Determinants of bacterial fitness and their impact on prokaryotic genome evolution

Inaugural-Dissertation

zur Erlangung des Doktorgrades der
Mathematisch-Naturwissenschaftlichen Fakultaet der
Heinrich-Heine-Universität Düsseldorf

Na Gao

Düsseldorf, 16. Mai 2019

aus dem Institut für Informatik

der Heinrich-Heine-Universität Düsseldorf

Gedruckt mit der Genehmigung der

Mathematisch-Naturwissenschaftlichen Fakultät der

Heinrich-Heine-Universität Düsseldorf

Berichterstatter:

1. Prof. Dr. Martin Lercher

2. Prof. Dr. William Martin

3. Prof. Dr. Wei-Hua Chen (HUST)

Tag der mündlichen Prüfung:

Erklärung

Ich versichere an Eides statt, dass die Dissertation von mir selbstständig und ohne unzulässige fremde Hilfe unter Beachtung der "Grundsätze zur Sicherung guter wissenschaftlicher Praxis an der Heinrich-Heine-Universität Düsseldorf" erstellt worden ist.

Düsseldorf, den 16. May 2019

_____

Na Gao

Contents

Summary

Prokaryotes thrive in all known habitats on earth. They have important industrial values and a significant impact on human health. In this thesis, I interrogate recently published (meta-)genomic sequences generated by high-throughput sequencing techniques to identify key factors that contribute to the fitness of prokaryotes and the impacts on their genomes.

We first explored intrinsic factors. A previous study of our group identified that efficient resource usage shapes nucleotide usage in coding regions of prokaryotic genomes. In this study, we further revealed that efficient resource usage could also drive genes to be preferably located on the leading strand, an observation known as strand-biased gene distribution (SGD). The leading strand is synthesized in the same direction as the movement of the replication fork, while the lagging strand is synthesized in the opposite direction. The transcription and replication machineries collide head-to-head on the lagging strand, leading to longer exposure time of single-stranded DNA to chemical modifications. Lagging strand genes thus accumulate more deleterious mutations. Mutational biases introduced energetically cheaper nucleotides on the lagging strand, resulting in more expensive protein products, which consequently drove genes to the leading strand. We tested our mutagenesis/energy efficiency model in 1,552 prokaryotic genomes and found that mutational biases in non-transcribed regions can explain ~71% of the variation in SGDs; consistently, the difference between averaged amino acid costs of proteins encoded by genes on the two strands explained ~50% of the variance in SGDs.

We next explored external factors such as bacteriophages. Phages invade microbes, accomplish host lysis, and are of vital importance in shaping the community structure of environmental microbiota. Phage-mediated horizontal gene transfer is known to have a significant impact on the formation, evolution,

and host range transition of virulence factors of pathogenic bacteria. We first identified 26,572 interactions between 18,608 viral clusters (complete and fragmented phage genomes) and 9,245 prokaryotes (i.e., bacteria and archaea). Based on these interactions, we calculated the host range for each of the phage clusters, and accordingly grouped them into subgroups such as species-, genus-, and family-specific phage clusters. We also calculated the size and *GC*-content of bacteria for the gut metagenome, which contains a variety of bacteriophages, plasmids, and CRISPRs. We found that both phages and plasmids contribute significantly to genome expansion, i.e., genomes with phages and/or plasmids are significantly larger than those without; the genome sizes were increased with increasing numbers of associated phages/plasmids. Conversely, we found that CRISPR systems have a negative impact on genome size, i.e., genomes with CRISPRs are significantly smaller in size than those without. These results confirmed that on an evolutionary timescale, phages and plasmids facilitated genome expansions while CRISPR impaired such processes in prokaryotes. Furthermore, our results also revealed a striking yet expected preference of CRISPR systems against phages over plasmids, consistent with the typical consequences of phage and plasmid infection to the hosts and the roles of CRISPR as a defence system.

Finally, we constructed an MVP database (microbe-phage interaction database) using the results of our microbe-phage interaction analysis. Phages can be used as antibiotic agents for pathogenic prokaryotes and/or a tool to specifically “knockdown” target prokaryotes without affecting others. Therefore, such a resource will be useful in (meta-)genomic studies and of potential clinical importance.

2 Introduction

Prokaryotes are everywhere around us. The microbiome plays crucial roles in human health (1-3), diseases (3-9), development (10-12), and in many other aspects of human life (5-8). In the work reported in this thesis, I interrogated recently published (meta-)genomic sequences generated by high-throughput sequencing techniques; my goal was to identify key factors that contribute to the genomic adaptation of prokaryotes.

2.1 Brief introduction to prokaryotes

2.1.1 Prokaryotes thrive in all known habitats with high abundance

Escherichia coli, the most widely studied prokaryote, has a genome about 700 times smaller than a human genome (13). Prokaryotes are considered to be the earliest organisms on earth (14). Their cells possess a cytoskeleton that is much more primitive than that of eukaryotes (15). Most prokaryotes are unicellular organisms. They lack a nuclear membrane, mitochondria, or any other membrane-bound organelles (16). However, some prokaryotes contain intracellular structures that could be seen as primitive organelles (17). All in all, the lack of a nuclear membrane makes it easier for prokaryotes to incorporate foreign DNA into their own genomes, a phenomenon known as horizontal gene transfer (HGT) (18-20). Recent analyses revealed that HGT may contribute more to the expansion of prokaryotic genomes than gene duplication (21,22). Prokaryotes frequently adapt to new environments by acquiring foreign genes, often from organisms living in the same habitats, through HGT (23,24). However, despite the adaptive advantages that may come with foreign DNA, the integration of foreign genetic material is risky: for this reason, more than half of all prokaryotic genomes encode CRISPR-CAS systems that can recognize and degrade invading

foreign DNA (25-27).

Prokaryotes are classified into two domains: bacteria and archaea (28). Both bacteria and archaea can thrive in practically all habitats on earth, including those that are cold, hot, salty, acidic, or alkaline (29). Prokaryotes can be found in human lungs and guts and on human skin (1,30,31). Even in rocks two miles below the surface of the earth, prokaryotes have been discovered (32).

Prokaryotes are highly abundant: their biomass has been estimated to outweigh that of all eukaryotes combined by at least tenfold (15). The total number of bacterial and archaeal cells in the human gut can be up to ten times more than that of the human cells (33).

2.1.2 Availability of large amounts of (meta-) genomic data facilitated large-scale comparative analyses of prokaryotic genomes

Next generation sequencing (NGS) has emerged as a cost-effective and convenient approach for addressing many microbiological questions, dramatically transforming this field. Metagenomic assembly of short sequencing reads facilitates functional insights. Compared to culture-based and single-cell methods, metagenomics provides a more convenient and unbiased way of obtaining genomic information for environmental microbes (34,35); accordingly, having access to genomic information has revolutionized fundamental research in microbiology (36).

With an increasing amount of sequencing data, the number of microbial species and genes discovered grows rapidly. This allowed me and my collaboration partners to use larger and more comprehensive samples than previous researchers to examine some controversial issues.

2.1.3 Aims of this dissertation and efforts towards their accomplishment

This thesis describes how I and my collaboration partners interrogated recently published metagenomic sequences to identify intrinsic and external key factors that contribute to the fitness of prokaryotes, and our examination of their impact on prokaryotic genomes. We first focused on intrinsic factors. We studied how basic cellular activities such as replication, transcription, and translation can change base composition, i.e., the relative frequencies of the four nucleotides of the genome. We found that this consequently drove protein-coding genes onto the leading strand, on which the DNA replication and the transcription machineries move in the same direction. We then looked at the external factors, studying how horizontal gene transfers, especially those facilitated by phages and plasmids, can drive genome expansions at evolutionary timescales. Prokaryotic cells may impair such processes using genome-encoded CRISPR-CAS systems, a wide-spread adaptive immune system of prokaryotes. Our results unveiled some interesting interactions between internal and external factors. Finally, taking advantage of the huge amount of data we collected for the two projects, especially the (pro)phage sequences and their interactions relationships with their host genomes, we constructed a microbe-phage interaction database. Phages can be used as antibiotic agents for pathogenic prokaryotes and/or a tool to specifically "knockdown" target prokaryotes without affecting others. Therefore, such a resource may contribute to (meta-)genomic studies and is of potential clinical importance.

2.2 Intrinsic and external factors shaping prokaryotic genomes

2.2.1 Mutational biases and selection for effective energy usage drives protein coding genes to the leading strand

Prokaryotes spend a substantial fraction of their cellular resources on making nucleotides – about 13% of glucose consumption in *Escherichia coli* is used to make nucleotides (37). Efficient energy usage is a trait under strong selection (38), and thus parsimonious resource usage has been observed in various genomic aspects: for example, highly expressed proteins are shorter than lowly expressed proteins (39,40) and preferentially use cheaper amino acids (37,41-43), microbes predominantly use energetically cheaper amino acids in secreted proteins (44), and prokaryotic genomes tend to use cheaper nucleotides in transcribed than in untranscribed sequences, as the former are often amplified thousand-fold compared with the latter (45). Moreover, transcription-related selection generally favours the cheaper nucleotides *U* and *C* at synonymous sites (45).

In most prokaryotes, protein-coding gene locations are biased to the leading strand (46), on which replication is continuous (47,48). Over 90% of 1,552 analysed prokaryotic genomes located their coding genes preferentially on the leading strand (49), a phenomenon called strand-biased gene distribution (SGD) (50). It has long been suspected that SGDs are caused by natural selection favouring the avoidance of collisions between the replication and transcription machineries (46,50-52). These two machineries share the same DNA template but move with different speed (53) and, importantly, in different directions on the lagging strand. Thus, collisions could happen either co-directionally (on the leading strand) or head-on (on the lagging strand) (54). Some results suggest that collisions are deleterious (55), and that head-on collisions are more deleterious

than co-directional collisions (56). The elevated deleterious effects of genes encoded on the lagging strand were believed to be stronger for highly expressed and for functionally important genes (e.g., essential genes), consistent with the observation that these two types of genes are underrepresented on the lagging strand (50,57).

Despite the mechanistic insights, a quantitative model that explains the variation of SGDs across different species is still lacking. For example, the expression-driven (50) and essentiality-driven (57) hypotheses are not quantitative; it is difficult to quantify their contributions to the SGD, i.e., they offer no explanation why in different genomes the SGDs are different, and how much of the variations can be explain by essential or highly expressed genes. Recently, Mao *et al.* (49) proposed a very sophisticated model to explain ~74% of the variance of the SGDs in 725 prokaryotic genomes. Although their work represents arguably one of the best quantitative models so far, no causal relationship has been inferred from their results.

In our study, we proposed a mutagenesis/energy efficiency model for SGDs and tested it on a dataset of 1,552 prokaryotic genomes. We showed that due to elevated mutational biases on the lagging strand (48), the energetically cheaper nucleotides *T* is introduced over *G*, so is *C* over *A* and *C* over *G*; proteins encoded by lagging-strand genes are slightly more expensive than those encoded by leading-strand genes, and subsequently drive genes to the leading strand. Consequently, genes, especially those that are highly expressed, are preferably located on the leading strand. Highly expressed genes code for cheaper products, even when they are located on the lagging strand; thus not all highly expressed genes, and certainly not all genes, are expected to be moved to the leading strand. Our model is quantitative, compatible with many existing hypotheses (37,41-43,45), and can explain more than two thirds (~71%) of the variance in SGDs.

2.2.2 Genome expansion is facilitated by the plasmids and phages, and is impaired by CRISPR-CAS systems

Gene duplication and/or horizontal gene transfer (HGT) play important roles in functional innovation and species adaptation, and are the main sources of genome expansions (23,24,58-60). While many prokaryotic genes have been acquired by horizontal transfer at some point in their evolutionary history, not all genes are equally likely to be transferred (61-63). HGT is also the one of main sources of genome expansions (23,24,58-60). Mobile DNA elements such as phages and plasmids can infect their hosts and introduce foreign DNAs into the host genomes. HGT occurs through three main mechanisms: transformation, conjugation, and transduction. The latter two mechanisms are related to plasmids and phages, respectively.

Mobile DNA elements such as bacteriophages (referred to as phages below) and plasmids can infect their hosts and introduce novel DNA into the host genomes (64-67). Plasmids that contain resistance genes from resistant donors can make previously susceptible bacteria express resistance, encoded by these newly acquired resistance genes (68). The acquisition of foreign DNA can have diverse fitness consequences: many adaptations are facilitated by HGT (69), but in other cases the DNA being shared is neutral or even harmful (70). Phages are pathogens that often lead to the lysis of their hosts (71). In transduction, the transfer of bacterial DNA is under the control of the phage's genes rather than bacterial genes (72). Phages often have a very narrow range of hosts; but under certain conditions, such as antibiotic stress, phages and plasmids can expand their host ranges (73). Overall, phages and plasmids are important sources of HGT and of prokaryotic innovations, and consequently contribute to bacterial evolution and adaptation (19,20,73). Accordingly, we hypothesized that the number of plasmids/phages may be related to the genome size of their host.

Over the evolution history of prokaryotes, they developed various defence

systems against phages and other invading genetic elements (74). CRISPR (clustered regularly interspaced short palindromic repeats), the adaptive immune system of prokaryotes, is a recently recognized player in the ongoing arms race between viruses and hosts, and plays an important role in the dynamic process by which the genomes of prokaryotes and mobile elements coevolve (75). CRISPR is wide-spread in prokaryotes, present on chromosomes, genomic islands, plasmids, and even mimiviruses (76), and has been distributed via HGT between different prokaryotic taxa (25-27). In 1987, a CRISPR-Cas system was first recognized in *Escherichia coli*; such systems are now known to occur in 90% of archaea and 40% of bacteria (25-27,77). CRISPR loci continuously acquire new spacers; this facilitates a partial reconstruction of the history of past selfish-element infections (78-81). In the absence of parasitic elements, spacers could be easily lost because of the deletion bias of prokaryotic genome evolution (82) and the presumed cost of maintaining CRISPR systems (83). The balance between spacer gain and loss could thus be affected by the relative selective pressures exerted (84). It is reasonable to speculate that over the course of evolution, phages and plasmids – as sources of HGT – may contribute to the expansion of prokaryotic genomes, while CRISPR systems – which prevent HGT – may impair such a process.

However, controversial observations on this issue have been reported recently. For example, Gophna and colleagues did not observe the expected negative correlation between CRISPR activity in microbes with three independent measures of recent HGT, leading them to conclude that the inhibitory effect of CRISPR against HGT is undetectable (85). Furthermore, a recent study revealed that CRISPR-mediated phage resistance can even enhance HGT by increasing the resistance of transductants against subsequent phage infections (86). These observations appear surprising, as the restricted acquisition of foreign genetic material is believed to be one of the sources of the maintenance fitness cost of CRISPR systems and may be one of the reasons for the patchy distribution of CRISPR among bacteria (87,88). Thus, it is currently unclear what long-term

effects CRISPR, phages, and plasmids have on genome expansions.

In this study, we first collected a comprehensive dataset of prokaryotes and their associations with phages, plasmids, and CRISPR systems. We then applied a generalized linear model to evaluate the contributions of phages, plasmids, and CRISPR to genome size. After controlling for genome *GC* (guanine+cytosine) content, which is known to correlate significantly with genome size (45,89), we found that both phages and plasmids are associated with larger genomes, while the presence of a CRISPR system is associated with small genome sizes. Genome sizes increase with increasing numbers of associated phages and plasmids. Our results thus indicate that in the long run, phages and plasmids facilitate genome expansions, while CRISPR impairs such a process in prokaryotes. Furthermore, our results also reveal a striking preference of CRISPR systems for targeting phages rather than plasmids, consistent with the typical consequences of phage and plasmid infections to the hosts and the roles of CRISPR as a defence system.

2.2.3 A comprehensive catalog of phage-microbe interactions

It has been increasingly recognized that the microbiome can play crucial roles in human health (1-3), diseases (3-9), responses to drugs and treatments (90,91), and other processes (10,12,92,93). However, due to limited experimental conditions and the lack of general purpose tools, it is difficult to directly infer causal relationships from the correlated alterations in microbial community structures and host phenotypes (e.g., health statuses) under different conditions (20-23), or to even directly pinpoint the causal species.

During the course of our data collection for the previous two projects, we assembled a large set of phage-microbe interactions. It is known that phages are key members of the environmental microbiota and could play important roles in shaping the population structure. Most importantly, they tend to have specific hosts (96) and are able to decrease the fitness of their host prokaryotes. Therefore phages can be used as a tool to specifically “knock-down” prokaryotes from an

environmental microbiota without affecting others in the same environment, providing us with an ideal tool to precisely manipulate prokaryotes of interests at the species level. Recently, Yen *et al.* successfully reduced *Vibrio cholerae* infection and colonization in the intestinal tract and prevented cholera-like diarrhea, by orally administrating *V. cholera*-specific phages in model animals (94). Therefore, knowledge about phage-microbe interactions can be particularly useful for researchers who are interested in environmental microbiota studies. We thus want to provide researchers with a comprehensive catalogue of phage-microbe interactions and to assist them to select phages that can target (and thus help to manipulate) specific microbes of interest.

In addition to experimental methods, microbe-phage interactions can be identified by taking advantage of large-scale genomic and metagenomic sequencing efforts. For example, it is known that many phages insert their genomes into that of their hosts; the integrated phages are known as prophages (95,96). Many computational tools exist and are able to identify prophages from complete prokaryotic genomes and/or assembled metagenomic contigs (97-99).

In this study, we first collected 50,782 viral sequences from published datasets, public databases, and re-analysis of genomic and metagenomic sequences, and clustered them into 33,097 unique viral clusters based on sequence similarity. We then identified 26,572 interactions between 18,608 viral clusters and 9,245 prokaryotes; we established these interactions based on 30,321 evidence entries that we collected from various sources. Based on these interactions, we calculated the host range for each of the phage clusters, and accordingly grouped them into subgroups such as species-, genus-, and family-specific phage clusters. All results are integrated into the MVP, a microbe-phage interaction database, which allows users to effortlessly explore all contents and to efficiently find interactions of interest to them. We expect that this resource will be useful in (meta-)genomic studies, and will be of potential clinical importance.

References

1. Arumugam, M., Raes, J., Pelletier, E., Le Paslier, D., Yamada, T., Mende, D.R., Fernandes, G.R., Tap, J., Bruls, T., Batto, J.M. *et al.* (2011) Enterotypes of the human gut microbiome. *Nature*, **473**, 174-180.
2. Qin, J., Li, R., Raes, J., Arumugam, M., Burgdorf, K.S., Manichanh, C., Nielsen, T., Pons, N., Levenez, F., Yamada, T. *et al.* (2010) A human gut microbial gene catalogue established by metagenomic sequencing. *Nature*, **464**, 59-65.
3. Turnbaugh, P.J., Ley, R.E., Mahowald, M.A., Magrini, V., Mardis, E.R. and Gordon, J.I. (2006) An obesity-associated gut microbiome with increased capacity for energy harvest. *Nature*, **444**, 1027-1031.
4. Pedersen, H.K., Gudmundsdottir, V., Nielsen, H.B., Hyotylainen, T., Nielsen, T., Jensen, B.A., Forslund, K., Hildebrand, F., Prifti, E., Falony, G. *et al.* (2016) Human gut microbes impact host serum metabolome and insulin sensitivity. *Nature*, **535**, 376-381.
5. Qin, J., Li, Y., Cai, Z., Li, S., Zhu, J., Zhang, F., Liang, S., Zhang, W., Guan, Y., Shen, D. *et al.* (2012) A metagenome-wide association study of gut microbiota in type 2 diabetes. *Nature*, **490**, 55-60.
6. Noguera-Julian, M., Rocafort, M., Guillen, Y., Rivera, J., Casadella, M., Nowak, P., Hildebrand, F., Zeller, G., Parera, M., Bellido, R. *et al.* (2016) Gut Microbiota Linked to Sexual Preference and HIV Infection. *EBioMedicine*, **5**, 135-146.
7. Frye, R.E., Slattery, J., MacFabe, D.F., Allen-Vercoe, E., Parker, W., Rodakis, J., Adams, J.B., Krajmalnik-Brown, R., Bolte, E., Kahler, S. *et al.* (2015) Approaches to studying and manipulating the enteric microbiome to improve autism symptoms. *Microbial ecology in health and disease*, **26**, 26878.
8. Hsiao, E.Y., McBride, S.W., Hsien, S., Sharon, G., Hyde, E.R., McCue, T., Codelli, J.A., Chow, J., Reisman, S.E., Petrosino, J.F. *et al.* (2013) Microbiota modulate behavioral and physiological abnormalities associated with neurodevelopmental disorders. *Cell*, **155**, 1451-1463.
9. Li, J., Zhao, F., Wang, Y., Chen, J., Tao, J., Tian, G., Wu, S., Liu, W., Cui, Q., Geng, B. *et al.* (2017) Gut microbiota dysbiosis contributes to the development of hypertension. *Microbiome*, **5**, 14.
10. Backhed, F., Roswall, J., Peng, Y.Q., Feng, Q., Jia, H.J., Kovatcheva-Datchary, P., Li, Y., Xia, Y., Xie, H.L., Zhong, H.Z. *et al.* (2015) Dynamics and Stabilization of the Human Gut Microbiome during the First Year of Life. *Cell host & microbe*, **17**, 690-703.
11. Forsgren, M., Isolauri, E., Salminen, S. and Rautava, S. (2017) Late preterm birth has direct and indirect effects on infant gut microbiota development during the first six months of life. *Acta paediatrica*, **106**, 1103-1109.
12. Wall, R., Ross, R.P., Ryan, C.A., Hussey, S., Murphy, B., Fitzgerald, G.F. and Stanton, C. (2009) Role of gut microbiota in early infant development. *Clinical medicine. Pediatrics*, **3**, 45-54.
13. Tenaillon, O., Skurnik, D., Picard, B. and Denamur, E. (2010) The population genetics of commensal Escherichia coli. *Nature reviews. Microbiology*, **8**, 207-217.
14. Watson, A.J. (2008) Implications of an anthropic model of evolution for emergence of

complex life and intelligence. *Astrobiology*, **8**, 175-185.

15. Tellez, G. (2014) Prokaryotes Versus Eukaryotes: Who is Hosting Whom? *Frontiers in veterinary science*, **1**, 3.
16. Charles C. Tseng, X.Y. (2013) Cell Cycle and DNA Replication: How Does DNA Replicate in Preparation for Cell Division?
17. Shively, J.M. (2018) Complex Intracellular Structures in Prokaryotes.
18. Koonin, E.V. and Wolf, Y.I. (2008) Genomics of bacteria and archaea: the emerging dynamic view of the prokaryotic world. *Nucleic acids research*, **36**, 6688-6719.
19. Argov, T., Azulay, G., Pasechnek, A., Stadnyuk, O., Ran-Sapir, S., Borovok, I., Sigal, N. and Herskovits, A.A. (2017) Temperate bacteriophages as regulators of host behavior. *Current opinion in microbiology*, **38**, 81-87.
20. Nogueira, T., Rankin, D.J., Touchon, M., Taddei, F., Brown, S.P. and Rocha, E.P. (2009) Horizontal gene transfer of the secretome drives the evolution of bacterial cooperation and virulence. *Current biology : CB*, **19**, 1683-1691.
21. Pal, C., Papp, B. and Lercher, M.J. (2005) Adaptive evolution of bacterial metabolic networks by horizontal gene transfer. *Nature genetics*, **37**, 1372-1375.
22. Treangen, T.J. and Rocha, E.P. (2011) Horizontal transfer, not duplication, drives the expansion of protein families in prokaryotes. *PLoS genetics*, **7**, e1001284.
23. Smith, G., Macias-Munoz, A. and Briscoe, A.D. (2016) Gene Duplication and Gene Expression Changes Play a Role in the Evolution of Candidate Pollen Feeding Genes in Heliconius Butterflies. *Genome biology and evolution*, **8**, 2581-2596.
24. Tsai, Y.M., Chang, A. and Kuo, C.H. (2018) Horizontal Gene Acquisitions Contributed to Genome Expansion in Insect-Symbiotic Spiroplasma clarkii. *Genome biology and evolution*, **10**, 1526-1532.
25. Godde, J.S. and Bickerton, A. (2006) The repetitive DNA elements called CRISPRs and their associated genes: evidence of horizontal transfer among prokaryotes. *Journal of molecular evolution*, **62**, 718-729.
26. Makarova, K.S., Haft, D.H., Barrangou, R., Brouns, S.J., Charpentier, E., Horvath, P., Moineau, S., Mojica, F.J., Wolf, Y.I., Yakunin, A.F. *et al.* (2011) Evolution and classification of the CRISPR-Cas systems. *Nature reviews. Microbiology*, **9**, 467-477.
27. Seed, K.D., Lazinski, D.W., Calderwood, S.B. and Camilli, A. (2013) A bacteriophage encodes its own CRISPR/Cas adaptive response to evade host innate immunity. *Nature*, **494**, 489-491.
28. Schleifer, K.H. (2009) Classification of Bacteria and Archaea: past, present and future. *Systematic and applied microbiology*, **32**, 533-542.
29. Hugenholtz, P., Pitulle, C., Hershberger, K.L. and Pace, N.R. (1998) Novel division level bacterial diversity in a Yellowstone hot spring. *Journal of bacteriology*, **180**, 366-376.
30. Venkataraman, A., Bassis, C.M., Beck, J.M., Young, V.B., Curtis, J.L., Huffnagle, G.B. and Schmidt, T.M. (2015) Application of a neutral community model to assess structuring of the human lung microbiome. *mBio*, **6**.
31. Grice, E.A., Kong, H.H., Conlan, S., Deming, C.B., Davis, J., Young, A.C., Program, N.C.S., Bouffard, G.G., Blakesley, R.W., Murray, P.R. *et al.* (2009) Topographical and temporal diversity of the human skin microbiome. *Science*, **324**, 1190-1192.
32. Davila, A.F., Skidmore, M., Fairen, A.G., Cockell, C. and Schulze-Makuch, D. (2010) New

priorities in the robotic exploration of Mars: the case for in situ search for extant life. *Astrobiology*, **10**, 705-710.

33. Zhu, B., Wang, X. and Li, L. (2010) Human gut microbiome: the second genome of human body. *Protein & cell*, **1**, 718-725.
34. Hu, Y., Zhang, G., Li, A., Chen, J. and Ma, L. (2008) Cloning and enzymatic characterization of a xylanase gene from a soil-derived metagenomic library with an efficient approach. *Applied microbiology and biotechnology*, **80**, 823-830.
35. Frankel, A.E., Coughlin, L.A., Kim, J., Froehlich, T.W., Xie, Y., Frenkel, E.P. and Koh, A.Y. (2017) Metagenomic Shotgun Sequencing and Unbiased Metabolomic Profiling Identify Specific Human Gut Microbiota and Metabolites Associated with Immune Checkpoint Therapy Efficacy in Melanoma Patients. *Neoplasia*, **19**, 848-855.
36. Fraser-Liggett, C.M. (2005) Insights on biology and evolution from microbial genome sequencing. *Genome research*, **15**, 1603-1610.
37. Akashi, H. and Gojobori, T. (2002) Metabolic efficiency and amino acid composition in the proteomes of Escherichia coli and Bacillus subtilis. *Proceedings of the National Academy of Sciences of the United States of America*, **99**, 3695-3700.
38. Hoehler, T.M. and Jorgensen, B.B. (2013) Microbial life under extreme energy limitation. *Nature reviews. Microbiology*, **11**, 83-94.
39. Li, S.W., Feng, L. and Niu, D.K. (2007) Selection for the miniaturization of highly expressed genes. *Biochemical and biophysical research communications*, **360**, 586-592.
40. Castillo-Davis, C.I., Mekhedov, S.L., Hartl, D.L., Koonin, E.V. and Kondrashov, F.A. (2002) Selection for short introns in highly expressed genes. *Nature genetics*, **31**, 415-418.
41. Raiford, D.W., Heizer, E.M., Jr., Miller, R.V., Doom, T.E., Raymer, M.L. and Krane, D.E. (2012) Metabolic and translational efficiency in microbial organisms. *Journal of molecular evolution*, **74**, 206-216.
42. Swire, J. (2007) Selection on synthesis cost affects interprotein amino acid usage in all three domains of life. *Journal of molecular evolution*, **64**, 558-571.
43. Heizer, E.M., Jr., Raiford, D.W., Raymer, M.L., Doom, T.E., Miller, R.V. and Krane, D.E. (2006) Amino acid cost and codon-usage biases in 6 prokaryotic genomes: a whole-genome analysis. *Molecular biology and evolution*, **23**, 1670-1680.
44. Smith, D.R. and Chapman, M.R. (2010) Economical evolution: microbes reduce the synthetic cost of extracellular proteins. *mBio*, **1**.
45. Chen, W.H., Lu, G., Bork, P., Hu, S. and Lercher, M.J. (2016) Energy efficiency trade-offs drive nucleotide usage in transcribed regions. *Nature communications*, **7**, 11334.
46. Rocha, E.P. (2008) The organization of the bacterial genome. *Annual review of genetics*, **42**, 211-233.
47. Ogawa, T. and Okazaki, T. (1980) Discontinuous DNA replication. *Annual review of biochemistry*, **49**, 421-457.
48. Gao, N., Lu, G., Lercher, M.J. and Chen, W.H. (2017) Selection for energy efficiency drives strand-biased gene distribution in prokaryotes. *Scientific reports*, **7**, 10572.
49. Mao, X., Zhang, H., Yin, Y. and Xu, Y. (2012) The percentage of bacterial genes on leading versus lagging strands is influenced by multiple balancing forces. *Nucleic acids research*, **40**, 8210-8218.
50. Brewer, B.J. (1988) When polymerases collide: replication and the transcriptional

organization of the E. coli chromosome. *Cell*, **53**, 679-686.

51. Hu, J., Zhao, X. and Yu, J. (2007) Replication-associated purine asymmetry may contribute to strand-biased gene distribution. *Genomics*, **90**, 186-194.
52. Price, M.N., Alm, E.J. and Arkin, A.P. (2005) Interruptions in gene expression drive highly expressed operons to the leading strand of DNA replication. *Nucleic acids research*, **33**, 3224-3234.
53. Mirkin, E.V. and Mirkin, S.M. (2005) Mechanisms of transcription-replication collisions in bacteria. *Molecular and cellular biology*, **25**, 888-895.
54. Liu, B. and Alberts, B.M. (1995) Head-on collision between a DNA replication apparatus and RNA polymerase transcription complex. *Science*, **267**, 1131-1137.
55. Sankar, T.S., Wastuwidyaningtyas, B.D., Dong, Y., Lewis, S.A. and Wang, J.D. (2016) The nature of mutations induced by replication-transcription collisions. *Nature*, **535**, 178-181.
56. Million-Weaver, S., Samadpour, A.N., Moreno-Habel, D.A., Nugent, P., Brittnacher, M.J., Weiss, E., Hayden, H.S., Miller, S.I., Liachko, I. and Merrikh, H. (2015) An underlying mechanism for the increased mutagenesis of lagging-strand genes in Bacillus subtilis. *Proceedings of the National Academy of Sciences of the United States of America*, **112**, E1096-1105.
57. Rocha, E.P. and Danchin, A. (2003) Essentiality, not expressiveness, drives gene-strand bias in bacteria. *Nature genetics*, **34**, 377-378.
58. Nyvltova, E., Stairs, C.W., Hrdy, I., Ridl, J., Mach, J., Paces, J., Roger, A.J. and Tachezy, J. (2015) Lateral gene transfer and gene duplication played a key role in the evolution of Mastigamoeba balamuthi hydrogenosomes. *Molecular biology and evolution*, **32**, 1039-1055.
59. Isambert, H. and Stein, R.R. (2009) On the need for widespread horizontal gene transfers under genome size constraint. *Biology direct*, **4**, 28.
60. Schonknecht, G., Chen, W.H., Ternes, C.M., Barbier, G.G., Shrestha, R.P., Stanke, M., Brautigam, A., Baker, B.J., Banfield, J.F., Garavito, R.M. *et al.* (2013) Gene transfer from bacteria and archaea facilitated evolution of an extremophilic eukaryote. *Science*, **339**, 1207-1210.
61. Smith, J.M., Smith, N.H., O'Rourke, M. and Spratt, B.G. (1993) How clonal are bacteria? *Proceedings of the National Academy of Sciences of the United States of America*, **90**, 4384-4388.
62. Feng, D.F., Cho, G. and Doolittle, R.F. (1997) Determining divergence times with a protein clock: update and reevaluation. *Proceedings of the National Academy of Sciences of the United States of America*, **94**, 13028-13033.
63. Jain, R., Rivera, M.C. and Lake, J.A. (1999) Horizontal gene transfer among genomes: the complexity hypothesis. *Proceedings of the National Academy of Sciences of the United States of America*, **96**, 3801-3806.
64. Malachowa, N. and DeLeo, F.R. (2010) Mobile genetic elements of Staphylococcus aureus. *Cellular and molecular life sciences : CMLS*, **67**, 3057-3071.
65. Yamaguchi, T., Hayashi, T., Takami, H., Ohnishi, M., Murata, T., Nakayama, K., Asakawa, K., Ohara, M., Komatsuzawa, H. and Sugai, M. (2001) Complete nucleotide sequence of a Staphylococcus aureus exfoliative toxin B plasmid and identification of a novel ADP-

ribosyltransferase, EDIN-C. *Infection and immunity*, **69**, 7760-7771.

66. Jensen, S.O. and Lyon, B.R. (2009) Genetics of antimicrobial resistance in Staphylococcus aureus. *Future microbiology*, **4**, 565-582.
67. Lindsay, J.A. (2010) Genomic variation and evolution of Staphylococcus aureus. *International journal of medical microbiology : IJMM*, **300**, 98-103.
68. Dahl, K.H., Mater, D.D., Flores, M.J., Johnsen, P.J., Midtvedt, T., Corthier, G. and Sundsfjord, A. (2007) Transfer of plasmid and chromosomal glycopeptide resistance determinants occurs more readily in the digestive tract of mice than in vitro and exconjugants can persist stably in vivo in the absence of glycopeptide selection. *The Journal of antimicrobial chemotherapy*, **59**, 478-486.
69. Pang, T.Y. and Lercher, M.J. (2019) Each of 3,323 metabolic innovations in the evolution of E. coli arose through the horizontal transfer of a single DNA segment. *Proceedings of the National Academy of Sciences of the United States of America*, **116**, 187-192.
70. Hikosaka, A. and Konishi, S. (2018) Multiple massive domestication and recent amplification of Kolobok superfamily transposons in the clawed frog Xenopus. *Zoological letters*, **4**, 17.
71. Jassim, S.A. and Griffiths, M.W. (2007) Evaluation of a rapid microbial detection method via phage lytic amplification assay coupled with Live/Dead fluorochromic stains. *Letters in applied microbiology*, **44**, 673-678.
72. Deuschle, U., Pepperkok, R., Wang, F.B., Giordano, T.J., McAllister, W.T., Ansorge, W. and Bujard, H. (1989) Regulated expression of foreign genes in mammalian cells under the control of coliphage T3 RNA polymerase and lac repressor. *Proceedings of the National Academy of Sciences of the United States of America*, **86**, 5400-5404.
73. Modi, S.R., Lee, H.H., Spina, C.S. and Collins, J.J. (2013) Antibiotic treatment expands the resistance reservoir and ecological network of the phage metagenome. *Nature*, **499**, 219-222.
74. Makarova, K.S., Grishin, N.V., Shabalina, S.A., Wolf, Y.I. and Koonin, E.V. (2006) A putative RNA-interference-based immune system in prokaryotes: computational analysis of the predicted enzymatic machinery, functional analogies with eukaryotic RNAi, and hypothetical mechanisms of action. *Biology direct*, **1**, 7.
75. Gomez-Valero, L., Rusniok, C., Jarraud, S., Vacherie, B., Rouy, Z., Barbe, V., Medigue, C., Etienne, J. and Buchrieser, C. (2011) Extensive recombination events and horizontal gene transfer shaped the Legionella pneumophila genomes. *BMC genomics*, **12**, 536.
76. Sharma, V., Colson, P., Pontarotti, P. and Raoult, D. (2016) Mimivirus inaugurated in the 21st century the beginning of a reclassification of viruses. *Current opinion in microbiology*, **31**, 16-24.
77. Huang, Q., Luo, H., Liu, M., Zeng, J., Abdalla, A.E., Duan, X., Li, Q. and Xie, J. (2016) The effect of Mycobacterium tuberculosis CRISPR-associated Cas2 (Rv2816c) on stress response genes expression, morphology and macrophage survival of Mycobacterium smegmatis. *Infection, genetics and evolution : journal of molecular epidemiology and evolutionary genetics in infectious diseases*, **40**, 295-301.
78. Tyson, G.W. and Banfield, J.F. (2008) Rapidly evolving CRISPRs implicated in acquired resistance of microorganisms to viruses. *Environmental microbiology*, **10**, 200-207.
79. Denef, V.J., Mueller, R.S. and Banfield, J.F. (2010) AMD biofilms: using model

communities to study microbial evolution and ecological complexity in nature. *The ISME journal*, **4**, 599-610.

80. Held, N.L., Herrera, A., Cadillo-Quiroz, H. and Whitaker, R.J. (2010) CRISPR associated diversity within a population of Sulfolobus islandicus. *PloS one*, **5**.
81. Stern, A., Keren, L., Wurtzel, O., Amitai, G. and Sorek, R. (2010) Self-targeting by CRISPR: gene regulation or autoimmunity? *Trends in genetics : TIG*, **26**, 335-340.
82. Kuo, C.H. and Ochman, H. (2009) Deletional bias across the three domains of life. *Genome biology and evolution*, **1**, 145-152.
83. Weinberger, A.D., Sun, C.L., Plucinski, M.M., Denef, V.J., Thomas, B.C., Horvath, P., Barrangou, R., Gilmore, M.S., Getz, W.M. and Banfield, J.F. (2012) Persisting viral sequences shape microbial CRISPR-based immunity. *PLoS computational biology*, **8**, e1002475.
84. Jiang, W., Maniv, I., Arain, F., Wang, Y., Levin, B.R. and Marraffini, L.A. (2013) Dealing with the evolutionary downside of CRISPR immunity: bateria and beneficial plasmids. *PLoS genetics*, **9**, e1003844.
85. Gophna, U., Kristensen, D.M., Wolf, Y.I., Popa, O., Drevet, C. and Koonin, E.V. (2015) No evidence of inhibition of horizontal gene transfer by CRISPR-Cas on evolutionary timescales. *The ISME journal*, **9**, 2021-2027.
86. Watson, B.N.J., Staals, R.H.J. and Fineran, P.C. (2018) CRISPR-Cas-Mediated Phage Resistance Enhances Horizontal Gene Transfer by Transduction. *mBio*, **9**.
87. Baltrus, D.A. (2013) Exploring the costs of horizontal gene transfer. *Trends in ecology & evolution*, **28**, 489-495.
88. Frost, L.S., Leplae, R., Summers, A.O. and Toussaint, A. (2005) Mobile genetic elements: the agents of open source evolution. *Nature reviews. Microbiology*, **3**, 722-732.
89. Chen, W.H., van Noort, V., Lluch-Senar, M., Hennrich, M.L., Wodke, J.A., Yus, E., Alibes, A., Roma, G., Mende, D.R., Pesavento, C. *et al.* (2016) Integration of multi-omics data of a genome-reduced bacterium: Prevalence of post-transcriptional regulation and its correlation with protein abundances. *Nucleic acids research*, **44**, 1192-1202.
90. Yu, T., Guo, F., Yu, Y., Sun, T., Ma, D., Han, J., Qian, Y., Kryczek, I., Sun, D., Nagarsheth, N. *et al.* (2017) Fusobacterium nucleatum Promotes Chemoresistance to Colorectal Cancer by Modulating Autophagy. *Cell*, **170**, 548-563 e516.
91. Forslund, K., Hildebrand, F., Nielsen, T., Falony, G., Le Chatelier, E., Sunagawa, S., Prifti, E., Vieira-Silva, S., Gudmundsdottir, V., Pedersen, H.K. *et al.* (2015) Disentangling type 2 diabetes and metformin treatment signatures in the human gut microbiota. *Nature*, **528**, 262-266.
92. Forsgren, M., Isolauri, E., Salminen, S. and Rautava, S. (2017) Late preterm birth has direct and indirect effects on infant gut microbiota development during the first six months of life. *Acta paediatrica*, **106**, 1103-1109.
93. Komaroff, A.L. (2017) The Microbiome and Risk for Obesity and Diabetes. *Jama*, **317**, 355-356.
94. Yen, M., Cairns, L.S. and Camilli, A. (2017) A cocktail of three virulent bacteriophages prevents Vibrio cholerae infection in animal models. *Nature communications*, **8**, 14187.
95. Krupovic, M., Prangishvili, D., Hendrix, R.W. and Bamford, D.H. (2011) Genomics of bacterial and archaeal viruses: dynamics within the prokaryotic virosphere.

Microbiology and molecular biology reviews : MMBR, **75**, 610-635.

96. Fortier, L.C. and Sekulovic, O. (2013) Importance of prophages to evolution and virulence of bacterial pathogens. *Virulence*, **4**, 354-365.
97. Fouts, D.E. (2006) Phage_Finder: automated identification and classification of prophage regions in complete bacterial genome sequences. *Nucleic acids research*, **34**, 5839-5851.
98. Arndt, D., Grant, J.R., Marcu, A., Sajed, T., Pon, A., Liang, Y. and Wishart, D.S. (2016) PHASTER: a better, faster version of the PHAST phage search tool. *Nucleic acids research*, **44**, W16-21.
99. Roux, S., Enault, F., Hurwitz, B.L. and Sullivan, M.B. (2015) VirSorter: mining viral signal from microbial genomic data. *PeerJ*, **3**, e985.

3 Manuscripts

3.1 Manuscript 1. Selection for energy efficiency drives strand-biased gene distribution in prokaryotes

Na L Gao [1, 3 *], Guanting Lu [2, *], Martin J Lercher [3], Wei-Hua Chen [1, §]

[1] Key Laboratory of Molecular Biophysics of the Ministry of Education, Hubei Key Laboratory of Bioinformatics and Molecular-imaging, Department of Bioinformatics and Systems Biology, College of Life Science and Technology, Huazhong University of Science and Technology (HUST), 430074 Wuhan, Hubei, China

[2] Department of Blood Transfusion, Tangdu Hospital, the Fourth Military Medical University, No 1, Xinsi Road, Chanba District, 710000 Xi'an, China

[3] Institute for Computer Science and Cluster of Excellence on Plant Sciences CEPLAS, Heinrich Heine University, 40225 Düsseldorf, Germany

[§]Correspondence should be addressed to Wei-Hua Chen. Tel: +862787542127; Fax: +862787542527; Email: weihuachen@hust.edu.cn

* These authors contribute equally to this work

This paper has been published in the journal *Scientific Reports* (Published online 2017 Sep 5. doi: [10.1038/s41598-017-11159-3]; PMID: 28874819). As the first author, I conceived the study together with Prof. Chen; I collected the data from public databases with help from Prof. Chen and other members of the lab; I performed the data analyses, interpreted the results together with Prof. Chen and Prof. Lercher, wrote the first draft of the manuscript, and edited the final manuscript together with Prof. Chen and Prof. Lercher.

www.nature.com/scientificreports

OPEN

Selection for energy efficiency drives strand-biased gene distribution in prokaryotes

Received: 5 June 2017
Accepted: 18 August 2017
Published online: 05 September 2017

Na Gao[1,3], Guanting Lu[2], Martin J. Lercher[3] & Wei-Hua Chen[1]

Lagging-strand genes accumulate more deleterious mutations. Genes are thus preferably located on the leading strand, an observation known as strand-biased gene distribution (SGD). Despite of this mechanistic understanding, a satisfactory quantitative model is still lacking. Replication-transcription-collisions induce stalling of the replication machinery, expose DNA to various attacks, and are followed by error-prone repairs. We found that mutational biases in non-transcribed regions can explain ~71% of the variations in SGDs in 1,552 genomes, supporting the mutagenesis origin of SGD. Mutational biases introduce energetically cheaper nucleotides on the lagging strand, and result in more expensive protein products; consistently, the cost difference between the two strands explains ~50% of the variance in SGDs. Protein costs decrease with increasing gene expression. At similar expression levels, protein products of leading-strand genes are generally cheaper than lagging-strand genes; however, highly-expressed lagging genes are still cheaper than lowly-expressed leading genes. Selection for energy efficiency thus drives some genes to the leading strand, especially those highly expressed and essential, but certainly not all genes. Stronger mutational biases are often associated with low-GC genomes; as low-GC genes encode expensive proteins, low-GC genomes thus tend to have stronger SGDs to alleviate the stronger pressure on efficient energy usage.

In most prokaryotic genomes, protein-coding genes are preferably located on the leading strand[1], on which the replication is continuous[2]. For example, in contrast to randomly expected 50% if there were no strand preferences, over 90% of the 1,552 bacterial and archaeal genomes we surveyed in this study show preferred location of their coding genes on the leading strand (see also[3]). This phenomenon, which is known as biased-strand gene distribution (SGD), has been intensively investigated in the past decades and many hypotheses have been proposed[4–15].

It has long been suspected that SGDs are caused by collisions between the replication and transcription machineries[1,4,9,11,14–17]. The latter two share the same DNA template but move with different speed[6]; in addition, they move in different directions on the lagging strand of the genome. Thus, collision can happen either co-directionally (on leading strand) or head-on (on lagging strand)[16]. Collisions can cause replication stalling, abortive transcription, and expose single-stranded DNAs to chemical modifications and other damages[18]. Collisions are thus deleterious. Recent experimental results suggest that genes on the lagging strand accumulate more mutations than those on the leading strand[19], due to head-on collisions or the discontinuous nature of the DNA synthesis of the lagging strand, or both. This indicates that head-on collisions are more deleterious than co-directional collisions. The elevated deleterious effects on the lagging strand are believed to cause a higher burden on fitness for highly expressed genes and functionally important genes (*e.g.*, essential genes), consistent with the observations that these two types of genes are underrepresented on the lagging strand[9,12].

Despite the mechanistic insights, a quantitative model that explains the variation of SGDs in different species is still lacking. For example, the expression-driven[9] and essentiality-driven[12] hypotheses are not quantitative; more importantly, after highly expressed and essential genes were removed, SGDs were decreased but not completely removed (see Figs 1 and 2). In addition, it is difficult to quantify their contributions to SGD: it is unclear

[1]Key Laboratory of Molecular Biophysics of the Ministry of Education, Hubei Key Laboratory of Bioinformatics and Molecular-imaging, Department of Bioinformatics and Systems Biology, College of Life Science and Technology, Huazhong University of Science and Technology (HUST), 430074, Wuhan, Hubei, China. [2]Department of Blood Transfusion, Tangdu Hospital, the Fourth Military Medical University, No 1, Xinsi Road, Chanba District, 710000, Xi'an, China. [3]Institute for Computer Science and Cluster of Excellence on Plant Sciences CEPLAS, Heinrich Heine University, 40225, Düsseldorf, Germany. Na Gao and Guanting Lu contributed equally to this work. Correspondence and requests for materials should be addressed to W.-H.C. (email: weihuachen@hust.edu.cn)

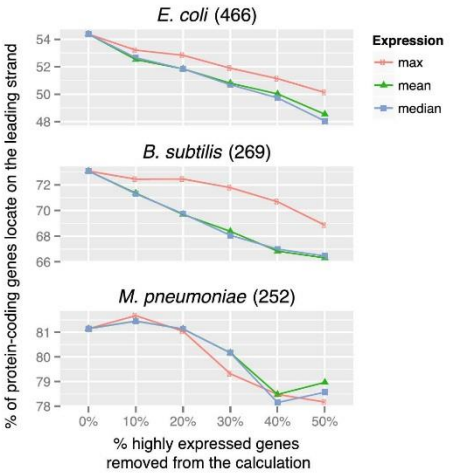


Figure 1. Removing highly expressed genes does not eliminate strand-biased gene distribution in selected species. Gene expression data were downloaded from NCBI GEO database[36] for the three model bacteria, *Escherichia coli*[37], *Bacillus subtilis*[38] and *Mycoplasma pneumoniae*[34]; the number of datasets for each species is indicated in the parenthesis of the panel title. For each gene in a genome, we calculated the max, mean and median expression values across the expression datasets we collected, and then ranked all genes in a genome accordingly.

why SGDs are different in different genomes, and how much of the variations can be explained by essential or highly expressed genes. Recently, Mao *et al.*[3] proposed a very sophisticated model; using data on the enrichment and depletion of genes in 25 Gene Ontology (GO) categories on the leading strand, they were able to explain ~74% of the variance of SGDs across 725 prokaryotic genomes; the authors argue that genes of certain functions prefer different strands and consequently drive SGD. Although it represents arguably one of the best quantitative models so far, ref. 3 blurs the cause and consequences of this issue. For example, one may argue that it is the head-on collisions between replication and transcription machineries that drive the highly-expressed and essential genes to the leading strand, and consequently cause the biased functional categories in the genes on the leading strand, rather than the other way round.

Here, we propose a mutagenesis/energy efficiency model for SGDs and test it on 1,552 prokaryotic genomes. In previous work, we showed that strand-specific mutational biases, observed as nucleotide compositional biases in inter-operonic regions, can be recapitulated using coding sequences from leading and lagging strands[20]. These results suggested that mutational biases in coding regions are of similar nature to that in non-transcribed regions but are inflated, likely due to the longer exposure time of single-stranded DNA during transcription[20], which causes increased DNA damage and error-prone repair. Mutational biases introduce the energetically cheaper nucleotides *T* and *C* over their complementary nucleotides *A* and *G*, respectively, as well as *C* over *G* on the lagging strand. Due to a trade-off between nucleotide and amino acid costs inherent in the codon translation table, the bias towards cheaper nucleotides results in more expensive protein products for genes on the lagging strand, driving genes to the leading strand.

Our model – which we develop in quantitative form below – makes the following predictions. First, strand-specific mutational biases observed in interoperonic regions should be able to predict the extent of SGD in a given genome: stronger mutational biases should lead to stronger SGD. Second, previous studies have shown that costs per protein decrease with increasing gene expression[20–24]; therefore, highly expressed genes on the lagging strand should still be cheaper than lowly expressed genes on the leading strand. We thus expect selection for energy efficiency to drive some genes to the leading strand, especially those highly expressed and essential, but not all genes.

Results and Discussion

Removing highly expressed or essential genes does not eliminate SGD. Avoidance of head-on collisions between replication and transcription machineries could drive (some) highly-expressed and/or essential genes to the leading strand. However, we hypothesized that other factors such as mutagenesis could also contribute significantly to SGDs. We thus removed highly expressed or essential genes from selected species and recalculated SGDs. As expected, SGDs remain in most species, especially in genomes with strong SGDs to begin with, suggesting that highly expressed or essential genes could only explain a small part of SGD (Figs 1 and 2).

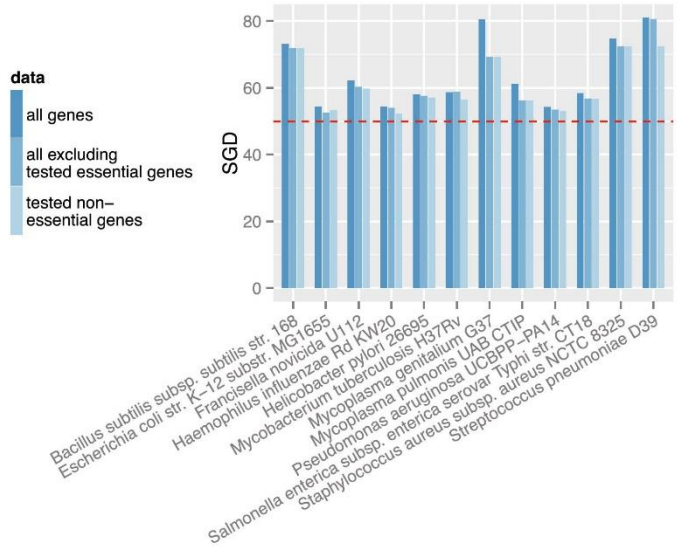


Figure 2. Removing essential genes does not eliminate strand-biased gene distribution in selected species. Tested essential and nonessential genes were obtained from OGEE - an online gene essentiality database[25]. "all genes" (dark blue bar): when all genes were used to calculate the SGD; "all excluding tested essential genes" (blue bar): when genes that were tested as nonessential genes and those were not tested in gene essentiality experiments were used; "tested non-essential genes" (light blue bar): when only genes that were tested as nonessential were used.

Gene expression abundances vary between different experimental conditions. We thus also tested whether the same trend could be observed in individual gene expression experiments. From each of the expression datasets we collected for the selected organisms, we ranked genes according to their expression abundance, removed the highly-expressed ones and recalculated the SGD. Figure 3 summarized the results as boxplots; as expected, we observed the same trend that SGDs decrease but remain after removing highly expressed genes.

Gene essentiality statuses can also be environment-/experiment-dependent. We thus further tested our hypothesis in species whose essential genes had been tested under different experimental conditions. As shown in Supplementary Figure 1, in all four bacteria (namely *Salmonella enterica subsp. enterica serovar Typhimurium str. SL1344*, *Pseudomonas aeruginosa UCBPP-PA14*, *Escherichia coli K12* and *Mycobacterium tuberculosis H37Rv*) for which multiple essentiality datasets are available in OGEE v2[25], removing essential genes did not eliminate SGD.

Gene essentiality can also be measured quantitatively (*e.g.*, as Fitness scores) instead of qualitatively; it has been previously shown that quantitatively measured gene essentiality contributes significantly to SGD in bacterial species[26]. To further test the robustness of hypotheses on this type of data, we obtained predicted "fitness scores" for 2,074 species from IFIM, a database of Integrated Fitness Information for Microbial genes[27]. Fitness scores in IFIM were predicted using Geptop[28] based on orthology and phylogeny; the scores range from 0 to 1, with lower scores representing greater fitness decreases and thus higher likelihood of being essential. A cutoff of 0.65 was recommended to classify genes into essential (those with fitness scores <= 0.65) and non-essential[27,28]. In total, 1,410 genomes overlapped with the 1,552 genomes used in this study. As shown in Fig. 4, when all genes were included, ~94.18% of the 1,410 genomes had SGDs larger than 50; excluding genes with lower fitness scores could reduce this percentage, but only to a very limited extend. For example, after excluding genes with fitness scores less than 0.7 from all genomes and re-calculating SGD, 92.62% of the genomes still had SGDs larger than 50.

Together, these results further confirmed that highly expressed or essential genes could only explain part of SGD in prokaryotes.

Replication skews can explain ~71% of the variance in SGDs in 1,552 prokaryotic genomes. Our previous results showed that mutational biases, i.e. strand-specific usage of *A* versus *T*, and of *G* versus *C* (also known as AT and GC skews respectively; see Methods) observed in interoperonic regions can be recapitulated using coding sequences from leading and lagging strands, with a certain inflation[20]. For example, mutational skews estimated by contrasting genes on the leading strand and on the lagging strand correlate significantly with the interoperonic skews, with correlation coefficients of 0.78 and 0.90 for AT and GC skews,

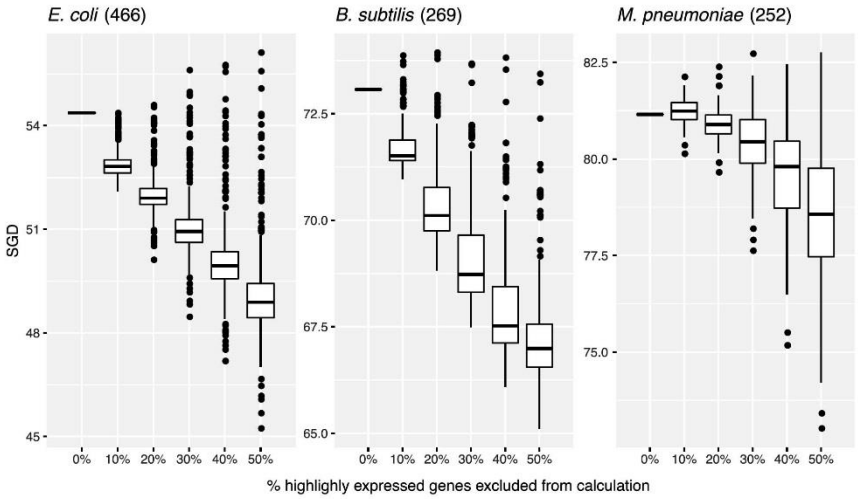


Figure 3. SGDs decrease but remain after removing highly expressed genes in selected species. The same data from Fig. 1 were also used here. For each expression dataset, we ranked genes according to their expression abundances, removed the highly-expressed ones and recalculated the SGD. We summarized the results as boxplots.

respectively. Interoperonic regions are either non-transcribed or only casually transcribed[29], and their skews are thus predominantly due to mutational biases and not to natural selection (see also ref. 20). These results indicate that mutational biases in coding regions are of a similar nature as those in non-transcribed regions; the inflation was likely due to the prolonged exposure time of single-stranded DNA during transcription and replication-transcription-collisions[20], followed by increased DNA damage and error-prone repair.

It has long been suspected that there is a connection between SGDs and the mutational biases[4,30]. For example, Hu and colleagues found that the nucleotide skews at fourfold-synonymous (4 s) sites of the coding regions and in intergenic regions correlate significantly with SGD (Pearson's correlation coefficients $R > 0.7$ in both cases)[4]. One problem with this calculation is the inclusion of transcribed regions. It is known that the overall nucleotide skews of the transcribed regions consists of at least two parts, one part is attributed to replication (i.e. mutational biases), while the other is attributed to transcription[20]. The replication skews in transcribed regions are proportional to that in interoperonic regions but slightly inflated, with the inflation rate being proportional to expression abundance[20]. Genes on the leading strand are often more abundantly expressed; the stronger the SGDs, the stronger the differences in expression abundances between strands, and the stronger the differences in nucleotide skews. Therefore, the inclusion of coding/transcribed regions in Hu's calculation will inflate the correlation by partially correlating SGD with its consequences (Methods).

By using a simple nonlinear regression model (Multivariate adaptive regression splines, MARS; Methods) on the interdependence of SGD and mutational bias (Fig. 5), we estimated that ~71% of the variation in SGDs in 1,552 prokaryotic genomes can be explained by the nucleotide skews from interoperonic regions that are presumably only subjected to replication (we hence refer them as replication skews; see also the discussions below) (Fig. 5). Our model has similar predictive power as the model proposed by Mao and colleagues (Pearson's R^2 71% versus 74%) but uses much fewer variables as input (2 versus 28)[3]; more importantly, SGD and replication skews in our model were derived from non-overlapping datasets. Our model thus clearly indicates that SGD and replication skews may have a common origin, *i.e.*, the factors that drive replication skews also drive SGD; the stronger the replication skews, the stronger the SGD (Fig. 5). Consistent with our expectations, the inclusion of coding / transcribed regions into the calculation indeed inflated the correlation: we estimate that over ~78% of variations in SGDs could be explained by the overall nucleotide skews (Supplementary Figure 3).

Mutational biases cause the use of slightly more expensive amino-acids in genes on the lagging strand. The synthesis of the four nucleotides *A, C, G, T* requires different amounts of energy: *de-novo* production costs are $A > T$, $G > C$, and $G + C > A + T$[20]. Replication skews are strand-specific; the leading strand is biased towards the more expensive nucleotide *G* over *C* in almost all prokaryotic genomes (93.9%), while on the lagging strand the opposite is found. Although only a small proportion of prokaryotes (36.1%) preferentially use the more expensive nucleotide *A* over *T*, a majority (87.6%) of the collected genomes prefer the use of the

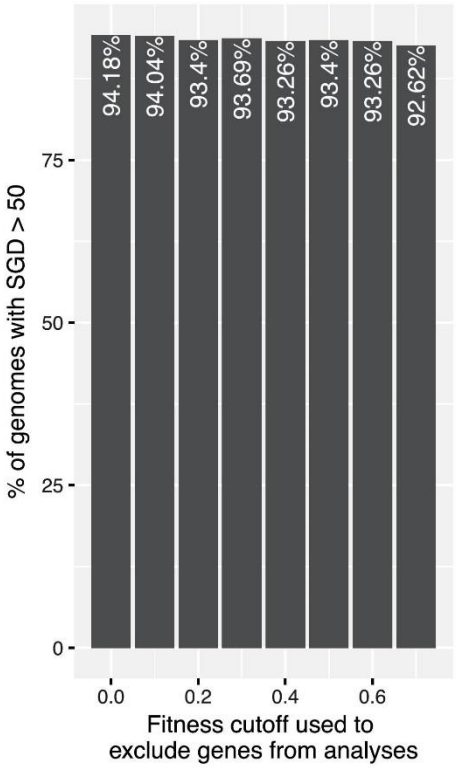


Figure 4. Excluding essential genes does not eliminate SGDs using quantitative measurements of gene essentiality (Fitness scores) obtained from IFIM, a database of Integrated Fitness Information for Microbial genes[27]. Genes with lower fitness scores more likely to be essential.

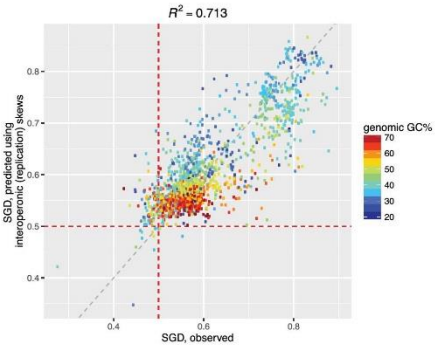


Figure 5. Predicted SGDs (y-axis) in 1,552 bacterial genomes using interoperonic skews and their correlation with the observed SGDs (x-axis). Each dot represents a genome, color-coded by genomic GC-content.

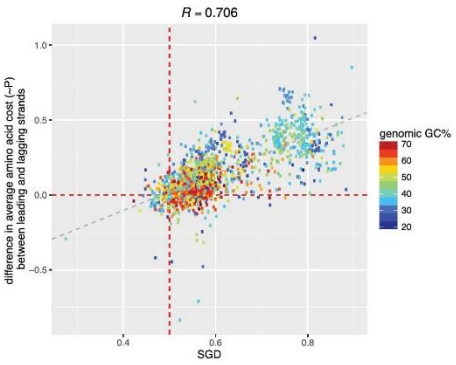


Figure 6. correlation between strand-biased gene distribution (SGD; x-axis) and the difference of average costs per amino acid of gene products encoded by genes on the lagging and leading strand.

more expensive purines (*G* and *A*) over pyrimidines (*T* and *C*) on the lagging strand in interoperonic regions (Supplementary Table 1).

Replication skews also exist in coding regions, where they are inflated as a function of expression abundance[20]. Due to an intrinsic tradeoff in the codon table, more expensive nucleotides code for cheaper amino acids and *vice versa*[20]; we thus expect that the replication skews would cause slightly cheaper protein products on the leading strand. This is indeed the case: we found that 91% of the genomes with positive purine skews (that is, purines are preferred over pyrimidines) encode cheaper protein products on their leading strand; interestingly, 62.5% of genomes with negative skews (that is, pyrimidines are preferred over purines) also encode cheaper protein products on their leading strand, indicating that additional factors such as GC-content also contribute to these observations. In addition, we found that the protein cost differences between lagging and leading strands (*i.e.*, average cost per amino acid of the lagging strand minus that of the leading strand) correlate significantly with replication skews (Pearson's $R = 0.56$, 0.47 and 0.61 for AT, GC, and the overall Purine-skews, respectively; see Methods) as well as with SGD ($R = 0.701$, Fig. 6).

Mutations are also known to be biased towards *AT* in bacteria[31]. Recent experimental results suggested that due to head-on collisions, lagging-strand genes tend to accumulate more mutations than leading-strand genes[19] and thus have lower GC-contents and code for more expensive proteins than leading-strand genes. A nonlinear regression analysis using MARS revealed that both the replication skews and the overall differences in GC-content between leading and lagging strand genes contribute significantly to the amino acid differences, with the replication skews as the most important factor, followed by GC-differences. Similarly, a linear regression model implemented in the R package 'relaimpo' reported that the replication skews contributed twice as much as the GC-differences (Methods). These results suggest that the protein cost difference between the two strands can be mostly attributed to replication skews.

Selection for energy efficiency drives some, but not all highly expressed genes to the leading strand. As shown in Fig. 7, when expression abundances (proxied by tAI, tRNA adaptation index[32,33]) are similar, protein products are always slightly more expensive on the lagging strand; however, as the per protein costs decrease with increasing expression abundance due to increasing skews[20] and GC-contents (see also Supplementary Figure 4), the protein products of lowly expressed leading strand genes could be more expensive than those of highly expressed lagging strand genes. These results have two important implications. First, for the purpose of energy efficiency, there is a tendency for highly expressed genes, especially those that are also universally expressed, to move to the leading strand through the fixation of local chromosomal inversions. This would explain why genes such as those involved in transcription, translation, and replication are preferably located on the leading strand; this would also increase the ratio of essential genes on the leading strand because these genes are more likely to be essential. Second, there is no need to move all genes to the leading strand. In fact, it might be beneficial to distribute genes onto different strands, *e.g.*, to avoid possible "transcriptional leakage" if transcription termination fails accidentally. This is consistent with a previous observation that more "unbalanced genomes", *i.e.*, those with strong SGDs, tend to have longer intergenic regions[3] in order to give more space or harbor necessary *cis*-regulatory elements and sequence signatures for the transcription machinery to terminate properly.

Relationships between mutational bias, GC-content, and genome size. Interestingly, we found that the genomic GC-content correlates significantly with both AT and GC replication skews ($R = -0.32$ and -0.54 for AT and GC skews, respectively, $P < 2.2 \times 10^{-16}$; AT and GC skews are also significantly correlated with each other, consistent with recent studies[30]). Because $G + C$ are more expensive than $A + T$ and encode cheaper amino acids, high-GC genomes spend proportionally more energy on nucleotide production than low-GC genomes, while the latter spend relatively more energy on the production of amino acids; in other words, genomic

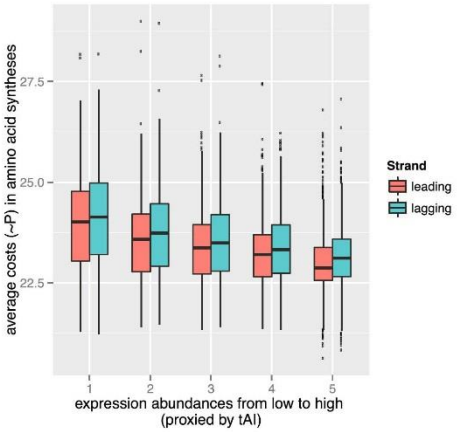


Figure 7. average costs in amino acid synthesis as a function of leading/lagging strand and expression abundance. Genes in each genome were ranked according to their expression abundance (proxied by tAI, tRNA adaption index) from low to high, divided into five equal-sized bins (so that each bin contains roughly the same number of genes) and then divided into two sub-groups according to their strand (leading versus lagging).

GC-content is an indicator of relative energy investment into nucleotides and amino acids[20]. GC content also correlates with genome size[20,34]. As amino acids are relatively more expensive than nucleotides (Supplementary Table 2, see also ref. 20), the selection for energy efficiency is stronger in low-GC genomes. The negative correlation between the replication skews and genomic GC indicates that stronger (more positive) replication skews are preferentially found in low-GC genomes and could result in cheaper encoded amino acids, thus partially alleviating the strong selection pressure due to low GC. These results suggest that replication skews are also influenced by selection for energy efficiency.

Intracellular pathogens and symbionts spend their entire life cycle inside the cells of other organisms that are often much larger in size; in other words, they live in extremely nutrient-rich environments and thus experience weaker selection on efficient resource usage[20]. Excluding 126 previously identified intracellular pathogens and symbionts (Table S2) from our analyses improved the correlation between genome-GC and replication skews ($R = -0.35$ and -0.57 for AT and GC skews respectively). These results further supported our conclusion that selection for energy efficiency constrain replication skews.

Relationship between our model and existing theories. Our model is compatible with many existing hypotheses. For example, similar to the head-on collision model, our model predicts that highly-expressed and essential genes are to be over-represented on the leading strand, consistent with previous observations[9,12]. However, although the head-on collision model is not quantitative, it also predicts that important non-coding genes such as tRNA and rRNA genes should be preferably located on the leading strand. In addition, the head-on collisions alone could drive genes to the leading strand, by either causing abortive transcription of genes that should be stably expressed at all times (*e.g.*, ribosomal genes), or introducing more deleterious mutations into the regulatory regions of genes, or both. Our model does not explicitly cover these situations.

A recent study by Paul *et al.* proposed that some lagging-strand genes take advantage of the increased mutagenesis resulting from the head-on collisions and are thus adaptively encoded on the lagging strand[17]. This model is the opposite to our model, and has been recently rebutted by Chen and Zhang[35]. Chen and Zhang reanalyzed the data in ref. 17 and found no evidence for adaptive evolution of the lagging-strand genes; instead, they argue that SGD can be explained by a mutation-selection balance model, where deleterious chromosomal inversions move genes from the leading to the lagging strand and purifying selection purges such mutants[35], a view compatible with our model.

In this study, we proposed an energy efficiency theory for strand-biased gene distributions (SGD) and tested it on prokaryotic genomes. We showed that due to elevated mutational biases on the lagging strand, proteins encoded by lagging-strand genes are slightly more expensive than those encoded by leading-strand genes. Consequently, genes, especially those that are highly expressed, are preferentially located on the leading strand. Highly expressed genes code for cheaper products, even when they are located on the lagging strand; thus not all highly expressed genes, and certainly not all genes would be moved to the leading strand. Our model is compatible with many existing hypotheses and can explain more than two-third (~71%) of the variance in SGDs.

Methods

Gene expression data were downloaded from NCBI GEO database[36] for the three model bacteria *Escherichia coli*[37], *Bacillus subtilis*[38] and *Mycoplasma pneumoniae*[34]. Gene essentiality data for selected model organisms were downloaded from OGEE – an online gene essentiality database[25].

Genome sequences and annotation for all completely sequenced prokaryotes were downloaded from NCBI Genbank[39]. Genomic coordinates for replication starts were downloaded from DoriC[40]; replication ends were obtained by adding ½ genome lengths to the starts. This working definition of replication termination was inferred from the work of Hendrickson and Lawrence[41], in which the authors found that replication in *E. coli* is more likely to terminate near the ½ genome length to the oriC site, instead of the multiple *Ter* sites in the genome (Fig. 1 of ref. 41). 1,552 genomes covered by all three databases were used in this study (Table S1). The division of a genome into leading and lagging strands is shown in Supplementary Figure 2. Coding genes located on the first half of the plus strand (blue solid line) and on the second half of the complementary strand (purple solid line) were assigned to the leading strand, as their transcription proceeds in the same direction as the replication fork; the remaining genes were assigned to the lagging strand.

Operon predictions were downloaded from DOOR[42]. Because the predictions only cover coding regions, we added other annotated regions including tRNAs and rRNAs from the GFF (General Feature Format) annotations downloaded from NCBI, so that we could extract interoperonic regions, which are presumably non-transcribed. To extract regions that are presumably only subject to replication, interoperonic sequences longer than 100 base-pairs were retained after removing 60 bp from the regions adjacent to the 5′-end of genes/operons. If an inter-operonic region was located in the second half of the genome (blue dashed line in Supplementary Figure 2), its sequence was reverse-complemented. Replication skews are denoted as γ_{AT} (for AT skew) and γ_{GC} (for GC skew) and were calculated using extracted interoperonic regions using the equations below:

$$\gamma_{AT} = \frac{A - T}{A + T} \tag{1}$$

and

$$\gamma_{GC} = \frac{G - C}{G + C} \tag{2}$$

where A, T, G, C are the numbers of the corresponding bases. The overall purine skews were also calculated similarly using the equation below:

$$\gamma_{purine} = \frac{A - T + G - C}{A + T + G + C} \tag{3}$$

The costs of *de novo* amino acid synthesis were obtained from[21] (Table S2). The costs of *de novo* nucleotide synthesis were obtained from[20] and are 21.12, 13.42, 20.37, 15.77 ATPs for *A*, *T/U*, *G*, *C* respectively; please note these numbers were calculated for *E. coli* and might be different for other organisms.

tAI (tRNA adaptation index)[32,33] was used as a proxy for gene expression level. For each protein-coding gene in a given genome, tAI is defined as the average of tRNA availability values over all its codons. The availability of tRNAs for a codon considers not only the copy number of perfectly matched anticodons in the corresponding genome, but also that of imperfectly matched anticodons; the contribution of the imperfectly matched anticodons will be weighted accordingly. For more details on the definition of tAI see refs 32, 33. For each of the selected 1,552 genomes, we obtained a list of tRNA genes using the tRNAscan-SE[43] program on the genome sequences. The tRNA genes were sorted into 61 groups according to their anticodons. We then used the R scripts for tAI calculation written by the authors of refs 32, 33 (obtained from http://people.cryst.bbk.ac.uk/~fdosr01/tAI/, without modifications) to calculate tAI scores for all protein-coding genes in this genome. Higher tAI scores indicate higher expression levels.

Within each genome, coding genes were ranked according to their tAI scores from low to high and then divided into five equal-sized bins (quantiles), denoted 1 to 5; 1 contains the genes with the lowest, and 5 contains the genes with the highest tAI scores. Genes in each bin were then further divided into two groups according to the strands (leading versus lagging) they are located on.

Fitness scores (i.e. quantitative measurements of gene essentiality) for 2,074 prokaryotic genomes were downloaded from IFIM, a database of Integrated Fitness Information for Microbial genes[27]. Fitness scores in IFIM were predicted using Geptop[28] based on orthology and phylogeny; the scores range from 0 to 1, with lower scores representing greater fitness decreases and thus the corresponding genes are highly likely to be essential. A cutoff of 0.65 was recommended to classify genes into essential (those with fitness scores $<= 0.65$) and non-essential[27,28]. In total, 1,410 genomes overlapped with the 1,552 genomes used in this study.

All data was analyzed in R[44]. Non-linear regression analyses were carried out using the MARS (multivariate adaptive regression splines) function implemented in the 'earth' package of R (available at: https://cran.r-project.org/web/packages/earth/index.html); linear modeling was done with the 'relaimpo' package[45]. All plots were generated using the ggplot2[46] package.

References

1. Rocha, E. P. The organization of the bacterial genome. *Annual review of genetics* **42**, 211–233, doi:10.1146/annurev.genet.42.110807.091653 (2008).
2. Ogawa, T. & Okazaki, T. Discontinuous DNA replication. *Annual review of biochemistry* **49**, 421–457, doi:10.1146/annurev.bi.49.070180.002225 (1980).

3. Mao, X., Zhang, H., Yin, Y. & Xu, Y. The percentage of bacterial genes on leading versus lagging strands is influenced by multiple balancing forces. *Nucleic Acids Res* **40**, 8210–8218, doi:10.1093/nar/gks605 (2012).
4. Hu, J., Zhao, X. & Yu, J. Replication-associated purine asymmetry may contribute to strand-biased gene distribution. *Genomics* **90**, 186–194, doi:10.1016/j.ygeno.2007.04.002 (2007).
5. Omont, N. & Képès, F. Transcription/replication collisions cause bacterial transcription units to be longer on the leading strand of replication. *Bioinformatics* **20**, 2719–2725, doi:10.1093/bioinformatics/bth317 (2004).
6. Mirkin, E. V. & Mirkin, S. M. Mechanisms of transcription-replication collisions in bacteria. *Mol Cell Biol* **25**, 888–895, doi:10.1128/MCB.25.3.888-895.2005 (2005).
7. Wu, H. *et al.* Strand-biased Gene Distribution in Bacteria Is Related to both Horizontal Gene Transfer and Strand-biased Nucleotide Composition. *Genomics, Proteomics & Bioinformatics* **10**, 186–196, doi:10.1016/j.gpb.2012.08.001 (2012).
8. Wang, J. D., Berkmen, M. B. & Grossman, A. D. Genome-wide coorientation of replication and transcription reduces adverse effects on replication in Bacillus subtilis. *Proc Natl Acad Sci USA* **104**, 5608–5613, doi:10.1073/pnas.0608999104 (2007).
9. Brewer, B. J. When polymerases collide: Replication and the transcriptional organization of the *E. coli* chromosome. *Cell* **53**, 679–686, doi:10.1016/0092-8674(88)90086-4 (1988).
10. McLean, M. J., Wolfe, K. H. & Devine, K. M. Base composition skews, replication orientation, and gene orientation in 12 prokaryote genomes. *J Mol Evol* **47**, 691–696 (1998).
11. Price, M. N., Alm, E. J. & Arkin, A. P. Interruptions in gene expression drive highly expressed operons to the leading strand of DNA replication. *Nucleic Acids Res* **33**, 3224–3234, doi:10.1093/nar/gki638 (2005).
12. Rocha, E. P. C. & Danchin, A. Essentiality, not expressiveness, drives gene-strand bias in bacteria. *Nat Genet* **34**, 377–378, doi:http://www.nature.com/ng/journal/v34/n4/suppinfo/ng1209_S1.html (2003).
13. Rocha, E. P. & Danchin, A. Gene essentiality determines chromosome organisation in bacteria. *Nucleic Acids Res* **31**, 6570–6577 (2003).
14. Rocha, E. P. C. Is there a role for replication fork asymmetry in the distribution of genes in bacterial genomes? *Trends in Microbiology* **10**, 393–395, doi:10.1016/S0966-842X(02)02420-4 (2002).
15. de Carvalho, M. O. & Ferreira, H. B. Quantitative determination of gene strand bias in prokaryotic genomes. *Genomics* **90**, 733–740, doi:10.1016/j.ygeno.2007.07.010 (2007).
16. Bin, L. & Alberts, B. M. Head-on collision between a DNA replication apparatus and RNA polymerase transcription complex. *Science* **267**, 1131–1137 (1995).
17. Paul, S., Million-Weaver, S., Chattopadhyay, S., Sokurenko, E. & Merrikh, H. Accelerated gene evolution through replication-transcription conflicts. *Nature* **495**, 512–515, doi:10.1038/nature11989 (2013).
18. Sankar, T. S., Wastuwidyaningtyas, B. D., Dong, Y., Lewis, S. A. & Wang, J. D. The nature of mutations induced by replication-transcription collisions. *Nature* **535**, 178–181, doi:10.1038/nature18316 (2016).
19. Million-Weaver, S. *et al.* An underlying mechanism for the increased mutagenesis of lagging-strand genes in Bacillus subtilis. *Proc Natl Acad Sci USA* **112**, E1096–1105, doi:10.1073/pnas.1416651112 (2015).
20. Chen, W. H., Lu, G., Bork, P., Hu, S. & Lercher, M. J. Energy efficiency trade-offs drive nucleotide usage in transcribed regions. *Nat Commun* **7**, 11334, doi:10.1038/ncomms11334 (2016).
21. Akashi, H. & Gojobori, T. Metabolic efficiency and amino acid composition in the proteomes of Escherichia coli and Bacillus subtilis. *Proc Natl Acad Sci USA* **99**, 3695–3700, doi:10.1073/pnas.062526999 (2002).
22. Raiford, D. W. *et al.* Metabolic and translational efficiency in microbial organisms. *J Mol Evol* **74**, 206–216, doi:10.1007/s00239-012-9500-9 (2012).
23. Swire, J. Selection on synthesis cost affects interprotein amino acid usage in all three domains of life. *J Mol Evol* **64**, 558–571, doi:10.1007/s00239-006-0206-8 (2007).
24. Heizer, E. M. Jr. *et al.* Amino acid cost and codon-usage biases in 6 prokaryotic genomes: a whole-genome analysis. *Mol Biol Evol* **23**, 1670–1680, doi:10.1093/molbev/msl029 (2006).
25. Chen, W. H., Lu, G., Chen, X., Zhao, X. M. & Bork, P. OGEE v2: an update of the online gene essentiality database with special focus on differentially essential genes in human cancer cell lines. *Nucleic Acids Res* **45**, D940–D944, doi:10.1093/nar/gkw1013 (2017).
26. Zheng, W. X., Luo, C. S., Deng, Y. Y. & Guo, F. B. Essentiality drives the orientation bias of bacterial genes in a continuous manner. *Scientific reports* **5**, 16431, doi:10.1038/srep16431 (2015).
27. Wei, W. *et al.* IFIM: a database of integrated fitness information for microbial genes. *Database: the journal of biological databases and curation* **2014**, 10.1093/database/bau052 (2014).
28. Wei, W., Ning, L. W., Ye, Y. N. & Guo, F. B. Geptop: a gene essentiality prediction tool for sequenced bacterial genomes based on orthology and phylogeny. *PLoS One* **8**, e72343, doi:10.1371/journal.pone.0072343 (2013).
29. Llorens-Rico, V. *et al.* Bacterial antisense RNAs are mainly the product of transcriptional noise. *Sci Adv* **2**, e1501363, doi:10.1126/sciadv.1501363 (2016).
30. Zhang, G. & Gao, F. Quantitative analysis of correlation between AT and GC biases among bacterial genomes. *PLoS One* **12**, e0171408, doi:10.1371/journal.pone.0171408 (2017).
31. Hershberg, R. & Petrov, D. A. Evidence that mutation is universally biased towards AT in bacteria. *PLoS Genet* **6**, e1001115, doi:10.1371/journal.pgen.1001115 (2010).
32. dos Reis, M., Wernisch, L. & Savva, R. Unexpected correlations between gene expression and codon usage bias from microarray data for the whole Escherichia coli K-12 genome. *Nucleic Acids Research* **31**, 6976–6985, doi:10.1093/nar/gkg897 (2003).
33. dos Reis, M., Savva, R. & Wernisch, L. Solving the riddle of codon usage preferences: a test for translational selection. *Nucleic Acids Res* **32**, 5036–5044, doi:10.1093/nar/gkh834 (2004).
34. Chen, W. H. *et al.* Integration of multi-omics data of a genome-reduced bacterium: Prevalence of post-transcriptional regulation and its correlation with protein abundances. *Nucleic Acids Res*. doi:10.1093/nar/gkw004 (2016).
35. Chen, X. & Zhang, J. Why are genes encoded on the lagging strand of the bacterial genome? *Genome Biol Evol* **5**, 2436–2439, doi:10.1093/gbe/evt193 (2013).
36. Barrett, T. *et al.* NCBI GEO: archive for functional genomics data sets—update. *Nucleic Acids Research* **41**, D991–D995, doi:10.1093/nar/gks1193 (2013).
37. Faith, J. J. *et al.* Large-scale mapping and validation of Escherichia coli transcriptional regulation from a compendium of expression profiles. *PLoS Biol* **5**, e8, doi:10.1371/journal.pbio.0050008 (2007).
38. Nicolas, P. *et al.* Condition-dependent transcriptome reveals high-level regulatory architecture in Bacillus subtilis. *Science* **335**, 1103–1106, doi:10.1126/science.1206848 (2012).
39. Benson, D. A. *et al.* GenBank. *Nucleic Acids Res* **41**, D36–42, doi:10.1093/nar/gks1195 (2013).
40. Gao, F., Luo, H. & Zhang, C. T. DoriC 5.0: an updated database of oriC regions in both bacterial and archaeal genomes. *Nucleic Acids Res* **41**, D90–93, doi:10.1093/nar/gks990 (2013).
41. Hendrickson, H. & Lawrence, J. G. Mutational bias suggests that replication termination occurs near the dif site, not at Ter sites. *Mol Microbiol* **64**, 42–56, doi:10.1111/j.1365-2958.2007.05596.x (2007).
42. Mao, F., Dam, P., Chou, J., Olman, V. & Xu, Y. DOOR: a database for prokaryotic operons. *Nucleic Acids Res* **37**, D459–463, doi:10.1093/nar/gkn757 (2009).
43. Lowe, T. M. & Eddy, S. R. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res* **25**, 955–964 (1997).

44. Team, R. C. R: A Language and Environment for Statistical Computing. (2017).
45. Grömping, U. Relative Importance for Linear Regression in R: The Package relaimpo. *Journal of Statistical Software* **17**, 1–27 (2006).
46. Wickham, H. *ggplot2: elegant graphics for data analysis*. (Springer New York, 2009).

Author Contributions

N.G., G.L. and W.H.C. conceived the study through iterative discussions, collected and analyzed the data and wrote the manuscript; M.J.L. helped with the revision.

Additional Information

Supplementary information accompanies this paper at doi:10.1038/s41598-017-11159-3

Competing Interests: The authors declare that they have no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

3.2 Manuscript 2. Prokaryotic genome expansion is facilitated by phages and plasmids but impaired by CRISPR

Na L. Gao [1, 2, †], Jingchao Chen [3, †], Martin J Lercher [2, §], Wei-Hua Chen [1, 3,4,§]

[1] Key Laboratory of Molecular Biophysics of the Ministry of Education, Hubei Key Laboratory of Bioinformatics and Molecular-imaging, Department of Bioinformatics and Systems Biology, College of Life Science and Technology, Huazhong University of Science and Technology, 430074 Wuhan, Hubei, China

[2] Institute for Computer Science and Dept. of Biology, Heinrich Heine University, 40225 Duesseldorf, Germany

[3] College of Life Science, HeNan Normal University, 453007 Xinxiang, Henan, China

[4] Huazhong University of Science and Technology Ezhou Industrial Technology Research Institute, 436044 Ezhou, Hubei, China

[§]To whom correspondence should be addressed； Wei-Hua Chen, Tel: +862787542127, Fax: +862787542527, Email: weihuachen@hust.edu.cn or Martin J Lercher, Tel: +492118110546, Email: martin.lercher@hhu.de.

[†] These authors contributed equally to the paper as first authors.

This manuscript is currently under review at *Frontiers in Microbiology*. As the first author, I conceived the study with help from Prof. Chen; I assembled the data from databases of our lab and public domains. I performed all the data analyses, interpreted the data together with Prof. Chen and Prof. Lercher, wrote the first draft of the manuscript, and edited the final manuscript together with Prof. Chen and Prof. Lercher.

Abstract

Bacteriophages and plasmids can introduce novel DNA into bacterial cells, thereby creating an opportunity for genome expansion; conversely, CRISPR, the prokaryotic adaptive immune system, which targets and eliminates foreign DNAs, may impair genome expansions. Recent studies presented conflicting results over the impact of CRISPR on genome expansion. In this study, we constructed a comprehensive dataset of prokaryotic genomes and identified their associations with phages and plasmids. We found that genomes associated with phages and/or plasmids were significantly larger than those without, indicating that both phages and plasmids contribute to genome expansion. Genomes were increasingly larger with increasing numbers of associated phages or plasmids. Conversely, genomes with CRISPR systems were significantly smaller than those without, indicating that CRISPR has a negative impact on genome size. These results confirmed that on evolutionary timescales, bacteriophages and plasmids facilitate genome expansion, while CRISPR impairs such a process in prokaryotes. Furthermore, our results also revealed that CRISPR systems show a preference for targeting phages over plasmids.

Keywords: Prokaryotic genome expansion, Bacteriophages, Plasmids, CRISPR, Horizontal gene transfer

Introduction

Gene duplication and/or horizontal gene transfer (HGT) play important roles in functional innovation and species adaptation, and are the main sources of genome expansions (Isambert and Stein, 2009;Schonknecht et al., 2013;Nyvltova et al., 2015;Smith et al., 2016;Tsai et al., 2018). In prokaryotes, it has been shown that the importance of HGT for genome expansions can even outweigh that of gene duplication (Pal et al., 2005;Treangen and Rocha, 2011).

Mobile DNA elements such as bacteriophages (viruses that infect archaea and bacteria (8)(8), referred to as phages below) and plasmids can introduce novel DNAs into the host genomes (Yamaguchi et al., 2001;Jensen and Lyon, 2009;Lindsay, 2010;Malachowa and DeLeo, 2010). They often have a very narrow range of hosts; but under certain conditions, such as antibiotic stress, phages and plasmids can expand their host ranges (Modi et al., 2013). Therefore, phages and plasmids are important sources of HGT and of prokaryotic innovations, and consequently drive bacterial evolution and adaptation (Koonin and Wolf, 2008;Nogueira et al., 2009;Argov et al., 2017).

Phages and plasmids are widely distributed in prokaryotes. Unlike plasmids, phages are pathogens that often lead to lysis of their hosts (Deresinski, 2009;Wernicki et al., 2017). Over the course of prokaryotic evolution, bacteria and archaea developed various defense systems against phages, plasmids, and other invading genetic elements (Luk et al., 2014). CRISPR (clustered regularly interspaced short palindromic repeats), the adaptive immune system of prokaryotes, is a recently recognized player in the ongoing arms race between prokaryotic viruses and hosts, and plays an important role in the dynamic process by which the genomes of prokaryotes and mobile elements coevolve. CRISPR systems are widespread in prokaryotes, exists in about 40% of bacteria and 90% of archaea (Godde and Bickerton, 2006;Makarova et al., 2011;Seed et al., 2013;Huang et al., 2016), or ~10% as revealed by a recent study (Burstein et al.,

2016). CRISPR systems can also target plasmids (Marraffini and Sontheimer, 2008), although plasmids are not necessarily detrimental to their host's fitness but instead often carry a diverse range of antimicrobial and biocide resistance genes that may help their hosts to survive under certain conditions (McCarthy and Lindsay, 2012;Shabbir et al., 2016).

Based on the above observations, it is reasonable to speculate that over the course of evolution, phages and plasmids may contribute to the expansion of prokaryotic genomes, while CRISPR systems may impair such a process. These speculations are consistent with recent observations that CRISPR limits HGT by targeting foreign DNAs (Marraffini and Sontheimer, 2008;Bikard et al., 2012). However, controversial observations have also been reported recently. For example, Gophna and colleagues did not observe the expected negative correlation between CRISPR activity in microbes with three independent measures of recent HGT, leading them to conclude that the inhibitory effect of CRISPR against HGT is undetectable (Gophna et al., 2015). Furthermore, a recent study revealed that CRISPR-mediated phage resistance can even enhance HGT by increasing the resistance of transductants against subsequent phage infections (Watson et al., 2018). These observations appear surprising, as the restricted acquisition of foreign genetic material is believed to be one of the sources of the maintenance fitness cost of CRISPR systems and may be one of the reasons for the patchy distribution of CRISPR among bacteria (Frost et al., 2005;Baltrus, 2013). Thus, it is currently unclear what long-term effects CRISPR, phages, and plasmids have on genome expansion.

In this study, we first collected a comprehensive dataset of prokaryotes and their associations with phages, plasmids, and CRISPR systems. We then evaluated the contributions of phages, plasmids, and CRISPR to genome size. After controlling for genome GC (guanine+cytosine) content, which is known to correlate significantly with genome size (Chen et al., 2016a;Chen et al., 2016b), we found that both phages and plasmids are associated with larger genomes, while the presence of a CRISPR system is associated with small genome size. Genome

sizes increase with increasing numbers of associated phages and plasmids. Our results clearly indicate that in the long run, phages and plasmids facilitate genome expansions, while CRISPR impairs such a process in prokaryotes. Furthermore, our results also reveal a striking preference of CRISPR systems for targeting phages rather than plasmids, consistent with the typical consequences of phage and plasmid infections to the hosts and the roles of CRISPR as a defense system.

Materials and Methods

Data

We obtained data from three sources. Microbe-phage interaction data was collected from the MVP database, which we described in a previous publication (Gao et al., 2018). MVP is one of the latest and largest databases about microbe-phage interactions, which supplied 26,572 interactions between 9,245 prokaryotes and 18,608 viral clusters based on 30,321 evidence entries (Gao et al., 2018).

The basic genome information from complete archaeal and bacterial genomes, including the number of associated plasmids, was downloaded from the NCBI Genome database (https://www.ncbi.nlm.nih.gov/genome/; accessed on June 28, 2018) (Coordinators, 2018). The genome size and GC-content from 10,686 complete prokaryotic genomes (287 archaeal and 10,279 bacterial genomes) were identified. 2,827 prokaryotes were associated with plasmids.

The CRISPRs data was obtained from the CRISPRdb database (Grissa et al., 2007) (http://crispr.i2bc.paris-saclay.fr/; last update May 09, 2017). 202 archaeal and 3,059 bacterial genomes were associated with CRISPR systems. 77 of these encode CRISPR on both plasmids and genome, while only 36 encode CRISPR exclusively on plasmids. The 77 genomes which contained plasmid-encoded CRISPR systems were removed from all analyses.

In total, 5,994 prokaryotes were found in both of the first two datasets; among these, 1,950 contained plasmids, 2,758 contained phages, and 2,056 contained CRISPRs on their chromosomes. Detailed information on the dataset can be found in Supplementary Table 1.

Statistical analysis

All data were analyzed using R v3.4 (Team, 2017). All pair-wise comparisons between two groups of numeric data (genome sizes or genomic GC-contents) were performed by Wilcoxon rank-sum tests. Linear model (LM) analysis was performed with the R function glm(). Relative importance analysis was performed with the calc.relimp() function available from the R package 'relatimpo' (U, 2006).

Results

Prokaryotic genomes and their associations with phages, plasmids and CRISPRs

To systematically investigate the impacts of phages, plasmids, and CRISPRs on genome expansion, we constructed a list of 5,994 completely sequenced prokaryotic genomes and obtained their associations with phages, plasmids, and CRISPRs; for details please consult the Materials and Methods section and Supplementary Table 1

As shown in Figure 1A, we found that 53.98% of prokaryotes had no known associations with infecting phages. 14.88%, 16.68%, and 14.46% of prokaryotes were associated with one, two to three, and more than three phages, respectively (Figure 1A). In addition, we found that 67.46% of prokaryotes did not associate with plasmids, while 14.75%, 11.68%, and 6.12% of the genomes associated with one, two to three, and more than three plasmids, respectively (Figure 1B). Previous studies suggested that the genomic GC-contents as well as nucleotide frequencies of phages and plasmids often closely resembles that of their hosts

(Hiroshi Nakashima1*, 2015;Ahlgren et al., 2017;Ren et al., 2017); consistent with these previous observations, we obtained correlation coefficient values of 0.972 and 0.970 between the GC-contents of the host genomes and their associated phages and plasmids, respectively (Supplementary Figure 1), confirming the high quality of our association data. We found that in total 42.58% of genomes collected in this study contained either phages or plasmids but not both, while 17.98% of genomes contained both phages and plasmids.

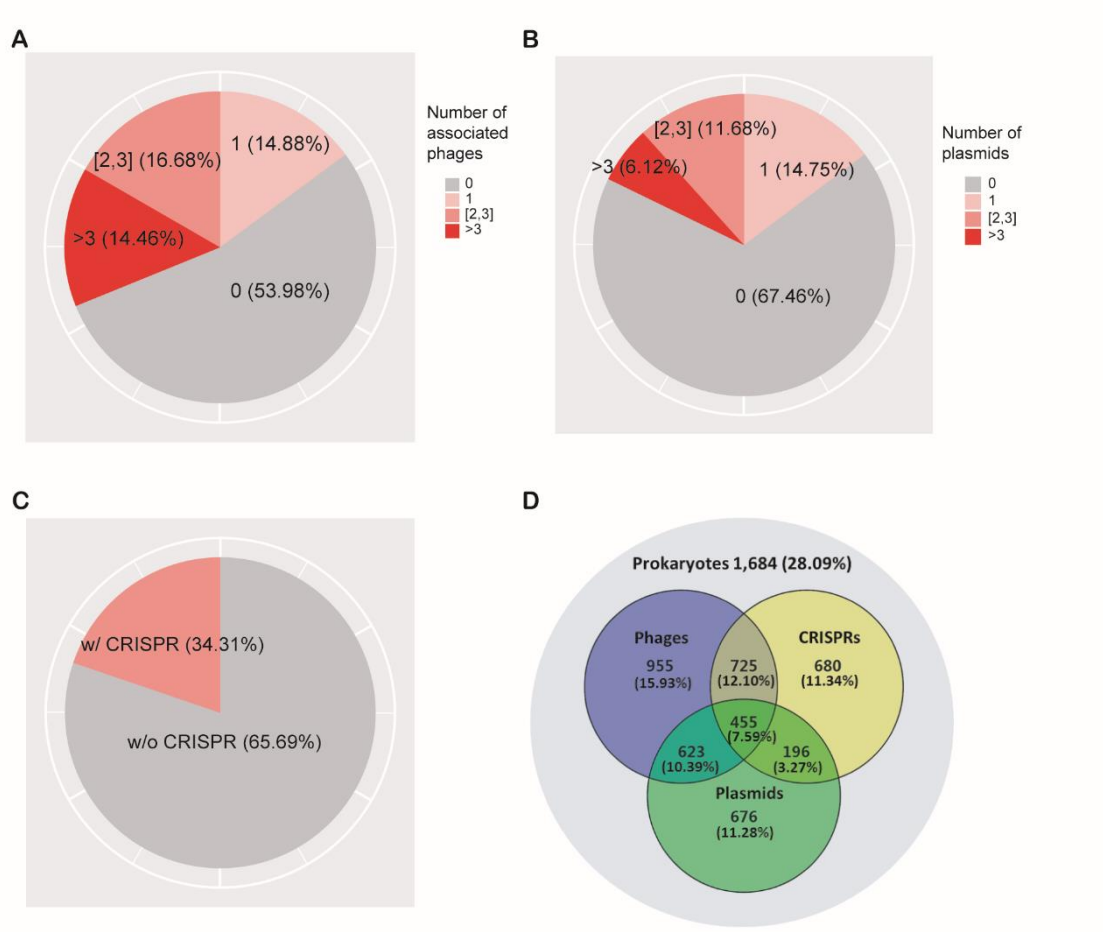


Figure 1. 5,994 prokaryotic genomes and their associations with phages (A), plasmids (B), and CRISPRs (C). The Venn diagram (D) shows the overlap of their distributions in prokaryotes. 1,684 genomes (28.09%) were not found to be associated with phages, plasmids, or CRISPRs; 455 (7.59%) genomes were associated with all three elements.

As shown in Figure 1C, we found CRISPR systems in 34.31% of the prokaryotic genomes (Figure 1C); this percentage is within the range of previously reported numbers (Godde and Bickerton, 2006;Makarova et al., 2011;Seed et al., 2013;Burstein et al., 2016;Huang et al., 2016). We found that CRISPRs were significantly enriched in phage-associated compared to non-phage-associated genomes (odds ratio OR=1.43, P=1.7x10^{-14} from Fisher's exact test) but not in plasmid-associated compared to non-plasmid-associated genomes (OR=0.96, P=0.47). In addition, we found that CRISPRs were more enriched in phage-associated compared to plasmid-associated genomes (OR=2.62, P= 9.0x10^{-26}, excluding genomes containing both phages and plasmids), suggesting a strong target preferences of CRISPRs toward phages (Table 1).

Table 1. Estimated enrichment of CRISPR in phage-associated and plasmid-associated genomes compared to other genomes, and enrichment of CRISPR in phage-associated compared to plasmid-associated genomes.

Comparison	Odds ratio[b]	*P*-value[c]
Phage-associated vs. others	1.43	1.75x10^{-14}
Plasmid-associated vs. others	0.96	0.47
Phage- vs. plasmid-associated	1.49	6.62x10^{-11}
Phage-associated vs. others[a]	1.45	3.77x10^{-11}
Plasmid-associated vs. others [a]	0.56	1.01x10^{-12}
Phage- vs. Plasmid-associated[a]	2.62	8.97x10^{-26}

[a] excluding genomes contained both phages and plasmids.

[b] odds ratio OR > 1 indicates enrichment of CRISPR in the first group, while OR < 1 indicates depletion.

[c] *P*-values indicate whether CRISPR is significantly enriched or depleted in the first group as compared with the second according to Fisher' exact test.

Phages and plasmids are associated with larger genomes, while CRISPR is associated with smaller genomes

We next investigated which factors contribute significantly to genome size. Previous results have shown a strong correlation between genomic GC content

and genome size (Chen et al., 2016a); GC content may even play a causal role in shaping genome size (Chen et al., 2016b). Applying a linear model (LM, see Materials and Methods for details), we found that GC content was indeed the strongest predictor of genome size (Table 2). The LM analysis also revealed that the presence/absence of phages, plasmids, and CRISPR all significantly influenced genome size; the presences of phages and of plasmids were associated with increased genome sizes, while CRISPR was associated with decreased genome sizes (Table 2). We estimated that the relative importance of these factors for genome size were 89% for GC-content, 5.8% for phage presence, 4.4% for plasmid presence, and 0.38% for CRISPR presence. Interestingly, we found that the presence of both phages and plasmids in the same genome was associated with a smaller genome size than expected (*i.e.*, the interaction term phages*plasmids was negative, Table 2). Unless stated otherwise, we thus limit our further analyses to prokaryotes that contained either phages or plasmids but not both. Note that our conclusions on the influence of phages, plasmids, and CRISPR systems on genome size remain unchanged if we perform separate analyses on genomes containing no phages and on genomes containing no plasmids (Table 2).

Increasing numbers of phages and plasmids are associated with increased genome sizes

We next investigated the impact of the numbers of phages and plasmids on genome size. Phages and plasmids often have very narrow host ranges (Haruo Suzuki, 2014;Gao et al., 2018); the number of known associations with phages may indicate the ability of the prokaryotic host to acquire external novel DNA. Consistent with our expectation, we found that genomes associated with more phages had larger overall genomes (Figure 2A). We observed similar results with plasmids (Figure 2B).

Table 2. Relative importance of various factors for genome size in a linear model (LM).

Dataset	Factor	Coefficient	*P*-value	Relative importance
All	GC%	0.089	$< 2x10^{-16}$	89.51%
	plasmid	0.754	$< 2x10^{-16}$	5.77%
	phage	0.598	$< 2x10^{-16}$	4.35%
	CRISPR	-0.158	$2.0x10^{-4}$	0.38%
	phage*plasmid	-0.216	0.012	-
No plasmids	GC%	0.09	$< 2x10^{-16}$	93.93%
	phage	0.596	$< 2x10^{-16}$	5.65%
	CRISPR	-0.164	$1.1x10^{-3}$	0.41%
No phages	GC%	0.093	$< 2x10^{-16}$	93.76%
	plasmid	0.743	$< 2x10^{-16}$	6.01%
	CRISPR	-0.145	0.024	0.28%

Note: The equation of "All" dataset used in the linear model (LM) is size ~ GC% + plasmid + phage + CRISPR + phage*plasmid. Here, size represents the genome size; GC% represents the genomic GC-content of the host genome; plasmid, phage, and CRISPR represent whether the host genomes are associated with plasmids, phages, and CRISPR, respectively. The "Coefficient" column contains estimated regression coefficients calculated by ordinary least squares. Relative importance was calculated using the 'relaimpo' package ; the equation of "No plasmids" dataset is size ~ GC% + phage + CRISPR; and the equation of "No phages" dataset is size ~ GC% + plasmid + CRISPR.

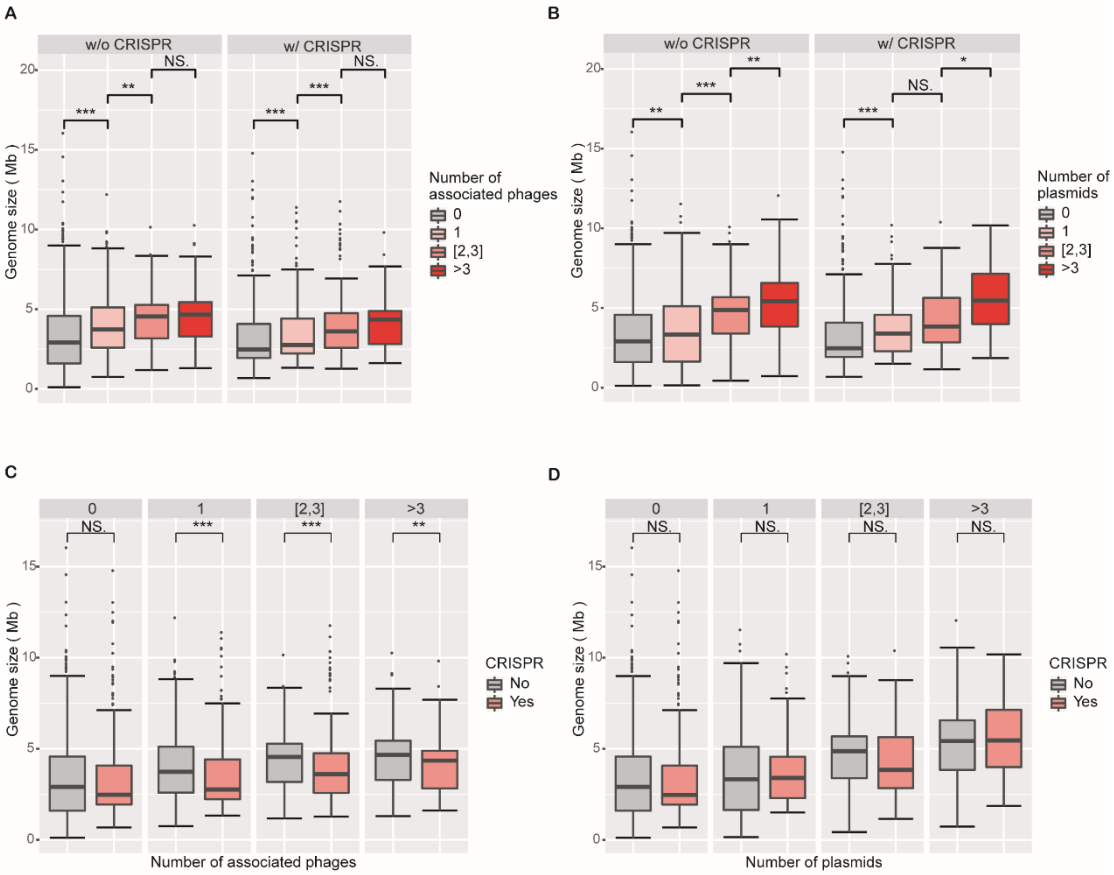


Figure 2. Increasing numbers of phages and plasmids are associated with increased genome sizes, while phage-associated genomes with CRISPR systems are smaller than those without CRISPR systems. A) Boxplot of genomes size as a function of the number of associated phages. Genome sizes are larger with increasing numbers of associated phages, regardless of whether genomes encode CRISPR systems. B) Boxplot of genomes size as a function of the number of associated plasmids. The impact of plasmids on genome size is similar to that of bacteriophages. C) Boxplot of genome size as a function of the presence/absence of CRISPRs in genomes associated with phages. Phage-associated genomes with CRISPR systems are significantly smaller in size than those without CRISPR, regardless of the number of phages they are associated with. D) Boxplots of genome sizes in genomes associated with plasmids as a function of the presence/absence of CRISPRs. CRISPRs have no significant impact on genome sizes in genomes associated with plasmids. Wilcoxon rank sum tests were used to compare between groups. Level of significance: *** $P<0.001$; ** $P<0.01$; * $P<0.05$; NS. $P\geq0.05$.

Consistent with the results from the LM analysis, we found that phage-associated genomes are statistically significantly smaller when they encode a CRISPR system compared to when they do not (Figure 2C). However, we did not find a corresponding trend in plasmid-associated genomes (Figure 2D). These results are consistent with the different fitness consequences of phage and plasmid invasions to the prokaryotic hosts. Both phages and plasmids can bring exogenous DNA to prokaryotes and decrease the fitness of their hosts, for example by increasing the burden on the host's transcription and translation apparatus. However, phages typically cause substantial additional fitness decreases through virion production and assembly and eventually host lysis, while plasmids often carry genes that are beneficial to the survival of their hosts under certain circumstances (Dionisio et al., 2005;Jiang et al., 2013). It is thus likely that the CRISPR systems in prokaryotes are more sensitive to phages than to plasmids. This line of argument is also consistent with our results that the presence of CRISPRs is more enriched in phage-associated than in plasmid-associated genomes.

The influence of associated phages, plasmids, and CRISPR on genome GC-content

We then investigated which factors contribute significantly to genome GC-content. Consistent with our previous results (LM analysis, Table 2), we found that genome size was indeed the most significant predictor of GC-content, with a relative importance of almost 99% (LM analysis, Table 3). The presence of plasmids also had a significant influence on GC-content, with a relative importance of 1% (Table 3). The presence/absence of phages and CRISPR had no significant influence on GC-content by themselves; surprisingly, however, the presence of phages reduced the influence of plasmid presence on GC content.

We also investigated whether these factors contribute significantly to GC-content when genomes contain no phages/plasmids. As expected, genome size remained the most significant factor for the prediction of genome GC-content, as

shown in Table 3, with a relative importance of around 99%. Analysis of genomes without phage-associations confirmed an important influence of plasmid presence on GC content (Table 3). In addition, analysis of genomes without plasmid-associations revealed a small but no statistically significant influence of phage presence on GC-content (Table 3).

As shown in Supplementary Table 2, we find that the number of associated phages and plasmids contribute significantly to GC-content, but we don't find clear and consistent trends in GC-content as a function of the number of associated phages or plasmids (Supplementary Figure 3).

Table 3. Relative importance of various factors for GC-content (GC%) in a LM.

Dataset	Factors	Coefficient	*P*-value	Relative importance
All	size	4.12	$< 2x10^{-16}$	98.91%
	plasmid	-1.135	$7.29x10^{-3}$	1.03%
	phage	-0.007	0.983	0.06%
	CRISPR	-0.056	0.846	0.00%
	phage*plasmid	-1.247	0.033	-
No plasmids	size	4.17	$< 2x10^{-16}$	100%
	phage	-0.074	0.829	0.00%
	CRISPR	0.066	0.847	0.00%
No phages	size	4.16	$< 2x10^{-16}$	99.39%
	plasmid	-1.108	0.017	0.28%
	CRISPR	1.108	0.011	0.32%

Note: The equation of "All" dataset used in the linear model (LM) is GC% ~ size + plasmid + phage + CRISPR + phage*plasmid; the equation of "No plasmids" dataset is GC% ~ size + phage + CRISPR; and the equation of "No phages" dataset is GC% ~ size + plasmid + CRISPR.

Discussion

We expected that phages and plasmids could facilitate genome expansions because they can bring novel DNAs (genes or fragments) into prokaryotic cells

that can be integrated into the host genome, while CRISPR immune systems could impair such a process by targeting and eliminating foreign DNAs. However, recent studies presented inconsistent results regarding this topic (Marraffini and Sontheimer, 2008;Makarova et al., 2011;Bikard et al., 2012;Gophna et al., 2015;Watson et al., 2018).

To address this issue, we constructed a comprehensive dataset of prokaryotic genomes and their associations with phages and plasmids. By dividing genomes into distinct groups according to whether they associated with phages and/or plasmids and/or contained CRISPRs, we revealed that genomes with phages or with plasmids were significantly larger than those without, and genome sizes increased with increasing numbers of associated phages/plasmids. Conversely, phage-associated (but not plasmid-associate) genomes with CRISPRs were significantly smaller in size than those without, regardless of the number of associated phages. These results confirm that in the long run, bacteriophages and plasmids facilitate genome expansions while CRISPR impairs phage-driven genome expansions.

Genome size evolution has previously been reported to be associated with that of genomic GC-content (Gao et al., 2017). Thus, it appeared possible that phage- and/or plasmid-association has a direct effect not only on genome size but also on GC-content. However, in this study, we found only minor influences of phages and plasmids on genomic GC-content (Table 3 and Supplementary Table 1).

Our results also imply that CRISPR immune systems might be more sensitive towards invading phages than plasmids, consistent with the differential fitness burdens brought by the two types of foreign invaders to the hosts (Canchaya et al., 2004;Weinberger et al., 2012;Jiang et al., 2013;Pleska and Guet, 2017).

Our results differ significantly from several previous studies (Gophna et al., 2015;Watson et al., 2018). For example, Gophna *et al.* reported that the inhibitory effect of CRISPR against HGT is undetectable using three independent measures of recent HGT (Gophna et al., 2015). However, it is known that CRISPR spacers

– which were used by Gophna *et al*. to assess CRISPR activity (Gophna et al., 2015) – have very high turnover rates, on the time-scale of days (Deveau et al., 2008;Horvath et al., 2008;Tyson and Banfield, 2008), while HGT genes may take a very long time to be incorporated into existing gene networks (Lercher and Pal, 2008), suggesting that it is only possible to look at the impacts of CRISPRs on HGTs at evolutionary scales. Interestingly, Gophna *et al.* also studied spacer acquisition and concluded there was a bias toward frequently encountered invasive exogenous genetic elements, especially infecting viruses (Gophna et al., 2015); this is consistent with our conclusion that CRISPRs tend to be more sensitive towards invading phages than plasmids. Recently, Watson *et al.* reported that the CRISPR system of the bacterium *Pectobacterium atrosepticum* enabled the host to resist phage infection, but that this enhanced rather than impeded HGT by transduction (Watson et al., 2018). However, it is yet to be seen whether or not this phenomenon is unique to *P. atrosepticum*.

Acknowledgments

This manuscript has been released as a Pre-Print at BioRxiv https://www.biorxiv.org/content/10.1101/474767v1 (Na L. Gao, 2019).

References:

Bacteriophage. *Wikipedia*, https://en.wikipedia.org/wiki/Bacteriophage.

Ahlgren, N.A., Ren, J., Lu, Y.Y., Fuhrman, J.A., and Sun, F. (2017). Alignment-free d_2^* oligonucleotide frequency dissimilarity measure improves prediction of hosts from metagenomically-derived viral sequences. *Nucleic Acids Res* 45, 39-53.

Argov, T., Azulay, G., Pasechnek, A., Stadnyuk, O., Ran-Sapir, S., Borovok, I., Sigal, N., and Herskovits, A.A. (2017). Temperate bacteriophages as regulators of host behavior. *Curr Opin Microbiol* 38, 81-87.

Baltrus, D.A. (2013). Exploring the costs of horizontal gene transfer. *Trends Ecol Evol* 28, 489-495.

Bikard, D., Hatoum-Aslan, A., Mucida, D., and Marraffini, L.A. (2012). CRISPR interference can prevent natural transformation and virulence acquisition during in vivo bacterial infection. *Cell Host Microbe* 12, 177-186.

Burstein, D., Sun, C.L., Brown, C.T., Sharon, I., Anantharaman, K., Probst, A.J., Thomas, B.C., and Banfield, J.F. (2016). Major bacterial lineages are essentially devoid of CRISPR-Cas viral defence systems. *Nat Commun* 7, 10613.

Canchaya, C., Fournous, G., and Brussow, H. (2004). The impact of prophages on bacterial chromosomes. *Mol Microbiol* 53**,** 9-18.

Chen, W.-H., Van noort, V., Lluch-Senar, M., Hennrich, M.L., H. wodke, J.A., Yus, E., Alibés, A., Roma, G., Mende, D.R., Pesavento, C., Typas, A., Gavin, A.-C., Serrano, L., and Bork, P. (2016a). Integration of multi-omics data of a genome-reduced bacterium: Prevalence of post-transcriptional regulation and its correlation with protein abundances. *Nucleic Acids Research* 44**,** 1192-1202.

Chen, W.H., Lu, G., Bork, P., Hu, S., and Lercher, M.J. (2016b). Energy efficiency trade-offs drive nucleotide usage in transcribed regions. *Nat Commun* 7**,** 11334.

Coordinators, N.R. (2018). Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* 46**,** D8-D13.

Deresinski, S. (2009). Bacteriophage therapy: exploiting smaller fleas. *Clin Infect Dis* 48**,** 1096-1101.

Deveau, H., Barrangou, R., Garneau, J.E., Labonte, J., Fremaux, C., Boyaval, P., Romero, D.A., Horvath, P., and Moineau, S. (2008). Phage response to CRISPR-encoded resistance in Streptococcus thermophilus. *J Bacteriol* 190**,** 1390-1400.

Dionisio, F., Conceicao, I.C., Marques, A.C., Fernandes, L., and Gordo, I. (2005). The evolution of a conjugative plasmid and its ability to increase bacterial fitness. *Biol Lett* 1**,** 250-252.

Frost, L.S., Leplae, R., Summers, A.O., and Toussaint, A. (2005). Mobile genetic elements: the agents of open source evolution. *Nat Rev Microbiol* 3**,** 722-732.

Gao, N., Lu, G., Lercher, M.J., and Chen, W.H. (2017). Selection for energy efficiency drives strand-biased gene distribution in prokaryotes. *Sci Rep* 7**,** 10572.

Gao, N.L., Zhang, C., Zhang, Z., Hu, S., Lercher, M.J., Zhao, X.-M., Bork, P., Liu, Z., and Chen, W.-H. (2018). MVP: a microbe–phage interaction database. *Nucleic Acids Research* 46**,** D700-D707.

Godde, J.S., and Bickerton, A. (2006). The repetitive DNA elements called CRISPRs and their associated genes: evidence of horizontal transfer among prokaryotes. *J Mol Evol* 62**,** 718-729.

Gophna, U., Kristensen, D.M., Wolf, Y.I., Popa, O., Drevet, C., and Koonin, E.V. (2015). No evidence of inhibition of horizontal gene transfer by CRISPR-Cas on evolutionary timescales. *ISME J* 9**,** 2021-2027.

Grissa, I., Vergnaud, G., and Pourcel, C. (2007). The CRISPRdb database and tools to display CRISPRs and to generate dictionaries of spacers and repeats. *BMC Bioinformatics* 8**,** 172.

Haruo Suzuki, C.J.B., Eva M. Top (2014). Genomic Signature Analysis to Predict Plasmid Host Range. *Molecular Life Sciences*.

Hiroshi Nakashima1*, K.H.a.K.M. (2015). Relationship of Genomic G+C Content between Phages/Plasmids and Their Hosts. *British Biotechnology Journal***,** 9(1): 1-9.

Horvath, P., Romero, D.A., Coute-Monvoisin, A.C., Richards, M., Deveau, H., Moineau, S., Boyaval, P., Fremaux, C., and Barrangou, R. (2008). Diversity, activity, and evolution of CRISPR loci in Streptococcus thermophilus. *J Bacteriol* 190**,** 1401-1412.

Huang, Q., Luo, H., Liu, M., Zeng, J., Abdalla, A.E., Duan, X., Li, Q., and Xie, J. (2016). The effect of Mycobacterium tuberculosis CRISPR-associated Cas2 (Rv2816c) on stress response genes expression, morphology and macrophage survival of Mycobacterium smegmatis. *Infect Genet Evol* 40**,** 295-301.

Isambert, H., and Stein, R.R. (2009). On the need for widespread horizontal gene transfers under genome size constraint. *Biol Direct* 4**,** 28.

Jensen, S.O., and Lyon, B.R. (2009). Genetics of antimicrobial resistance in Staphylococcus aureus. *Future Microbiol* 4**,** 565-582.

Jiang, W., Maniv, I., Arain, F., Wang, Y., Levin, B.R., and Marraffini, L.A. (2013). Dealing with the evolutionary downside of CRISPR immunity: bacteria and beneficial plasmids. *PLoS Genet* 9**,** e1003844.

Koonin, E.V., and Wolf, Y.I. (2008). Genomics of bacteria and archaea: the emerging dynamic view of the prokaryotic world. *Nucleic Acids Res* 36**,** 6688-6719.

Lercher, M.J., and Pal, C. (2008). Integration of horizontally transferred genes into regulatory interaction networks takes many million years. *Mol Biol Evol* 25**,** 559-567.

Lindsay, J.A. (2010). Genomic variation and evolution of Staphylococcus aureus. *Int J Med Microbiol* 300**,** 98-103.

Luk, A.W., Williams, T.J., Erdmann, S., Papke, R.T., and Cavicchioli, R. (2014). Viruses of haloarchaea. *Life (Basel)* 4**,** 681-715.

Makarova, K.S., Haft, D.H., Barrangou, R., Brouns, S.J., Charpentier, E., Horvath, P., Moineau, S., Mojica, F.J., Wolf, Y.I., Yakunin, A.F., Van Der Oost, J., and Koonin, E.V. (2011). Evolution and classification of the CRISPR-Cas systems. *Nat Rev Microbiol* 9**,** 467-477.

Malachowa, N., and Deleo, F.R. (2010). Mobile genetic elements of Staphylococcus aureus. *Cell Mol Life Sci* 67**,** 3057-3071.

Marraffini, L.A., and Sontheimer, E.J. (2008). CRISPR interference limits horizontal gene transfer in staphylococci by targeting DNA. *Science* 322**,** 1843-1845.

Mccarthy, A.J., and Lindsay, J.A. (2012). The distribution of plasmids that carry virulence and resistance genes in Staphylococcus aureus is lineage associated. *BMC Microbiol* 12**,** 104.

Modi, S.R., Lee, H.H., Spina, C.S., and Collins, J.J. (2013). Antibiotic treatment expands the resistance reservoir and ecological network of the phage metagenome. *Nature* 499**,** 219-222.

Na L. Gao, J.C., Martin J Lercher, Wei-Hua Chen (2019). Prokaryotic genome expansion is facilitated by phages and plasmids but impaired by CRISPR. *BioRxiv*.

Nogueira, T., Rankin, D.J., Touchon, M., Taddei, F., Brown, S.P., and Rocha, E.P. (2009). Horizontal gene transfer of the secretome drives the evolution of bacterial cooperation and virulence. *Curr Biol* 19**,** 1683-1691.

Nyvltova, E., Stairs, C.W., Hrdy, I., Ridl, J., Mach, J., Paces, J., Roger, A.J., and Tachezy, J. (2015). Lateral gene transfer and gene duplication played a key role in the evolution of Mastigamoeba balamuthi hydrogenosomes. *Mol Biol Evol* 32**,** 1039-1055.

Pal, C., Papp, B., and Lercher, M.J. (2005). Adaptive evolution of bacterial metabolic networks by horizontal gene transfer. *Nat Genet* 37**,** 1372-1375.

Pleska, M., and Guet, C.C. (2017). Effects of mutations in phage restriction sites during escape from restriction-modification. *Biol Lett* 13.

Ren, J., Ahlgren, N.A., Lu, Y.Y., Fuhrman, J.A., and Sun, F. (2017). VirFinder: a novel k-mer based tool for identifying viral sequences from assembled metagenomic data. *Microbiome* 5**,** 69.

Schönknecht, G., Chen, W.H., Ternes, C.M., Barbier, G.G., Shrestha, R.P., Stanke, M., Brautigam, A., Baker, B.J., Banfield, J.F., Garavito, R.M., Carr, K., Wilkerson, C., Rensing, S.A., Gagneul,

D., Dickenson, N.E., Oesterhelt, C., Lercher, M.J., and Weber, A.P. (2013). Gene transfer from bacteria and archaea facilitated evolution of an extremophilic eukaryote. *Science* 339**,** 1207-1210.

Seed, K.D., Lazinski, D.W., Calderwood, S.B., and Camilli, A. (2013). A bacteriophage encodes its own CRISPR/Cas adaptive response to evade host innate immunity. *Nature* 494**,** 489-491.

Shabbir, M.A., Hao, H., Shabbir, M.Z., Wu, Q., Sattar, A., and Yuan, Z. (2016). Bacteria vs. Bacteriophages: Parallel Evolution of Immune Arsenals. *Front Microbiol* 7**,** 1292.

Smith, G., Macias-Munoz, A., and Briscoe, A.D. (2016). Gene Duplication and Gene Expression Changes Play a Role in the Evolution of Candidate Pollen Feeding Genes in Heliconius Butterflies. *Genome Biol Evol* 8**,** 2581-2596.

Team, R.C. (2017). R: A Language and Environment for Statistical Computing.

Treangen, T.J., and Rocha, E.P. (2011). Horizontal transfer, not duplication, drives the expansion of protein families in prokaryotes. *PLoS Genet* 7**,** e1001284.

Tsai, Y.M., Chang, A., and Kuo, C.H. (2018). Horizontal Gene Acquisitions Contributed to Genome Expansion in Insect-Symbiotic Spiroplasma clarkii. *Genome Biol Evol* 10**,** 1526-1532.

Tyson, G.W., and Banfield, J.F. (2008). Rapidly evolving CRISPRs implicated in acquired resistance of microorganisms to viruses. *Environ Microbiol* 10**,** 200-207.

U, G. (2006). Relative importance for linear regression in R: The package relaimpo. *J. Stat. Softw***,** 17:11–27.

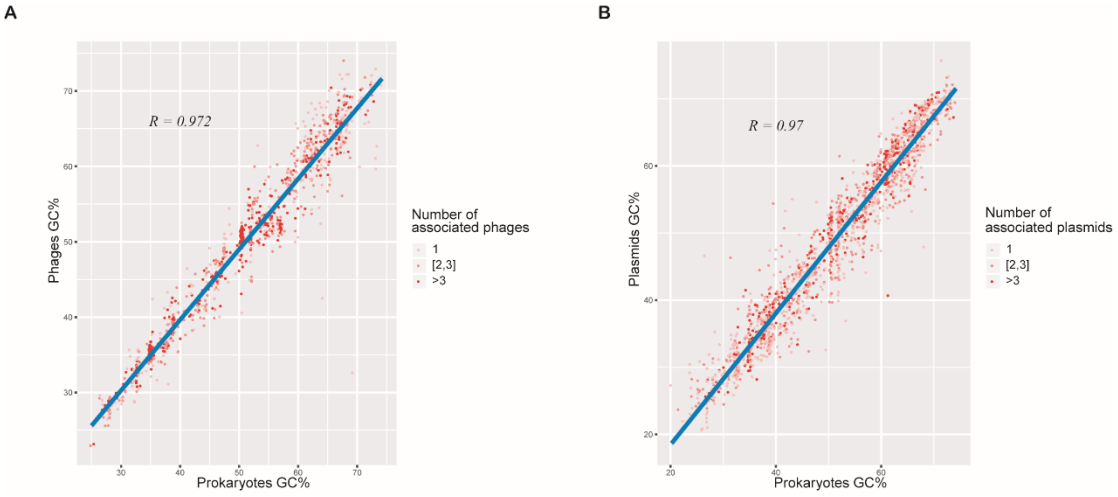
Watson, B.N.J., Staals, R.H.J., and Fineran, P.C. (2018). CRISPR-Cas-Mediated Phage Resistance Enhances Horizontal Gene Transfer by Transduction. *MBio* 9.

Weinberger, A.D., Sun, C.L., Plucinski, M.M., Denef, V.J., Thomas, B.C., Horvath, P., Barrangou, R., Gilmore, M.S., Getz, W.M., and Banfield, J.F. (2012). Persisting viral sequences shape microbial CRISPR-based immunity. *PLoS Comput Biol* 8**,** e1002475.

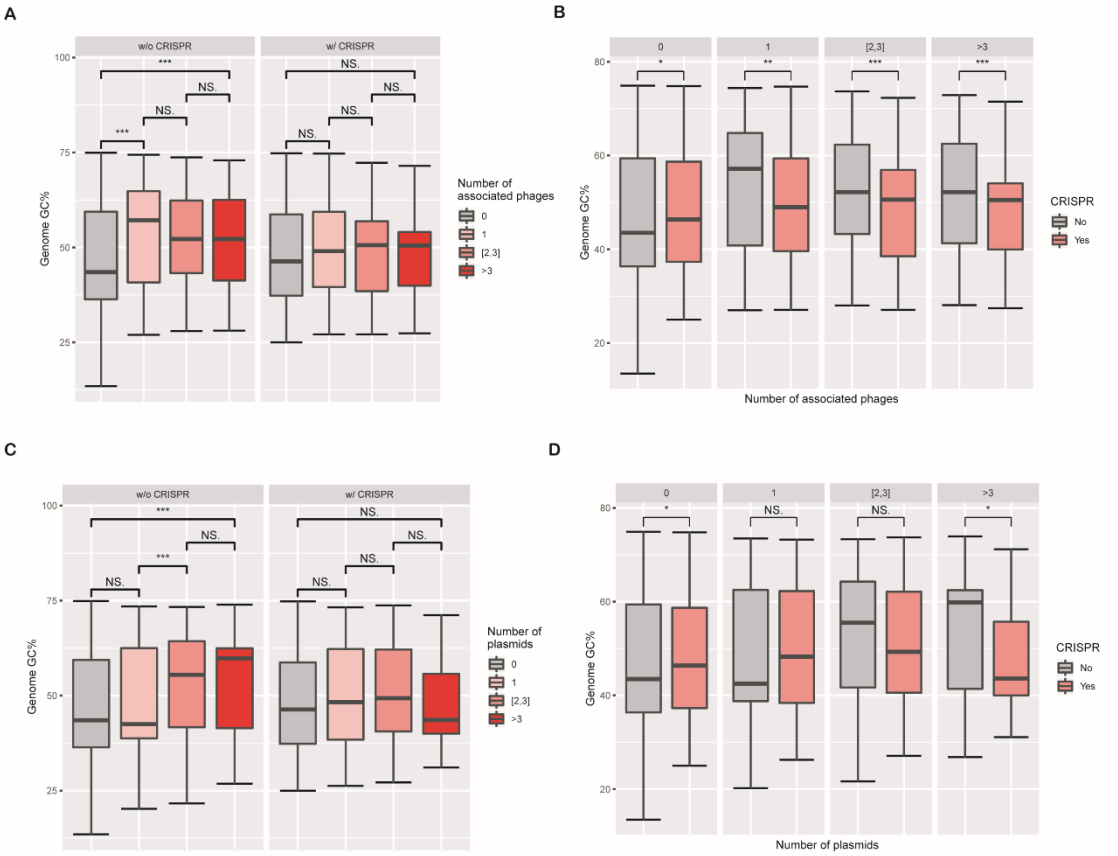
Wernicki, A., Nowaczek, A., and Urban-Chmiel, R. (2017). Bacteriophage therapy to combat bacterial infections in poultry. *Virol J* 14**,** 179.

Yamaguchi, T., Hayashi, T., Takami, H., Ohnishi, M., Murata, T., Nakayama, K., Asakawa, K., Ohara, M., Komatsuzawa, H., and Sugai, M. (2001). Complete nucleotide sequence of a Staphylococcus aureus exfoliative toxin B plasmid and identification of a novel ADP-ribosyltransferase, EDIN-C. *Infect Immun* 69**,** 7760-7771.

Supplementary Material



Supplementary Figure 1. Correlation of the GC-contents of the host genomes and their associated phages (A) and plasmids (B).



Supplementary Figure 2. No clear trends could be observed in genome GC-content as a function of the number of associated phages and plasmids. A) Boxplot of genome GC-content as a function of the number of associated phages.

B) Boxplots of genome GC-content as a function of the presence/absence of CRISPRs in genomes associated with phages. The GC-content of phage-associated genomes with CRISPRs are significantly lower than without, regardless of the number of associated phages. In contrast, in genomes without phage-associations, CRISPR-containing genomes are significantly higher in GC-content than genomes without CRISPRs. C) Boxplot showing the genome GC-content as a function of associated plasmids. D) Boxplots of genome GC-content in genomes associated with plasmids as a function of the presence/absence of CRISPRs. Wilcoxon rank sum tests were used to compare between groups. Level of significance: *** $P<0.001$; ** $P<0.01$; * $P<0.05$; NS. $P\geq0.05$.

Supplementary Table 1. The dataset of prokaryotic genomes and their associations with phages and plasmids (part).

Taxon	Size (mb)	GC%	Phage	Plasmid	Crispr
9	0.456703	28.258	No	Yes	No
24	4.38446	44.4	Yes	No	No
43	12.3497	68.5	No	No	No
48	12.4894	69.4	No	No	Yes
52	11.3881	68.7	Yes	No	Yes
56	14.5576	71.7	No	No	No
63	3.78755	63.4916	No	Yes	No
69	6.09602	69.4	No	No	No
114	8.99889	67.4	No	No	No
139	1.30155	28.5082	No	Yes	No
159	3.03465	27.0438	No	Yes	No

Note: Taxon represents NCBI taxon ID of prokaryotes; Size is the genomic size of prokaryote; GC% represents the genomic GC-content of the host genome; Phage, Plasmid and CRISPR represent whether the host genomes are associated with plasmids, phages, and CRISPR, respectively.

Supplementary Table 2. Relative importance of various factors for GC-content (GC%) in a LM.

Dataset	Factors	Coefficient	*P*-value	Relative importance
All	size	4.134	$< 2x10^{-16}$	98.10%
	plasmidNumber	-0.638	$4.3x10^{-15}$	1.67%
	phageNumber	-0.037	$3.4x10^{-3}$	0.23%
	CRISPR	-0.131	0.646	0.01%
No plasmids	size	4.179	$< 2x10^{-16}$	99.84%
	phageNumber	-0.029	0.042	0.16%
	CRISPR	0.097	0.774	0.00%
No phages	size	4.219	$< 2x10^{-16}$	98.57%
	plasmidNumber	-0.636	$9.7x10^{-7}$	1.15%
	CRISPR	1.048	0.015	0.28%

Note: The equation used in the linear model (LM) is GC% ~ size + plasmidNumber + phageNumber + CRISPR. Here, plasmidNumber and phageNumber represent the number of plasmids and phages in genomes, respectively.

3.3 Manuscript 3. MVP: a microbe-phage interaction database

Na L Gao [1,2†], Chengwei Zhang [3,4†], Zhanbing Zhang [1], Songnian Hu [3], Martin J Lercher[2], Xing-Ming Zhao [5], Peer Bork [6, 7, 8, 9§], Zhi Liu1[§], Wei-Hua Chen [1§]

[1] Key Laboratory of Molecular Biophysics of the Ministry of Education, Hubei Key Laboratory of Bioinformatics and Molecular-imaging, Department of Bioinformatics and Systems Biology, College of Life Science and Technology, Huazhong University of Science and Technology (HUST), 430074 Wuhan, Hubei, China

[2] Institute for Computer Science and Cluster of Excellence on Plant Sciences CEPLAS, Heinrich Heine University, 40225 Düsseldorf, Germany.

[3] CAS Key Laboratory of Genome Sciences and Information, Beijing Institute of Genomics (BIG), Chinese Academy of Sciences (CAS), No.7 Beitucheng West Road, Chaoyang District, 100029 Beijing, PR China,

[4] University of Chinese Academy of Sciences, Beijing 100049, China

[5] Institute of Science and Technology for Brain-Inspired Intelligence (ISTBI), Fudan University, Office 2304, East Main Building of Guanghua Towers, 220 Handan Road, Shanghai 200433, China

[6] European molecular biology laboratory (EMBL), Meyerhofstrasse 1, 69117 Heidelberg, Germany

[7] Molecular Medicine Partnership Unit, University of Heidelberg and European Molecular Biology Laboratory, 69120 Heidelberg, Germany

[8] Max-Delbrück-Centre for Molecular Medicine, Robert-Rössle-Straße 10, 13125 Berlin, Germany

[9] Department of Bioinformatics, Biocenter, University of Würzburg, 97074 Würzburg, Germany.

[§] To whom correspondence should be addressed to Wei-Hua Chen. Tel: +862787542127; Fax: +862787542527; Email: weihuachen@hust.edu.cn; correspondence should also be addressed to: Zhi Liu (zhiliu@hust.edu.cn) or Peer Bork (bork@embl.de).

† These authors contributed equally to the paper as first authors.

This paper has been published in the journal *Nucleic Acids Research* (Published online 2017 Nov 21. doi: [10.1093/nar/gkx1124]; PMID: 29177508). As the first author, I assembled data from public databases with help from Prof. Chen and other members of the lab; I performed all the data analyses, interpreted the data together with Prof. Chen and Prof. Lercher, wrote the first draft of the manuscript, and edited the final manuscript together with Prof. Chen.

Nucleic Acids Research, 2017 **1**
doi: 10.1093/nar/gkx1124

MVP: a microbe–phage interaction database

Na L. Gao[1,2,†], Chengwei Zhang[3,4,†], Zhanbing Zhang[1], Songnian Hu[3], Martin J. Lercher[2], Xing-Ming Zhao[5], Peer Bork[6,7,8,9,*], Zhi Liu[1,*] and Wei-Hua Chen[1,*]

[1]Key Laboratory of Molecular Biophysics of the Ministry of Education, Hubei Key Laboratory of Bioinformatics and Molecular-imaging, Department of Bioinformatics and Systems Biology, College of Life Science and Technology, Huazhong University of Science and Technology (HUST), 430074 Wuhan, Hubei, China, [2]Institute for Computer Science and Cluster of Excellence on Plant Sciences CEPLAS, Heinrich Heine University, 40225 Düsseldorf, Germany, [3]CAS Key Laboratory of Genome Sciences and Information, Beijing Institute of Genomics (BIG), Chinese Academy of Sciences (CAS), No.7 Beitucheng West Road, Chaoyang District, 100029 Beijing, PR China, [4]University of Chinese Academy of Sciences, Beijing 100049, China, [5]Institute of Science and Technology for Brain-Inspired Intelligence (ISTBI), Fudan University, Office 2304, East Main Building of Guanghua Towers, 220 Handan Road, Shanghai 200433, China, [6]European molecular biology laboratory (EMBL), Meyerhofstrasse 1, 69117 Heidelberg, Germany, [7]Molecular Medicine Partnership Unit, University of Heidelberg and European Molecular Biology Laboratory, 69120 Heidelberg, Germany, [8]Max-Delbrück-Centre for Molecular Medicine, Robert-Rössle-Straße 10, 13125 Berlin, Germany and [9]Department of Bioinformatics, Biocenter, University of Würzburg, 97074 Würzburg, Germany

Received August 15, 2017; Revised October 5, 2017; Editorial Decision October 22, 2017; Accepted November 19, 2017

ABSTRACT

Phages invade microbes, accomplish host lysis and are of vital importance in shaping the community structure of environmental microbiota. More importantly, most phages have very specific hosts; they are thus ideal tools to manipulate environmental microbiota at species-resolution. The main purpose of *MVP* (*M*icrobe *V*ersus *P*hage) is to provide a comprehensive catalog of phage–microbe interactions and assist users to select phage(s) that can target (and potentially to manipulate) specific microbes of interest. We first collected 50 782 viral sequences from various sources and clustered them into 33 097 unique viral clusters based on sequence similarity. We then identified 26 572 interactions between 18 608 viral clusters and 9245 prokaryotes (i.e. bacteria and archaea); we established these interactions based on 30 321 evidence entries that we collected from published datasets, public databases and re-analysis of genomic and metagenomic sequences. Based on these interactions, we calculated the host range for each of the phage clusters and accordingly grouped them into subgroups such as 'species-', 'genus-' and 'family-' specific phage clusters. *MVP* is equipped with a modern, responsive and intuitive interface, and is freely available at: http://mvp.medgenius.info.

INTRODUCTION

It has been increasingly recognized that microbiome can play crucial roles in human health (1–3), diseases (4–10), responses to drugs and treatments (11,12), development (13–15) and many other aspects of human life (16–19). However, due to limited availability of tools that enable researchers to manipulate microbiome, it is often difficult to directly infer causal relationships from the correlated alterations in microbial community structures and host phenotypes (e.g. health statuses) under different conditions (20–23). Experimental procedures such as fecal microbiota transplantation (24,25) and/or the use of germ-free mice (3,26) can be used to identify and validate causal factors, but they are neither easy nor cheap. Furthermore, due to the lack of general purpose tools that could manipulate microbiota at species level, it is difficult to directly pinpoint the causal species.

Phages are known to be key players in microbial communities; they could invade microbes, accomplish host lysis and are of vital importance in shaping the community structure of human and environmental microbiota (27–29). More importantly, phages could provide potential tools for the precision manipulation of environmental microbiota: it is known that phages have rather narrow host ranges, mostly at the species or genes levels (30); they are thus ideal

*To whom correspondence should be addressed. Tel: +86 278 754 2127; Fax: +86 278 754 2527; Email: weihuachen@hust.edu.cn
Correspondence may also be addressed to Zhi Liu. Email: zhiliu@hust.edu.cn
Correspondence may also be addressed to Peer Bork. Email: bork@embl.de
†These authors contributed equally to the paper as first authors.

tools to target (and eliminate) specific microbes at species-resolution while avoid potential 'off-target' effects. A recent study provided us with a great example for such an application; Yen *et. al.* successfully reduced *Vibrio cholerae* infection and colonization in the intestinal tract and prevents cholera-like diarrhea, by orally administrating *V. cholera*-specific phages in model animals (31).

We thus developed *MVP*—a microbe-phage interaction database (*MVP* stands for *M*icrobe *V*ersus *P*hage), with the main aims being to provide researchers with a comprehensive catalog of phage–microbe interactions and assist them to select phage(s) that can target (and potentially to manipulate) specific microbes of interest.

In addition to experimental methods, microbe–phage interactions can be identified by taking advantage of the large-scale genomic- and metagenomic sequencing efforts. For example, it is known that many phages insert their genomes into that of their hosts; the integrated phages are known as prophages (32,33). Many computational tools exist and are able to identify prophages from complete prokaryotic genomes and/or assembled metagenomic contigs (34–36). In addition, CRISPR spacer sequences can also be used to infer host–phage interactions (37,38), although their short lengths (usually 24–50 bp) in nature make it difficult to reliably determine their source phages (27,37).

In this study, we obtained in total 50 782 viral sequences from various sources and assembled them into 33 097 unique viral clusters. We identified 26 572 interactions between 18 608 viral clusters and 9245 prokaryotes, and calculated the host range for each of the phage clusters accordingly. We presented these data and related information in an online database MVP (Microbe Versus Phage); we designed MVP to be a modern website with a responsive and intuitive interface, and incorporated many widgets (i.e. functional elements of a web page that serve specific purposes) that enables users to effortlessly explore all contents and find what they are interested in.

DATA GENERATION

Viral sequences and clustering them into viral clusters

We obtained viral sequences from the following four sources.

First, we downloaded all available viral sequences from the NCBI viral genomes resource (39).

Second, we identified putative prophage sequences from complete bacterial and archaeal genomes downloaded from the NCBI prokaryotic reference genome database (40) and EMBL proGenomes database (41).

Third, we identified putative prophage sequences from assembled metagenomic sequences derived from the human gut. We included in the current version of *MVP* two human gut metagenomic datasets containing 124 (1) and 1267 (42) human fecal samples respectively that we downloaded from the EBI metagenomic database (43). Prophage identification was carried out using a phage_finder (34) tool v2.1 (last updated: 26 Oct 26 2011) with default parameters.

Last, we included viral and prophage sequences from several published datasets (44,45), including those from a 'Uncovering Earth's virome' project, and the International Committee on Taxonomy of Viruses (https://talk.ictvonline.org; ICTV). Worth to mention is the recent work by Roux *et al.*; by using a virus/prophage identification tool VirSorter that they developed (36), they identified in total 12 498 high-confidence viral genomes by scanning the publicly available bacterial and archaeal genomic sequences. These newly identified viral sequences were either prophages or un-incorporated viral sequences that were previously annotated as plasmids (45).

In total we collected 50 782 viral sequences from these sources. We next used a cd-hit-est program (46) to cluster them into clusters based on sequence similarities. As previously suggested (27), the following options of cd-hit-est were used: -c 0.95 and -aS 0.85. The '-c' option specifies the sequence identity threshold and is calculated as the number of identical nucleotides in alignment divided by the full length of the shorter sequence, while the '-aS' option specifies alignment coverage threshold and is defined as the proportion of shorter sequence covered by the alignment. Sequences in alignments with measurements above these thresholds are clustered; the longest sequences in a cluster is chosen as representative of the cluster. Please note that the much relaxed parameter '–aS 0.85' for clustering may not be used as a general-purpose threshold for viral studies because it could result in very inclusive cluster, but it suits our purpose nicely: with MVP we aimed to facilitate users to select phages that can specifically target a bacterium, therefore any phages with (putative) broad host-ranges should be marked and removed from the candidate list. A further relaxed threshold of '-c 0.8 –aS 0.85' was also tested and resulted in ∼3% few clusters, suggesting that the viral clusters we obtained in this study were relatively stable.

In sum, we obtained 33 097 clusters from the 50 782 viral sequences.

We checked the overlap in phages from different sources. We found only a small proportion (∼19.5%) of phages were covered by multiple evidence (i.e. the same prophage sequence can be identified from multiple (meta-) genomic sequences); even lower proportion (∼9%) of the total phage clusters were covered by multiple data-sources. However, within a data source, the phage overlap ratios vary significantly; more importantly, they seem to correlate with the number of samples taken from the same niche environment (Table 1). For example, 57.4% of the identified phages are covered multiple times in the 'Uncovering Earth's virome' (44), which collected over 3000 samples around the world; this ratio is followed by 18.67% in the human gut, which in total ∼1700 samples were used to identified the phages (1,42). Conversely, the overlap ratio in the EMBL proGenomes database is only ∼0.6%, mainly due to the fact that only 'representative' genomes were presented in the dataset we used and the 'redundant' genomes were excluded (41). Thus the low overlap ratios in some data sources are mainly because of the diverse environments from which the genomes were sampled. These results further confirmed that phages indeed could have very narrow host range.

Interactions between viral clusters and microbes

In this study we focused on prokaryotes (i.e. bacteria and archaea), and used prokaryotes and microbes interchangeably, although the latter can also include eukary-

Table 1. Overlaps in phages within data-sources

Data source	# clusters	% overlap *	Notes
'Earth's virome' project (44)	5412	57.4%	Over 3000 samples were sequenced; most are environmental samples
Predicted prophages in human gut (1,42)	1505	18.67%	~1700 fecal samples from two gut metagenomic studies (1,42)
Predicted viral and prophage sequences from complete and draft genomes (36)	7117	18.07%	
Predicted prophages from NCBI complete genomes (40)	6964	15.4%	All available complete prokaryotic genomes (as of May 2017)
NCBI reference viral genome database (39)	776	0.64%	
Predicted prophages from EMBL proGenomes database (41)	3275	0.61%	Representative complete prokaryotic genomes (as of May 2017)
ICTV	668	0	Data obtained from the International Committee on Taxonomy of Viruses (https://talk.ictvonline.org; ICTV)

* within each data-source, the overlap ratio is defined as proportion of phage clusters containing multiple sequences from the data source, out of the total phage clusters containing any number of sequences from the same data source.

Table 2. Overlaps in host prokaryotes

Data source	# hosts	% overlap with other data sources*
ICTV	11	100%
'Earth's virome' project (44)	1247	79.4%
Predicted prophages from EMBL proGenomes database (41)	2549	78.6%
Predicted prophages from NCBI complete genomes (40)	4398	68.18%
Predicted prophages in human gut (1,42)	210	67.61%
NCBI reference viral genome database (39)	282	56.73%
Predicted viral and prophage sequences from complete and draft genomes (36)	6388	56.6%

* the overlap ratio is defined as proportion of hosts in a data source that could also found in any of the other data sources.

otic microbes. We also used viral- and phage- clusters interchangeably, under the circumstances that a virus invades a prokaryotic microbe.

We inferred interactions between viral-/phage- clusters and microbes from the following four sources.

First, we established phage-host relationships by extracting the 'host' fields from the annotation files downloaded from the NCBI reference viral genome database (39).

Second, we could easily establish the phage-host relationships for prophages identified in reference prokaryotic genomes.

Third, for prophages identified from assembled metagenomic contigs, their host information are not readily available. Therefore for each of the identified prophages, we first extracted the two flanking sequences from the contig, and submitted them as queries for BLAST searches (47) against prokaryotic reference genomes. We required that each flanking sequence should be at least 200 bp in size and at least 50 bp apart from the putative prophage. Predicted phages with flanking sequences shorter than 250 bp on either sides were discarded. We filtered out BLAST hits that had sequence similarity less than 0.95 or covered <80% of the query sequences. If there was only one hit left for a query, we used the corresponding species of the hit sequence as the putative host. For queries that matched multiple hits above the thresholds, we calculated the last common ancestor (LCA) of all hits in the NCBI taxonomic database using an in-house Perl script; we kept LCAs that had taxonomic ranking of genus or species according to the NCBI taxonomy database (40). Metagenomic sequences are a mixture of multiple species and are often highly fragmented. In addition, lateral gene transfers frequently occur and contribute significantly to the expansion of gene repertoire in prokaryotes (48). Together these factors make it technically challenging to accurately assemble metagenomic sequences (49–51). Therefore to reduce possible false-positive results, at the end we only kept the host–phage relationships if the identified hosts met the two following criteria: (i) both flanking sequences should match to some reference genomes, and (ii) the taxonomy ranks of the BLAST hits of the two flanking sequences should be the same.

To determine the error rate in host species identification using metagenomic data, we run the following simulations: we took randomly two fragments from a host genome, searched them against the NCBI prokaryotic sequence database using BLAST (47), and run the above analysis pipeline to determine the their species identity. We did this ten times for each of the complete prokaryotic genomes. At the species level, we obtained an overall accuracy rate of 95% with ~90% sensitivity. However, when we removed the 'source' genome (i.e. the genome from which the two fragments were taken) from the analysis, the overall accuracy rate dropped to ~79% at the species level with ~50% sensitivity (i.e. about half of the queries were removed because of no significant BLAST hits in the genome, or the species assignment was ambiguous).

Last, we also obtained phage-host associations from published datasets (44,45) and databases such as the International Committee on Taxonomy of Viruses (ICTV; https://talk.ictvonline.org).

In total, we identified 30 321 host–phage associations, corresponding to 26 572 unique interactions between 18 608 viral clusters and 9245 prokaryotes. We summarized in Figure 1 the distribution of the 9245 prokaryotic hosts across

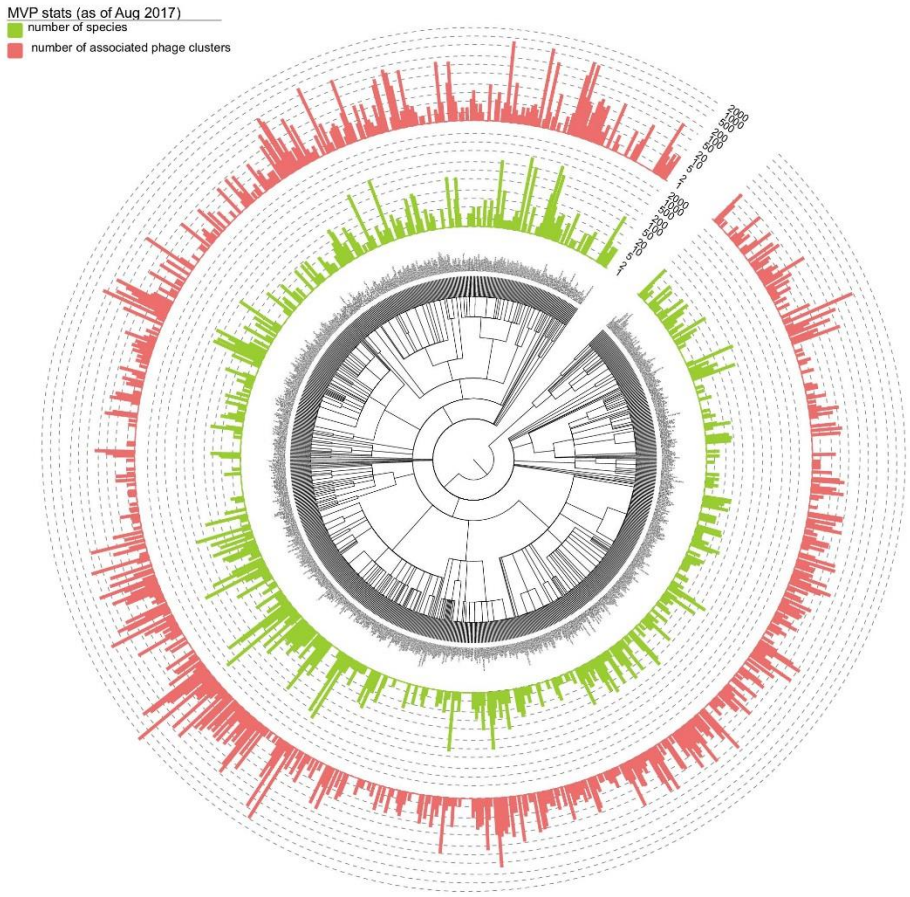


Figure 1. Distribution of the 9245 prokaryotic hosts across the bacterial and archaeal phylogeny at the genus level according to NCBI taxonomy and their associated phage clusters. For each bacterial and archaeal genus-level group, the number daughter species collected in *MVP* and the corresponding number of associated virial clusters (unique count) are indicated with light-green and red bars. Bacterial and archaeal species that are not collected in *MVP* are not shown. Bar heights are log-transformed. The tree and the datasets were visualized using Evolview, an online visualization and management tool for customized and annotated phylogenetic trees (55). An interactive version of the tree can be found at: http://www.evolgenius.info/evolview/#shared/mvp2017_stats/462.

the bacterial and archaeal phylogeny at the genus level and their associated phage clusters.

We also check the overlap of prokaryotic hosts among different data sources. We found that 44.35% of the hosts were found in at least two data sources. We summarized in Table 2 the overlaps between each data source with all others.

In addition, 61.09% hosts associate with multiple phage clusters.

Calculation of host ranges of phage clusters

One of the main aim of *MVP* is to provide researchers with a list of phages that can specifically target certain bacteria of interests while avoid any 'off-target' effects. To achieve this, we calculated the host range for each of the phage clus-

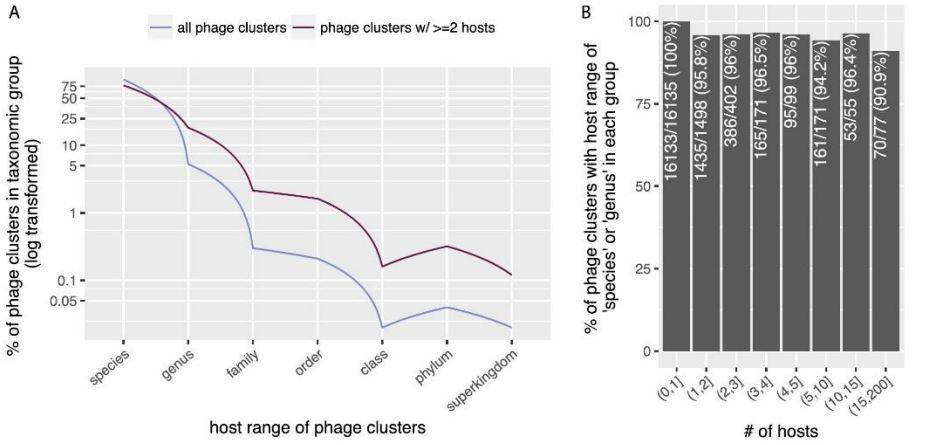


Figure 2. Most phage clusters have rather narrow host ranges. For phage clusters with at least two hosts, their host ranges were calculated as the LCAs in the NCBI taxonomic database (see 'Data Generation' for more details). (**A**) *X*-axis: host range of phage clusters, *Y*-axis: percentage of phage clusters (out of total) with their LCAs in the taxonomic groups. The *Y*-axis has been log-transformed. (**B**) *X*-axis: number of hosts (i.e. phage clusters were grouped into bins according to the numbers of hosts they have); '(5,10]' specifies a subgroup in which phage clusters have >5 and ≤10 hosts. *Y*-axis, percentage of phage clusters (in each bin) that have host ranges at the 'species' or 'genus' levels in each subgroup.

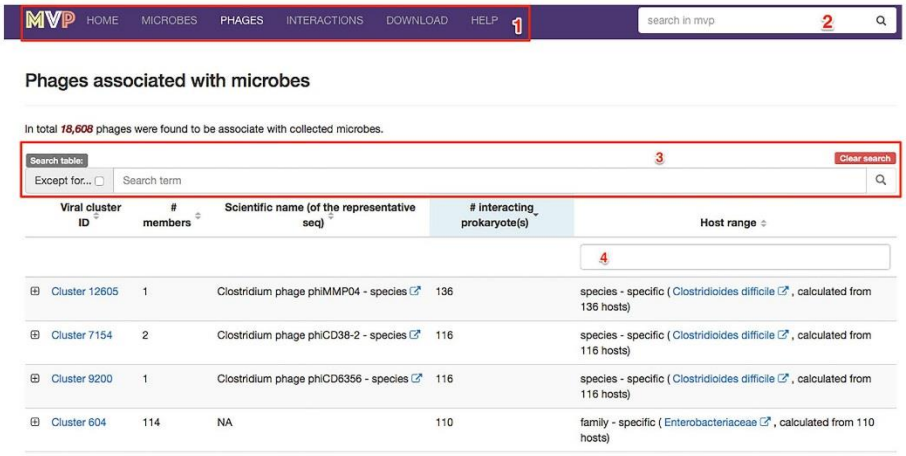


Figure 3. A screenshot of the 'Phages' page; highlighted are built-in widgets (i.e. functional elements of a web page that serve specific purposes) that enables users to easily find what they are interested. (1) a navigation toolbar that floats on top of the page, allowing users to access our data in pre-organized categories (i.e. 'microbes', 'phages' and 'interactions' and etc.); (2) a global search widget that enables uses to search for microbes and virial clusters with any information, including the taxonomy IDs, scientific names and taxonomic ranks, and then redirect to the corresponding page that the users choose; (3) a set of widgets allowing users to search for (or filter out when the 'Except for… ' checkbox is selected) the contents of the table below (a list of phages in *MVP* in this case) with any keywords; (4) a widget allowing users to filter for phage clusters according to the values in the column of 'Host range'.

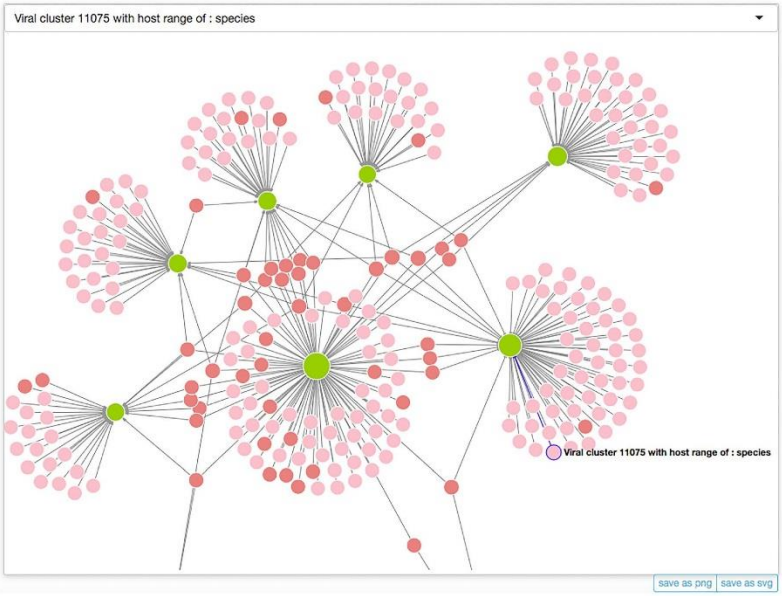


Figure 4. A screenshot of the interaction network (only partial) visualized with our built-in visualization tool. Microbes and phage clusters are visualized as light green and pink/reddish circles, respectively, with their sizes (diameters) being propositional to the numbers of the interacting partners (including also those that may not be shown in the visualization). Two colors, namely pink and reddish are used for phages, in order to distinguish those that infect only one host (pink) from those that infect multiple hosts (reddish). Click the text-labels next to the circles, users will be redirect the page for the corresponding microbe or phage cluster. In addition to the canvas, two additional widgets are also provided. The first is the selector at the top of the canvas, from which users can browse or search for a node of interests, select it from the drop-down menu and highlight it and bring it into the middle of the canvas. The other includes two buttons that can be used to export the visualization to an external file in either SVG or PNG format. For more information please consult the Interactions page (http://mvp.medgenius.info/interactions).

ters collected in *MVP*. For a phage cluster that infects only one host, we defined the host range as the taxonomic rank of the host in the NCBI taxonomy database; for a cluster that infects multiple hosts, we defined the host range as the taxonomic rank of the LCA of all its hosts in the NCBI taxonomic database.

As shown in Figure 2, we found that more than 99% phage clusters have host range at the 'species' or 'genus' levels. Excluding those with only one host (Figure 2A), or considering phage clusters with certain numbers of hosts (Figure 2B), the results remained largely the same, i.e. more than 90% of the remaining clusters have host range at the 'species' or 'genus' levels. These results are consistent with previous findings that phages often have very narrow host range (30), and further confirmed the high-quality of our data.

WEB INTERFACE OF *MVP*

We provided *MVP* with a modern, responsive and intuitive interface. As explained in Figure 3, the design of the web pages, especially the use of a few powerful search widgets would allow users to easily find what they are interested in.

We also incorporated into *MVP* a powerful network visualization tool that allows users to interactively visualize, interact and explore phage-host associations collected in our database. Please consult the Interactions page (http://mvp.medgenius.info/interactions) for details; shown in Figure 4 is a screenshot of the interaction network.

DATA ACCESS

All data are freely accessible to all academic users. This work is licensed under a Creative Commons Attribution 3.0 Unported License (CC BY 3.0). Users can download combined data from the 'DOWNLOAD' page. Users can also download data for individual viral clusters from the 'PHAGES' page.

FUTURE DIRECTIONS

During the development of *MVP* we came across numerous resources and tools that would make our database

more complete and better. Also due to limitations of current methods, we wish to thoroughly test and benchmark existing tools/analysis pipeline before we include their results into MVP. Therefore our plans for the near future will include: (i) to use more tools, especially those that were recently developed for the identification of prophage and viral sequences, including virFinder (52), PHASTER (35) and VirSorter (36); (ii) to include more metagenomics datasets from the EBI Metagenomic database (43), (iii) to infer and include putative host–phage interactions from CRISPR-spacer sequences; the latter can also be used to infer bacterial-/archaeal- resistance to phages, and is a vitally important player in the phage-host interaction network and (iv) to compile sets of microbes according to their niche environments (i.e. soil or human gut), and recalculate host-ranges for phage clusters that could interact with them. Finally, it has been shown that virus and their host genomes often share certain similar genomic features such as oligonucleotide frequency patterns (53,54). We will thus also include such measurements for the phage–host interactions in *MVP* calculated from existing tools such as VirHostMatcher (54).

FUNDING

National Natural Science Foundation of China [31770132, 81572050 to Z.L.].
Conflict of interest statement. None declared.

REFERENCES

1. Arumugam,M., Raes,J., Pelletier,E., Le Paslier,D., Yamada,T., Mende,D.R., Fernandes,G.R., Tap,J., Bruls,T., Batto,J.M. *et al.* (2011) Enterotypes of the human gut microbiome. *Nature*, **473**, 174–180.
2. Qin,J., Li,R., Raes,J., Arumugam,M., Burgdorf,K.S., Manichanh,C., Nielsen,T., Pons,N., Levenez,F., Yamada,T. *et al.* (2010) A human gut microbial gene catalogue established by metagenomic sequencing. *Nature*, **464**, 59–65.
3. Turnbaugh,P.J., Ley,R.E., Mahowald,M.A., Magrini,V., Mardis,E.R. and Gordon,J.I. (2006) An obesity-associated gut microbiome with increased capacity for energy harvest. *Nature*, **444**, 1027–1031.
4. Qin,N., Yang,F., Li,A., Prifti,E., Chen,Y., Shao,L., Guo,J., Le Chatelier,E., Yao,J., Wu,L. *et al.* (2014) Alterations of the human gut microbiome in liver cirrhosis. *Nature*, **513**, 59–64.
5. Pedersen,H.K., Gudmundsdottir,V., Nielsen,H.B., Hyotylainen,T., Nielsen,T., Jensen,B.A., Forslund,K., Hildebrand,F., Prifti,E., Falony,G. *et al.* (2016) Human gut microbes impact host serum metabolome and insulin sensitivity. *Nature*, **535**, 376–381.
6. Qin,J., Li,Y., Cai,Z., Li,S., Zhu,J., Zhang,F., Liang,S., Zhang,W., Guan,Y., Shen,D. *et al.* (2012) A metagenome-wide association study of gut microbiota in type 2 diabetes. *Nature*, **490**, 55–60.
7. Noguera-Julian,M., Rocafort,M., Guillen,Y., Rivera,J., Casadella,M., Nowak,P., Hildebrand,F., Zeller,G., Parera,M., Bellido,R. *et al.* (2016) Gut microbiota linked to sexual preference and HIV infection. *EBioMedicine*, **5**, 135–146.
8. Frye,R.E., Slattery,J., MacFabe,D.F., Allen-Vercoe,E., Parker,W., Rodakis,J., Adams,J.B., Krajmalnik-Brown,R., Bolte,E., Kahler,S. *et al.* (2015) Approaches to studying and manipulating the enteric microbiome to improve autism symptoms. *Microb. Ecol. Health Dis.*, **26**, 26878.
9. Hsiao,E.Y., McBride,S.W., Hsien,S., Sharon,G., Hyde,E.R., McCue,T., Codelli,J.A., Chow,J., Reisman,S.E., Petrosino,J.F. *et al.* (2013) Microbiota modulate behavioral and physiological abnormalities associated with neurodevelopmental disorders. *Cell*, **155**, 1451–1463.
10. Li,J., Zhao,F., Wang,Y., Chen,J., Tao,J., Tian,G., Wu,S., Liu,W., Cui,Q., Geng,B. *et al.* (2017) Gut microbiota dysbiosis contributes to the development of hypertension. *Microbiome*, **5**, 14.
11. Yu,T., Guo,F., Yu,Y., Sun,T., Ma,D., Han,J., Qian,Y., Kryczek,I., Sun,D., Nagarsheth,N. *et al.* (2017) Fusobacterium nucleatum promotes chemoresistance to colorectal cancer by modulating autophagy. *Cell*, **170**, 548–563.
12. Forslund,K., Hildebrand,F., Nielsen,T., Falony,G., Le Chatelier,E., Sunagawa,S., Prifti,E., Vieira-Silva,S., Gudmundsdottir,V., Pedersen,H.K. *et al.* (2015) Disentangling type 2 diabetes and metformin treatment signatures in the human gut microbiota. *Nature*, **528**, 262–266.
13. Backhed,F., Roswall,J., Peng,Y., Feng,Q., Jia,H., Kovatcheva-Datchary,P., Li,Y., Xia,Y., Xie,H., Zhong,H. *et al.* (2015) Dynamics and stabilization of the human gut microbiome during the first year of life. *Cell Host Microbe*, **17**, 690–703.
14. Forsgren,M., Isolauri,E., Salminen,S. and Rautava,S. (2017) Late preterm birth has direct and indirect effects on infant gut microbiota development during the first six months of life. *Acta Paediatr.*, **106**, 1103–1109.
15. Wall,R., Ross,R.P., Ryan,C.A., Hussey,S., Murphy,B., Fitzgerald,G.F. and Stanton,C. (2009) Role of gut microbiota in early infant development. *Clin. Med. Pediatr.*, **3**, 45–54.
16. Komaroff,A.L. (2017) The microbiome and risk for obesity and diabetes. *JAMA*, **317**, 355–356.
17. Mayer,E.A., Tillisch,K. and Gupta,A. (2015) Gut/brain axis and the microbiota. *J. Clin. Invest.*, **125**, 926–938.
18. Alcock,J., Maley,C.C. and Aktipis,C.A. (2014) Is eating behavior manipulated by the gastrointestinal microbiota? Evolutionary pressures and potential mechanisms. *Bioessays*, **36**, 940–949.
19. Fujimura,K.E. and Lynch,S.V. (2015) Microbiota in allergy and asthma and the emerging relationship with the gut microbiome. *Cell Host Microbe*, **17**, 592–602.
20. Harley,I.T. and Karp,C.L. (2012) Obesity and the gut microbiome: striving for causality. *Mol. Metab.*, **1**, 21–31.
21. Zhao,L. (2013) The gut microbiota and obesity: from correlation to causality. *Nat. Rev. Microbiol.*, **11**, 639–647.
22. Fritz,J.V., Desai,M.S., Shah,P., Schneider,J.G. and Wilmes,P. (2013) From meta-omics to causality: experimental models for human microbiome research. *Microbiome*, **1**, 14.
23. Saraswati,S. and Sitaraman,R. (2014) Aging and the human gut microbiota-from correlation to causality. *Front. Microbiol.*, **5**, 764.
24. Li,S.S., Zhu,A., Benes,V., Costea,P.I., Hercog,R., Hildebrand,F., Huerta-Cepas,J., Nieuwdorp,M., Salojarvi,J., Voigt,A.Y. *et al.* (2016) Durable coexistence of donor and recipient strains after fecal microbiota transplantation. *Science*, **352**, 586–589.
25. Borody,T.J., Paramsothy,S. and Agrawal,G. (2013) Fecal microbiota transplantation: indications, methods, evidence, and future directions. *Curr. Gastroenterol. Rep.*, **15**, 337.
26. Charbonneau,M.R., O'Donnell,D., Blanton,L.V., Totten,S.M., Davis,J.C., Barratt,M.J., Cheng,J., Guruge,J., Talcott,M., Bain,J.R. *et al.* (2016) Sialylated milk oligosaccharides promote microbiota-dependent growth in models of infant undernutrition. *Cell*, **164**, 859–871.
27. Waller,A.S., Yamada,T., Kristensen,D.M., Kultima,J.R., Sunagawa,S., Koonin,E.V. and Bork,P. (2014) Classification and quantification of bacteriophage taxa in human gut metagenomes. *ISME J.*, **8**, 1391–1402.
28. Roux,S., Brum,J.R., Dutilh,B.E., Sunagawa,S., Duhaime,M.B., Loy,A., Poulos,B.T., Solonenko,N., Lara,E., Poulain,J. *et al.* (2016) Ecogenomics and potential biogeochemical impacts of globally abundant ocean viruses. *Nature*, **537**, 689–693.
29. Ogilvie,L.A. and Jones,B.V. (2015) The human gut virome: a multifaceted majority. *Front. Microbiol.*, **6**, 918.
30. Hyman,P. and Abedon,S.T. (2010) Bacteriophage host range and bacterial resistance. *Adv. Appl. Microbiol.*, **70**, 217–248.
31. Yen,M., Cairns,L.S. and Camilli,A. (2017) A cocktail of three virulent bacteriophages prevents Vibrio cholerae infection in animal models. *Nat. Commun.*, **8**, 14187.
32. Krupovic,M., Prangishvili,D., Hendrix,R.W. and Bamford,D.H. (2011) Genomics of bacterial and archaeal viruses: dynamics within the prokaryotic virosphere. *Microbiol. Mol. Biol. Rev.*, **75**, 610–635.
33. Fortier,L.C. and Sekulovic,O. (2013) Importance of prophages to evolution and virulence of bacterial pathogens. *Virulence*, **4**, 354–365.

34. Fouts,D.E. (2006) Phage_Finder: automated identification and classification of prophage regions in complete bacterial genome sequences. *Nucleic Acids Res.*, **34**, 5839–5851.
35. Arndt,D., Grant,J.R., Marcu,A., Sajed,T., Pon,A., Liang,Y. and Wishart,D.S. (2016) PHASTER: a better, faster version of the PHAST phage search tool. *Nucleic Acids Res.*, **44**, W16–W21.
36. Roux,S., Enault,F., Hurwitz,B.L. and Sullivan,M.B. (2015) VirSorter: mining viral signal from microbial genomic data. *PeerJ*, **3**, e985.
37. Stern,A., Mick,E., Tirosh,I., Sagy,O. and Sorek,R. (2012) CRISPR targeting reveals a reservoir of common phages associated with the human gut microbiome. *Genome Res.*, **22**, 1985–1994.
38. Wang,J., Gao,Y. and Zhao,F. (2016) Phage-bacteria interaction network in human oral microbiome. *Environ. Microbiol.*, **18**, 2143–2158.
39. Brister,J.R., Ako-Adjei,D., Bao,Y. and Blinkova,O. (2015) NCBI viral genomes resource. *Nucleic Acids Res.*, **43**, D571–D577.
40. Coordinators,N.R. (2017) Database resources of the national center for biotechnology information. *Nucleic Acids Res.*, **45**, D12–D17.
41. Mende,D.R., Letunic,I., Huerta-Cepas,J., Li,S.S., Forslund,K., Sunagawa,S. and Bork,P. (2017) proGenomes: a resource for consistent functional and taxonomic annotations of prokaryotic genomes. *Nucleic Acids Res.*, **45**, D529–D534.
42. Li,J., Jia,H., Cai,X., Zhong,H., Feng,Q., Sunagawa,S., Arumugam,M., Kultima,J.R., Prifti,E., Nielsen,T. *et al.* (2014) An integrated catalog of reference genes in the human gut microbiome. *Nat. Biotechnol.*, **32**, 834–841.
43. Mitchell,A., Bucchini,F., Cochrane,G., Denise,H., ten Hoopen,P., Fraser,M., Pesseat,S., Potter,S., Scheremetjew,M., Sterk,P. *et al.* (2016) EBI metagenomics in 2016–an expanding and evolving resource for the analysis and archiving of metagenomic data. *Nucleic Acids Res.*, **44**, D595–D603.
44. Paez-Espino,D., Eloe-Fadrosh,E.A., Pavlopoulos,G.A., Thomas,A.D., Huntemann,M., Mikhailova,N., Rubin,E., Ivanova,N.N. and Kyrpides,N.C. (2016) Uncovering Earth's virome. *Nature*, **536**, 425–430.
45. Roux,S., Hallam,S.J., Woyke,T. and Sullivan,M.B. (2015) Viral dark matter and virus-host interactions resolved from publicly available microbial genomes. *Elife*, **4**, doi:10.7554/eLife.08490.
46. Fu,L., Niu,B., Zhu,Z., Wu,S. and Li,W. (2012) CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics*, **28**, 3150–3152.
47. Camacho,C., Coulouris,G., Avagyan,V., Ma,N., Papadopoulos,J., Bealer,K. and Madden,T.L. (2009) BLAST+: architecture and applications. *BMC Bioinformatics*, **10**, 421.
48. Treangen,T.J. and Rocha,E.P. (2011) Horizontal transfer, not duplication, drives the expansion of protein families in prokaryotes. *PLoS Genet.*, **7**, e1001284.
49. Ji,P., Zhang,Y., Wang,J. and Zhao,F. (2017) MetaSort untangles metagenome assembly by reducing microbial community complexity. *Nat. Commun.*, **8**, 14306.
50. Nielsen,H.B., Almeida,M., Juncker,A.S., Rasmussen,S., Li,J., Sunagawa,S., Plichta,D.R., Gautier,L., Pedersen,A.G., Le Chatelier,E. *et al.* (2014) Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes. *Nat. Biotechnol.*, **32**, 822–828.
51. Albertsen,M., Hugenholtz,P., Skarshewski,A., Nielsen,K.L., Tyson,G.W. and Nielsen,P.H. (2013) Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes. *Nat. Biotechnol.*, **31**, 533–538.
52. Ren,J., Ahlgren,N.A., Lu,Y.Y., Fuhrman,J.A. and Sun,F. (2017) VirFinder: a novel k-mer based tool for identifying viral sequences from assembled metagenomic data. *Microbiome*, **5**, 69.
53. Edwards,R.A., McNair,K., Faust,K., Raes,J. and Dutilh,B.E. (2016) Computational approaches to predict bacteriophage-host relationships. *FEMS Microbiol. Rev.*, **40**, 258–272.
54. Ahlgren,N.A., Ren,J., Lu,Y.Y., Fuhrman,J.A. and Sun,F. (2017) Alignment-free d_2^* oligonucleotide frequency dissimilarity measure improves prediction of hosts from metagenomically-derived viral sequences. *Nucleic Acids Res.*, **45**, 39–53.
55. He,Z., Zhang,H., Gao,S., Lercher,M.J., Chen,W.H. and Hu,S. (2016) Evolview v2: an online visualization and management tool for customized and annotated phylogenetic trees. *Nucleic Acids Res.*, **44**, W236–W241.

Acknowledgements

Many thanks to all members of Professor Martin Lercher's lab and special thanks to Guang-zhong Wang, Sabine Thuß, Thomas Laubach, Gabriel Gelius-Dietrich, Janina Maß and Xiaopan Hu. With their help, I had a great time in Düsseldorf and learned so much.

Many thanks to all members of Professor Wei-Hua Chen's lab: Puzi Jiang, Die Dai, Xiaowen Hao, Teng Wang and Sicheng Wu. They provided a creative scientific environment.

Many thanks to Martin Lercher for his kind and constant support.

Many thanks to Wei-Hua Chen. Without his help, it would be difficult to keep going.

Finally, I owe my parents, family and friends more gratitude for their love and constant support.