Bacteriophage 4, e979664; January-December 2014; Published with license by Taylor & Francis Group, LLC

Metagenomic ventures into outer sequence space

Bas E Dutilh[1,2,3,*]

[1]Theoretical Biology and Bioinformatics; Utrecht University; Utrecht, The Netherlands; [2]Centre for Molecular and Biomolecular Informatics; Radboud Institute for Molecular Life Sciences, Radboud University Medical Centre; Nijmegen, The Netherlands; [3]Department of Marine Biology, Institute of Biology; Federal University of Rio de Janeiro; Rio de Janeiro, Brazil

Sequencing DNA or RNA directly from the environment often results in many sequencing reads that have no homologs in the database. These are referred to as "unknowns," and reflect the vast unexplored microbial sequence space of our biosphere, also known as "biological dark matter." However, unknowns also exist because metagenomic datasets are not optimally mined. There is a pressure on researchers to publish and move on, and the unknown sequences are often left for what they are, and conclusions drawn based on reads with annotated homologs. This can cause abundant and widespread genomes to be overlooked, such as the recently discovered human gut bacteriophage crAssphage. The unknowns may be enriched for bacteriophage sequences, the most abundant and genetically diverse component of the biosphere and of sequence space. However, it remains an open question, what is the actual size of biological sequence space? The de novo assembly of shotgun metagenomes is the most powerful tool to address this question.

Keywords: biological dark matter, crAssphage, human gut, human virome, metagenomics, metagenome assembly, unknowns

*Correspondence to: Bas E Dutilh; Email: bedutilh@gmail.com

Submitted: 08/23/2014

Revised: 10/14/2014

Accepted: 10/17/2014

http://dx.doi.org/10.4161/21597081.2014.979664

Addendum to: Dutilh BE, et al. A highly abundant bacteriophage discovered in the unknown sequences of human faecal metagenomes," Nature Communications 2014; 5:4498.

Metagenomics is the untargeted sequencing of genetic material isolated from communities of micro-organisms and viruses. These communities may be derived from bioreactors, environmental, clinical, or industrial samples; in short, from anywhere in our unsterile biosphere. The classical questions in metagenomics that are asked about the sampled microbial community are "Who is there?" and "What are they doing?."[1] Originally an approach to answer these classical questions, metagenomics as a field has made great progress in the past decade. Applications include the use of metagenomics for the discovery of novel genetic functionality,[2] for describing microbial ecosystems and tracking their variation,[3] in untargeted medical diagnostics and forensics,[4] and as a powerful tool to determine the genome sequences of rare, uncultivable microbes.[5]

Powered by advances in next-generation sequencing technology, metagenomics has the potential to venture beyond the limits of currently explored sequence space by sampling environmental microbes and viruses at an unprecedented scale and resolution. Quite literally, sequence space is defined as the multi-dimensional space of all possible nucleotide (or protein) sequences.[6] Sequence space contains n dimensions; one dimension per residue that can take one of 4 (or 20, for proteins) states, with a total volume of $\Sigma 4^n$ sequences when summed over all possible sequence lengths n. Evolution may have largely explored this space,[7] but it remains an open question how large the current biological sequence space is, i.e. the fraction occupied by extant life. Figuratively, and within the context of this paper, "outer sequence space" is the remainder of this biological sequence space waiting to be explored by science.

Metagenomics has traditionally addressed the 2 classical questions listed above by aligning the sequencing reads in metagenomic data sets to a reference database containing known, annotated sequences. This allows the taxonomic and functional diversity of the sampled microbes to be described in terms of existing knowledge, allowing for straightforward interpretation of the results. However, a persistent concern in the analysis of metagenomes has been the unknown fraction, consisting of the reads

that cannot be annotated by using database searches. The level of unknowns can range up to 99% of the metagenomic reads, depending on the sampled environment, the protocols used for nucleotide isolation and sequencing, the homology search algorithm, and the reference database.[8]

Unknowns exist for 4 reasons that are not unrelated. The first reason is technical. Due to limitations of some next-generation sequencing platforms and library preparation protocols, spurious sequences may be generated that do not reflect true biological molecules. These artificial sequences include artifacts due to the sequencing technology[9] and chimeras, i.e., sequences generated from separate genetic molecules derived from different organisms. Since chimeras frequently arise during PCR amplification, they are expected to be more abundant in environmental amplicon sequencing than in shotgun metagenomics, and can be detected using bioinformatic tools.[10]

The second reason that unknowns exist is biological, as they reflect the enormous natural diversity of microorganisms that we are only beginning to unveil with metagenomics. This is both overwhelming and exciting, highlighting how much remains to be discovered in biology. This genetic diversity has been referred to as biological "dark matter,"[11,12] and is especially pronounced in viral metagenomes.[8] This issue can only be resolved by expanding reference databases, as exemplified by recent studies of one of the most studied microbial ecosystems: the human gut. The first metagenomic snapshots of the microbiota in the human gut were taken from 2 healthy adults, and revealed a high inter-individual diversity and many unknowns.[13] To a large extent, these unknowns were resolved when a reference catalog was created based on the sequences in the gut metagenomes themselves, decreasing the percentage of unknowns from ~85% to ~20%.[14] Moreover, subsequent large scale sequencing efforts revealed that in fact, many people share a similar intestinal flora, regardless of whether these similarities are viewed as discrete enterotypes[15] or as gradients.[16] These results illustrate how unknowns can be depleted by expanding the databases with appropriate reference sequences. This not only requires increased sequencing effort of phylogenetically diverse isolates[17] or single cells,[11] but also mining of draft genomes from metagenomes,[18] sampled from microbial environments around the globe.[19] Thus, by mapping the global sequence space, we can provide reassurance that at least some level of sampling saturation can be achieved. For viruses, and particularly for bacteriophages, efforts to provide a denser sampling of sequence space are still lacking.

The third reason that unknowns exist is methodological. Because the advances in DNA sequencing technology have greatly outpaced improvements in computer power,[20] bioinformatic approaches to analyze metagenomes often cut corners. For example, reference databases may be reduced to include only those references that are expected in the sample a priori. Moreover, read annotation may be limited to identifying almost exact sequence matches, as this can be computed much faster than if sequence variations needs to be taken into consideration in a permissive homology search. These issues lead to an inherent blind spot for discovering true novelty, such as sequences that are not expected in the sample, or organisms that have not been observed before. One way to, at least partially resolve this issue is by de novo assembly of the metagenome. Depending on the diversity of the sample, assembly can combine many short sequences (individual reads) into fewer, longer ones (assembled contigs). Reducing the number, and increasing the length of the sequences allows homology searches to be performed with more sensitive, computationally more expensive algorithms such as translated homology searches or profile searches, leading to more specific annotation and improved biological interpretation. Moreover, larger and more comprehensive reference databases can be used, allowing unexpected hits to be found.

The fourth reason that unknowns exist is logistical. Most research projects that generate metagenomic sequencing datasets deposit the read files in large repositories, provide an accession number in the associated publication, and move on. It is not unlikely that many of these data sets, consisting of files sometimes gigabytes in size, are never looked at again. Thus, while a certain sequence may have been "seen" in a metagenome and is thus strictly no longer "dark matter," it will still not be recognized when it is observed again. Re-identification of this sequence would only be possible if the publishing researcher identified it as an interesting sequence in his or her (assembled) metagenome, and submitted it to a searchable database like Genbank.[21] Because GenBank maintains very high standards for the sequences it accepts, submission can be a tedious process that is rarely worthwhile for unknown metagenomic contigs. An in depth investigation of the unknowns is rarely within the scope of a research project, and those sequences are thus first ignored and later forgotten. This is a waste of valuable resources: time, money, and work. The metagenomes available in public databases should be better exploited and mined for common sequences. To facilitate this, it is critical that metadata annotations of the metagenomes include a detailed description of the samples and sequencing protocol.[22] Exploiting these datasets will allow us to create more comprehensive maps of sequence space, and greatly improve our understanding and interpretation of metagenomes.

In the short term, ignoring the unknowns can facilitate the interpretation of a metagenome. Because a taxonomic or functional description cannot be provided, the classical questions in metagenomics are left unanswered for the unknown fraction of the metagenome, and concentrating on the annotated sequences leads to a more straightforward answer. However, unexpected or novel sequences are quickly overlooked, even if they represent highly abundant or widespread organisms. Thus, in the long term, stockpiling the unknown sequencing reads in badly accessible bulk sequence repositories can severely slow down research, the discovery of novel species, and the charting of biological sequence space.

One striking example of a novel genome discovered among the unknown sequences is crAssphage, a bacteriophage whose genome uniquely aligned sequencing reads from 73% of the 466 analyzed human gut metagenomes, and constituted

a total of 1.68% of those metagenomic reads.[23] Like many bacteriophages, its genome sequence is highly divergent from everything that was present in the annotated part of the Genbank database, which is why it was not observed before. It has been suggested that the unknown fraction of metagenomes is enriched for viral sequences,[8,24] because viral genomes are thought to evolve more rapidly than the genomes of cellular organisms, allowing them to explore a larger region of sequence space in the same amount of time.

To summarize, unknowns are genetic sequences that are difficult to identify using standard methods, such as by alignment to an annotated reference database. Unknowns remain a persistent elephant in the room in most metagenomics research projects, and exist for technical, biological, methodological, and logistical reasons. The most promising option to resolve the unknowns is by creating improved reference databases that chart biological sequence space, including the outer realms that remain unexplored by science (also known as dark matter). Besides sequencing reference strains or single cells, it may be expected that metagenomic sequencing, assembly, and binning will greatly add to improving these reference databases, for example by identifying common sequences in many metagenomes, and prioritizing them for targeted characterization. Characterizing unknowns will be vital to fully exploit the increasingly available metagenomic data sets from all ecosystems, toward understanding the roles of microbes and viruses in the biosphere. It remains an open question what is the actual size of biological sequence space, but the untargeted, shotgun nature of metagenomics makes it the most powerful tool to address this question.

Acknowledgments

I thank my collaborators for their contributions in the crAssphage project, and the anonymous reviewers of this manuscript for valuable suggestions.

References

1. Handelsman J. Metagenomics: application of genomics to uncultured microorganisms. Microbiol Mol Biol Rev 2004; 68:669-685; PMID:15590779; http://dx.doi.org/10.1128/MMBR.68.4.669-685.2004.
2. Wrighton KC, Thomas BC, Sharon I, Miller CS, Castelle CJ, VerBerkmoes NC, Wilkins MJ, Hettich RL, Lipton MS, Williams KH, et al. Fermentation, hydrogen, and sulfur metabolism in multiple uncultivated bacterial phyla. Science 2012; 337:1661-1665; PMID:23019650; http://dx.doi.org/10.1126/science.1224041
3. Yatsunenko T, Rey FE, Manary MJ, Trehan I, Dominguez-Bello MG, Contreras M, Magris M, Hidalgo G, Baldassano RN, Anokhin AP, et al. Human gut microbiome viewed across age and geography. Nature 2012; 486:222-227; PMID:22699611
4. Zaki AM, van Boheemen S, Bestebroer TM, Osterhaus AD, Fouchier RA. Isolation of a novel coronavirus from a man with pneumonia in Saudi Arabia. N Engl J Med 2012; 367:1814-1820; PMID:23075143; http://dx.doi.org/10.1056/NEJMoa1211721
5. Albertsen M, Hugenholtz P, Skarshewski A, Nielsen KL, Tyson GW, Nielsen PH. Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes. Nat Biotechnol 2013; 31:533-538; PMID:23707974; http://dx.doi.org/10.1038/nbt.2579
6. Smith JM. Natural selection and the concept of a protein space. Nature 1970; 225:563-564; PMID:5411867; http://dx.doi.org/10.1038/225563a0
7. Dryden DT, Thomson AR, White JH. How much of protein sequence space has been explored by life on Earth?. J R Soc Interface 2008; 5:953-956; PMID:18426772; http://dx.doi.org/10.1098/rsif.2008.0085
8. Mokili JL, Rohwer F, Dutilh BE. Metagenomics and future perspectives in virus discovery. Curr Opin Virol 2012; 2:63-77; PMID:22440968; http://dx.doi.org/10.1016/j.coviro.2011.12.004
9. Lassmann T, Hayashizaki Y, Daub CO. TagDust–a program to eliminate artifacts from next generation sequencing data. Bioinformatics 2009; 25:2839-2840; PMID:19737799; http://dx.doi.org/10.1093/bioinformatics/btp527
10. Edgar RC, Haas BJ, Clemente JC, Quince C, Knight R. UCHIME improves sensitivity and speed of chimera detection. Bioinformatics 2011; 27:2194-2200; PMID:21700674; http://dx.doi.org/10.1093/bioinformatics/btr381
11. Rinke C, Schwientek P, Sczyrba A, Ivanova NN, Anderson IJ, Cheng JF, Darling A, Malfatti S, Swan BK, Gies EA, et al. Insights into the phylogeny and coding potential of microbial dark matter. Nature 2013; 499:431-437; PMID:23851394; http://dx.doi.org/10.1038/nature12352
12. Youle M, Haynes M, Rohwer F. 2012 Scratching the surface of biology's dark matter. In Witzany G. (ed), Viruses: Essential Agents of Life. pp. 61–81.
13. Gill, SR, Pop M, Deboy RT, Eckburg PB, Turnbaugh PJ, Samuel BS, Gordon JI, Relman DA, Fraser-Liggett CM, Nelson KE. Metagenomic analysis of the human distal gut microbiome. Science 2006; 312:1355-1359; PMID:16741115; http://dx.doi.org/10.1126/science.1124234
14. 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; PMID:20203603; http://dx.doi.org/10.1038/nature08821
15. Arumugam M., Raes J, Pelletier E, Le Paslier D, Yamada T, Mende DR, Fernandes GR, Tap J, Bruls T, Batto JM, et al. Enterotypes of the human gut microbiome. Nature 2011; 473:174-180; PMID:21508958; http://dx.doi.org/10.1038/nature09944
16. Koren O, Knights D, Gonzalez A, Waldron L, Segata N, Knight R, Huttenhower C, Ley RE. A guide to enterotypes across the human body: meta-analysis of microbial community structures in human microbiome datasets. PLoS Comput Biol 2013; 9:e1002863; PMID:23326225; http://dx.doi.org/10.1371/journal.pcbi.1002863
17. Wu D, Hugenholtz P, Mavromatis K, Pukall R, Dalin E, Ivanova NN, Kunin V, Goodwin L, Wu M, Tindall BJ, et al. A phylogeny-driven genomic encyclopaedia of Bacteria and Archaea. Nature 2009; 462:1056-1060; PMID:20033048; http://dx.doi.org/10.1038/nature08656
18. Sharon I, Banfield JF. Microbiology. Genomes from metagenomics. Science 2013; 342:1057-1058; PMID:24288324; http://dx.doi.org/10.1126/science.1247023
19. Gilbert JA, Meyer F, Antonopoulos D, Balaji P, Brown CT, Brown CT, Desai N, Eisen JA, Evers D, Field D, et al. Meeting report: the terabase metagenomics workshop and the vision of an Earth microbiome project. Stand Genomic Sci 2010; 3:243-248; PMID:21304727; http://dx.doi.org/10.4056/sigs.1433550
20. Carlson R. The pace and proliferation of biological technologies. Biosecur Bioterror 2013; 1, 203-214.; http://dx.doi.org/10.1089/153871303769201851
21. Benson DA, Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW. GenBank. Nucleic Acids Res 2014; 42:D32-37; PMID:24217914; http://dx.doi.org/10.1093/nar/gkt1030
22. 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. Nat Biotechnol 2008; 26:541-547; PMID:18464787; http://dx.doi.org/10.1038/nbt1360
23. Dutilh BE, Cassman N, McNair K, Sanchez SE, Silva GG, Boling L, Barr JJ, Speth DR, Seguritan V, Aziz RK, et al. A highly abundant bacteriophage discovered in the unknown sequences of human faecal metagenomes. Nat Commun 2014; 5:4498; PMID:25058116; http://dx.doi.org/10.1038/ncomms5498
24. Li SC, Chan WC, Lai CH, Tsai KW, Hsu CN, Jou YS, Chen HC, Chen CH, Lin WC. UMARS: Un-MAppable Reads Solution. BMC Bioinformatics 2011; 12 Suppl 1:S9; PMID:21342592; http://dx.doi.org/10.1186/1471-2105-12-S1-S9