THEME ARTICLE: SCIENTIFIC IMPACT OF THE EXASCALE COMPUTING PROJECT (ECP)

Exabiome: Advancing Microbial Science through Exascale Computing

Steven Hofmeyr, Aydin Buluç, Robert Riley, Rob Egan, Oguz Selvitopi, Leonid Oliker and Katherine Yelick, *Lawrence Berkeley National Laboratory, Berkeley, CA, 94720, USA*

Migun Shakya and Brett Youtsey, *Los Alamos National Laboratory, Los Alamos, NM, 87545, USA*

Ariful Azad, *Indiana University, Bloomington, IN, 47405, USA*

Abstract—The Exabiome project seeks to improve the understanding of microbiomes through the development of methods for accelerating metagenomic science using exascale computing. This article gives an overview of scientific impact of the three components of the project: metagenome assembly, protein family detection and comparative analysis of metagenomes. Exabiome developed MetaHipMer, the only metagenome assembler capable of scaling to full exascale systems. MetaHipMer has enabled ground-breaking assemblies on the Frontier supercomputer, with many scientific benefits, such as the discovery of rare species and viral genomes. To investigate protein families, Exabiome developed two exascale tools, PASTIS and HipMCL. Together, these can utilize exascale resources to understand the functional diversity of billions of "dark matter" proteins and novel protein families. For comparative analysis, Exabiome developed kmerprof, a tool that can be used to compare huge metagenomes for many different scientific purposes, for example, grouping human microbiomes according to body location.

Microbiomes are profoundly important in many different areas, such as human health, farming, the environment and bio-manufacturing. Microbiomes are comprised of communities of thousands of microbial species, organisms too small to be seen by the naked eye, such as bacteria, protozoa, fungi and viruses. Much of a deeper understanding of these communities comes from an analysis of the genetics of the various species, a task that is made difficult by the fact that less than one percent of species in a microbiome are individually culturable in a laboratory. This led to metagenomics, where samples that include most of the genomes in the community are sequenced and analyzed to identify new species, characterize taxonomic diversity, provide functional annotation, and generally enhance understanding of microbiomes.

The collection of sequencing data is increasing exponentially, leading to growing analysis challenges that are intractable for traditional tools. The Exabiome project developed tools to harness the power of exascale computing to address these challenges. Most bioinformatics analysis pipelines use single shared-memory computers or small clusters; being able to run analysis pipelines on supercomputers enables new scientific discoveries, some of which will be presented in this article.

We will discuss three areas where new tools from Exabiome have opened the way to deeper scientific understanding. First, we will describe some of the scientific findings that result from the largest metagenome assemblies ever done, using Exabiome's MetaHipMer assembler running on almost all of the Frontier supercomputer. Metagenome assembly is the first stage in the processing of microbiomes, and enables us to find species, genes and their relative abundances in microbial communities. Second, we will describe how Exabiome's PASTIS tool for protein sequence similarity search combined with Exabiome's HipMCL tool for

XXXX-XXX © IEEE
Digital Object Identifier 10.1109/XXX.0000.0000000

protein clustering uses exascale resources to understand the functional diversity of billions of “dark matter” proteins, and to explore the diversity of novel protein families. Finally, we will discuss how Exabiome tools for comparative analysis, including *kmerprof*, have been used to compare metagenomes for different scientific purposes, for example, grouping human microbiomes correctly into different body locations.

METAGENOME ASSEMBLY

The first phase in metagenomic analysis is the sequencing and assembly of environmental samples, such as water, soil or animal gut. Each sample contains thousands to tens of thousands of microbial genomes of varying abundance, all mixed together. Modern sequencing technology can only extract the DNA in these samples in disconnected short fragments, called *reads*. Typical metagenome sequencing results in reads that are 100 to 150 nucleotides (bases) long[1]. The picture is further complicated by the fact that the reads are error prone; typically there is about a 0.25% chance per base of an error. To counteract these errors, a sample is sequenced multiple times (called the sequencing depth) to ensure that each part of the sample is covered with at least some error-free sequences. A typical single sample will contain between 40 and 400 million reads, and projects could potentially comprise hundreds or even thousands of samples.

The challenge of metagenome assembly is to take this mix of error-prone genome fragments of varying abundances of many different species and stitch the reads together to form contiguous sequences (called *contigs*[2]) that are as long as possible, with as few errors as possible, separating out the various species correctly. To manage the exponential number of combinations of reads, metagenome assemblers typically use the de Bruijn graph approach [1], which is highly computationally intensive with very large memory demands. Most assemblers (including the most used, state-of-the-art, MEGAHIT and metaSPAdes) only work on shared-memory systems, due to the challenge of developing scalable algorithms for the irregular computation patterns of genome assembly, which make extensive use of hash tables. Consequently, these state-of-the-art assemblers can only assemble a few samples together at a time (called *coassembly*); to assemble larger sets of samples they have to use *multiassembly*, where samples are assembled separately and the results combined, which then have to be deduplicated (removing identical copies).

The MetaHipMer Assembler

The MetaHipMer assembler [2], developed by the Exabiome project under funding from the Exascale Computing Project, is the only metagenome assembler that scales efficiently on distributed memory systems, and hence is the only assembler that is able to coassemble very large datasets. MetaHipMer is a complex pipeline that is written in UPC++ [3] and is very different from most scientific applications. It makes extensive use of distributed hash tables and performs primarily string manipulations, with negligible floating point operations. Although these computational patterns are not well-suited to GPU acceleration, MetaHipMer has GPU support for some of the most computationally intensive stages in the pipeline, which confers a significant speedup (around 2x on the Frontier supercomputer at scale). MetaHipMer exhibits efficient strong scaling, as shown in Figure 1. In addition, MetaHipMer also produces assemblies that are of comparable quality to the leading shared-memory assemblers [2]; in the CAMI2 contest, an independent assessment of assembly quality, MetaHipMer was ranked overall the best out of 20 different assemblers: "The best ranking method across metrics and all datasets was [Meta]HipMer..." [4].

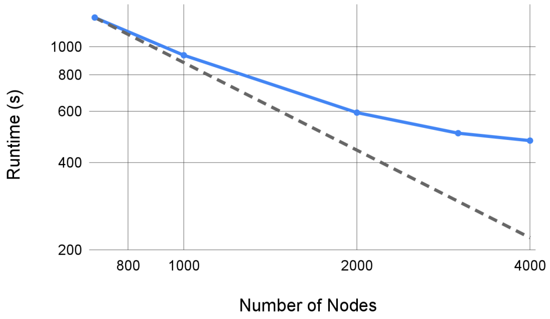

FIGURE 1. MetaHipMer strong scaling. Assembling the 8.1TB Indian ocean subset of the Tara Oceans on increasing numbers of nodes on the Frontier supercomputer. The dashed line shows perfect scaling. This dataset requires at least 700 nodes to assemble due to memory requirements, and scaling drops off at higher node counts because communication costs start to dominate.

When many samples taken across time or space are available, there are significant advantages to

[1] These are known as short reads; long read technology produces reads up to 20 thousand bases, but long reads are not commonly used for metagenome sequencing.

[2] For some assemblers, the final sequences are called *scaffolds*.

coassembling them all together, compared with multiassembly. The driving reason is that coassembly can effectively exploit the combined information from all the samples, increasing the frequency of reads for every species, effectively increasing sequencing depth. This means that coassembly can link together many more reads, producing more contiguous assemblies with longer sequences. Furthermore, the aggregation of samples facilitates the identification of rare genomes. Assemblers must discard infrequent subsequences[3] because these are indistinguishable from errors; combining multiple samples increases the frequency of reads covering rare genomes and enables the discovery of species that could never be found through multiassembly.

MetaHipMer has been used to coassemble the largest datasets to date. These include the Tara Oceans, which comprises 84TB of data from 1,213 samples collected from all the world's oceans over a four year voyage [5]. The exascale resources of the Frontier supercomputer were used to coassemble all 1,213 samples in 94 minutes on 9,000 compute nodes (containing 36,000 GPUs). The largest dataset assembled at the time of writing is a collection of 15,863 samples from the Human Microbiome Project (HMP) [6], for the human gut, amounting to 98TB worth of data. This assembly took 38 minutes on 9,000 Frontier nodes (36,0000 GPUs). The time taken for the HMP was less than for the Tara Oceans because the human gut microbiome is a much less complex environment.

Although we are still evaluating the scientific implications of these exceptional assemblies, there are many smaller (but still large scale) projects that are demonstrating the value of coassembly to the scientific community. In particular, MetaHipMer is used at the Joint Genome Institute (JGI) for large coassemblies, and the benefits are such that the demand is steadily increasing, with over 60 projects pending in 2023 that can only be done by MetaHipMer on large-scale computer clusters. These projects cover diverse environments, from mountainous watersheds, to ponds to wetlands and coastal mangroves. We illustrate some of the benefits of large coassembly by describing some of the results with a particular dataset: the Great Redox Experiment.

The Great Redox Experiment

The Great Redox Experiment (GRE) is a study of the impacts of fluctuating redox reactions on microbiomes in humid tropical soils. The data consist of 95 samples totalling 7.7TB, gathered from soil in the Luquillo Experimental Forest in Puerto Rico. A complete analysis of the full coassembly is reported in Riley et al. [7]; we highlight some of these results here. The full coassembly was performed using MetaHipMer, and the multiassembly is comprised of individual samples each assembled using metaSPAdes, which has traditionally been the standard metagenome assembler for individual samples at the JGI.

There are many ways of measuring contiguity for assemblies; one of the most informative is the number of long contigs (those exceeding 50,000 bases in length). The coassembly contains 26,818 such contigs (3.25% of the total), whereas the median for the individual sample assemblies is only 3 (0.015% of the total), and the combination across all 95 samples (the full multiassembly) is thus about 100x less than for the coassembly. Furthermore, many of the longer contigs in the multiassembly are full or partial duplicates, which is not the case for the coassembly. Furthermore, 87.5% of the 22 billion reads align to the coassembly contigs, whereas only 51.5% align to the individual assembly contigs. This implies that the coassembly successfully extracts almost all of the sequence data, whereas the multiassembly misses around half of it.

The superior contiguity and genome recovery in the coassembly have significant impacts on the downstream analysis. In the typical metagenomic analysis process, the contigs in an assembly are clustered into bins representing taxonomic units, such as individual species. These bins are called *metagenome assembled genomes* (MAGs), which can be separated into low, medium and high quality [8]. There were 307 high and medium quality (HQ/MQ) MAGs in the coassembly, representing 23 distinct phyla, including three candidate phyla[4], and two of the MAGs in the *Eremiobacterota* candidate phylum were of such high quality that they can be used as reference genomes for unculturable microbes. Although the multiassembly yielded almost as many HQ/MQ MAGs (294), clustering those MAGs at the species level to elimate duplications resulted in only 38 MAGs, representing nine phyla. The same clustering procedure does not reduce the number of coassembly MAGs at all. Figure 2 shows the phyla and the number of deduplicated MAGs found

[3] In the de Bruijn graph approach, short subsequences of reads of length k, called k-mers, are discarded if they occur fewer than m times in the reads being assembled, where typically $m = 2$ for metagenome assembly.

[4] A phylum is labelled as a candidate if the members are uncultivated, and hence only known from metagenomics.

with coassembly and multiassembly.

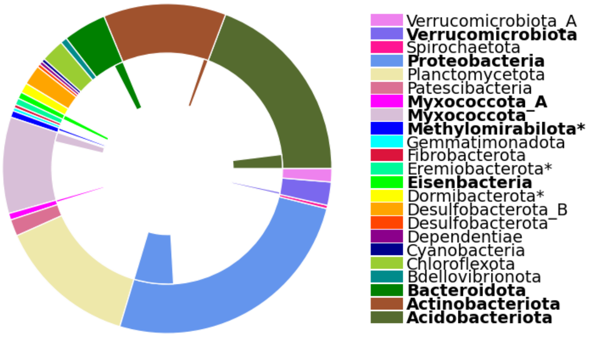

FIGURE 2. Phyla found in the GRE assemblies. The outer circle shows the coassembly phyla and the inner circle shows the multiassembly phyla. The size of the slices represents the number of clustered MAGs within that phyla. The legend lists all 23 phyla found, with the ones that occur in both coassembly and multiassembly in bold, and the candidate phyla marked with an asterix (*).

Most of metagenomic analysis focuses on HQ/MQ bins, because these contain the most complete and uncontaminated microbial genomes. However, low quality MAGs and unbinned contigs offer the opportunity to resolve genomes for the rarest microbes in the environment, those occurring in less than 1% of the total abundance. Analysis of the coassembly found at least three rare biosphere microbes that have under five genomes in the NCBI database, which maintains genomes from all known microbes. These microbes have under 1x coverage in individual samples, and so can only be recovered in the coassembly. In addition to recovery of bacterial genomes, eukaryotic[5] genomes can also be found in the lower quality bins. Over 30 partial eukaryotic MAGs were found, with one of them being a fairly complete fungal genome. None of these could be recovered from the multiassembly. Another area where coassembly proves extremely beneficial is in identifying viral genomes. A total of 7615 viral families was identified in the coassembly, of which only 861, or 11% were also found in the multiassembly. Many of these coassembly-only genomes occur at low abundance, as is shown in figure 3. Identifying rare viruses in environmental samples could be valuable in monitoring for novel disease outbreaks.

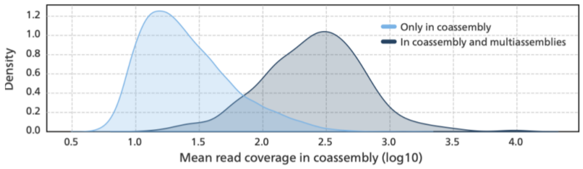

FIGURE 3. Read coverage of viruses in assemblies. Low coverage genomes are only recovered in the coassembly. Reproduced from Riley et al [7].

PROTEIN FAMILY DETECTION

A protein family is a group of proteins that descended from a common ancestor. In other words, members of a protein family are homologous with one another. Detecting true homology with high confidence is often not possible, but we can infer homology through protein sequence or protein structure similarity. Protein structure data is scarce, labor intensive and expensive to produce. Protein sequence data, on the other hand, is much more abundant and cheaper. Therefore, sequence data is often used to find protein families.

Evolution favors preservation of certain protein letters and subsequences over others, which let biologists develop statistical models, such as the BLOSUM matrix, that capture the amino acid sequence substitution frequencies. These models are used to compute sequence similarities between pairs of proteins to develop a *protein sequence similarity graph*.

Due to the immense number of proteins in metagenomic samples, we cannot compare all pairs of proteins. Instead, we use *k*-mers to find candidate pairs of proteins that are statistically likely to align, a process called *many-to-many protein sequence similarity search*. PASTIS, developed by Exabiome in collaboration with the ExaGraph co-design center (https://exagraph.lbl.gov), is a high performance implementation of this process. PASTIS is the first part of the pipeline that starts from (metagenomic) protein sequences and ends with protein families, as shown in Figure 5.

PASTIS [9] uses distributed sparse matrices to encode protein/*k*-mer presence information and sparse matrix-matrix multiplication to find candidate pairs. We use concepts from BLOSUM matrices to perform higher accuracy search using the concept of substitute *k*-mers, all within the same sparse matrix machinery. This allows PASTIS to utilize the most scalable sparse matrix algorithms and codes offered by the Combinatorial BLAS library [10].

PASTIS recently has been ported to run on GPU-

[5] Organisms with a cell nucleus.

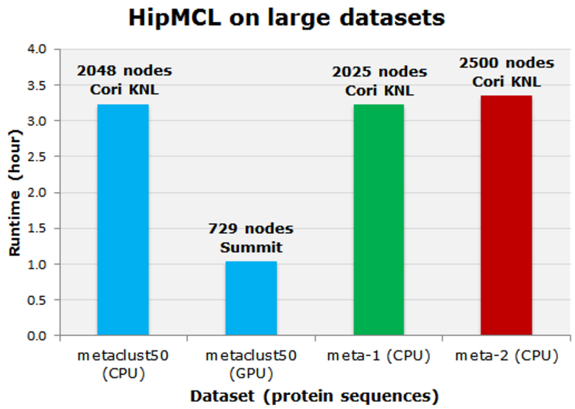

FIGURE 4. Some of the large-scale runs with HipMCL; details of the networks are included in the main text

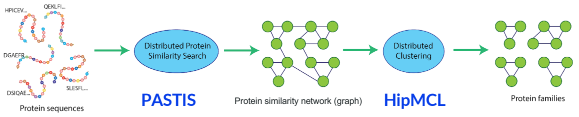

FIGURE 5. Protein family identification pipeline developed by Exabiome. PASTIS is the many-to-many sequence aligner, HipMCL is the graph clustering. Both run on CPU and GPU equipped supercomputers.

equipped supercomputers such as Summit, Perlmutter and Frontier. This porting, along with many optimizations, led to an order of magnitude performance improvement over the initial results [9]. The resulting impressive performance of 691 million alignments per second was a finalist for the ACM Gordon Bell Prize in 2022.

The protein similarity graph obtained via PASTIS needs to be clustered via a high-performance graph-based clustering algorithm. For that, we developed HipMCL [11], a scalable distributed-memory implementation of the flow-based Markov cluster (MCL) algorithm. HipMCL removes the compute power and memory size limitations of MCL and enables clustering of extreme-scale datasets.

We have also ported HipMCL to run on GPU-equipped supercomputers, overcoming various challenges related to the low-memory capacity of GPUs and the bottlenecks of the CPU-GPU memory interface. We solved the memory-capacity problem using randomized algorithms that predict memory consumption cheaply yet accurately. The CPU-GPU memory-bandwidth bottlenecks were overcome by pipelining. The results show significant performance boosts compared to CPU only performance.

Various large scale runs of HipMCL are shown in Figure 4. The MetaClust50 graph has 282 million vertices and 37 billion edges. The Meta-2 protein similarity graph has 510 million vertices and 5 billion edges, and the Meta-1 dataset is similarly sized to Meta-2.

Meta-2 is the network clustered to find novel protein families from **all** publicly available protein data from environmental samples [12]. The results of this large-scale study showed that publicly available metagenome datasets contain 106, 198 novel protein families. These protein families are novel in the sense that they are sufficiently large (more than 100 members), and they contain no members that are sufficiently similar to any known protein sequences. Furthermore, analysis by AlphaFold showed that these novel protein families fold into 1, 215 unique folds, 141 of which are novel.

COMPARATIVE ANALYSIS

Microbial communities from different locations and times are often compared to characterize variations and dynamics of communities along these gradients. The comparison is usually done by assigning reads to taxonomic groups (such as species or phyla) based on their similarity to a database of reference genomes. However, reference databases consist of sequences from cultured and dominant organisms, whereas metagenomes potentially contain sequences from all the organisms in a sample, even the rare and uncultured ones. Consequently, in many cases, only a fraction of sequences from metagenomes end up as part of the comparative analysis.

An alternative approach is to compare metagenomes without reference to databases of known organisms. Metgenomes can be compared based on their *k*-mer content, which can provide more accurate distances between metagenomes as these comparisons will likely use all the data in metagenome samples. With a record of *k*-mers and their abundances in each sample, any pairwise distance (e.g Jaccard similarity, Bray-Curtis, etc.) can be calculated to estimate similarity between samples. These methods can find near neighbors of a genome, classify genomes into phylogenetically relevant clusters, and cluster metagenomes into bins that match with their phenotypes and functional contents.

Counting the abundance of all the *k*-mers present across all the samples in a metagenome is computationally expensive, and requires distributed memory for the larger samples. Exabiome has developed a comparative *k*-mer analysis tool called *kmerprof*, which is based on MetaHipMer's *k*-mer counting stage, with several extensions needed for carrying out comparisons. With *kmerprof*, we have the ability to compare datasets with billions of *k*-mers and calculate distances based on Jaccard Index and Bray-Curtis. We have used *kmerprof* to compare 8.47 TB of reads from the Human Microbiome Project (HMP) [6] containing metagenomes from 882 samples in 19 different body sites (Figure 6). By calculating pairwise *k*-mer distances between samples, *kmerprof* recapitulates the clustering we would expect from body sampling location. As shown here, comparisons of every single unique *k*-mer sequence can lead to finding samples that are similar to each other, ones that might have been missed by traditional reference based analyses.

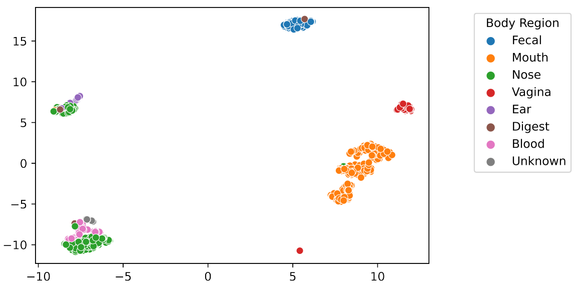

FIGURE 6. UMAP projection of Human Microbiome Project metagenomes. Pairwise distances calculated from Jaccard Index at *k*-mer length of 21. Metagenomes are colored by approximate body region sampled.

Conclusion

The Exabiome project, under funding from the Exascale Computing Project, has developed multiple tools for using exascale resources to investigate metagenomes. These tools are uniquely able to exploit exascale resources, and have already demonstrated that there is potential for new scientific discoveries. The MetaHipMer metagenome assembler has produced record breaking assemblies that are only now starting to be analyzed for scientific impact. We discussed one of them, the Great Redox Experiment, where the MetaHipMer coassembly yielded 307 HQ/MQ MAGs, representing 23 distinct phyla, compared to multiassembly, which only produced 38 unique MAGs, representing 9 phyla. Two other Exabiome tools, PASTIS and HipMCL, were used to find over 100,000 novel protein families from all publicly available environmental sample proteins, a huge dataset. Another tool, *k*merprof, was used on 882 metagenome samples from the human body, and demonstrated that it could cluster them effectively according to body location. Metagenomic data is growing exponentially, and these results from the various tools developed by Exabiome point at the potential for new scientific breakthroughs in metagenomics made possible by exascale computing.

Acknowledgements

This research was supported by the Exascale Computing Project (17-SC-20-SC), a collaborative effort of the U.S. Department of Energy Office of Science and the National Nuclear Security Administration. Authors from Lawrence Berkeley National Laboratory were supported by the Applied Mathematics and Computer Science Programs of the DOE Office of Advanced Scientific Computing Research and the DOE Office of Biological and Environmental Research under contract number DE-AC02-05CH11231. This research used resources of the National Energy Research Scientific Computing Center, which is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231. The work conducted by the U.S. Department of Energy Joint Genome Institute, a DOE Office of Science User Facility, is supported under Contract No. DE-AC02-05CH11231. This research was also supported by Los Alamos National Laboratory's internal Laboratory Directed Research and Development grant 20220585ECR.

REFERENCES

1. P. Compeau, P. Pevzner, and G. Tesler, "How to apply de bruijn graphs to genome assembly," *Nature Biotechnology*, no. 29, 2011. [Online]. Available: https://doi.org/10.1038/nbt.2023
2. S. Hofmeyr, R. Egan, E. Georganas, A. C. Copeland, R. Riley, A. Clum, E. Eloe-Fadrosh, S. Roux, E. Goltsman, A. Buluç, D. Rokhsar, L. Oliker, and K. Yelick, "Terabase-scale metagenome coassembly with MetaHipMer," *Scientific Reports*, vol. 10, no. 1, Jul. 2020. [Online]. Available: https://doi.org/10.1038/s41598-020-67416-5
3. J. Bachan, S. Baden, S. Hofmeyr, M. Jacquelin, A. Kamil, D. Bonachea, P. Hargrove, and H. Ahmed, "UPC++: A high-performance communication framework for asynchronous computation," in *33rd IEEE International Parallel and Distributed Processing Symposium (IPDPS'19)*, 2019.
4. F. Meyer, A. Fritz, Z.-L. Deng, D. Koslicki, T. R. Lesker, A. Gurevich, G. Robertson, M. Alser, D. Antipov, F. Beghini, D. Bertrand, J. J. Brito, C. T.

Brown, J. Buchmann, A. Buluç, B. Chen, R. Chikhi, P. T. L. C. Clausen, A. Cristian, P. W. Dabrowski, A. E. Darling, R. Egan, E. Eskin, E. Georganas, E. Goltsman, M. A. Gray, L. H. Hansen, S. Hofmeyr, P. Huang, L. Irber, H. Jia, T. S. Jørgensen, S. D. Kieser, T. Klemetsen, A. Kola, M. Kolmogorov, A. Korobeynikov, J. Kwan, N. LaPierre, C. Lemaitre, C. Li, A. Limasset, F. Malcher-Miranda, S. Mangul, V. R. Marcelino, C. Marchet, P. Marijon, D. Meleshko, D. R. Mende, A. Milanese, N. Nagarajan, J. Nissen, S. Nurk, L. Oliker, L. Paoli, P. Peterlongo, V. C. Piro, J. S. Porter, S. Rasmussen, E. R. Rees, K. Reinert, B. Renard, E. M. Robertsen, G. L. Rosen, H.-J. Ruscheweyh, V. Sarwal, N. Segata, E. Seiler, L. Shi, F. Sun, S. Sunagawa, S. J. Sørensen, A. Thomas, C. Tong, M. Trajkovski, J. Tremblay, G. Uritskiy, R. Vicedomini, Z. Wang, Z. Wang, Z. Wang, A. Warren, N. P. Willassen, K. Yelick, R. You, G. Zeller, Z. Zhao, S. Zhu, J. Zhu, R. Garrido-Oter, P. Gastmeier, S. Hacquard, S. Häußler, A. Khaledi, F. Maechler, F. Mesny, S. Radutoiu, P. Schulze-Lefert, N. Smit, T. Strowig, A. Bremges, A. Sczyrba, and A. C. McHardy, “Critical assessment of metagenome interpretation: the second round of challenges,” *Nature Methods*, vol. 19, no. 4, pp. 429–440, Apr 2022. [Online]. Available: https://doi.org/10.1038/s41592-022-01431-4

5. Tara Oceans. https://fondationtaraocean.org/en/expedition/tara-oceans. Accessed: 2023-10-11.
6. NIH Human Microbiome Project. https://www.hmpdacc.org. Accessed: 2023-10-11.
7. R. Riley, R. M. Bowers, A. P. Camargo, A. Campbell, R. Egan, E. A. Eloe-Fadrosh, B. Foster, S. Hofmeyr, M. Huntemann, M. Kellom, J. A. Kimbrel, L. Oliker, K. Yelick, J. Pett-Ridge, A. Salamov, N. J. Varghese, and A. Clum, “Terabase-scale coassembly of a tropical soil microbiome,” *Microbiology Spectrum*, vol. 11, no. 4, Aug. 2023. [Online]. Available: https://doi.org/10.1128/spectrum.00200-23
8. J. C. Setubal, “Metagenome-assembled genomes: concepts, analogies, and challenges,” *Biophysical Reviews*, vol. 13, no. 6, pp. 905–909, Nov. 2021. [Online]. Available: https://doi.org/10.1007/s12551-021-00865-y
9. O. Selvitopi, S. Ekanayake, G. Guidi, M. G. Awan, G. A. Pavlopoulos, A. Azad, N. Kyrpides, L. Oliker, K. Yelick, and A. Buluç, “Extreme-scale many-against-many protein similarity search,” in *SC22: International Conference for High Performance Computing, Networking, Storage and Analysis*. IEEE, 2022, pp. 1–12.
10. A. Azad, O. Selvitopi, M. T. Hussain, J. R. Gilbert, and A. Buluç, “Combinatorial BLAS 2.0: Scaling combinatorial algorithms on distributed-memory systems,” *IEEE Transactions on Parallel and Distributed Systems*, vol. 33, no. 4, pp. 989–1001, 2021.
11. A. Azad, G. A. Pavlopoulos, C. A. Ouzounis, N. C. Kyrpides, and A. Buluç, “HipMCL: a high-performance parallel implementation of the markov clustering algorithm for large-scale networks,” *Nucleic acids research*, vol. 46, no. 6, pp. e33–e33, 2018.
12. G. A. Pavlopoulos, F. A. Baltoumas, S. Liu, O. Selvitopi, A. P. Camargo, S. Nayfach, A. Azad, S. Roux, L. Call, N. N. Ivanova, I. M. Chen, D. Paez-Espino, E. Karatzas, Novel Metagenome Protein Families Consortium, I. Iliopoulos, K. Konstantinidis, J. M. Tiedje, J. Pett-Ridge, D. Baker, A. Visel, C. A. Ouzounis, S. Ovchinnikov, A. Buluç, and N. C. Kyrpides, “Unraveling the functional dark matter through global metagenomics,” *Nature*, vol. 622, no. 7983, pp. 594–602, 2023. [Online]. Available: https://doi.org/10.1038/s41586-023-06583-7

Steven Hofmeyr is a senior engineer at Lawrence Berkeley National Laboratory in Berkeley, California. His current research interests include high performance metagenome assembly, disease modeling, and operating systems. He received a Ph.D. in Computer Science from UNM. Contact him at shofmeyr@lbl.gov.

Aydin Buluç is a Senior Scientist at Lawrence Berkeley National Laboratory in Berkeley, California and an Adjunct Faculty at EECS department of UC Berkeley in California. His research interests include parallel computing, high performance graph analysis and machine learning, and computational genomics. He received his Ph.D. in Computer Science from the University of California, Santa Barbara. He is a member of the IEEE Computer Society. Contact him at abuluc@lbl.gov.

Robert Riley is a data scientist at the DOE Joint Genome Institute in Berkeley, California, where he works on the assembly and analysis of fungal, algal, and metagenomic data. He received his Ph.D. in Human Genetics from UCLA. Contact him at rwriley@lbl.gov.

Rob Egan is a software developer at the DOE Joint Genome Institute in Berkeley, California. His research is in genome analysis and computational methods to analyze large metagenomes. Rob received his B.A. in biochemistry from Cornell University. Contact him at

rsegan@lbl.gov.

Oguz Selvitopi is a research scientist at Lawrence Berkeley National Laboratory in Berkeley, California. His research interests include high performance computing, parallel graph algorithms, and bioinformatics. He got his Ph.D. in Computer Engineering from Bilkent University. Contact him at roselvitopi@lbl.gov.

Leonid Oliker is a Computer Senior Scientist and lead of the Performance and Algorithms research group at Lawrence Berkeley National Laboratory in Berkeley, California. His research interests focus on performance optimization, evaluation, and modeling of high-performance computing systems. Leonid received his PhD in computer science from the University of Colorado in Boulder and is a member of the IEEE Computer Society. Contact him at loliker@lbl.gov.

Katherine Yelick is the Vice Chancellor for Research and the Robert S. Pepper Distinguished Professor of Electrical Engineering and Computer Sciences at the University of California, Berkeley. She is also a Faculty Senior Scientist at Lawrence Berkeley National Laboratory in Berkeley, California. Her current research interests are in parallel computing and computational genomics. She is a Senior Member of the IEEE Computer Society. Contact her at yelick@berkeley.edu.

Migun Shakya is a researcher at the Bioscience Division of Los Alamos National Laboratory in Los Alamos, New Mexico. His current research interest is in using multi-omics approaches to develop novel methods for biosurveillance and understanding the roles of viruses in environments. He received his Ph.D in Genomics from the University of Tennessee, Knoxville. Contact him at migun@lanl.gov.

Brett Youtsey is a technologist at the Bioscience Division of Los Alamos National Laboratory in Los Alamos, New Mexico. His interests include utilizing techniques in Big Data, Machine Learning, and Metagenomics for biosurveillance. He received a Master's Degree in Bioinformatics and Genomics from the University of Oregon. Contact at brettyoutsey@gmail.com.

Ariful Azad is an Assistant Professor in the department of Intelligent Systems Engineering at Indiana University Bloomington in Bloomington, Indiana. His current research interests include high-performance computing, graph algorithms, and bioinformatics. He received his Ph.D. in Computer Science from Purdue University. Contact him at azad@iu.edu.