RESEARCH HIGHLIGHT OPEN

Gut Phage Database: phage mining in the cave of wonders

Magdalena Unterer[1], Mohammadali Khan Mirzaei[1] and Li Deng[1]

Signal Transduction and Targeted Therapy (2021)6:193 ; https://doi.org/10.1038/s41392-021-00615-2

The diversity of the viruses in the human gut remains mostly unexplored. To shed some light on this known unknown, Camarillo-Guerrero et al.[1] developed the Gut Phage Database (GPD), which includes 142,809 non-redundant gut phages obtained from analyzing 28,060 shotgun metagenome datasets.

The human intestinal tract is one of the most diverse microhabitats known, harboring billions of microorganisms including bacteria, viruses, fungi and archaea.[2] Bacteria and their viruses, called phages, are the most abundant microbes in the human gut in an approximate ratio of 1:1 and a total abundance of about 10e13.[1–3] Yet, we currently have little more than anecdotal data about gut phages relative to what we know about their bacterial hosts.[2] The gut bacteria play central roles in human metabolism, immune modulation, and protection against pathogens.[2,4] In addition, imbalances in their community contribute to human diseases or conditions such as Inflammatory Bowel Disease (IBD), allergies, obesity, and more.[2,3] Similarly, changes in phage composition have been observed in IBD, type 2 diabetes, stunting, and Parkinson's disease.[2,3]

Despite the renowned effects of phages on bacterial communities in other ecosystems, their function in the human gut is, for the most part, still unclear.[1–3] This is due to the specific characteristics of phage genomes and the limited toolkit for studying them. Phage genomes are highly diverse, relatively small, and represent high levels of genetic mosaicism.[2,3] They also lack universal gene markers unlike bacteria. Thus, their identification requires shotgun sequencing, which is prone to high background noise from human and bacterial genomes, which asks for extensive decontamination in the downstream analysis.[2,3] Some experimental approaches like enriching VLPs before metagenome extractions can decrease contaminations by hosts' genomes, but they also have some limitations.[5] The public databases also do not sufficiently represent phage diversity resulting in most gut phages showing no significant homology to known reference sequences.[1–3] Taken together, these challenges significantly increase the complexity of studying gut phages. As the result, earlier metagenomic studies found the majority of identified gut phages share no homology to public databases with a high variability between different studies, 14–87 %.[2,3]

To address the limitations with the public databases and expand our understanding of gut viral diversity Camarillo-Guerrero et al.[1] developed the GPD which includes 142,809 non-redundant phage genomes obtained from mining 28,060 human gut metagenomes, and 2,898 genomes of cultured gut bacteria Fig. 1. Among phage genomes assembled, 9.4% were classified as complete, and 19.6% as high quality, as estimated by CheckV. The median phage genome size in GPD is ~31 kb, which is twice or three times longer than in other phage databases.[1] For this, they used a rigorous quality control approach combined with an inhouse machine learning method to filter out the high background noise of metagenomic data and obtain complete viral genomes Fig. 1. Their machine learning approach uses gene density and k-mer frequency to distinguish and remove contamination with integrative and conjugative elements (ICEs). It also recognizes prophages that rarely enter the lytic cycle and mobile genetic elements. In addition, they predicted the host range of the GPD phages by screening them for CRISPR spacers, linking prophages to bacterial genomes, and then analyzing their co-occurrence with the predicted host to validate their predictions.[1] Using this approach, 28.66% of phages could be assigned to 2157 strains of the host bacteria, while the highest diversity among phages linked to the Firmicutes phylum. In addition, about 36% of gut phages identified had broad host range and can potentially infect more than one bacterial species, which suggests that broad host range phages are more common in the human gut than previously hypothesized.[1] Moreover, the epidemiological analysis of the GPD phages revealed 280 viral clusters (VCs) that are globally distributed over five different continents, at least. Phages that were found to be globally distributed had a broader host range compared to those seen only in one region. They could also see a clear separation of gut phages based on the geographical area they originated from and the human host lifestyles— living in rural vs. urban area.[1] Finally, they discovered a new phage clade called Gut Bacteroidales phage or Gubaphage Fig. 1, which was found to be the second most prevalent VC in the human gut after the crAssphages— with no homology with these phages.[1]

Taken together, Camarillo-Guerrero et al. have significantly improved our understanding of the diversity, host range, and geographical distribution of unknown gut phages, as well as identified a new phage clade— highly common in the human gut. In addition, the high-quality, large-scale phage database of gut phages developed in this study will be a valuable resource for studying the role of phages in regulating human health. Yet, the full diversity of gut phages remains unexplored; no taxonomy could be assigned to the majority of reconstructed phages; the function of most phage proteins is still unknown, and diversity of RNA phages is not represented in the GPD database Fig. 1.

ACKNOWLEDGEMENTS

The authors thank Jinlong Ru for his constructive comments, and Sophie E. Smith for proofreading the manuscript. Kawtar Tiamani has created the figures. Magnifier used in the figure was taken from BioRender.com. This work was funded by the German Research Foundation (DFG Emmy Noether program, Proj. No. 273124240, SFB 1371, Proj. No. 395357507), and the European Research Council Starting grant (ERC StG 803077) awarded to L.D.

[1]Institute of Virology, Helmholtz Center Munich and Technical University of Munich, Neuherberg, Bavaria, Germany
Correspondence: Li Deng (li.deng@helmholtz-muenchen.de)

Received: 7 March 2021 Revised: 31 March 2021 Accepted: 13 April 2021
Published online: 17 May 2021

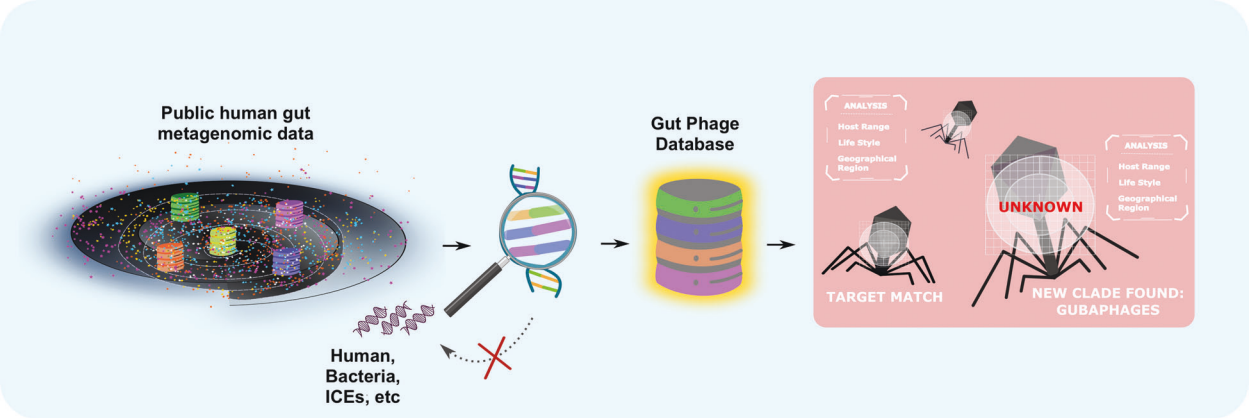

Fig. 1 Schematic representation of key steps followed to develop the Gut Phage Database. The black halo hints at the viral dark matter. Databases, in colors, represent metagenomic data from different continents. The magnifier suggests the rigorous quality controls done. Phage scanning implies viral profiling and discovery. ICEs integrative and conjugative elements

FUNDING

Open Access funding enabled and organized by Projekt DEAL.

ADDITIONAL INFORMATION

Conflict of interest: The authors declare no competing interests.

REFERENCES

1. Camarillo-Guerrero, L. F., Almeida, A., Rangel-Pineros, G., Finn, R. D. & Lawley, T. D. Massive expansion of human gut bacteriophage diversity. *Cell* **184**, 1098–1109 (2021). e9.
2. Mirzaei, M. K. & Maurice, C. F. Ménage à trois in the human gut: Interactions between host, bacteria and phages. *Nat. Rev. Microbiol.* **15**, 397–408 (2017).
3. Shkoporov, A. N. & Hill, C. Bacteriophages of the human gut: the 'known unknown' of the microbiome. *Cell Host Microbe.* **25**, 195–209 (2019).
4. Ducarmon, Q. R. et al. Gut microbiota and colonization resistance against bacterial enteric infection. *Microbiol. Mol. Biol. Rev.* **83**, e00007–19 (2019).
5. Hall, R. J. et al. Evaluation of rapid and simple techniques for the enrichment of viruses prior to metagenomic virus discovery. *J. Virol. Methods* **195**, 194–204 (2014).