequate for reporting pan-genome variation. Even reporting more or less complete microbial genomes extracted from metagenomic data sets will proof to be a difficult task given the current community standard operating procedures.

Acknowledgment

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

References

1. Tyson GW, Chapman J, Hugenholtz P, Allen EE, Ram RJ, et al. (2004) Community structure and metabolism through reconstruction of microbial genomes from the environment.
2. Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, et al. Accurate whole human genome sequencing using reversible terminator chemistry. Nature 428: 37-43.. 2008;456(7218):53-9. PMCID: 2581791.
3. Venter JC, Remington K, Heidelberg JF, Halpern AL, Rusch D, Eisen JA, et al. Environmental genome shotgun sequencing of the Sargasso Sea. Science. 2004;304(5667):66-74
4. Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, et al. Genome sequencing in microfabricated high-density picolitre reactors. Nature. 2005;437(7057):376-80.
5. Williamson SJ, Rusch DB, Yooseph S, Halpern AL, Heidelberg KB, et al. The Sorcerer II Global Ocean Sampling Expedition: metagenomic characterization of viruses within aquatic microbial samples. PLoS ONE 2008:3: e1456.
6. McHardy AC, Martin HG, Tsirigos A, Hugenholtz P, Rigoutsos I. Accurate phylogenetic classification of variable-length DNA fragments. Nature methods. 2007;4.
7. Yooseph S, Sutton G, Rusch DB, Halpern AL, Williamson SJ, et al. (2007) The Sorcerer II Global Ocean Sampling expedition: expanding the universe of protein families. PLoS Biol (1):63-72.
8. Wilkening J, Wilke A, Desai N, Meyer F, editors. Using Clouds for Metagenomics: A Case Study IEEE Cluster; 2009; New Orleans: IEEEE.
.9. Stein LD. The case for cloud computing in genome informatics. Genome biology. 2010;11(5.):207. PMCID: 2898083.
10. Qin J, Li R, Raes J, Arumugam M, Burgdorf KS, Manichanh C, et al. A human gut microbial gene catalogue established by metagenomic sequencing. Nature. 2010;464(7285):59-65.
11. Meyer F, Paarmann D, D'Souza M, Olson R, Glass EM, Kubal M, et al. The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes. BMC Bioinformatics BMC bioinformatics [electronic resource]. 2008;9: 386. PMCID: 2563014.
12. Seshadri R, Kravitz SA, Smarr L, Gilna P, Frazier M. CAMERA: a community resource for metagenomics. PLoS Biol. 2007;5(3):e75. PMCID: 1821059.
13. Markowitz VM, Ivanova NN, Szeto E, Palaniappan K, Chu K, Dalevi D, et al. IMG/M: a data management and analysis system for metagenomes. Nucleic Acids Res . 2008;36: (Database issue):D534-538.8. PMCID: 2238950.
14. Liolios K, Mavromatis K, Tavernarakis N, Kyrpides NC. The Genomes On Line Database (GOLD) in 2007: status of genomic and metagenomic projects and their associated metadata. Nucleic Acids Res. 2008;36(Database issue):D475-9. PMCID: 2238992.
15. Field D, Garrity G, Gray T, Morrison N, Selengut J, Sterk P, et al. The minimum information about a genome sequence (MIGS) specification. Nature biotechnology. 2008;26(5):541-7. PMCID: 2409278.
16. Kottmann R, Gray T, Murphy S, Kagan L, Kravitz S, Lombardot T, et al. A standard MIGS/MIMS compliant XML Schema: toward the development of the Genomic Contextual Data Markup Language (GCDML). Omics. 2008;12(2):115-21.
17. Huse SM, Welch DM, Morrison HG, Sogin ML. Ironing out the wrinkles in the rare biosphere through improved OTU clustering. Environmental microbiology. 2010;12(7):1889-98. PMCID: 2909393.
18. Sogin ML, Morrison HG, Huber JA, Mark Welch D, Huse SM, Neal PR, et al. Microbial diversity in the deep sea and the underexplored "rare biosphere". Proc Natl Acad Sci U S A. 2006;103(32):12115-20. PMCID: 1524930.

19. Reeder J, Knight R (2009). The 'rare biosphere': a reality check. Nat Methods Nature methods. 2009;6: (9):636-637.
20. Martin-Laurent F, Philippot L, Hallet S, Chaussod R, Germon JC, et al. (2001) DNA extraction from soils: old bias for new microbial diversity analysis methods. Appl Environ Microbiol 67: 2354-2359.
21. Lauber CL, Zhou N, Gordon JI, Knight R, Fierer N (2010) Effect of storage conditions on the assessment of bacterial community structure in soil and human-associated samples. FEMS Microbiol Lett 307: 80-86.
22. Bentley S (2009) Sequencing the species pan-genome. Nat Rev Microbiol 7: 258-259.