INfrastructure for a **PHA**ge **RE**ference **D**atabase: Identification of large-scale biases in the current collection of phage genomes

Ryan Cook[1], Nathan Brown[2], Tamsin Redgwell[3], Branko Rihtman[4], Megan Barnes[2], Martha Clokie[2], Dov J. Stekel[5], Jon Hobman[5], Michael A. Jones[1], Andrew Millard[2*]

[1] School of Veterinary Medicine and Science, University of Nottingham, Sutton Bonington Campus, College Road, Loughborough, Leicestershire, LE12 5RD, UK

[2] Dept Genetics and Genome Biology, University of Leicester, University Road, Leicester, Leicestershire, LE1 7RH, UK

[3] COPSAC, Copenhagen Prospective Studies on Asthma in Childhood, Herlev and Gentofte Hospital, University of Copenhagen, Copenhagen, Denmark

[4] University of Warwick, School of Life Sciences, Coventry, UK

[5] School of Biosciences, University of Nottingham, Sutton Bonington Campus, College Road, Loughborough, Leicestershire, LE12 5RD,

Corresponding author: adm39@le.ac.uk

Abstract

Background

With advances in sequencing technology and decreasing costs, the number of bacteriophage genomes that have been sequenced has increased markedly in the last decade.

Materials and Methods

We developed an automated retrieval and analysis system for bacteriophage genomes, INPHARED (https://github.com/RyanCook94/inphared), that provides data in a consistent format.

Results

As of January 2021, 14,244 complete phage genomes have been sequenced. The data set is dominated by phages that infect a small number of bacterial genera, with 75% of phages isolated only on 30 bacterial genera. There is further bias with significantly more lytic phage genomes than temperate within the database, resulting in ~54% of temperate phage genomes originating from just three host genera. Within phage genomes, putative antibiotic resistance genes were found in higher frequencies in temperate phages than lytic phages.

Conclusion

We provide a mechanism to reproducibly extract complete phage genomes and highlight some of the biases within this data, that underpins our current understanding of phage genomes.

Keywords: phage genomes, antibiotic resistance genes, virulence genes, jumbo-phages

Introduction

Bacteriophages (hereafter phages), are viruses that specifically infect bacteria and are thought to be the most abundant biological entities in the biosphere (1). In the oceans they are important in diverting the flow of carbon into dissolved and particulate organic matter via the lysis of their hosts (1), or directly halting the fixation of CO_2 carried out by their cyanobacterial hosts (2). In the human microbiome, it is becoming increasingly clear that phages play a role in a range of different diseases. Many recent studies have shown disease-specific alterations to the gut virome community in both gastrointestinal and systemic conditions, including irritable bowel disease (3), AIDs (4), malnutrition (5), and diabetes (6).

Phages alter the physiology of their hosts such that their bacterial hosts display increased virulence, a notable example being phage CTX into the genome of *Vibrio cholerae*, resulting in cholera (7). However, there are many cases where the expression of phage-encoded toxins cause an otherwise harmless commensal bacterium to convert into a pathogen, including multi-drug resistant ST11 strains of *Pseudomonas aeruginosa* (8, 9), and the Shiga-toxin encoding *Escherichia coli* (10). As well as increasing the virulence of the host bacteria, phages can also utilise parts of their genomes known as auxiliary metabolic genes (AMGs), homologues of host metabolic genes, to modulate their hosts metabolism (11).

Our understanding of how phages alter host metabolism has increased in conjunction with the number of phage genomes that have been sequenced, following sequencing of the first phage genome in 1977 (12). Since then, the number of phages that are isolated and the relative ease of high-throughput sequencing has led to a rapid increase in the number of sequenced bacteriophage genomes (13). The relatively simple nature of phage genomes

means that the vast majority of isolated phage genomes can be completely assembled using short-read next generation sequencing (14). The greater number of phage genomes available results in common analyses, including comparative genomic analyses (15, 16), taxonomic classification of phages (17–20), forming the basis of software to predict new phages (21–26), and as is often the first step in analysis of viromes, the comparison of sequences to a known database.

To do all of the above requires a comprehensive set of complete phage genomes from cultured isolates that can be used to build databases for further analyses. It also raises the question of how many complete phage genomes are currently available. While this should be relatively trivial question to answer, it is not very simple to do so, as there is currently no such database of all complete phage genomes. Therefore, the aim of this work was to provide a reproducible and automated way to extract complete phage genomes from GenBank and identify general properties within the data and limitations.

Materials and Methods

Bacteriophage genomes were download using the "PHG" identifier along with minimum and maximum length cut-offs. Genomes were then filtered based on several parameters to identify complete and near complete phage genomes. This includes initial searching for the term "Complete" & "Genome" in the phage description, followed by "Complete" & ("Genome" or "Sequence") & a genome length of greater than 10 kb. The list of genomes was then manually curated to identify obviously incomplete phage genomes, with the process on going. The accessions of these are then excluded in future iterations by the use of an exclusion list, which can be added to by the community via GitHub. Whilst this process

is not perfect, we thank numerous people that have identified genomes within this list that are obviously incomplete. The initial search term for downloading genomes was: *esearch -db nucleotide -query "gbdiv_PHG[prop]" | efilter -query "1417:800000 [SLEN] " | efetch -format gb > $phage_db.gb*. An exclusion list of phage genomes that are automatically called "complete", yet when manually checked are not is continually being updated.

After filtering, genes are called using Prokka with the –noanno flag, with a small number of phages using –gcode 15 (27, 28). Gene calling was repeated to provide consistency across all genomes, which is essential for comparative genomics. A database is provided so that this process does not continually have to be rerun and only new genomes are added. The original GenBank files are used to gather useful metadata including taxa and bacterial host, and the Prokka output files are used to gather data relating to genomic features. The gathered data are summarised in a tab-delimited file that includes the following: accession number, description of the phage genome, GenBank classification, genome length (bp), molGC (%), modification date, number of CDS, proportion of CDS on positive sense strand (%), proportion of CDS on negative sense strand (%), coding capacity (%), number of tRNAs, bacterial host, viral genus, viral sub-family, viral family, and the lowest viral taxa available (from genus, sub-family and family). Coding capacity was calculated by comparing the genome length to the sum length of all coding features within the Prokka output, and tRNAs were identified by the use of tRNA tag. Other outputs include a fasta file of all phage genomes, a MASH index for rapid comparison of new sequences, vConTACT2 input files, and various annotation files for IToL and vConTACT2. The vConTACT2 input files produced from the script were processed using vConTACT2 v0.9.13 with --rel-mode Diamond --db 'None' --

pcs-mode MCL --vcs-mode ClusterONE --min-size 1 and the resultant network was visualised using Cytoscape v3.8.0 (29, 30).

To identify genes indicative of a temperate lifestyle within genomes, we used a set of PFAM HMMs as described previously (31). If a genome encoded one of these genes, it was assumed to be temperate. Antimicrobial resistance genes (ARGs) and virulence factors were identified using Abricate with the resfinder and VFDB databases using 95% identity and 75% coverage cut-offs (32–34).

The phylogeny of "jumbo-phages" was constructed from the amino acid sequence of the TerL protein, extracted from 313/314 of the "jumbo-phage" genomes. Sequences were queried against a database of proteins from non-"jumbo-phages" using Blastp and the top five hits were extracted (35) with redundant sequences being removed. Sequences were aligned with MAFFT, with a phylogenetic tree being produced using IQ-Tree with "-m WAG -bb 1000" which was visualised using IToL (36–38). Additional information was overlaid using IToL templates that are generated via INPHARED.

Rarefaction analysis was carried out for phage genomes from the top ten most common hosts (70% ID over 95% length) and species (95% ID over 95% length) using ClusterGenomes v5.1 (39). An additional set of these genomes pooled together was included. Rarefaction curves and species richness estimates were produced using Vegan in R (40, 41).

133 All data from January 2021 is available at Figshare
134 https://doi.org/10.25392/leicester.data.14242085 and the script used for downloading and
135 analysing genomes is available on GitHub https://github.com/RyanCook94/ .

Results

The output of the INPHARED script provides as set of complete phage genomes, whereby genes have been called in a consistent manner that allows comparative genomics and phylogenetic analysis. In addition, it provides a MASH database to allow rapid comparison of new phage genomes against to identify close relatives. Along with formatted databases for input into vConTACT2 to allow identification of more distant relatives. The host data (Genus) for each phage is extracted along with summary information for each genome, which is reformatted to allow overlay onto trees in IToL (See Supplementary Figure 1 for full details).

For this study, we used a lenient definition of "complete" for the identification of complete phage genomes. Strictly speaking a complete phage genome would include the terminal ends of the phage genome. As many phages are sequenced using a transposon based library preparation (16, 42), the genome can never be complete as the terminal bases can never be sequenced, unless it is circularly permuted (14). For phage genomes with long terminal repeats, if the length of the repeat is larger than the library insert size, these cannot be resolved. As this information is not included in every GenBank file, automated retrieval is not possible.

We next set about identifying how many phage genomes have been sequenced to date. The extraction of genomes from the nucleotide database of GenBank results in 18,134 genomes. Of these, 3,890 phage genomes are REFSEQ entries which are derived from primary submissions, resulting in 14,244 putative complete phage genomes. Current recommendations are that phages are uniquely named (43), if this assumption is true then number of unique phage genomes is 12,127 if phages with the same name are truly

identical. However, there are multiple examples of phages with the same name. In some cases this is the same phage being re-sequenced due to experimental evolution studies such as *E. coli* phiX (44) . In other instances, phages with the same name are not genetically identical. Thus, using a phage name as a means to identify different phages is not a suitable method for determining the number of unique phage genomes. As an alternative, de-duplication of genomes at 100%, 97% and 95% identity results in 13,830, 12,845 and 12,770 genomes respectively.

Having established a dataset of "complete" phage genomes, we then analysed this data to look at general phage genomic properties. First, we looked at the increase in the number of phage genomes that are sequenced over time. Whilst the number of phage genomes has rapidly increased over the last 20 years, the rate of increase has slowed in the last decade (Figure 1), with the number of phage genomes doubling every 2-3 years.

Bacteriophage Hosts and Predicted Gene Function

Utilising the database of complete genomes, we extracted the hosts and predicted number of hypothetical proteins for each phage. Across all phages, the majority of genes which encode proteins with unknown function (hypothetical) was mean 56% (+/- 20), supporting the truism that the majority of genes encode proteins within unknown function.

The host of 12,403 phages were extracted with the remainder unknown as the host was not clear from the information contained within the GenBank file alone. The genomes of phages infecting 234 hosts have been sequenced. However, there is a clear bias in the isolation of phages against the same host (Figure 2a). Phages that infect *Mycobacterium* spp. are the

most commonly deposited genomes (~13%), largely due to the pioneering work of the SEA-PHAGES program (45), followed by *Escherichia* spp., *Streptococcus* spp., and *Pseudomonas* spp. (Figure 2a). Phages isolated on just 30 different bacterial genera accounts for ~75% of all phage genomes in the database (Supplementary Table 1). For genomes isolated against the top ten hosts, we used rarefaction analysis to gain an understanding of the diversity of phage genomes isolated to date and determine redundancy of phages isolated on a particular host. Using a cut-off of 95% identity to define a species, it was clearly observed the number of phage species continues to increase with the number of genomes sequenced, a pattern also observed at the level of genus (70% identity) (Figure 3). Using the current data, it was possible to estimate how many different species of phage might infect these different hosts (Supplementary Table 4). For *Mycobacterium*, for which most phages have isolated on, there are 695 observed species with an estimated 2132-2282 total species. Thus, demonstrating even for hosts where thousands of phages have been isolated, we are only just scratching the surface of the diversity of total phage diversity. We are also likely under estimating the total number of different phage species. In the case of phages infecting *Mycobacterium*, the majority of these have been isolated on only a single strain as part of the SEA-PHAGES program. Increasing the diversity of the host *Mycobacterium*, is likely to lead to higher estimates.

Lytic and Temperate phages

To identify if the phage is lytic or temperate, we searched for genes that facilitate a temperate lifestyle (e.g., integrase and recombinase) that have been used in previous studies. This process is not perfect, as the presence of an identifiable gene linked to temperate phages does not mean it will access a lysogenic cycle. However, it does allow

large scale comparative analyses, compared to the manual searching of literature of every phage compared to determine if it has been experimentally tested. Within the dataset, 4,258 (~30%) phages have the potential to access a lysogenic lifecycle. The frequency of putative temperate phages was highly variable depending on the host (Supplementary Figure 2). The number of putative temperate phages is also biased towards a small number of hosts with 1,217, 846 and 214 isolated on *Mycobacterium, Streptococcus* and *Gordonia* respectively. Collectively these three hosts account for ~54% of all putative temperate phage genomes sequenced to date (Supplementary Figure 2).

Genomic Properties

Phage genomes ranged from 3.1 kb to 642.4 kb in size, with a clear distribution in the size of genomes with the most prominent peaks at 5-10 kb, 40 kb, 50 kb and ~165 kb (Figure 2b). The mean and median coding capacity was found to be 90.45% and 91.52%, respectively (Supplementary Figure 2). Of the 14,244 genomes, 5,731 (~40%) were found to have ≥ 90% of coding features on one strand and 3,293 (~23%) of these had coding features entirely on one strand (Supplementary Figure 2). The number of phages with genes encoding tRNAs was 4,590 (~32%). For those phages encoding tRNAs, the range was 1 to 62 with a median of 3. Whilst there is much literature on the presence of tRNAs in phages, it is still not clear entirely what role they provide to phages and why they absent in some phages and not others (46).

Phages with genomes greater than 200 kb are often referred to as “jumbo-phages” and are reported to be rarely isolated (47). 314 genomes (~2.2%) greater than 200 kb in length were identified, suggesting that they are rare. To further investigate if “jumbo-phages” are as rare

as is thought, we looked at the distribution in the context of the previously identified host bias. "Jumbo-phages" have only been isolated on 31 of 234 identifiable bacterial hosts (Supplementary Table 1) and are far more commonly isolated on some hosts than others. Noticeably absent are any "jumbo-phages" that infect *Mycobacterium, Gordonia, Lactococcus, Arthrobacter,* and *Streptococcus,* with >4,000 phages having been sequenced from these bacterial hosts (Figure 2c). For host bacteria that have had far fewer phages isolated on them such as *Caluobacter, Sphingomonas, Erwinia, Areomonas, Dickeya* and *Ralstonia*, the frequency of "jumbo-phage" isolation is far higher (Figure 2c). Due to the small sampling depth of some of these hosts (e.g., *Photobacterium* and *Tenacibaclum*), it is not possible to determine whether the high proportion of genomes is merely a result of the low number of genomes sequenced. However, for other hosts such as *Aeromonas, Erwinia* and *Caulobacter* from which more than 20 phages have been isolated, ~26%, ~44% and ~63% are categorised as "jumbo" respectively. Therefore suggesting "jumbo-phages" are not always rare on particular hosts.

We further investigated the phylogeny of "jumbo-phages" using the translated sequence of the *terL* gene. The "jumbo-phages" are well distributed across the tree and do not form a single monophyletic clade, suggesting that they have arisen on multiple occasions, with multiple clades of phages having representatives of "jumbo-phages" within them. Not all "jumbo-phages" are equal, with "jumbo" cyanophages infecting the cyanobacteria *Synechococcus* and *Procholorococcus* only marginally larger than there non-jumbo cyanophages relatives. These "jumbo-phages" are also more closely related to their non-jumbo cyanophages relatives than other "jumbo-phages" (Figure 4). This is not limited to the cyanophages, with many other "jumbo-phages" more closely related to a non-jumbo

256 phage. A similar pattern of grouping non-jumbo with "jumbo-phages" is observed when a
257 reticulate approach is used to look at the relatedness of phage genomes using vConTACT2
258 (Supplementary Figure 3).
259
260 **Virulence Factors and Antimicrobial Resistance Genes**
261 The presence of ARGs and virulence factors is major concern for phage therapy, as the use
262 of phages carrying such genes may make the populations of bacteria they are intended to
263 kill more virulent or resistant to antibiotics. We therefore used this database to integrate
264 the frequency and diversity of phage-encoded virulence factors and ARGs. 235 genomes
265 (~1.6%) were found to encode a virulence factor and 43 genomes (~0.3%) to encode an
266 ARG. The most common virulence genes were the stx_{2A} (72 genomes) and stx_{2B} (71
267 genomes) genes that encode subtypes of the Shiga toxin (Supplementary Table 2). The most
268 common ARGs were the *mef*(A) (14 genomes) and *msr*(D) genes which confer resistance to
269 macrolide antibiotics (Supplementary Table 3) (48). Most genomes encoding a virulence
270 factor were predicted to be from temperate phages (222/235), and were found to infect six
271 bacterial genera, with the three most abundant hosts being *Streptococcus*, *Staphylococcus*
272 and *Escherichia* respectively. The hosts for many genomes could not be determined
273 (55/235). The virulence factor encoding genomes were widely distributed over 26 putative
274 genera (Supplementary Figure 3). All genomes encoding an ARG were predicted to be
275 temperate and were found to be isolated from eight bacterial genera, with the majority of
276 phages linked to *Streptococcus* spp. (27/43).
277
278 **Discussion**

Defining how many different complete phage genomes have been sequenced is not a simple question as it might appear. Based on accession numbers, there are 14,244 phage genomes, once RefSeq duplicates have been removed. Using unique names results in 12,127 phages, however using names alone does not give an accurate estimate of the number of different phages, as genomically different phages have the same name. The use of de-duplication at 100% identity suggests 13,830 unique phage genomes (January 2021) from cultured isolates. This assumes that the genome submissions are from isolates and not predictions of prophages from bacterial genomes. For the vast majority of phages, this appears to be case, although not easily discernible for all phage genomes.

The data reveals clear patterns in phage genomes and biases in the selection of phage genomes that are currently available, but not always discussed in the analysis of genomes. The first is the number of phage genomes is relatively small. Even for hosts where the highest number of phages have been isolated on, our estimates suggest 1000s of new phage species remain to isolated and sequenced. If we consider there are now more than 300,000 assembled representative bacterial genomes in GenBank, with many hundreds of thousands more for particular genera e.g., >300,000 *Salmonella* and *Escherichia* genomes alone (49). The representation of phage genomes to date is tiny compared to their bacterial hosts. Furthermore, the rate at which phage genomes are being sequenced is slowing down rather than increasing. Given the renewed interest in phages and increased accessibility of sequencing, the decrease in the rate over time was surprising.

The second point of note is the bias in phage genomes. With a clear bias in both the hosts phages are isolated on and for lytic phages over temperate phages. Thus, these phages are

303 representative of these particular hosts, rather than phages in their entirety. Due to the
304 enormous success of the SEA-PHAGES program, many phages have been isolated on
305 *Mycobacterium* and *Gordonia* (50). This in turn results in ~1/3rd of all temperate phage
306 genomes being isolated on these two bacterial genera, whereas the remaining 2/3^{rds} are
307 distributed across 142 different hosts.
308
309 The overrepresentation of phages infecting particular hosts can lead to truisms that may not
310 be correct. For instance, "jumbo-phages", those that have genomes >200 kb, are rarely
311 isolated (47). Analysis of the complete dataset suggests ~2.2% of genomes fall into this
312 category. However, this needs to be viewed in the context of the large bias in the hosts used
313 for isolation, with ~75% of phages isolated on only ~16% of bacterial hosts that could be
314 identified. When the number of "jumbo-phages" is expressed as a percentage of all phage
315 genomes, their isolation is clearly rare. For some hosts, such as *Mycobacterium,* many
316 hundreds of phages isolated on the same host strain have been sequenced without the
317 isolation of a "jumbo-phage", suggesting they are truly rare for this host (45). However, for
318 other hosts such as *Prochlorococcus*, *Synechococcus, Caulobacter,* and *Erwinia,* the
319 isolation of "jumbo-phages" is not a rare event. While methodological adjustments of
320 decreasing agar viscosity and large pore size filters may increase the number of phages
321 isolated that have larger genome sizes (47), we suggest that using a wider variety of hosts
322 may increase the number of "jumbo-phages" isolated. Phylogenetic analysis demonstrated
323 many "jumbo-phages" are more closely related to non-jumbo phages than other "jumbo-
324 phages". Thus, as the number of phage genomes has increased an arbitrary descriptor or
325 "jumbo" for phages with genomes over 200 kb in length has less meaning. Recent
326 comparative analysis of 224 "jumbo-phages", used proteome size and analysis of protein

length to determine a cut-off of 180 kb to separate "jumbo-phages", from other phages. From this using a clustering-based approach, three major clades of "jumbo-phages" were identified (51). In this study using *terL* as a phylogenetic marker to determine the phylogeny of 313 "jumbo-phages" and their closely related phages, suggests they have arisen on multiple occasions, as has been demonstrated previously (51). "Jumbo-phages" are clearly not monophyletic and what applies to one "jumbo-phage" does not hold true for many others (51). As the number and diversity of "jumbo-phages" increases, the use of the term seems to have less meaning.

With the increasing interest and use of phages for therapy, the isolation of phages that do not contain known virulence factors or ARGs is imperative. How frequently phages encode antibiotic resistance genes is a topic of much debate (52, 53). A previous study of 1,181 phage genomes found that they are rarely encoded by phages with only 13 candidate genes, of which four where experimentally tested and found to have no functional antibiotic activity (47). We estimate ~0.3% of phage genomes encode a putative ARG (none have been experimentally tested), a finding that is consistent with previous reports of low-level carriage in phage genomes (52) in a dataset that is ~10x larger using similarly stringent cut-offs. Critically, all of these ARGs were found in phages that are predicted to be temperate or have been engineered to carry ARGs as a marker for selection. With the frequency of carriage in temperate phages being ~1% overall. However, this data is still biased by the majority of temperate phages being isolated on only three bacterial genera. Notably no ARGs were detected on phages of *Mycobacterium*, which accounts for ~28 % of temperate phages. In comparison, ~2.6% (27/1055) of temperate phages of *Streptococcus* carry putative ARGs and 50% of phages from *Erysipelothrix* (1/2). Clearly a much deeper sampling

of temperate phages from a broader range of hosts is required to get an accurate understanding of the role of phage in the carriage of ARGs. Based on the skewed data available to date, it seems unlikely there will be issues in the isolation of lytic phages for therapeutic use that contain known ARG within their genomes. However, we cannot determine whether these lytic phages cannot spread ARGs via transduction, or through carriage of as-yet uncharacterised ARGs.

Whilst there is much debate on the presence and importance of ARGs in phage genomes, the role of genes encoding virulence factors is well studied and the process of lysogenic conversion well known (7–10). However, how widespread known virulence genes are in phages is not widely reported. We estimate 1.6% of phages encode at least one putative virulence factor, with the frequency of carriage far higher in temperate phages (5.5%) than lytic phages (0.13%). Again, these overall percentages are skewed by host bias with no known virulence factors detected in *Mycobacterium* temperate phages (0/1217), in comparison 72% of temperate phages of *Shigella* (5/7) and 7% (61/846) of *Streptococcus* contain virulence factors. It is currently impossible to determine if the higher proportion of ARGs and virulence factors in phages of known pathogens is a feature of their biology, or a skew in the database towards phages of clinically relevant isolates.

Given the biases in the dataset, it is not clear if the general phage patterns we observe (e.g., jumbo-phages are rarely isolated, more temperate phages on particular hosts, and the carriage of ARGs and virulence genes) are linked to biology or chronic under sampling of phage genomes. We speculate that currently is most likely the latter, which distorts some generalisations about phages. It clear that jumbo-phages are not rare on some hosts and

375 putative ARGs are far more abundant on temperate phages. However, far deeper sampling
376 of phage diversity across different hosts is required at an increasing rate.
377
378 **Conclusions**
379 We have provided a simple method to automate the download of curated set of complete
380 genomes from cultured phage isolates, providing metadata in a format that can be used as a
381 starting point for many common analyses. Analysis of the current data highlights what we
382 know about phage genomes is skewed by the majority of phages having been isolated from
383 a small number of bacterial genera. Furthermore, the rate at which phage genomes are
384 being deposited is decreasing. Whilst understanding of genomic diversity is always
385 influenced by the data available, this is particularly acute for phage genomes with so many
386 phages isolated on smaller number of hosts. To obtain a greater understanding of phage
387 diversity, larger numbers of phages, in particular temperate phages, isolated from a broader
388 range of bacteria need to be sequenced.
389
390
391 **Acknowledgments**
392
393 **Authorship Confirmation Statement**
394 **Competing Interests.**
395 **Funding**
396 R.C is supported by a scholarship from the Medical Research Foundation National PhD Training
397 Programme in Antimicrobial Resistance Research (MRF-145-0004-TPG-AVISO). A.M, D.J.S, M.J, and
398 J.H were supported by NERC (NE/N019881/1). A.M was supported by MRC
399 (MR/T030062/1and MR/L015080/1)

Figure 1

Number of complete phage genomes in GenBank over time. Dates were estimated based on date of submission (* for 235 genomes, the date of update was used as no submission date was available). The reference lines showing doubling rates (dashed) begin in 1989, as this is when the number of phage genomes increased beyond the first submission in 1982.

Figure 2

Overall properties of phages. A) Proportion of phages isolated on the top 30 most abundant hosts. B) Distribution of phage genome sizes. C) Proportion of "jumbo-phages" on top 30 hosts for which at least one "jumbo-phage" has been isolated.

Figure 3

Genomic diversity of phages on the top ten most abundant hosts. A) Rarefaction curve of phage species. Species were defined as 95% identity over 95% of genome length. B) Rarefaction curve of phage genera. Genera were defined as 70% identity over 95% of genome length.

Figure 4

Phylogenetic tree of translated terL gene for 313 "jumbo-phages" and their closest relatives. The alignment was produced using MAFFT (36) and tree produced using IqTree using WAG model with 1000 bootstrap repeats (37). Pink shaded regions indicate "jumbo-phages", coloured ring indicates viral genus, and blue bars indicate genome length. Bootstrap values indicated by black circles are shown with a minimum of 70%.

References

1. Suttle CA. 2007. Marine viruses--major players in the global ecosystem. Nat Rev Microbiol 5:801–12.

2. Puxty RJ, Millard AD, Evans DJ, Scanlan DJ. 2016. Viruses inhibit CO 2 fixation in the most abundant phototrophs on Earth. Curr Biol 26:1585–1589.

3. Norman JM, Handley SA, Baldridge MT, Droit L, Liu CY, Keller BC, Kambal A, Monaco CL, Zhao G, Fleshner P, Stappenbeck TS, McGovern DPB, Keshavarzian A, Mutlu EA, Sauk J, Gevers D, Xavier RJ, Wang D, Parkes M, Virgin HW. 2015. Disease-specific alterations in the enteric virome in inflammatory bowel disease. Cell https://doi.org/10.1016/j.cell.2015.01.002.

4. Monaco CL, Gootenberg DB, Zhao G, Handley SA, Ghebremichael MS, Lim ES, Lankowski A, Baldridge MT, Wilen CB, Flagg M, Norman JM, Keller BC, Luévano JM, Wang D, Boum Y, Martin JN, Hunt PW, Bangsberg DR, Siedner MJ, Kwon DS, Virgin HW. 2016. Altered Virome and Bacterial Microbiome in Human Immunodeficiency Virus-Associated Acquired Immunodeficiency Syndrome. Cell Host Microbe https://doi.org/10.1016/j.chom.2016.02.011.

5. Reyes A, Blanton L V., Cao S, Zhao G, Manary M, Trehan I, Smith MI, Wang D, Virgin HW, Rohwer F, Gordon JI. 2015. Gut DNA viromes of Malawian twins discordant for severe acute malnutrition. Proc Natl Acad Sci U S A https://doi.org/10.1073/pnas.1514285112.

6. Ma Y, You X, Mai G, Tokuyasu T, Liu C. 2018. A human gut phage catalog correlates the gut phageome with type 2 diabetes. Microbiome https://doi.org/10.1186/s40168-018-0410-y.

7. Waldor MK, Mekalanos JJ. 1996. Lysogenic Conversion by a Filamentous Phage Encoding Cholera Toxin. Science (80-) https://doi.org/10.1126/science.272.5270.1910.

8. van Belkum A, Soriaga LB, LaFave MC, Akella S, Veyrieras J-B, Barbu EM, Shortridge D, Blanc B, Hannum G, Zambardi G, Miller K, Enright MC, Mugnier N, Brami D, Schicklin S, Felderman M, Schwartz AS, Richardson TH, Peterson TC, Hubby B, Cady KC. 2015. Phylogenetic Distribution of CRISPR-Cas Systems in Antibiotic-Resistant <span class="named-

content genus-species" id="named-content-1">Pseudomonas aeruginosa MBio 6:e01796-15.

9. Tsao Y-F, Taylor VL, Kala S, Bondy-Denomy J, Khan AN, Bona D, Cattoir V, Lory S, Davidson AR, Maxwell KL. 2018. Phage Morons Play an Important Role in Pseudomonas aeruginosa Phenotypes. J Bacteriol 200:e00189-18.

10. O'Brien AD, Newland JW, Miller SF, Holmes RK, Smith HW, Formal SB. 1984. Shiga-like toxin-converting phages from Escherichia coli strains that cause hemorrhagic colitis or infantile diarrhea. Science (80-) https://doi.org/10.1126/science.6387911.

11. Breitbart M, Thompson LR, Suttle CA, Sullivan MB. 2007. Exploring the vast diversity of marine viruses. Oceanography 20:135–139.

12. Sanger F, Air GM, Barrell BG, Brown NL, Coulson AR, Fiddes CA, Hutchison CA, Slocombe PM, Smith M. 1977. Nucliotide sequence of bacteriophage phi X174 DNA. Nature.

13. Perez Sepulveda B, Redgwell T, Rihtman B, Pitt F, Scanlan DJ, Millard A. 2016. Marine phage genomics: the tip of the iceberg. FEMS Microbiol Lett 363:fnw158.

14. Rihtman B Clokie MRJ, Koskella B, Millard AD. MS. 2016. Assessing Illumina technology for the high-throughput sequencing of bacteriophage genomes. PeerJ.

15. Javan RR, Ramos-Sevillano E, Akter A, Brown J, Brueggemann A. 2018. Prophages and satellite prophages are widespread among *Streptococcus* species and may play a role in pneumococcal pathogenesis. bioRxiv 502740.

16. Michniewski S, Redgwell T, Grigonyte A, Rihtman B, Aguilo-Ferretjans M, Christie-Oleza J, Jameson E, Scanlan DJ, Millard AD. 2019. Riding the wave of genomics to investigate aquatic coliphage diversity and activity. Environ Microbiol2019/04/04. 21:2112–2128.

17. Barylski J, Enault F, Dutilh BE, Schuller MB, Edwards RA, Gillis A, Klumpp J, Knezevic P, Krupovic M, Kuhn JH, Lavigne R, Oksanen HM, Sullivan MB, Jang H Bin, Simmonds P, Aiewsakun P, Wittmann J, Tolstoy I, Brister JR, Kropinski AM, Adriaenssens EM. 2019. Analysis of Spounaviruses as a Case Study for the Overdue Reclassification of Tailed Phages. Syst Biol

0:1–14.

18. Chibani CM, Farr A, Klama S, Dietrich S, Liesegang H. 2019. Classifying the Unclassified: A Phage Classification Method. Viruses 11:195.

19. Rohwer F, Edwards R. 2002. The phage proteomic tree: A genome-based taxonomy for phage. J Bacteriol https://doi.org/10.1128/JB.184.16.4529-4535.2002.

20. Adriaenssens EM, Wittmann J, Kuhn JH, Turner D, Sullivan MB, Dutilh BE, Jang H Bin, van Zyl LJ, Klumpp J, Lobocka M, Moreno Switt AI, Rumnieks J, Edwards RA, Uchiyama J, Alfenas-Zerbini P, Petty NK, Kropinski AM, Barylski J, Gillis A, Clokie MRC, Prangishvili D, Lavigne R, Aziz RK, Duffy S, Krupovic M, Poranen MM, Knezevic P, Enault F, Tong Y, Oksanen HM, Rodney Brister J. 2018. Taxonomy of prokaryotic viruses: 2017 update from the ICTV Bacterial and Archaeal Viruses Subcommittee. Arch Virol 163:1125–1129.

21. Ren J, Ahlgren NA, Lu YY, Fuhrman JA, Sun F. 2017. VirFinder: a novel k-mer based tool for identifying viral sequences from assembled metagenomic data. Microbiome 5:69.

22. Ren J, Song K, Deng C, Ahlgren NA, Fuhrman JA, Li Y, Xie X, Sun F. 2018. Identifying viruses from metagenomic data by deep learning.

23. Roux S, Enault F, Hurwitz BL, Sullivan MB. 2015. VirSorter: mining viral signal from microbial genomic data. PeerJ 3:e985.

24. Akhter S, Aziz RK, Edwards R a. 2012. PhiSpy: A novel algorithm for finding prophages in bacterial genomes that combines similarity-and composition-based strategies. Nucleic Acids Res 40:1–13.

25. Arndt D, Marcu A, Liang Y, Wishart DS. 2017. PHAST, PHASTER and PHASTEST: Tools for finding prophage in bacterial genomes. Brief Bioinform 1–8.

26. Bolduc B, Jang H Bin, Doulcier G, You Z-Q, Roux S, Sullivan MB. 2017. vConTACT: an iVirus tool to classify double-stranded DNA viruses that infect *Archaea* and *Bacteria*. PeerJ 5:e3243.

27. Seemann T. 2014. Prokka: Rapid prokaryotic genome annotation. Bioinformatics 30:2068–2069.

502 28. Devoto AE, Santini JM, Olm MR, Anantharaman K, Munk P, Tung J, Archie EA, Turnbaugh PJ,
503 Seed KD, Blekhman R, Aarestrup FM, Thomas BC, Banfield JF. 2019. Megaphages infect
504 Prevotella and variants are widespread in gut microbiomes. Nat Microbiol
505 https://doi.org/10.1038/s41564-018-0338-9.
506 29. Bin Jang H, Bolduc B, Zablocki O, Kuhn JH, Roux S, Adriaenssens EM, Brister JR, Kropinski AM,
507 Krupovic M, Lavigne R, Turner D, Sullivan MB. 2019. Taxonomic assignment of uncultivated
508 prokaryotic virus genomes is enabled by gene-sharing networks. Nat Biotechnol 37:632–639.
509 30. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker
510 T. 2003. Cytoscape: A software environment for integrated models of biomolecular
511 interaction networks. Genome Res 13:2498–2504.
512 31. Cook R, Hooton S, Trivedi U, King L, Dodd CER, Hobman JL, Stekel DJ, Jones MA, Millard AD.
513 2021. Hybrid assembly of an agricultural slurry virome reveals a diverse and stable
514 community with the potential to alter the metabolism and virulence of veterinary pathogens.
515 Microbiome 9:65.
516 32. Chen L, Zheng D, Liu B, Yang J, Jin Q. 2016. VFDB 2016: hierarchical and refined dataset for big
517 data analysis--10 years on. Nucleic Acids Res2015/11/17. 44:D694–D697.
518 33. Zankari E, Hasman H, Cosentino S, Vestergaard M, Rasmussen S, Lund O, Aarestrup FM,
519 Larsen MV. 2012. Identification of acquired antimicrobial resistance genes. J Antimicrob
520 Chemother2012/07/10. 67:2640–2644.
521 34. Seemann T. Abricate. Github.
522 35. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. J
523 Mol Biol 215:403–410.
524 36. Nakamura T, Yamada KD, Tomii K, Katoh K. 2018. Parallelization of MAFFT for large-scale
525 multiple sequence alignments. Bioinformatics 34:2490–2492.
526 37. Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ. 2015. IQ-TREE: a fast and effective
527 stochastic algorithm for estimating maximum-likelihood phylogenies. Mol Biol

528 Evol2014/11/03. 32:268–274.

529 38. Letunic I, Bork P. 2019. Interactive Tree Of Life (iTOL) v4: recent updates and new

530 developments. Nucleic Acids Res 47:W256–W259.

531 39. GitHub - simroux/ClusterGenomes: Archive for ClusterGenomes scripts.

532 40. Team RC. 2018. R: A language and environment for statistical computing. R Foundation for

533 Statistical Computing, Vienna.

534 41. Oksanen J, Guillaume Blanchet F, Friendly M, Kindt R, Legendre P, McGlinn D, Minchin PR,

535 O'Hara RB, Simpson GL, Solymos P, Stevens M henry H, Szoecs E, Wagner H. 2020. vegan:

536 Community Ecology Package.

537 42. Rihtman B, Meaden S, Clokie MRJ, Koskella B, Millard AD, Rihtman B Clokie MRJ, Koskella B,

538 Millard AD. MS. 2016. Assessing Illumina technology for the high-throughput sequencing of

539 bacteriophage genomes. PeerJ 4:e2055.

540 43. Adriaenssens EM, Rodney Brister J. 2017. How to name and classify your phage: An informal

541 guide. Viruses 9:1–9.

542 44. Pepin KM, Wichman HA. 2008. Experimental evolution and genome sequencing reveal

543 variation in levels of clonal interference in large populations of bacteriophage ϕX174. BMC

544 Evol Biol https://doi.org/10.1186/1471-2148-8-85.

545 45. Hatfull GF, Pedulla ML, Jacobs-Sera D, Cichon PM, Foley A, Ford ME, Gonda RM, Houtz JM,

546 Hryckowian AJ, Kelchner VA, Namburi S, Pajcini K V., Popovich MG, Schleicher DT, Simanek

547 BZ, Smith AL, Zdanowicz GM, Kumar V, Peebles CL, Jacobs WR, Lawrence JG, Hendrix RW.

548 2006. Exploring the mycobacteriophage metaproteome: Phage genomics as an educational

549 platform. PLoS Genet https://doi.org/10.1371/journal.pgen.0020092.

550 46. Bailly-Bechet M, Vergassola M, Rocha E. 2007. Causes for the intriguing presence of tRNAs in

551 phages. Genome Res2007/09/04. 17:1486–1495.

552 47. Yuan Y, Gao M. 2017. Jumbo bacteriophages: An overview. Front Microbiol

553 https://doi.org/10.3389/fmicb.2017.00403.

554 48. Daly MM, Doktor S, Flamm R, Shortridge D. 2004. Characterization and prevalence of MefA,
555 MefE, and the associated msr(D) gene in Streptococcus pneumoniae clinical isolates. J Clin
556 Microbiol 42:3570–3574.
557 49. Zhou Z, Alikhan NF, Mohamed K, Fan Y, Achtman M. 2020. The EnteroBase user's guide, with
558 case studies on Salmonella transmissions, Yersinia pestis phylogeny, and Escherichia core
559 genomic diversity. Genome Res 30:138–152.
560 50. Hanauer DI, Graham MJ, Betancur L, Bobrownicki A, Cresawn SG, Garlena RA, Jacobs-Sera D,
561 Kaufmann N, Pope WH, Russell DA, Jacobs WR, Sivanathan V, Asai DJ, Hatfull GF. 2017. An
562 inclusive Research Education Community (iREC): Impact of the SEA-PHAGES program on
563 research outcomes and student learning. Proc Natl Acad Sci U S A 114:13531–13536.
564 51. M. Iyer L, Anantharaman V, Krishnan A, Burroughs AM, Aravind L. 2021. Jumbo Phages: A
565 Comparative Genomic Overview of Core Functions and Adaptions for Biological Conflicts.
566 Viruses .
567 52. Enault F, Briet A, Bouteille L, Roux S, Sullivan MB, Petit MA. 2017. Phages rarely encode
568 antibiotic resistance genes: A cautionary tale for virome analyses. ISME J 11:237–247.
569 53. Debroas D, Siguret C. 2019. Viruses as key reservoirs of antibiotic resistance genes in the
570 environment. ISME J 13:2856–2867.
571
572

573

Figure 1

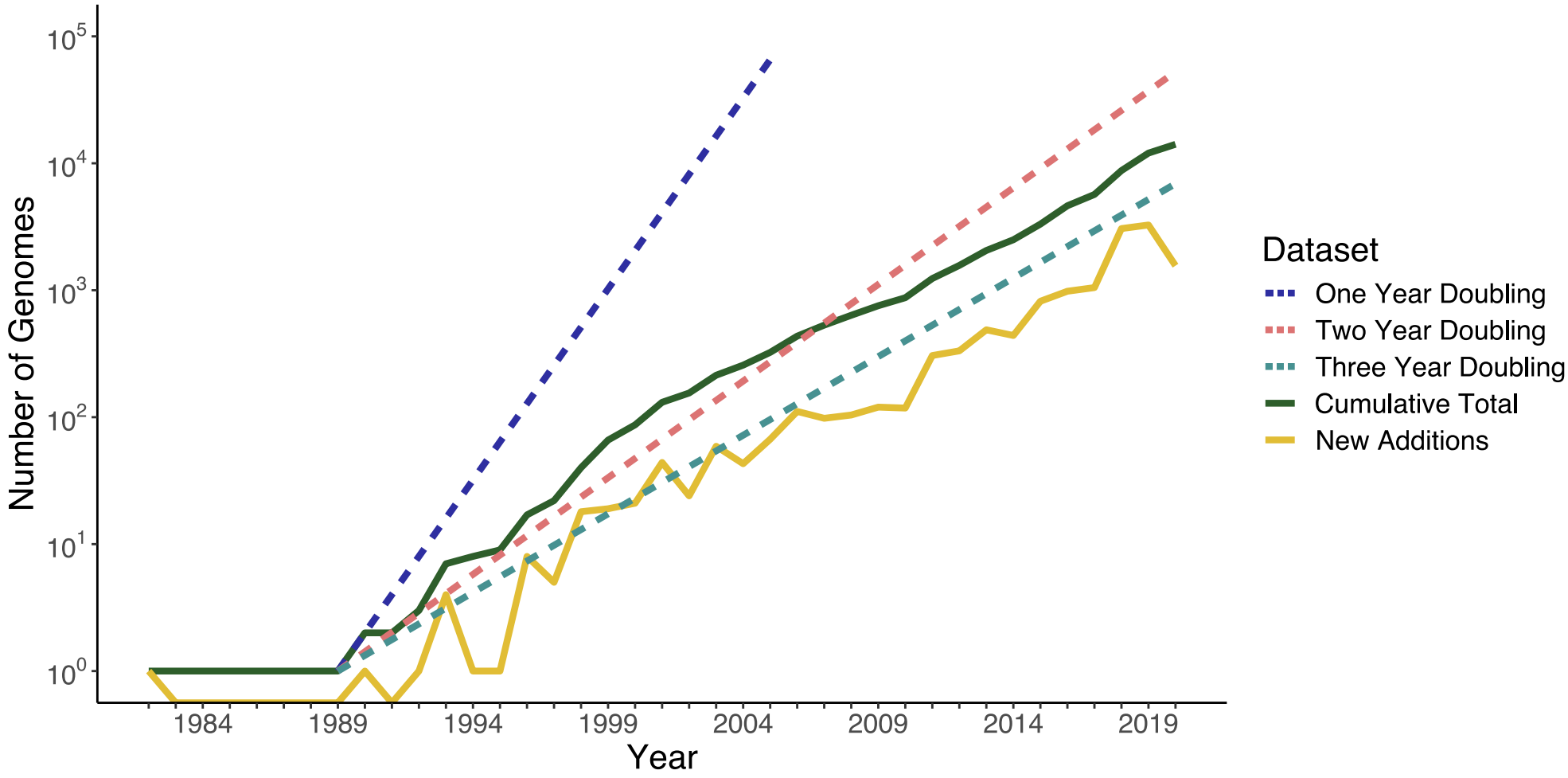
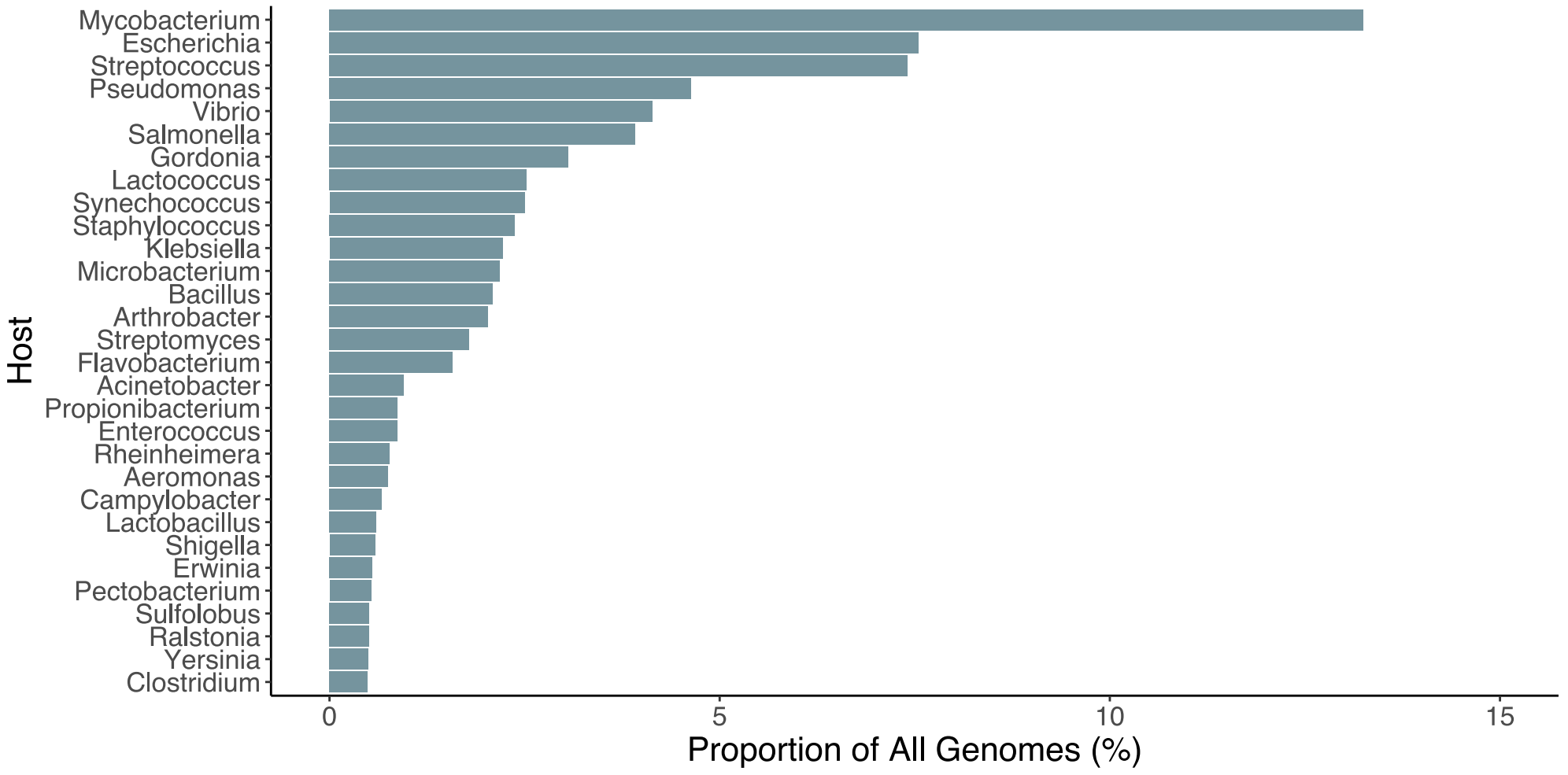
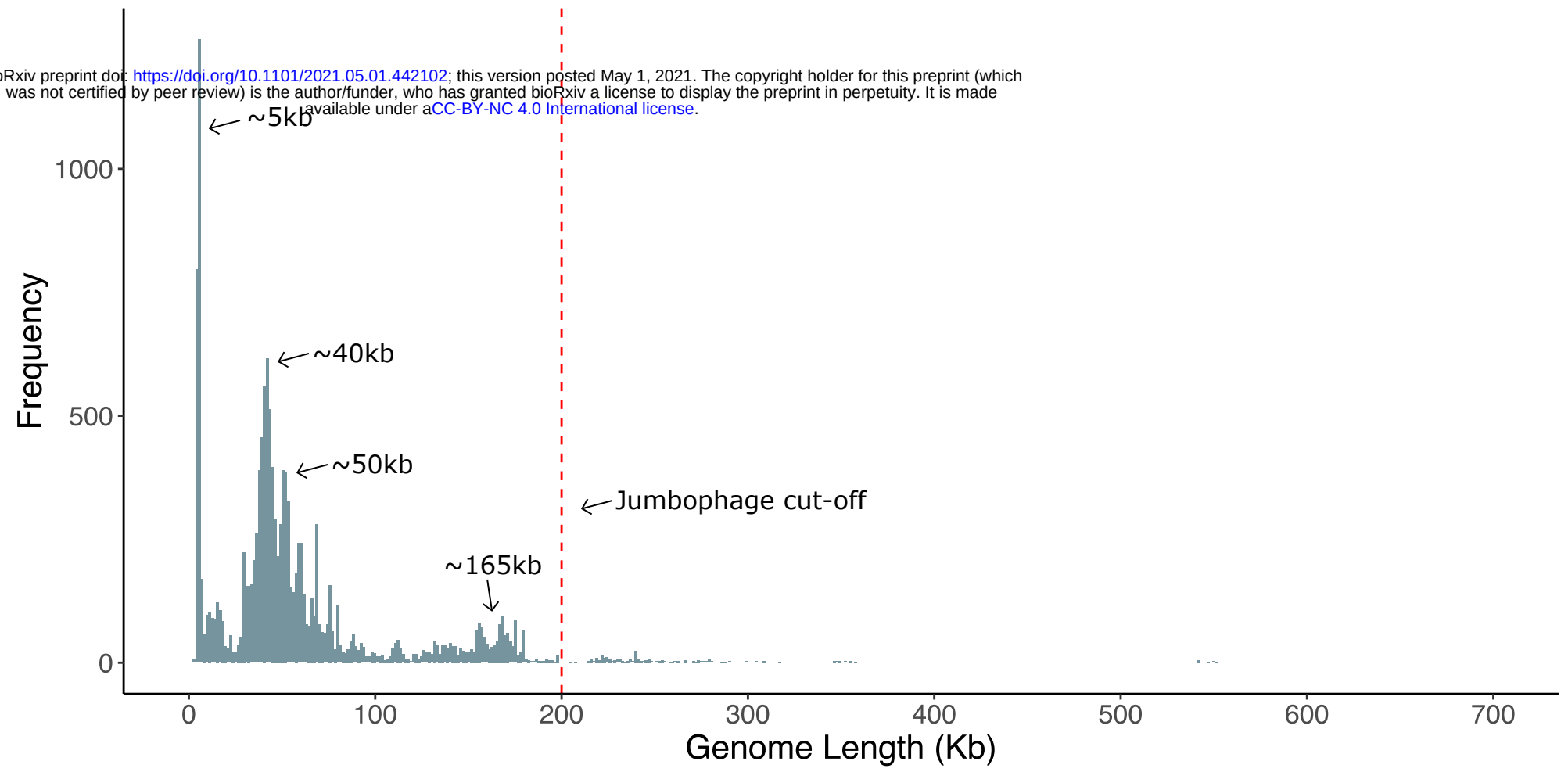

Number of Genomes
10^5
10^4
10^3
10^2
10^1
10^0
1984
1989
1994
1999
2004
2009
2014
2019
Year
Dataset
One Year Doubling
Two Year Doubling
Three Year Doubling
Cumulative Total
New Additions

Figure 2

A



B



C

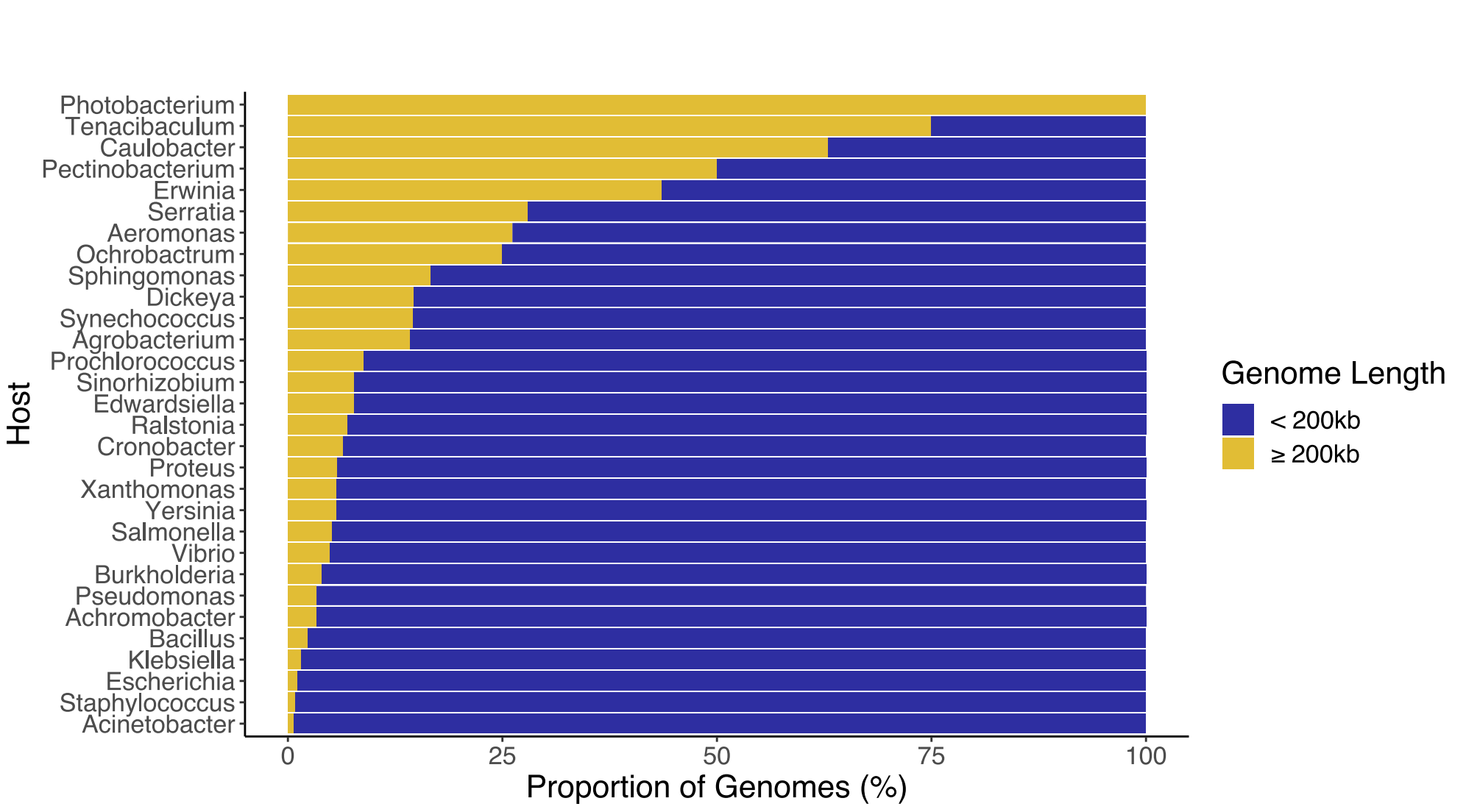
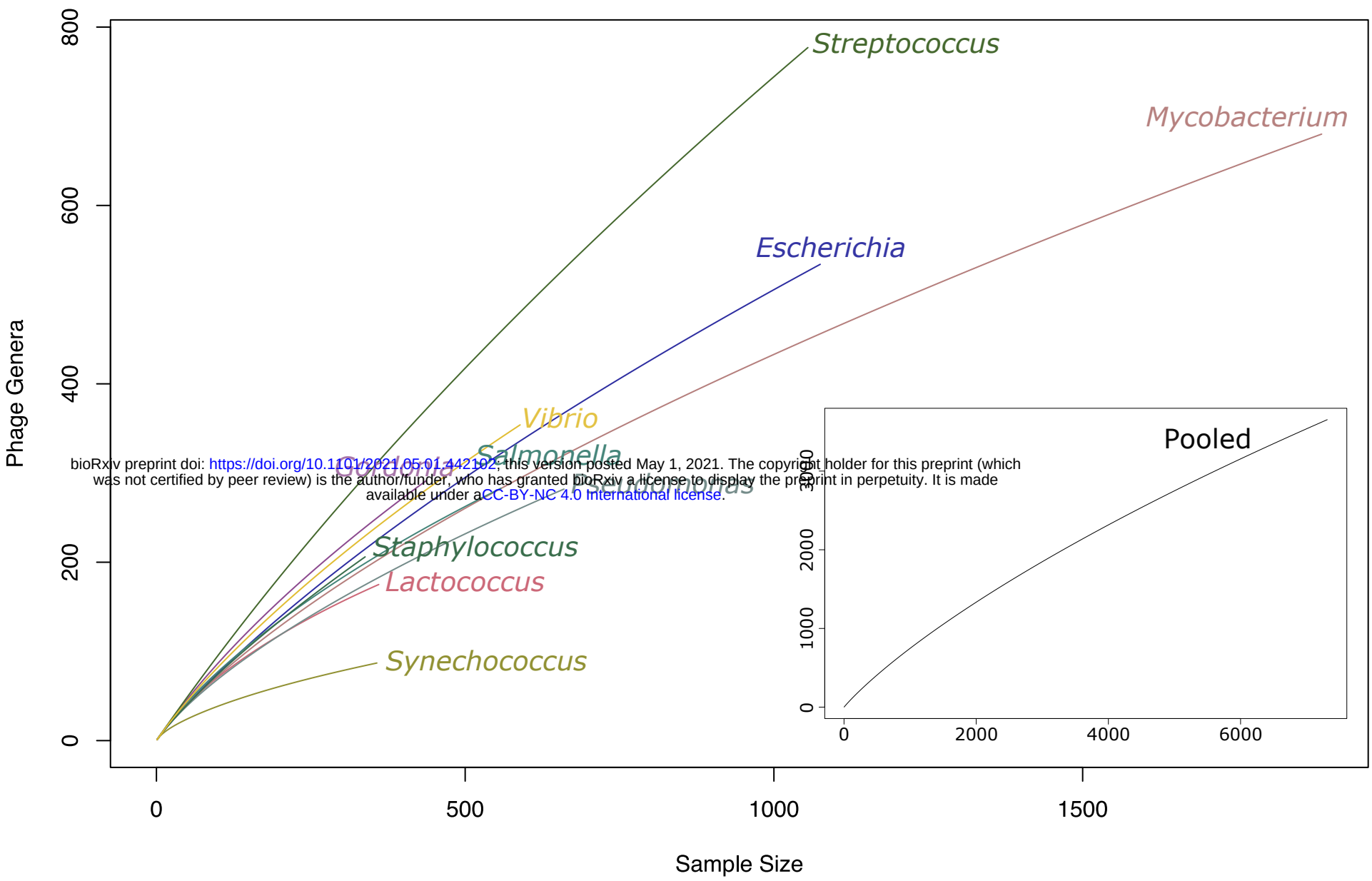


Figure 3

A



B

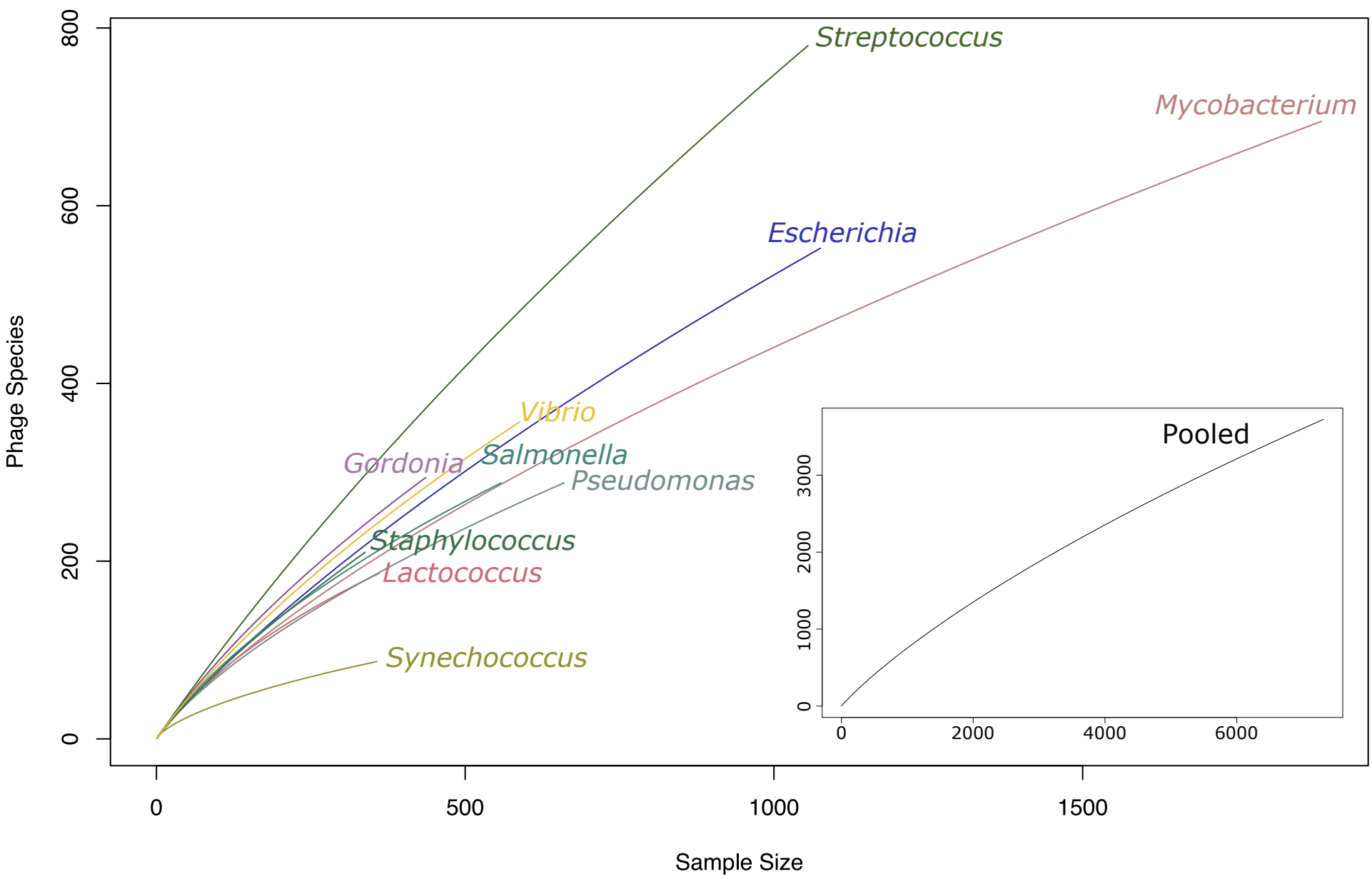


Figure 4

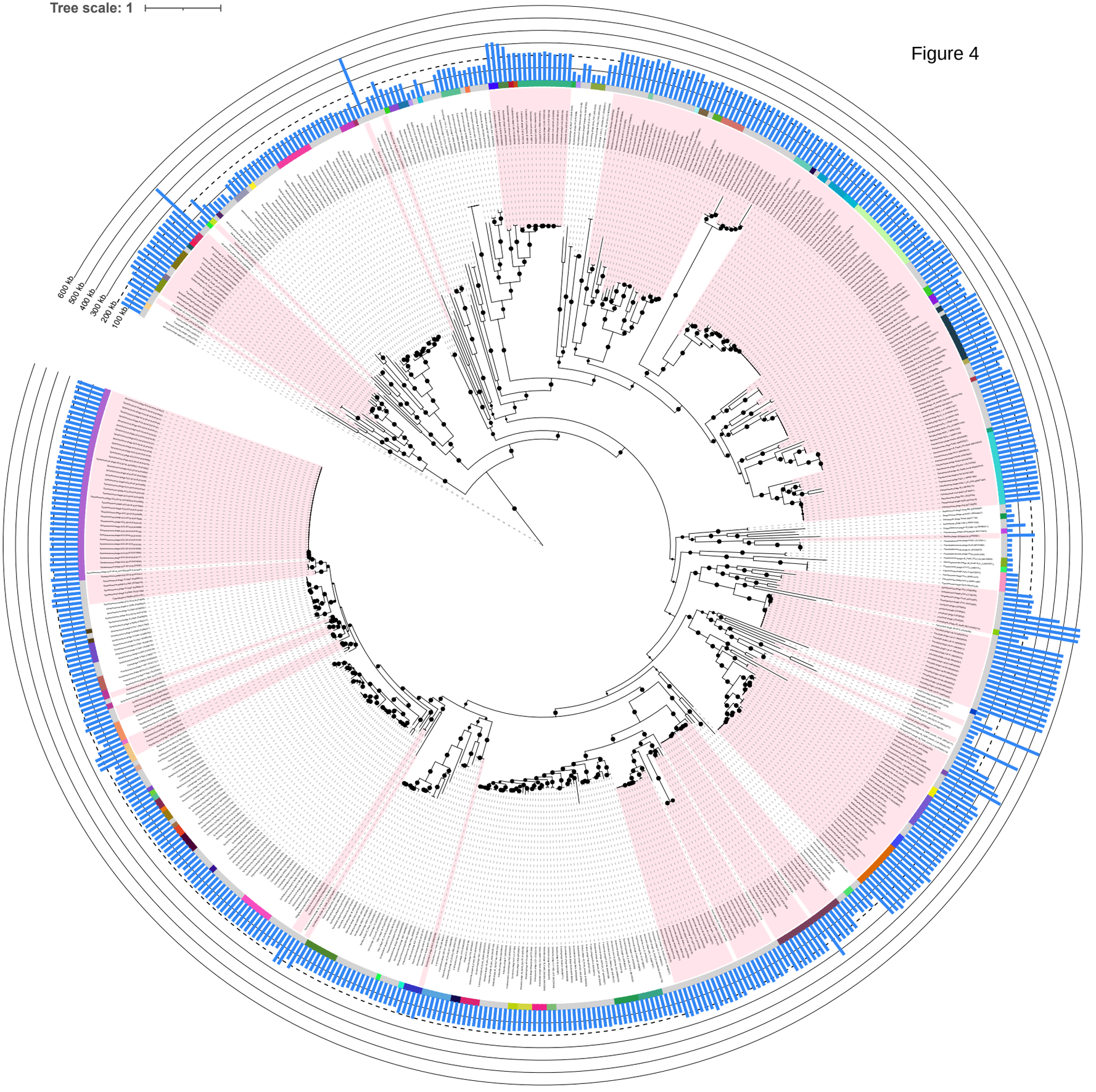


Figure S1

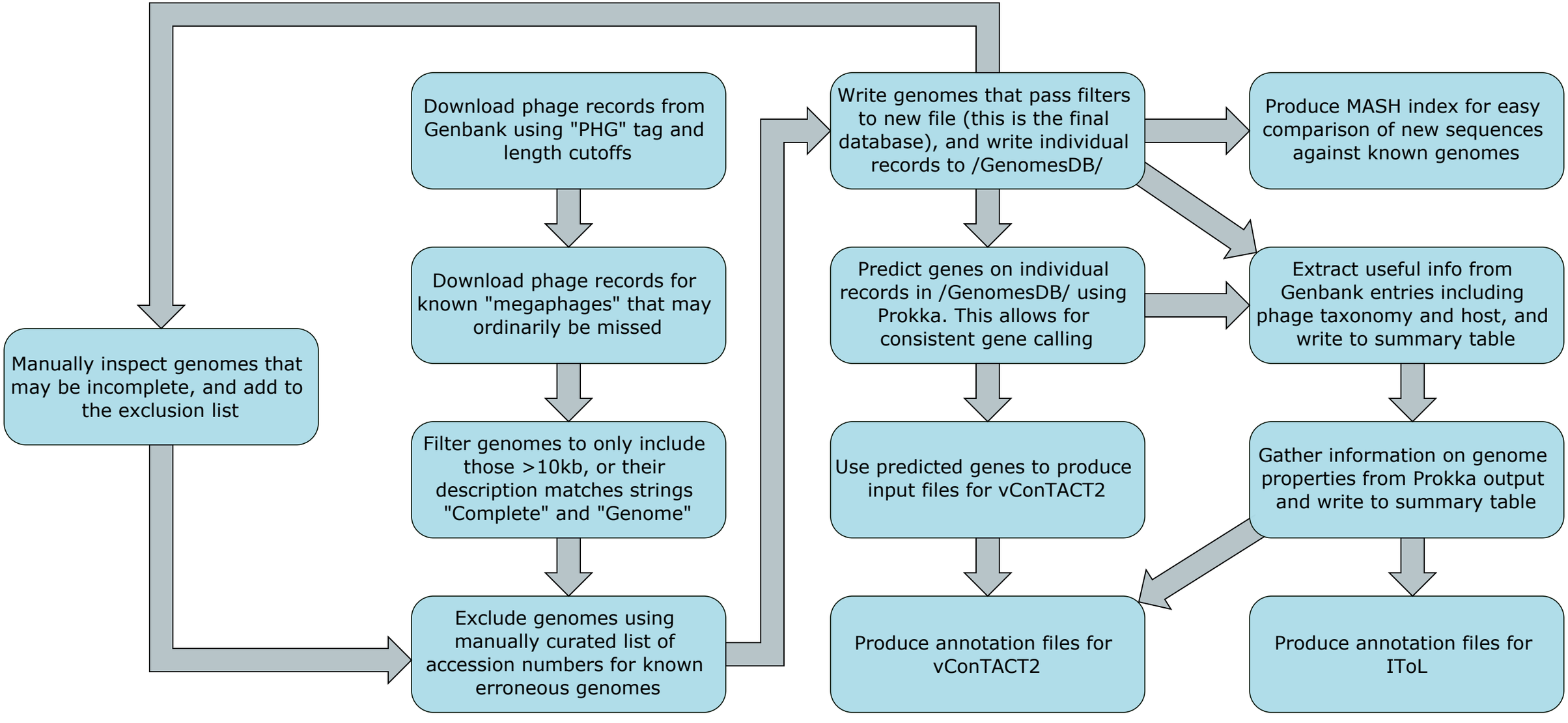
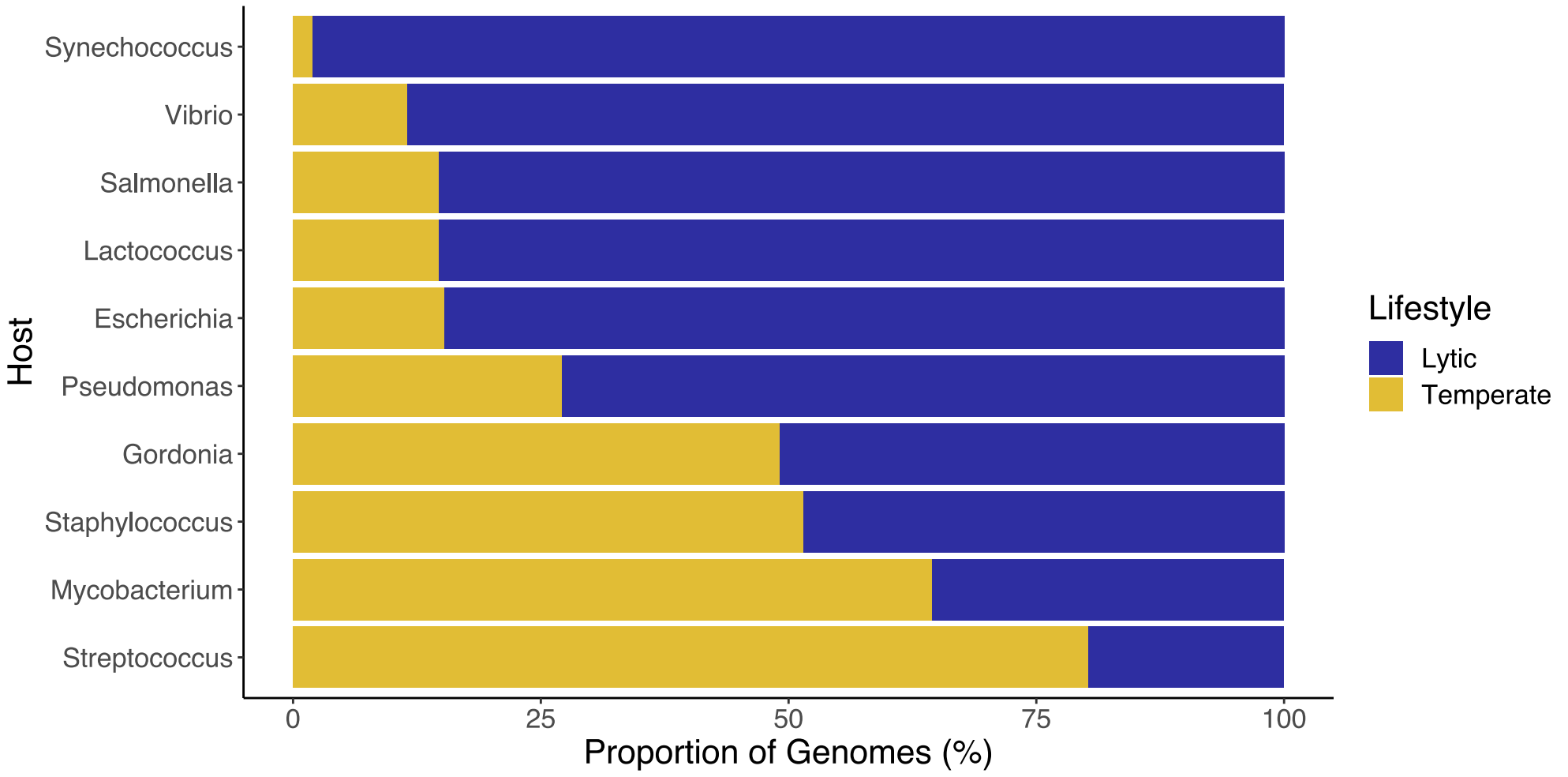
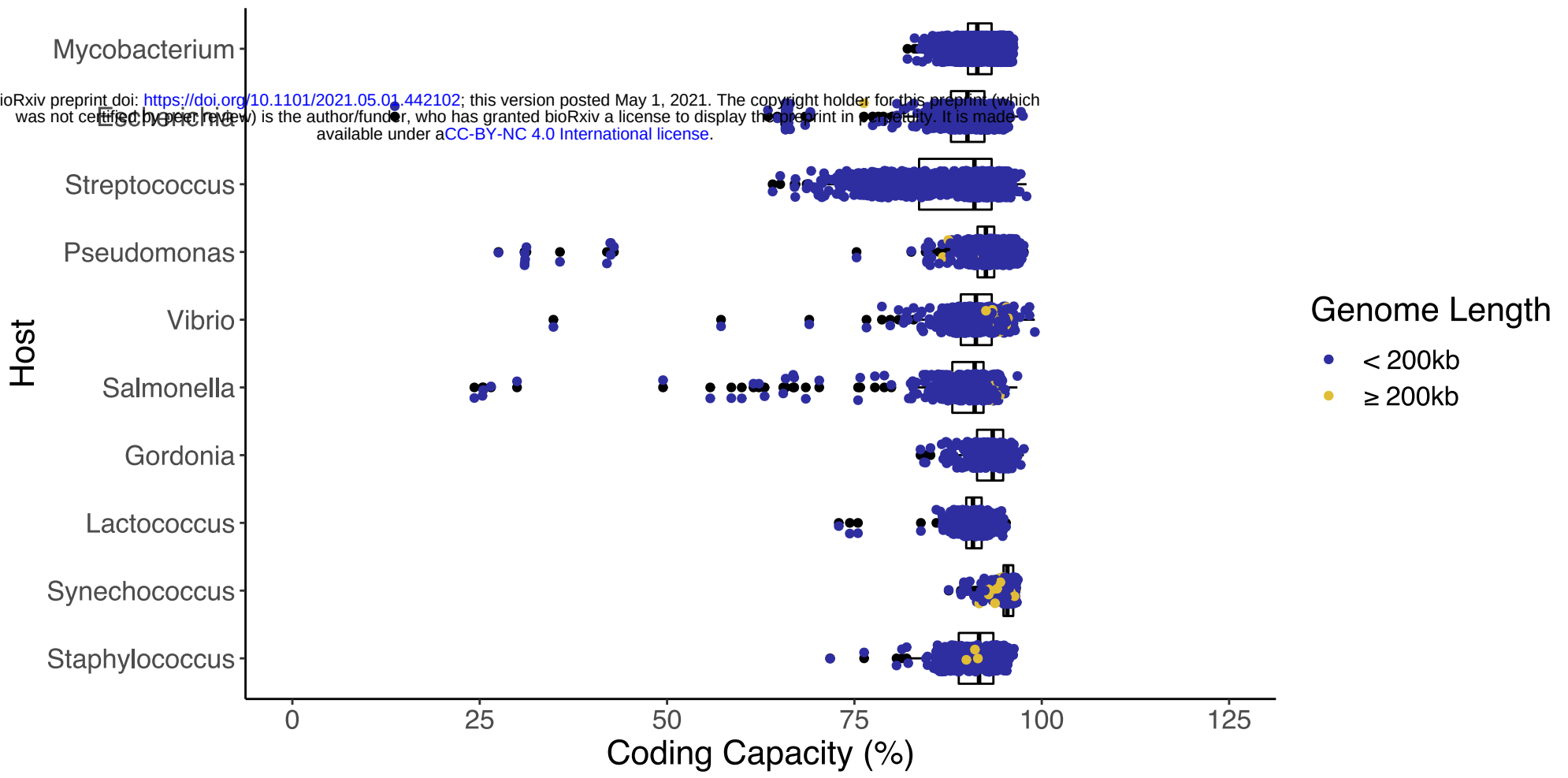

Download phage records from Genbank using "PHG" tag and length cutoffs
Download phage records for known "megaphages" that may ordinarily be missed
Filter genomes to only include those >10kb, or their description matches strings "Complete" and "Genome"
Exclude genomes using manually curated list of accession numbers for known erroneous genomes
Manually inspect genomes that may be incomplete, and add to the exclusion list
Write genomes that pass filters to new file (this is the final database), and write individual records to /GenomesDB/
Produce MASH index for easy comparison of new sequences against known genomes
Predict genes on individual records in /GenomesDB/ using Prokka. This allows for consistent gene calling
Extract useful info from Genbank entries including phage taxonomy and host, and write to summary table
Use predicted genes to produce input files for vConTACT2
Gather information on genome properties from Prokka output and write to summary table
Produce annotation files for vConTACT2
Produce annotation files for IToL

Figure S2

A



B



C

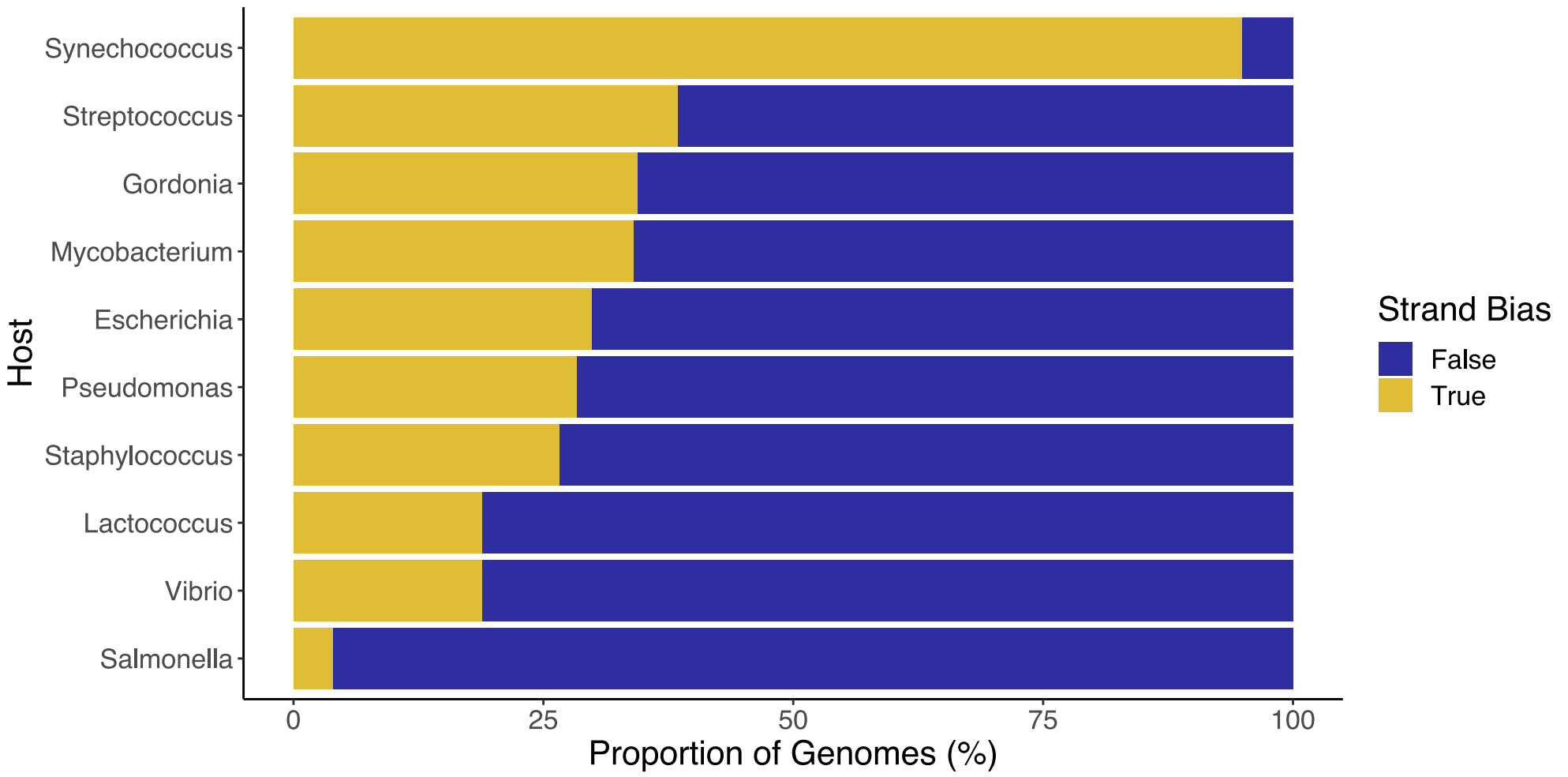
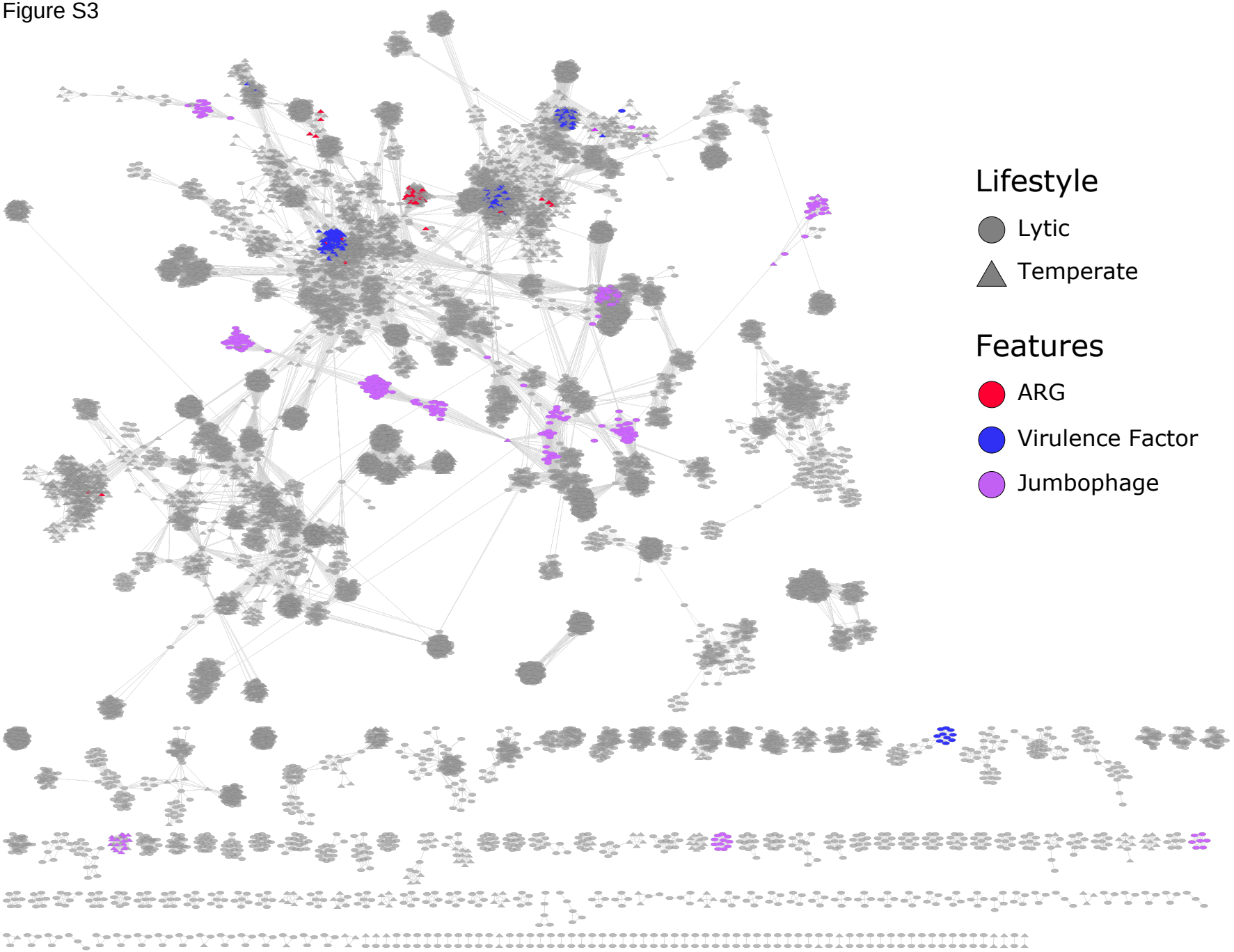


Figure S3


Lifestyle
Lytic
Temperate
Features
ARG
Virulence Factor
Jumbophage