A Genomic Blueprint of the Chicken Gut Microbiome

Rachel Gilroy
Quadram Institute Bioscience
Anuradha Ravi
Quadram Institute Bioscience
Maria Getino
University of Surrey School of Veterinary Medicine
Isabella Pursley
University of Surrey School of Veterinary Medicine
Daniel Horton
University of Surrey School of Veterinary Medicine
Nabil-Fareed Alikhan
Quadram Institute Bioscience
David Baker
Quadram Institute Bioscience
Karim Gharbi
Earlham Institute
Neil Hall
Earlham Institute
Mick Watson
The University of Edinburgh The Roslin Institute
Evelien M. Adriaenssens
Quadram Institute Bioscience
Ebenezer Foster-Nyarko
Quadram Institute Bioscience
Sheikh Jarju
MRC Laboratories The Gambia
Arss Secka
MRC Laboratories The Gambia
Martin Antonio
MRC Laboratories The Gambia
Aharon Oren
Hebrew University of Jerusalem - Edmond J Safra Campus
Roy Chaudhuri

The University of Sheffield Faculty of Science
Falk Hildebrand
Quadram Institute Bioscience
Mark Pallen (mark.pallen@quadram.ac.uk)
Quadram Institute Bioscience https://orcid.org/0000-0003-1807-3657

Research

Keywords: chickens, gut microbiome, biodiversity, metagenomics, culturomics

Posted Date: August 17th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-56027/v1

Abstract

Background: The chicken is the most abundant food animal in the world. However, despite its importance, the chicken gut microbiome remains largely undefined. Here, we exploit culture-independent and culture-dependent approaches to deliver a genomic blueprint of this complex microbial community.

Results: We performed metagenomic sequencing of fifty chicken faecal samples from two breeds and analysed these, alongside all (n=582) relevant publicly available chicken metagenomes, to cluster over 20 million non-redundant genes and to construct over 5,500 metagenome-assembled bacterial genomes. In addition, we recovered nearly 600 bacteriophage genomes This represents the most comprehensive view of the chicken gut associated microbiome to date, encompassing dozens of novel candidate bacterial genera and hundreds of novel candidate species. Keen to provide a stable, clear and memorable nomenclature for novel species, we devised a scalable combinatorial system for the creation of hundreds of well-formed Latin binomials. We cultured bacterial isolates from faeces to deliver 282 whole genome sequences, incorporating thirty novel species, together with three species from the genus *Escherichia*, including the newly named species *Escherichia whittamii*.

Conclusions: Our metagenomic and culture-based analyses provide new insights into the bacterial, archaeal and bacteriophage components of the chicken gut microbiome. The resulting datasets expand the known diversity of the chicken gut microbiome and provides a key resource for future high-resolution taxonomic and functional studies on the chicken gut microbiome.

Background

The domestic chicken is the most abundant bird and most abundant food animal on Earth, accounting for a larger fraction of the planet's biomass than all species of wild birds combined [1]. Consumption of chicken meat is growing faster than any other type of meat and is seen as a cheaper, healthier, low-carbon alternative to meat from mammalian livestock [2, 3]. Chicken eggs remain a nutritious, affordable food across the globe [4].

The chicken gastrointestinal tract is home to a complex community of microbes and their genes—the chicken gut microbiome—that underpins links between diet, health and productivity in poultry, as evidenced by the ability of antibiotics to promote growth in chicks [5]. This microbial community also acts as a source of pathogens associated with disease in birds or in humans—including *Campylobacter*, *Salmonella*, and *Escherichia coli*—as well as providing a reservoir of antimicrobial resistance (AMR) genes [6–8] .

Previous studies of this community have documented a rich variety of microorganisms (dominated by bacteria, but including viruses, archaea and microbial eukaryotes) and have shown that the taxonomic composition of this community varies with age, breed and disease status [9, 10]. However, these earlier efforts have largely relied on analyses of molecular barcodes (in particular short 16S sequences), which

fail to provide species-level resolution, are unable to detect viruses and reveal nothing about the genome sequences, population structures or functional repertoires of microbial species [11].

Two strategies have proven productive for exploring taxonomic and functional diversity in complex microbial communities [12, 13]. Culture-independent approaches rely on shotgun metagenomic sequencing of DNA extracted from relevant samples, followed by bioinformatics-based community profiling and analysis [12, 14, 15]. Culture-dependent approaches (often termed “culturomics”) combine large-scale isolation of microorganisms in pure culture with whole-genome sequencing and phylogenomic analysis [13, 16]. Keen to explore taxonomic novelty in the chicken gut microbiome, we generated phylogenetic profiles to document known and unknown diversity and then exploited culture-dependent and culture-independent approaches to create an unprecedented high-quality reference collection of microbial genes and genomes from the chicken gut, revealing and naming hundreds of new candidate species from this commonplace but important ecological setting.

Methods

Sample collection and storage

Faecal samples were collected in South-East of England from adult Lohmann Brown laying hens and adult Silkie hens in 2018. Birds were housed in a large outdoor run with a substrate of stone chippings and small turf enrichment beds during the day and kept in a coop overnight. They were fed a commercial layer feed and no antibiotics were used. Faecal sampling was approved by the University of Surrey’s NASPA ethics committee.

Sixty faecal samples were collected from the Lohmann Brown laying hens and thirty samples from the Silkie hens (six and three samples per day, respectively, for ten days). Freshly evacuated faeces from individual birds were collected in sterile containers and immediately stored at -20 °C. Samples were then transferred to the laboratory for culture or DNA extraction. DNA was extracted using DNeasy PowerSoil kit (Qiagen), following manufacturer’s instructions and then stored at -20°C.

Sequencing and subsequent workflow

Workflow from this point forward is summarised in Fig. 1.The fifty samples yielding >20 ng DNA were processed according to the Low Input, Transpose Enabled (LITE) library construction pipeline [17] before being subjected to paired-end (2x150bp) metagenomic sequencing on the Illumina Novaseq 6000 platform. Bioinformatics analyses were performed on the Earlham Institute’s High Performance Computing cluster and on the Cloud Infrastructure for Microbial Bioinformatics [18]. Sequences were assessed for quality using FastQC Version 0.11.8 and trimmed using Trimmomatic Version 0.36, configured to a minimum read length of 40, “leading” and “trailing” settings of 3 (SLIDINGWINDOW:4:20)

[19, 20]. Metagenomic sequences for all samples have been uploaded to the Sequence Read Archive under Bioproject ID PRJNA543206.

Reference-based metagenomic analysis

An initial analysis of our chicken faecal sequences using the Kraken 2 taxonomic classifier [21] was performed on custom databases representing the domestic chicken genome (GenBank assembly accession GCF_000002315.6) and the food plants *Triticum aestivum* (wheat), *Aegilops tauschii* (diploid progenitor of the D genome of hexaploid wheat) and *Glycine max* (soy bean): GenBank assembly accessions GCF_001957025.1, GCA_900519105.1, GCA_000004515.4. Kraken 2 revealed that 8% (±16%) of reads originated from the chicken and at least 19% (±21%) originated from the diet. These sequences were filtered from our dataset and excluded from subsequent analyses by keeping only reads 'Unclassified' by Kraken 2 after comparison with each database in turn.

The remaining dataset underwent taxonomic profiling using Kraken 2 against a microbial database built from all complete/representative archaeal, bacterial, fungal, protozoan, viral and UniVec_Core sequences in RefSeq [22] in January 2020. Bracken [23] was used to estimate taxon abundance from the Kraken2 profiles, accepting only those taxa with ≥1000 assigned reads. Bracken-database files were generated using "bracken-build" on our microbial database and visualised using KronaTools [24].

Metagenomic assembly

We searched the NCBI BioProjects database [25] in November 2019 with the term "chicken gut microbiome" and then selected eight publicly available projects that contained at least one metagenomic sequence dataset >1GByte in size (PRJEB33338, PRJNA193217, PRJNA291299, PRJNA375762, PRJNA415593, PRJNA417359, PRJEB22062, PRJNA543206, PRJNA417359, PRJNA385038). We also analysed an additional dataset derived from chicken caecal samples collected in the Gambia [26].

All shotgun metagenomic reads were quality-filtered by removing reads shorter than 70% of the maximum expected read length (100 bp, 250 bp for miSeq data), an estimated accumulated error >2.5 with a probability of ≥0.01 [27] or with an observed accumulated error >2, or >1 ambiguous position to assist assembly. If base quality dropped below 20 in a window of 15 bases at the 3' end, or if the accumulated error exceeded 2, reads were trimmed. All these filter steps are integrated in sdm [28]. Reads mapping to the chicken genome and diet were removed from the metagenomic data as described previously, classifying reads with Kraken 2 [29] against custom databases built on the aforementioned genomes.

Sequence datasets from our fifty samples—together with 582 samples from the selected BioProjects—were assembled using MegaHIT [30] under the option "–k-list 25,43,67,87,101,127". Keen to avoid artefacts that sometimes result from co-assembly of sequences from different samples and different sources, we generally performed individual assemblies on each sample. However, we noted that in BioProject PRJNA17359 multiple metagenomic samples had been sourced from different tissues of the same individual bird, so we co-assembled reads from the 120 BioSamples from that project.

Bacteriophage identification and characterisation

Scaffold sequences from the MegaHIT assemblies of our fifty samples that were ≥10kb were analysed with VirSorter v1.0.5 with the "-db 2" option to identify viral genomes [31]. VirSorter Category 1 and 2 scaffold sequences were collapsed at 95% nucleotide identity over 70% of the sequence length using CD-Hit Est v4.6.1 [32]. Classification of bacteriophage sequences relied on nucleotide searches using BLASTN against the NCBI NT database (Completed April 2020) and protein searches using Kaiju Version 1.7.3 against the RefSeq database (Completed April 2020) [33]. Only bacteriophage genomes with BLASTN hit E-Value <0.05, percentage identity >70% and query covering >50% were selected as reliable hits.

A taxonomic assignment was drawn from the highest scoring BLASTN (or in rare cases BLASTP) hit ranked by query cover and percentage ID. Synteny between predicted coliphages and their respective reference genomes were visualised using EasyFig [34]. *Escherichia* bacteriophage coverage per sample was determined using Anvi'o v6.1 [35] using default parameters and visualised in R using the Pheatmap package [36]. Remaining viral geomes were filtered for completeness, retaining those that were circular and encoded a complete terminase gene (as predicted by VirSorter). Taxonomic assignments to family were performed on viral genomes using Demovir [37].

Gene catalogue

Complete genes identified by Prodigal v2.6.1 [38] were clustered at 95% nucleotide identity using CD-HIT-Est v4.6.1 [32]. Incomplete genes were then mapped to this complete gene list using Bowtie2 v 2.3.4.1 [39] and any mapping at 95% nucleotide identity were incorporated into the relevant gene clusters. Finally, genes belonging to the forty conserved marker genes defined by Mende *et al* [40] were clustered separately and then merged with the existing set of gene clusters. We thus obtained a gene catalogue of >20 million genes, defined as non-redundant at 95% average nucleotide identity.

Abundance estimates of contigs and genes

Prodigal [38] was applied in metagenome-mode to all contigs from the MegaHIT assemblies. Unfiltered reads from each sample were mapped against their respective assembly to provide an estimate of contig and gene abundance using Bowtie2 [39] with the options "–no-unal–end-to-end -score-min L, -0.6,-0.6". Samtools 1.3.1 was used to sort and index all resulting Bam files [41]. Only reads with mapping quality >20, >95% nucleotide identity and >75% overall alignment length were retained. BEDTools v2.21.0 [42] was used to create depth profiles from the Bam files. These depth profiles were then translated with rdCover [43] into average coverage (in a 50 bp window) per contig or per gene predicted from each contig. Bam files were translated to abundances using the "jgi_summarize_bam_contig_depths" script from the MetaBAT 2 package [44].

Gene abundances were linked to their respective gene clusters and originating samples. Redundant genes representing the same orthologue were removed.

Binning

We identified metagenomic species (MGSs) using the combinatorial approach described by Hildebrand *et al* [45], incorporating single-assembly binning in the creation of metagenome-assembled genomes (MAGs), gene catalogue binning in the creation of canopy clusters [46] and hierarchical clustering of candidate genes using the R function hclust, method = complete. To start with, we used MetaBAT 2 v2.15 [44] to bin contigs ≥400bp. These were quality filtered using CheckM v1.0.11 [47] to obtain 5,695 bins at >80% completeness and <5% contamination.

Species-level clusters were formed using a combination of two distinct approaches. One approach removed redundancy between samples by pre-clustering bins if ≥30% of their genes overlapped with a higher-quality bin to create a set of pre-MGS bins. Lower-quality bins (>60% completeness and <10% contamination) were also included in the analysis but were not used to form new species clusters. To recover prokaryotic species usually obscured using single-sample assemblies and conventional binning techniques, we refined all species bins into "hcl-clusters" using gene correlations and hierarchical clustering, as described by Hildebrand *et al* [45]. We chose genes occurring in ≥10% of all associated MAGs as representatives for each pre-MGS bin and used these to fish for additional co-occurring genes from the gene catalogue, using a threshold of >0.75 Pearson correlation and >0.85 spearman rho to identify gene co-occurrences within this core gene set. We then merged MetaBAT 2 bins, canopy bins and co-occurring genes into our species bins. We used the presence of 40 known single-copy marker genes, without duplicates, as a quality criterion in selection of sub-clusters, before extracting the final set of MGS gene representatives using MATAFILER [48]. The final collection of MGS bins (canopy clusters + hcl-clusters) was re-assessed for contamination and completeness using CheckM [47], so that we could be confident that each bin represents a single species. A second approach dereplicated all MAGs at 95%

average nucleotide identity (ANI) (species-level) and 99% ANI (strain-level) using dRep Version 2.0 [49] and only species not identified in approach one were added to the resulting non-redundant species catalogue. A single representative MAG for each species cluster was uploaded to NCBI SRA under BioProject PRJNA543206. CompareM Version0.1.1 [47] was used to calculate average amino acid identity between novel genera.

Taxonomy of metagenomic species

We used the Genome Taxonomy Database Toolkit (GTDB-Tk Release 95) to perform taxonomic assignments on strain-level dereplicated MAGs [50]. In addition, genes from each MGS were analysed through GTDB-Tk (Release 95), proGenomes resource [51] and underwent k-mer-based taxonomic profiling using Kraken 2 [21]. In assigning taxonomy, we allowed GTDB assignments to take precedence—only when no GTDB taxonomy was available would we adopt taxonomies assigned by ProGenomes and Kraken 2 and, then, only where genus and family assignments from these sources matched. When exploiting the taxonomy assigned according genes from metagenomic species, we applied a least-common-ancestor approach to unplaced taxa at higher taxonomic levels. Species distribution analyses were conducted using the Vegan package in R [52], before visualisation using ggplot2 [53] and Pheatmap R packages [36]. Pan-genome analysis was conducted using Roary v3.11.2 and visualised using the roary2svg.pl script [54]. Comparison of our derived metagenomes with those of Glendinning et al. [15] was performed at 95% ANI using dRep and visualised using web-tool BioVenn [55] [49].

Bacterial culture

To estimate species richness and diversity, the Phyloseq package of R [52] was applied to the output from Bracken [23] on all of our chicken faecal metagenomic datasets. The six faecal samples that showed highest species richness and taxonomic diversity were selected for culture-based studies. Frozen faecal samples were thawed, vortexed and two 0.5g aliquots (once processed aerobically, the other anaerobically) from each sample were suspended in 5ml PBS. Since the culture work used here is solely exploratory and in no way used to define the chicken gut microbiome as a whole, impacts of sample storage on bacterial recovery rate will provide negligible influence the on final conclusions. Each aliquot was vortexed until homogenised, before performing serial dilutions in duplicate down to 1×10^{-5}. Processing of samples for aerobic and anaerobic culture was identical, except that, for anaerobic culture, all culture media, diluent and consumables were pre-reduced to anaerobic conditions for at least 24 hours before faecal samples were processed in a Whitley A95TG workstation.

For dilutions 10^{-3} – 10^{5}, 200µl was plated directly on to three agar plates of broad culture medium with or without vancomycin supplementation at a concentration of 6µg/ml (Additional File 1, Supplementary Table 1). Cultures were incubated at 37°C for 72 hours in their respective conditions before assessment of colony growth. Well-isolated colonies were picked according to colonial morphotype distinctive in colour, shape and size, before being re-streaked on to the growth medium from which they were sourced to confirm purity. Individual colonies were subsequently used to inoculate 2ml of broth based on the source culture medium, incubated at 37°C for a further 24 hours before bacterial DNA extraction. All isolates were archived at -80°C in glycerol at 20% concentration.

Genome sequencing and analysis

Genomic DNA was extracted using a DNeasy UltraClean DNA isolation kit according to the manufacturer's instructions (Qiagen, Hilden, Germany). DNA was quantified using a Qubit® fluorometer (Invitrogen, CA, USA) high-sensitivity assay, before dilution to the required concentration in RNase-free water and purification on AMPure XP beads (Beckman Coulter). Sequencing libraries were prepared from 0.5ng/µl of RNA free genomic DNA. A total of 282 isolates were included for genomic sequencing using the Nextera-XT DNA sample preparation kit (Illumina) and whole-genome sequencing performed using the Illumina NextSeq sequencing platform, generating paired-end reads (2 x 150bp). Reads were uploaded to the Sequence Read Archive under Bioproject ID PRJNA543206.

Paired-end reads were quality-assessed and trimmed using FastQC and Trimmomatic as described above [19, 20]. Trimmed reads were assembled into scaffolds using SPAdes version 3.13.1 [56]. Scaffolds shorter than 500 bp were discarded from analysis. Genome contamination and completeness was assessed using CheckM version 1.0.13 [47]. To confirm assembly quality, only genomes conforming to all the following criteria were included in further analysis: (i) scaffold N50 of >20 kbp (ii) 90% of assembled bases at > 5x read coverage (iii) completeness of > 95% (iv) contamination of < 5% (v) complete 16S rRNA gene sequence.

Genome sequence taxonomic assignment

Barrnap Version 0.9 [57] was applied to all genomes that passed the quality filters to extract full-length 16S rRNA gene sequences. These were then compared to NCBI 16S rRNA gene sequences from RefSeq genomes using the NCBI's web-based BLASTN facility [58]. 16S rRNA sequences that showed an identity of <98.7% to known sequences were assigned to novel species, using the conservative approach of [59]. We used ReferenceSeeker Version 1.6.2 [60] to determine average nucleotide identity (ANI) and conserved DNA values compared to RefSeq bacterial genomes (Completed March 2020) [22]. Genomes that showed

ANI ≤95% and conserved DNA ≤69% to the closest relative were designated novel species. The Genome Taxonomy Database Toolkit (GTDB-Tk Release 89) was used to perform taxonomic assignments on isolate genomes [50]. Genomes were clustered at 95% and 99% ANI before selection of a single representative isolate per species using dREP [49]. Where a genome previously designated as novel clustered with a genome of assigned taxonomy, this taxonomy was then applied to the previously designated 'novel' genome. Final taxonomic assignments were based on genome-based ANI values derived from RefSeq and GTDB – with GTDB assignments taking precedence.

Phylogenetic analysis

For phylogenetic analysis of all MGS and genome sequenced isolates we used Anvi'o v6.1 [61] to retrieve protein sequences associated with the set of 40 conserved marker genes for metagenomic species described above and representative genomes of dereplicated cultured species [40]. We performed a multiple sequence alignment on concatenated sequences of the marker gene products from each species using Muscle v3.8.31 [62]. We then built an approximate-maximum-likelihood phylogenetic tree using FastTree v2.1 [63] and visualised used the online iTOLv1.4 platform for visualisation and manual annotation [64]. This was used to confirm that species and genera were monophyletic.

To investigate the phylogenetic placement of cultured isolates designated as *Escherichia marmotae* and *Escherichia* sp001660175 by GTDB, we constructed a core genome phylogenetic tree. The genomes from cultured isolates were compared to genomes representing the full diversity of the genus *Escherichia* (Additional File 1, Supplementary Table 5). Three *Salmonella* genomes were included as an outgroup. The genome sequences were aligned using Mugsy [65], and alignment blocks conserved across all genomes were concatenated to produce a core genome alignment. A phylogenetic tree was constructed by maximum likelihood with 100 rapid bootstrap replicates, using the general time reversible model of nucleotide substitution with gamma correction for rate heterogeneity, as implemented in RAxML version 8.2.12 [66].

Results

Reference-based profiling documents novel diversity

We collected faecal samples from fifty chickens reared in the UK belonging to two breeds: Lohman Browns (n=30) and Silkies (n=20). Short-read sequencing of fifty faecal samples generated a metagenomic dataset in excess of a billion paired-end reads or three hundred billion base pairs (Additional File 1, Supplementary Table 2).

In recent years, bioinformatics methods have been developed for ultrafast, highly accurate phylogenetic profiling of metagenomic sequences that rely on matching short sequences (*k-mers*) to a reference

database built from sequenced genomes [21]. Such methods can assign sequences within a taxonomic hierarchy that extends from domains to species and can then quantify the representation of each taxon within a sample. Although these methods have been used in a very focused way to detect *Campylobacter* in chickens [67], they have yet to be applied to a global analysis of microbial diversity in the chicken gut microbiome. We therefore initially analysed the faecal samples using the k-mer-based program Kraken 2, followed by refined phylogenetic analysis using the allied program Bracken [23] (Additional File 1, Supplementary Table 3).

Unsurprisingly, Kraken 2 and Bracken assigned sequence reads from the faecal samples to all three domains of life, as well as to viruses (Fig. 2a, Additional File 1, Supplementary Table 4), although relative abundance assignments show that bacteria predominate in this environment. Sequences were assigned to a wide range of bacterial phyla, including the three expected as predominant in the vertebrate gut (Bacteroidetes, Firmicutes, Proteobacteria), but also including over twenty additional phyla (Fig. 2b). Searches of the PubMed database with each phylum name and the term “chicken” reveal that round half of these have been previously documented in the chicken gut. However, at least a dozen appear to be novel in this setting, including the *Aquificae, Balneolaeota, Calditrichaeota, Chlorobi, Dictyoglomi, Fibrobacteres, Gemmatimonadetes, Ignavibacteriae, Kiritimatiellaeota, Lentisphaerae, Nitrospirae,* and the *Thermodesulfobacteria*.

When we rank-ordered the species identified by Bracken according to maximum abundance in any one sample, we found, as expected, that species of *Lactobacillus* dominated among the top twenty most abundant organisms. However, we found that two species of *Escherichia*— *Escherichia coli* and *Escherichia marmotae*—accounted for ≥ 5% of reads in nearly half of the samples (23/50) and in two samples, accounted for more than 50%. Such monodominance of the gut microbiome by bacterial species has been described in diseased humans [45, 68], but is surprising in the context of healthy poultry. We also noted a high relative abundance of the recently described chicken pathogen *Gallibacterium anatis* [69] in most birds (with five birds showing >5% reads assigned to this organism), despite their healthy status. Similarly, *Fusobacterium mortiferum*— an opportunistic pathogen of humans [70]—accounted for >10% of sequences in ten birds, corroborating a recent report of high abundance of 16S sequences from this organism obtained from the chicken caecum [71].

Bracken assigned sequences to over a hundred bacteriophage genomes, predominately phages infecting members of the *Enterobacteriaceae* assigned to the families *Myoviridae* and *Podoviridae*. Particularly noteworthy was the high abundance of reads in some samples from two distinct bacteriophages that prey on *E. coli*: phiEcoM-GJ1—a lytic bacteriophage isolated in Canada from pig sewage [72]—which accounted for 6.7% reads in a single sample and phAPEC8—a lytic bacteriophage with a large 147kb genome, isolated from a Belgian poultry farm—which accounted for 10% of reads in a single sample and for >1% of reads in three others [73].

Although we were pleased to see that these k-mer-based analyses can provide interesting insights into taxonomic diversity within the chicken gut, we quickly realised that they provide an incomplete and

misleading picture of this important microbiome for several reasons: (1) they often report the presence of highly implausible organisms—for example, Kraken 2 reported the presence of human pathogens such as *Shigella flexneri* and *Plasmodium falciparum* that are simply not credible in this context on clinical grounds; (2) as with 16S studies, they fail to provide genomic data or insights into the functional diversity or population structure of the microbial species that they identify and; (3) they rely on a reference database and so can only report previously known organisms and can never uncover “unknown unknowns”.

The scale of the problem of unknown diversity is clear from the observation that nearly three quarters (73%) of sequence reads from our chicken samples cannot be confidently classified by Kraken 2 to species level and more than half of the reads (54%) cannot be classified at all and are simply designated as “Unassigned” (Fig. 2a). We therefore sought to extend our understanding of this community through two powerful reference-free approaches: assembly-based metagenome analyses and high-throughput culture.

Metagenomic assembly uncovers a wealth of viral diversity

Assembly of metagenomic sequences is a reference-free approach that involves aligning and merging short sequence reads into long contiguous sequences (contigs), which can then be ordered into larger scaffolds that include sequence gaps.

Keen to confirm the presence of bacteriophages inferred through the reference-based analysis and to identify novel viral genomes, we assembled sequence reads from our fifty chicken faecal samples into scaffolds. Scaffold sequences ≥10kb were analysed with VirSorter—a program designed to detect viral signals in microbial sequence data to find novel viruses [31].

VirSorter identified 184 of our chicken faecal scaffolds as Category 1 (“most confident”) bacteriophage sequences and identified an additional 1,840 scaffolds as Category 2 ("likely”) bacteriophage sequences. This was de-replicated to 1,455 genomes using similarity thresholds of 95% ANI over 70% of the genome (Additional File 2, Supplementary Table 1). BLASTN analysis revealed only ten of these bacteriophage genomes showed high similarity (percentage identity > 70%; query covering > 50%) to known phages at the nucleotide level (Additional File 2, Supplementary Table 2). These included close relatives of the two phages (phiEcoM-GJ1 and phAPEC8) found highly abundant in the Bracken analyses (Fig. 3). Interestingly, more than one genus of coliphage (e.g. *Jilinvirus, Phapecoctavirus,* or *Gamaleyavirus*) was often detected in the same sample, along with an abundance of reads from their predicted prey (*Escherichia*) suggesting interesting dynamics in phage-host and phage-phage interactions (Fig. 3; Additional File 2, Supplementary Table 3).

Of the remaining 1,445 unclassified bacteriophage genomes, nearly 600 encoded either an obvious terminase region or were circular and as such were suggested as being near-complete. Classification of these genomes revealed all genomes were predicted to belong to the order *Caudovirales* of tailed phages, with the majority belonging to the family *Siphoviridae* (n=429), but we also found representatives from the *Myoviridae* (n=87) and *Podoviridae* (n=27), plus some bacteriophages unclassified at family level (n=28) (Additional File 2, Supplementary Table 4).

Remarkable microbial genome diversity in the chicken gut

Next, we subjected our samples to computational binning—a process of grouping contigs/scaffolds on the basis of sequence composition and depth of coverage into discrete population bins representing metagenome-assembled genomes (MAGs). However, to carry out a definitive survey of bacterial and archaeal diversity in the chicken gut microbiome—in addition to analysing the fifty faecal samples mentioned and before we started the binning—we retrieved all publicly available chicken gut metagenomic datasets, to create an expansive dataset representing >630 samples, drawn from ten studies and twelve countries (Belgium, China, France, Germany, Italy, Malaysia, Netherlands, Poland, Spain, The Gambia, UK, USA) (Figure S1a/S1b; Additional File 3, Supplementary Table 1).

Sequence assembly and binning on all these samples generated 5,595 MAGs that passed our quality threshold of ≥80% completion and ≤5% contamination (Figure S1c). Of these 3,131 could be considered high-quality draft genomes, with >90% completion and <5% contamination, as judged by recently published criteria (Additional File 3, Supplementary Table 2) [74]. Genome sizes of the MAGs ranged from ~0.5 to 6.4 Mbp, while GC content ranged from 24% to 73%.

Then, we grouped the MAGs into metagenomic species (MGSs). Initially, this involved de-replicating MAGs at the widely accepted 95% average nucleotide identity (ANI) for defining bacterial and archaeal species and 99% ANI for defining bacterial and archaeal strains [75, 76]. De-replication of MAGs at 95% ANI resulted in 846 clusters representing bacterial and archaeal species, while de-replication at 99% ANI resulted in 2182 clusters, representing strains. However, to improve recovery of MAGs, MGSs and associated gene sets, we used gene correlations to identify species-representative genes and then applied hierarchical clustering to co-occurring genes across the samples. This allowed us to identify additional genes from the core genome of a species, even when they show divergent nucleotide compositions (such as genes from genomic islands and plasmids) [45]. Similarly, using canopy clustering [46], we could identify commonly occurring species of low abundance. Using these approaches, we were able to identify an additional seven MGSs (Additional File 3, Supplementary Table 3).

Analysis of bacterial metagenomic species, primarily using the Genome Taxonomy Database (GTDB) taxonomy [50], confirmed and extended the taxonomic novelty uncovered by reference-based community profiling (Fig. 4), recovering species spanning nineteen of the bacterial phyla defined by GTDB (Additional

File 3, Supplementary Table 4). These include *Cyanobacteria* (12 species, 32 strains); *Deferribacterota* (1 species, 1 strain) *Synergistota* (2 species; 5 strains) and the *Verrucomicrobiota* (7 species; 8 strains).

Of the 853 de-replicated bacterial metagenomic species, 321 represented previously delineated species catalogued in publicly available databases (Additional File 3, Supplementary Table 4). Following direct comparison, a further 165 metagenomic species had been previously identified by Glendenning *et al* [15], with these sequences not currently available in public archives. However, only 158 of our metagenomic species possess validly published names based on Latin binomials (Additional File 4, Supplementary Tables 1 and 2).

We performed a search of PubMed with the species name and "chicken", leaving aside the 33 species named by Glendenning *et al* [15]. This suggested that our study provides the first-evidence-in-chickens for the majority (81/125) of these species (Additional File 4, Supplementary Table 3). Examples include: *Jeotgalicoccus halophilus,* first isolated from the traditional fermented seafood, Jeotgal [77]; *Aliicoccus persicus*, first isolated from a hypersaline lake [78]; and *Bacteroides reticulotermitis,* first isolated from the gut of a termite [79]. A search of PubMed with the species name and the terms "humans" and infection" suggests that nearly 40% of these known species have been associated with human infection, highlighting the role of the chicken gut as a source of zoonotic pathogens.

We found that 310 of our metagenomic species could be assigned a taxonomy only at the level of genus and so represent novel candidate species. A further 56 species could be assigned a taxonomy only at the level of family and, after AAI clustering at 60%, were assigned to 36 novel candidate genera. One candidate bacterial species could be assigned a taxonomy only at the level of order (*Oscillospirales*) and so represent a new family.

Three MAGs were assigned to the domain Archaea. One represents the species *Methanobrevibacter woesei*—which is already known to inhabit the chicken gut [80]—while the other two represent novel species within the genera *Methanocorpusculum* and UBA71.

Linnaean binomials for hundreds of new candidate species

Linnaeus first proposed the assignment of Latin binomials to provide a universal nomenclature for biological species [81]. The International Code of Nomenclature of Prokaryotes (ICNP) sets the rules for naming prokaryotic species [82], but currently precludes the publication of names of uncultivated organisms, represented by MAGs or other sequences. Furthermore, high-throughput generation of MAGs and of sequence-based taxonomies for bacteria, such as the GTDB [50] is often assumed to preclude the detailed attention usually given to one-by-one construction of Linnaean binomials. As a result, most uncultured taxa, as well as many taxa defined on sequence-based criteria, have been assigned unstable, confusing and hard to-remember alphanumerical identifiers.

Keen to provide a stable, clear and memorable nomenclature for novel and/or previously unnamed bacterial and archaeal species from the chicken gut, we exploited the provision within the ICNP for naming uncultivated taxa via *Candidatus* assignments, which, although provisional, provide the scientific community with well-formed Latin binomials [83, 84]. However, this prompted us into an unprecedented effort to create hundreds of new names for the purpose of this single research study—an effort that required us to devise a scalable combinatorial system for the creation of binomials. Here, we made extensive combinatorial use of around twenty Latin and Greek roots pertaining to poultry (*avi-, galli-, pulli-, alektryo, ptero, kotto-, ornitho-*), intestines (*intestini- entero-*), faeces (*faec-, kakke, merd-, kopro-, excrement-*) or microbial life *(-monas, -bacterium, -microbium, -coccus, -bacillus, -bium, -cola*)—twinned with addition of these roots (singly or in tandem) and/or prefixes (*allo, hetero, meta-, para-, crypto-*) to existing genus names—to create over 200 *Candidatus* genus names. An additional source of diversity stemmed from repetitive use of around forty *Candidatus* species epithets built from similar roots, which when combined with genus names gave us a total of over 600 distinctive new binomials.

Taxonomic diversity of cultured bacterial isolates

To extend our metagenomics analyses, we applied culture-based methods to six faecal samples that appeared species-rich in Kraken 2 analyses and in so doing obtained 282 isolates from aerobic culture (~80% of isolates) and anaerobic culture (~20% of isolates) (Additional File 5; Supplementary Table 1). All isolates underwent genome sequencing on the Illumina platform and phylogenetic analysis to enable taxonomic assignment. The resulting chicken gut culture collection was found to contain 56 genera, 93 species and 162 strains drawn from five phyla. These included thirty novel species. As with the metagenomic species, all novel or previously unnamed genera and species from cultured isolates were assigned Linnaean binomials (Table 1; Additional File 5, Supplementary Table 2).

Interestingly, alongside ten cultured isolates of the well-characterised species *Escherichia coli,* we recovered three isolates from *Escherichia marmotae* (a species recently described in Himalayan marmots [85]). As previously reported [86], the *E. marmotae* strains cluster closely with the *Escherichia* Clade V [87] , so all members of this clade should be considered members of this species (Fig. 6, Additional File 5; Supplementary Table 3). Further analysis of the GTDB species designated *Escherichia* sp001660175 [88] confirmed that this species forms a monophyletic lineage that corresponds to the Clade II, among the cryptic environmental clades described by Whittam and his colleagues [89], which was subsequently documented in birds [90]. As Clade II is comparable in divergence to the other *Escherichia* spp. and cryptic clades, we have therefore assigned the Linnaean binomial *Escherichia whittamii* to designate a new species (Table 1), honouring the outstanding contribution of Thomas S. Whittam to the study of *Escherichia* spp. [91].

We found that only sixteen species were common to our cultured isolates and our MGSs. Subsequent sequence mapping allowed us to detect a further two cultured species at ≥1x coverage in at least one

metagenomic sample (Fig. 5; Additional File 5, Supplementary Table 4), The genomes from cultured isolates were on average 20% larger than the corresponding MAG sequences retrieved from the same source sample (Additional File 5, Supplementary Table 5), which is in line with the completeness threshold of 80% we adopted in quality assurance of the MAGs. However, when we performed detailed gene content analyses on three abundant species in both cultured and metagenomic datasets — *Lactobacillus reuteri (with the* synonym *Limosilactobacillus reuteri), Escherichia coli* (including the synonym *Escherichia flexneri*) and *Enterococcus faecium*—we found that >99% of the genes from the core genomes and nearly half of the genes in the accessory genomes of cultured species were represented in at least one MAG (Figure S2). These observations suggest that our high-quality MAGs are sufficiently complete to warrant *Candidatus* names.

We analysed our chicken faecal metagenomes with a Kraken 2 database derived from genomes representing our candidate metagenomic and cultured species, this yielded a considerable improvement in the number of reads that can be classified through rapid phylogenetic profiling (Figure S3).

Distribution of microbial species

An analysis of the distribution of 820 MGSs across the entire metagenomic dataset revealed marked variation between samples, with not a single species present at ≥1x coverage in all samples and only 39 species present in >90% of samples— although 441 species were present in >50% of samples at ≥1x coverage (Fig. 7; Additional File 5, Supplementary Table 6).

Among the species with high coverage, frequency is clearly linked to Bioproject. Although species quantification curves showed that the number of species identified increased rapidly with the number of samples, species discovery appeared to plateau at approximately 230 species after including only 50 metagenomes (Figure S4). Only two species appeared to be restricted (at ≥1x coverage) to just a single sample: *Aliarcobacter thereius* and *Candidatus* Avibacteroides faecavium. Correlation clustering confirmed structure in the data linked to BioProject (Figure S5) —for example, the BioProject from the study by Glendenning *et al* [15] clearly shows enhancement of clostridial species compared to other BioProjects, reflected samples sourced from chicks with no post-hatching contact with an adult bird. However, the BioSamples do not appear to cluster by country and unfortunately metadata for other potentially important factors, such as breed, age or diet, is not adequate enough to draw conclusions on how these might influence clustering.

Discussion

Given the dominance of chickens in the planetary biomass, the chicken gut microbiome ranks as one of the most abundant microbial communities on the planet. Here, we have exploited two complementary approaches—metagenomics and culture—to create an extensive catalogue of genes, genomes and

isolates from this important ecosystem. Our work illustrates the value of combining culture-dependent and culture-independent approaches in analysing microbiomes.

We have clearly demonstrated the advantages of shotgun metagenomic sequencing, and allied bioinformatics approaches [92], when applied to the chicken gut microbiome, providing catalogues of genes and genome sequences that takes us well beyond what can be achieved using 16S ribosomal RNA gene sequences. However, the limited overlap between bacterial species represented among our cultured isolates and in our MGS reinforces the utility of the combined approach. Nonetheless, the substantial co-linearity between genomes obtained by the two approaches—and with those from another similar metagenomic study [15]—confirms the reliability of our binning approaches.

We were surprised to find such a remarkable phylogenetic diversity within this commonplace livestock ecosystem—diversity that rivals that associated with the human gut. Our work has more than doubled the number of bacterial species known to reside in the chicken gut and has resulted in the creation of an unprecedented number of new *Candidatus* species. By including well-formed Latin binomials with the genomes we have uploaded into public repositories, we have ensured that the new proposed names and associated sequences will be integrated into commonly used online taxonomies and databases [22, 93] and will provide a stable taxonomic nomenclature for future studies. In addition, we have provided proof-of-principle for a scalable approach to Linnaean nomenclature that could be applied to species recovered from other metagenomic assembly projects [94].

Given that we did not recover by culture some of the organisms that appear most abundant by metagenomics, there is clearly scope for additional culture-based investigations, using a wider range of cultural conditions—perhaps drawing on the precedent of the Human Microbiome Project to create and target a list of the “most-wanted-for-culture” organisms documented by metagenomics [95]. The fact that novel metagenomic species are still being recovered from human gut datasets that include tens of thousands of metagenomes [12]—twinned with the promise of novel long-read and proximity-capture approaches to metagenome analyses [96]—make it clear that our attempts here to analyse all currently available chicken gut metagenomes provide far from the last word on microbial diversity in this abundant and important ecosystem. Nonetheless, the availability of so many novel genes, genome and species represents a great leap forward.

Conclusions

The extensive catalogue of genes, genomes and isolates we have created here substantially improves the coverage of the chicken gut microbiome in the public databases and will make it possible to profile sequences from the chicken gut much more rapidly, easily and comprehensively, providing a valuable resource that lays the ground work for future comparative and intervention studies. We also offer a provocative precedent—relevant not just to animal microbiomes, but to studies on all microbiomes—assigning well-formed Latin binomials to hundreds of metagenomic species in a scalable alternative to

the automated use of bland, unstable, user-unfriendly alphanumerical designations. We hope that others will agree that it is now time to bring Linnaeus right into the heart of microbiome studies.

Abbreviations

AAI: Average amino acid identity

AMR: Antimicrobial resistance genes

ANI: Average nucleotide identity

GTDB: Genome Taxonomy Database

ICNP: International Code of Nomenclature of Prokaryotes

LITE: Low Input, Transpose Enabled

MGS: Metagenomic species

MAG: Metagenome-assembled genome

PBS: Phosphate buffered saline

SRA: Sequence Read Archive

Declarations

Ethics approval and consent to participate

Faecal sampling conducted as part of this study was approved by the University of Surrey NASPA ethics committee.

Consent for publication

Not applicable.

Availability of data and materials

The datasets generated and/or analysed during the current study are available in the NCBI BioProject ID PRJNA543206 via this link https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA543206

Competing interests

The authors declare that they have no competing interests.

Funding

This research is supported by the Quadram Institute Bioscience BBSRC-funded Strategic Program: Microbes in the Food Chain (project no. BB/R012504/1) and its constituent project BBS/E/F/000PR10351 (Theme 3, Microbial Communities in the Food Chain) and by the Medical Research Council CLIMB grant (MR/L015080/1) and the British Egg Marketing Board Research and Education Trust. EMA and FH were funded by the BBSRC Institute Strategic Programme Gut Microbes and Health BB/r012490/1, its constituent project BBS/E/F/000Pr10353 and BBS/E/F/000PR10356.

Authors' contributions

RG, AR contributed to the study design, processing, analysis, and interpretation; and manuscript preparation. EFN, SJ, AS, MA contributed to sample collection and processing. NH, DB, KG contributed to sample processing. NFA, EMA contributed to the data analysis. FH contributed to the study design, data analysis interpretation, and manuscript preparation. MW contributed to the study design, data analysis and interpretation, and manuscript preparation. AO contributed to the naming of the candidate taxa and to the manuscript preparation. MJP conceived of the study in design, coordination and manuscript preparation. All authors read and approved the final manuscript. RML, DLH, IP and MG contributed to the collection and processing of the samples and metadata. RC contributed to data analysis and interpretation and manuscript preparation.

Acknowledgements

The authors thank the farmers for collecting the chicken faecal samples for the study.

References

1. Bennett CE, Thomas R, Williams M, Zalasiewicz J, Edgeworth M, Miller H, et al. The broiler chicken as a signal of a human reconfigured biosphere. R Soc Open Sci. 2018;5:180325.
2. Eshel G, Shepon A, Makov T, Milo R. Land, irrigation water, greenhouse gas, and reactive nitrogen burdens of meat, eggs, and dairy production in the United States. Proc Natl Acad Sci USA. 2014;111:11996–2001.
3. Willett W, Rockström J, Loken B, Springmann M, Lang T, Vermeulen S, et al. Food in the Anthropocene: the EAT-Lancet Commission on healthy diets from sustainable food systems. Lancet. 2019;393:447–92.
4. Réhault-Godbert S, Guyot N, Nys Y. The Golden Egg: Nutritional Value, Bioactivities, and Emerging Benefits for Human Health. Nutrients. 2019;11.

5. Bedford M. Removal of antibiotic growth promoters from poultry diets: implications and strategies to minimise subsequent problems. Worlds Poult Sci J. 2000;56:347–65.
6. Florez-Cuadrado D, Moreno MA, Ugarte-Ruíz M, Domínguez L. Antimicrobial Resistance in the Food Chain in the European Union. Adv Food Nutr Res. 2018;86:115–36.
7. Jørgensen SL, Stegger M, Kudirkiene E, Lilje B, Poulsen LL, Ronco T, et al. Diversity and Population Overlap between Avian and Human Escherichia coli Belonging to Sequence Type 95. mSphere. 2019;4.
8. Hermans D, Pasmans F, Messens W, Martel A, Van Immerseel F, Rasschaert G, et al. Poultry as a host for the zoonotic pathogen Campylobacter jejuni. Vector Borne Zoonotic Dis. 2012;12:89–98.
9. Rychlik I. Composition and Function of Chicken Gut Microbiota. Anim. 2020;10.
10. Shang Y, Kumar S, Oakley B, Kim WK. Chicken Gut Microbiota: Importance and Detection Technology. Front Vet Sci. 2018;5:254.
11. Hillmann B, Al-Ghalith GA, Shields-Cutler RR, Zhu Q, Gohl DM, Beckman KB, et al. Evaluating the information content of shallow shotgun metagenomics. mSystems. 2018;3.
12. Almeida A, Mitchell AL, Boland M, Forster SC, Gloor GB, Tarkowska A, et al. A new genomic blueprint of the human gut microbiota. Nature. 2019;568:499–504.
13. Forster SC, Kumar N, Anonye BO, Almeida A, Viciani E, Stares MD, et al. A human gut bacterial genome and culture collection for improved metagenomic analyses. Nat Biotechnol. 2019;37:186–92.
14. Sergeant MJ, Constantinidou C, Cogan TA, Bedford MR, Penn CW, Pallen MJ. Extensive microbial and functional diversity within the chicken cecal microbiome. PLoS One. 2014;9:e91941.
15. Glendinning L, Stewart RD, Pallen MJ, Watson KA, Watson M. Assembly of hundreds of novel bacterial genomes from the chicken caecum. Genome Biol. 2020;21:34.
16. Lagier JC, Drancourt M, Charrel R, Bittar F, La Scola B, Ranque S, et al. Many More Microbes in Humans: Enlarging the Microbiome Repertoire. Clin Infect Dis. 2017;65 suppl_1:S20–9.
17. Perez-Sepulveda BM, Heavens D, Pulford C V, Predeus A V, Low R, Webster H, et al. An accessible, efficient and global approach for the large-scale sequencing of bacterial genomes. bioRxiv. 2020;:2020.07.22.200840.
18. Connor TR, Loman NJ, Thompson S, Smith A, Southgate J, Poplawski R, et al. CLIMB (the Cloud Infrastructure for Microbial Bioinformatics): an online resource for the medical microbiology community. Microb Genom. 2016;2:e000086.
19. Andrews S. FastQC: a quality control tool for high throughput sequence data.
20. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30:2114–20.
21. Wood DE, Lu J, Langmead B. Improved metagenomic analysis with Kraken 2. Genome Biol. 2019;20:257.

22. O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res. 2016;44:D733-45.
23. Lu J, Breitwieser FP, Thielen P, Salzberg SL. Bracken: estimating species abundance in metagenomics data. PeerJ Comput Sci. 2017;3:e104.
24. Ondov BD, Bergman NH, Phillippy AM. Interactive metagenomic visualization in a Web browser. BMC Bioinformatics. 2011;12:385.
25. NCBI BioProjects. https://www.ncbi.nlm.nih.gov/bioproject/. Accessed 01 October 2019
26. Foster-Nyarko E, Alikhan N-F, Ravi A, Thomson NM, Jarju S, Kwambana-Adams BA, et al. Genomic diversity of Escherichia coli isolates from backyard chickens and guinea fowl in the Gambia. bioRxiv. 2020;:2020.05.14.096289.
27. Puente-Sanchez F, Aguirre J, Parro V, Puente-s F, Aguirre J. A novel conceptual approach to read-filtering in high-throughput amplicon sequencing studies. Nucleic Acids Res. 2015;44:4.
28. Hildebrand F, Tadeo R, Voigt A, Bork P, Raes J. LotuS: an efficient and user-friendly OTU processing pipeline. Microbiome. 2014;2:30.
29. Wood DE, Salzberg SL. Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome Biol. 2014;15:R46.
30. Li D, Luo R, Liu CM, Leung CM, Ting HF, Sadakane K, et al. MEGAHIT v1.0: A fast and scalable metagenome assembler driven by advanced methodologies and community practices. Methods. 2016;102:3–11.
31. Roux S, Enault F, Hurwitz BL, Sullivan MB. VirSorter: mining viral signal from microbial genomic data. PeerJ. 2015;3:e985.
32. Fu L, Niu B, Zhu Z, Wu S, Li W. CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics. 2012;28:3150–2.
33. Menzel P, Ng KL, Krogh A. Fast and sensitive taxonomic classification for metagenomics with Kaiju. Nat Commun. 2016;7:11257.
34. Sullivan MJ, Petty NK, Beatson SA. Easyfig: a genome comparison visualizer. Bioinformatics. 2011;27:1009–10.
35. Eren AM, Sogin ML, Morrison HG, Vineis JH, Fisher JC, Newton RJ, et al. A single genus in the gut microbiome reflects host preference and specificity. ISME J. 2015;9:90–100.
36. Pheatmap. https://www.rdocumentation.org/packages/pheatmap. Accessed 02 May 2020.
37. Demovir. https://github.com/feargalr/Demovir. Accessed 06 April 2020.
38. Hyatt D, LoCascio PF, Hauser LJ, Uberbacher EC. Gene and translation initiation site prediction in metagenomic sequences. Bioinformatics. 2012;28:2223–30.
39. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012;9:357–9.
40. Mende DR, Sunagawa S, Zeller G, Bork P. Accurate and universal delineation of prokaryotic species. Nat Methods. 2013;10:881–4.

41. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25:2078–9.
42. Quinlan AR. BEDTools: The Swiss-Army Tool for Genome Feature Analysis. Curr Protoc Bioinforma. 2014;47:11.12.1-34.
43. rdCover. https://github.com/hildebra/rdCover. Accessed 10 January 2020.
44. Kang DD, Li F, Kirton E, Thomas A, Egan R, An H, et al. MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. PeerJ. 2019;7:e7359.
45. Hildebrand F, Moitinho-Silva L, Blasche S, Jahn MTT, Gossmann TI, Heuerta Cepas J, et al. Antibiotics-induced monodominance of a novel gut bacterial order. Gut. 2019.
46. Nielsen HB, Almeida M, Juncker AS, Rasmussen S, Li J, Sunagawa S, et al. Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes. Nat Biotechnol. 2014;32:822–8.
47. Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Res. 2015;25:1043–55.
48. METAFILER. https://github.com/aaronwolen/metafiler. Accessed 10 January 2020.
49. Olm MR, Brown CT, Brooks B, Banfield JF. dRep: a tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication. ISME J. 2017;11:2864–8.
50. Chaumeil PA, Mussig AJ, Hugenholtz P, Parks DH. GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. Bioinformatics. 2019.
51. Mende DR, Letunic I, Huerta-Cepas J, Li SS, Forslund K, Sunagawa S, et al. proGenomes: a resource for consistent functional and taxonomic annotations of prokaryotic genomes. Nucleic Acids Res. 2017;45:D529–34.
52. R-Core-Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2018.
53. Wickham H. ggplot2: Elegant Graphics for Data Analysis. New York: Springer-Verlag; 2016.
54. Page AJ, Cummins CA, Hunt M, Wong VK, Reuter S, Holden MT, et al. Roary: rapid large-scale prokaryote pan genome analysis. Bioinformatics. 2015;31:3691–3.
55. Hulsen T, de Vlieg J, Alkema W. BioVenn - a web application for the comparison and visualization of biological lists using area-proportional Venn diagrams. BMC Genomics. 2008;9:488.
56. Nurk S, Bankevich A, Antipov D, Gurevich A, Korobeynikov A, Lapidus A, et al. Assembling Genomes and Mini-metagenomes from Highly Chimeric Reads. In: Deng M, Jiang R, Sun F, Zhang X, editors. Research in Computational Molecular Biology. Berlin, Heidelberg: Springer Berlin Heidelberg; 2013. p. 158–70.
57. Barrnap. https://github.com/tseemann/barrnap. Accessed 05 May 2020
58. NCBI 16S RefSeq records processing and curation.

59. Chun J, Oren A, Ventosa A, Christensen H, Arahal DR, da Costa MS, et al. Proposed minimal standards for the use of genome data for the taxonomy of prokaryotes. Int J Syst Evol Microbiol. 2018;68:461–6.
60. Schwengers O, Hain T, Chakraborty T, Goesmann A. ReferenceSeeker: rapid determination of appropriate reference genomes. bioRxiv. 2019;:863621.
61. Eren AM, Esen ÖC, Quince C, Vineis JH, Morrison HG, Sogin ML, et al. Anvi'o: an advanced analysis and visualization platform for 'omics data. PeerJ. 2015;3:e1319.
62. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 2004;32:1792–7.
63. Price MN, Dehal PS, Arkin AP. FastTree 2–approximately maximum-likelihood trees for large alignments. PLoS One. 2010;5:e9490.
64. Letunic I, Bork P. Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. Nucleic Acids Res. 2016;44:W242-5.
65. Angiuoli S V, Salzberg SL. Mugsy: fast multiple alignment of closely related whole genomes. Bioinformatics. 2010;27:334–42.
66. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. 2014;30:1312–3.
67. Andersen SC, Kiil K, Harder CB, Josefsen MH, Persson S, Nielsen EM, et al. Towards diagnostic metagenomics of Campylobacter in fecal samples. BMC Microbiol. 2017;17:133.
68. Ravi A, Halstead FD, Bamford A, Casey A, Thomson NM, van Schaik W, et al. Loss of microbial diversity and pathogen domination of the gut microbiota in critically ill patients. Microb Genom. 2019;5.
69. Narasinakuppe Krishnegowda D, Dhama K, Kumar Mariappan A, Munuswamy P, Iqbal Yatoo M, Tiwari R, et al. Etiology, epidemiology, pathology, and advances in diagnosis, vaccine development, and treatment of infection in poultry: a review. Vet Q. 2020;40:16–34.
70. Almohaya AM, Almutairy TS, Alqahtani A, Binkhamis K, Almajid FM. Fusobacterium bloodstream infections: A literature review and hospital-based case series. Anaerobe. 2020;62:102165.
71. Kollarcikova M, Kubasova T, Karasova D, Crhanova M, Cejkova D, Sisak F, et al. Use of 16S rRNA gene sequencing for prediction of new opportunistic pathogens in chicken ileal and cecal microbiota. Poult Sci. 2019;98:2347–53.
72. Jamalludeen N, Kropinski AM, Johnson RP, Lingohr E, Harel J, Gyles CL. Complete genomic sequence of bacteriophage phiEcoM-GJ1, a novel phage that has myovirus morphology and a podovirus-like RNA polymerase. Appl Env Microbiol. 2008;74:516–25.
73. Tsonos J, Adriaenssens EM, Klumpp J, Hernalsteens JP, Lavigne R, De Greve H. Complete genome sequence of the novel Escherichia coli phage phAPEC8. J Virol. 2012;86:13117–8.
74. Bowers RM, Kyrpides NC, Stepanauskas R, Harmon-Smith M, Doud D, Reddy TBK, et al. Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome

(MIMAG) of bacteria and archaea. Nat Biotechnol. 2017;35:725–31.

75. Jain C, Rodriguez-R LM, Phillippy AM, Konstantinidis KT, Aluru S. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. Nat Commun. 2018;9:5114.
76. Luo C, Rodriguez-R LM, Konstantinidis KT. MyTaxa: an advanced taxonomic classifier for genomic and metagenomic sequences. Nucleic Acids Res. 2014;42:e73.
77. Yoon J-H, Lee K-C, Weiss N, Kang KH, Park Y-H. Jeotgalicoccus halotolerans gen. nov., sp. nov. and Jeotgalicoccus psychrophilus sp. nov., isolated from the traditional Korean fermented seafood jeotgal. Int J Syst Evol Microbiol. 2003;53 Pt 2:595–602.
78. Amoozegar MA, Bagheri M, Makhdoumi-Kakhki A, Didari M, Schumann P, Nikou MM, et al. Aliicoccus persicus gen. nov., sp. nov., a halophilic member of the Firmicutes isolated from a hypersaline lake. Int J Syst Evol Microbiol. 2014;64 Pt 6:1964–9.
79. Sakamoto M, Lapidus AL, Han J, Trong S, Haynes M, Reddy TB, et al. High quality draft genome sequence of Bacteroides barnesiae type strain BL2(T) (DSM 18169(T)) from chicken caecum. Stand Genomic Sci. 2015;10:48.
80. Saengkerdsub S, Anderson RC, Wilkinson HH, Kim WK, Nisbet DJ, Ricke SC. Identification and quantification of methanogenic Archaea in adult chicken ceca. Appl Env Microbiol. 2007;73:353–6.
81. Linnaeus C. Systema Naturae (10th ed.). Stockholm: Laurentius Salvius; 1759.
82. Parker CT, Tindall BJ, Garrity GM. International Code of Nomenclature of Prokaryotes. Int J Syst Evol Microbiol. 2019;69:S1–111.
83. Oren A. A plea for linguistic accuracy - also for Candidatus taxa. Int J Syst Evol Microbiol. 2017;67:1085–94.
84. Oren A, Garrity GM, Parker CT, Chuvochina M, Trujillo ME. Lists of names of prokaryotic Candidatus taxa. Int J Syst Evol Microbiol. 2020.
85. Liu S, Jin D, Lan R, Wang Y, Meng Q, Dai H, et al. Escherichia marmotae sp. nov., isolated from faeces of Marmota himalayana. Int J Syst Evol Microbiol. 2015;65:2130–4.
86. Liu S, Feng J, Pu J, Xu X, Lu S, Yang J, et al. Genomic and molecular characterisation of Escherichia marmotae from wild rodents in Qinghai-Tibet plateau as a potential pathogen. Sci Rep. 2019;9:10619.
87. Walk ST. The “Cryptic” Escherichia. EcoSal Plus. 2015;6.
88. Parks DH, Chuvochina M, Chaumeil PA, Rinke C, Mussig AJ, Hugenholtz P. A complete domain-to-species taxonomy for Bacteria and Archaea. Nat Biotechnol. 2020.
89. Walk ST, Alm EW, Gordon DM, Ram JL, Toranzos GA, Tiedje JM, et al. Cryptic lineages of the genus Escherichia. Appl Environ Microbiol. 2009;75:6534–44.
90. Clermont O, Gordon DM, Brisse S, Walk ST, Denamur E. Characterization of the cryptic Escherichia lineages: rapid identification and prevalence. Environ Microbiol. 2011;13:2468–77.
91. Walk S, Feng P. No Title. In: Population Genetics of Bacteria. ASM Press, Washington, DC; 2011. p. 1–4.

92. Frioux C, Singh D, Korcsmaros T, Hildebrand F. From bag-of-genes to bag-of-genomes: metabolic modelling of communities in the era of metagenome-assembled genomes. Comput Struct Biotechnol J. 2020;18:1722–34.
93. Federhen S. The NCBI Taxonomy database. Nucleic Acids Res. 2012;40 Database issue:D136-43.
94. Hildebrand F, Pallen MJ, Bork P. Towards standardisation of naming novel prokaryotic taxa in the age of high-throughput microbiology. Gut. 2020;69:1358 LP – 1359.
95. Fodor AA, DeSantis TZ, Wylie KM, Badger JH, Ye Y, Hepburn T, et al. The “most wanted” taxa from the human microbiome for whole genome sequencing. PLoS One. 2012;7:e41294.
96. Stewart RD, Auffret MD, Warr A, Wiser AH, Press MO, Langford KW, et al. Assembly of 913 microbial genomes from metagenomic sequencing of the cow rumen. Nat Commun. 2018;9:870.
97. Patel S, Gupta RS. A phylogenomic and comparative genomic framework for resolving the polyphyly of the genus Bacillus: Proposal for six new genera of Bacillus species, Peribacillus gen. nov., Cytobacillus gen. nov., Mesobacillus gen. nov., Neobacillus gen. nov., Metabacillus gen. nov. and Alkalihalobacillus gen. nov. Int J Syst Evol Microbiol. 2020;70:406–38.
98. Zheng J, Wittouck S, Salvetti E, Franz C, Harris HMB, Mattarelli P, et al. A taxonomic note on the genus Lactobacillus: Description of 23 novel genera, emended description of the genus Lactobacillus Beijerinck 1901, and union of Lactobacillaceae and Leuconostocaceae. Int J Syst Evol Microbiol. 2020;70:2782–858.
99. Hespell RB. Serpens flexibilis gen. nov., sp. nov., an Unusually Flexible, Lactate-Oxidizing Bacterium. Int J Syst Bacteriol. 1977;27:371–81.

Tables

Table 1. Protologues for new taxa cultured from chicken faeces

DESCRIPTION OF *ACINETOBACTER PECORUM* SP. NOV.

(pe.co’rum M.L. gen. pl. *pecorum* of flocks of sheep, birds etc., as this species has also been isolated from sheep)

A bacterial species cultured from chicken faeces and assigned to this genus and delineated as a species by analysis of its genome sequence. GTDB [88] has given this species the alphanumerical designation sp001647535. The Type Strain is Sa1BUA6, which has been submitted for deposition in NCTC and DSMZ. The GC content of the Type Strain is 42.9% and the genome size is 3,209,341 base pairs. Further information can be found in the Methods and in Additional File 5.

DESCRIPTION OF *ARTHROBACTER GALLICOLA* SP. NOV.

(gal.li'co.la. L. masc. n. *gallus* a cock; N.L. suff. –*cola* an inhabitant of; N.L. masc. or fem. n. *gallicola* an inhabitant of the chicken)

A bacterial species cultured from chicken faeces and assigned to this genus and delineated as a species by analysis of its genome sequence. GTDB [88] has applied the designation *Arthrobacter_B*, to this genus. The Type Strain is Sa2CUA1, which has been submitted for deposition in NCTC and DSMZ. The GC content of the Type Strain is 65.5% and the genome size is 3,679,471 base pairs. Further information can be found in the Methods and in Additional File 5.

DESCRIPTION OF *ARTHROBACTER PULLICOLA* SP. NOV.

(pul.li'co.la. L. masc. n. *pullus* a young chicken; N.L. suff. –*cola* an inhabitant of; N.L. masc. or fem. n. *pullicola* an inhabitant of young chickens)

A bacterial species cultured from chicken faeces and assigned to this genus and delineated as a species by analysis of its genome sequence. GTDB [88] has applied the designation *Arthrobacter_B* to this genus. The Type Strain is Sa2BUA2, which has been submitted for deposition in NCTC and DSMZ. The GC content of the Type Strain is 65.7% and the genome size is 3,726,732 base pairs. Further information can be found in the Methods and in Additional File 5.

DESCRIPTION OF *BACILLUS NORWICHENSIS* SP. NOV.

(nor.wich.en'sis. N.L. masc. adj. *norwichensis* pertaining to English city of Norwich, where the organism was isolated)

A bacterial species cultured from chicken faeces and assigned to this genus and delineated as a species by analysis of its genome sequence. GTDB [88] has applied the designation *Bacillus_AM* to this genus. The Type Strain is Sa3CUA8, which has been submitted for deposition in NCTC and DSMZ. The GC content of the Type Strain is 40.2% and the genome size is 4,696,597 base pairs. Further information can be found in the Methods and in Additional File 5.

DESCRIPTION OF *BREVIBACTERIUM GALLINARUM* SP. NOV.

(gal.li.na'rum. L. pl. gen. n. *gallinarum* of hens)

A bacterial species cultured from chicken faeces and assigned to this genus and delineated as a species by analysis of its genome sequence. The Type Strain is Re57, which has been submitted for deposition in NCTC and DSMZ. The GC content of the Type Strain is 67.0% and the genome size is 3,231,168 base pairs. Further information can be found in the Methods and in Additional File 5.

DESCRIPTION OF *BREVUNDIMONAS GUILDFORDENSIS* SP. NOV.

(guild.ford.en'sis. N.L. fem. adj. *guildfordensis* pertaining to English town Guildford, home to the University of Surrey, where the samples were taken)

A bacterial species cultured from chicken faeces and assigned to this genus and delineated as a species by analysis of its genome sequence. The Type Strain is Sa3CVA3, which has been submitted for deposition in NCTC and DSMZ. The GC content of the Type Strain is 67.3% and the genome size is 2,875,662 base pairs. Further information can be found in the Methods and in Additional File 5.

DESCRIPTION OF *CELLULOMONAS AVISTERCORIS* SP. NOV.

(a.vi.ster'co.ris. L. fem. n. *avis* bird; L. neut. n. *stercus* dung; N.L. gen. n. *avistercoris* of bird faeces)

A bacterial species cultured from chicken faeces and assigned to this genus and delineated as a species by analysis of its genome sequence. The Type Strain is Sa3CUA2, which has been submitted for deposition in NCTC and DSMZ. 18 The GC content of the Type Strain is 74.5% and the genome size is 4,169,055 base pairs. Further information can be found in the Methods and in Additional File 5.

DESCRIPTION OF *CLOSTRIDIUM AVIUM* SP. NOV.

(a'vi.um. L. gen. pl. n. *avium* of birds)

A bacterial species cultured from chicken faeces and assigned to this genus and delineated as a species by analysis of its genome sequence. GTDB [88] has given this species the alphanumerical designation sp007115085. The Type Strain is Sa3CVN1, which has been submitted for deposition in NCTC and DSMZ. The GC content of the Type Strain is 29.8% and the genome size is 4,256,035 base pairs. Further information can be found in the Methods and in Additional File 5.

DESCRIPTION OF *CLOSTRIDIUM GALLINARUM* SP. NOV.

(gal.li.na'rum. L. pl. gen. n. *gallinarum* of hens)

A bacterial species cultured from chicken faeces and assigned to this genus and delineated as a species by analysis of its genome sequence. The Type Strain is Sa3CUN1, which has been submitted for deposition in NCTC and DSMZ. The GC content of the Type Strain is 27.2% and the genome size is 3,426,635 base pairs. Further information can be found in the Methods and in Additional File 5.

DESCRIPTION OF *CLOSTRIDIUM MERDIGALLINARUM* SP. NOV.

(mer.di.gal.li.na'rum. L. fem. n. *merda* faeces; L. fem. n. *gallina* a hen; N.L. gen. n. *merdigallinarum* of chicken faeces)

A bacterial species cultured from chicken faeces and assigned to this genus and delineated as a species by analysis of its genome sequence. GTDB [88] has applied the designation *Clostridium_J to this genus*. GTDB [88] has given this species the alphanumerical designation sp900547625. The Type Strain is N37, which has been submitted for deposition in NCTC and DSMZ. The GC content of the Type Strain is 28.7% and the genome size is 3,853,386 base pairs. Further information can be found in the Methods and in Additional File 5.

DESCRIPTION OF *COMAMONAS AVIUM* SP. NOV.

(a'vi.um. L. gen. pl. n. *avium* of birds)

A bacterial species cultured from chicken faeces and assigned to this genus and delineated as a species by analysis of its genome sequence. The Type Strain is Sa2BVA9, which has been submitted for deposition in NCTC and DSMZ. 58 The GC content of the Type Strain is 57.5% and the genome size is 3,926,881 base pairs. Further information can be found in the Methods and in Additional File 5.

DESCRIPTION OF *CORYNEBACTERIUM GALLINARUM* SP. NOV.

(gal.li.na'rum. L. pl. gen. n. *gallinarum* of hens)

A bacterial species cultured from chicken faeces and assigned to this genus and delineated as a species by analysis of its genome sequence. The Type Strain is Sa1YVA5, which has been submitted for deposition in NCTC and DSMZ. The GC content of the Type Strain is 63.1% and the genome size is 3,086,957 base pairs. Further information can be found in the Methods and in Additional File 5.

DESCRIPTION OF *CYTOBACILLUS STERCORIGALLINARUM* SP. NOV.

(ster.co.ri.gal.li.na'rum. L. neut. n. *stercus* faeces; L. fem. n. *gallina* a hen; N.L. gen. n. *stercorigallinarum* of chicken faeces)

A bacterial species cultured from chicken faeces and assigned to this genus and delineated as a species by analysis of its genome sequence. GTDB [88] has applied the designation *Bacillus_AA,* which cannot be incorporated into a well-formed binomial. However, according to Patel and Gupta [97]. *Cytobacillus* encompasses other species classified by GTDB [88] within this provisional genus designation and therefore is a synonym of GTDB *Bacillus_AA*. The Type Strain is Sa1BUA13, which has been submitted for deposition in NCTC and DSMZ. The GC content of the Type Strain is 36.8% and the genome size is 4,443,518 base pairs. Further information can be found in the Methods and in Additional File 5.

DESCRIPTION OF *ESCHERICHIA WHITTAMII* SP. NOV.

(whitt.am'i.i. N.L. gen. n. *whittamii* named in honour of American microbiologist Thomas S. Whittam)

A bacterial species cultured from chicken faeces and assigned to this genus and delineated as a species by analysis of its genome sequence. GTDB [88] has given this species the alphanumerical designation sp001660175. The Type Strain is Sa3CUN2, which has been submitted for deposition in NCTC and DSMZ. The GC content of the Type Strain is 50.6% and the genome size is 4,551,298 base pairs. Further information can be found in the Methods and in Additional File 5.

DESCRIPTION OF *FICTIBACILLUS NORFOLKENSIS* SP. NOV.

(nor.folk.en'sis. N.L. masc. adj. *norfolkensis* pertaining to the English county of Norfolk)

A bacterial species cultured from chicken faeces and assigned to this genus and delineated as a species by analysis of its genome sequence. GTDB [88] has applied the designation *Fictibacillus_B* to this genus. *Fictibacillus_B*. The Type Strain is Sa5YUA1, which has been submitted for deposition in NCTC and DSMZ. The GC content of the Type Strain is 39.5% and the genome size is 4,031,565 base pairs. Further information can be found in the Methods and in Additional File 5.

DESCRIPTION OF *KAISTELLA PULLORUM* SP. NOV.

(pul.lor'um. L. gen. pl. n. *pullorum* of chickens)

A bacterial species cultured from chicken faeces and assigned to this genus and delineated as a species by analysis of its genome sequence. The Type Strain is Sa1CVA4, which has been submitted for deposition in NCTC and DSMZ. The GC content of the Type Strain is 42.9% and the genome size is 2,562,418 base pairs. Further information can be found in the Methods and in Additional File 5.

DESCRIPTION OF *LIMOSILACTOBACILLUS AVISTERCORIS* SP. NOV.

(a.vi.ster'co.ris. L. fem. n. *avis* bird; L. neut. n. *stercus* dung; N.L. gen. n. *avistercoris* of bird faeces)

A bacterial species cultured from chicken faeces and assigned to this genus and delineated as a species by analysis of its genome sequence. GTDB [88] has applied the designation *Lactobacillus_H,* which cannot be incorporated into a well-formed binomial. However, according to Zheng et al [98] *Limosilactobacillus* encompasses other species classified by GTDB within this provisional genus

designation and therefore is a synonym for GDTB. The Type Strain is Sa5BUN4, which has been submitted for deposition in NCTC and DSMZ. The GC content of the Type Strain is 39.9% and the genome size is 1,779,587 base pairs. Further information can be found in the Methods and in Additional File 5.

DESCRIPTION OF *LUTEIMONAS COLNEYENSIS* SP. NOV.

(col.ney.en'sis. N.L. fem. adj. *colneyensis* pertaining to the English village of Colney, home to the Quadram Institute)

A bacterial species cultured from chicken faeces and assigned to this genus and delineated as a species by analysis of its genome sequence. The Type Strain is Sa2CVA6, which has been submitted for deposition in NCTC and DSMZ. The GC content of the Type Strain is 71.0% and the genome size is 3,019,767 base pairs. Further information can be found in the Methods and in Additional File 5.

DESCRIPTION OF *MICROBACTERIUM COMMUNE* SP. NOV.

(com.mu’ne. L. neut. adj. *commune* common, referring to diverse habitats as this species has been isolated from mosquitos and chicken)

A bacterial species cultured from chicken faeces and assigned to this genus and delineated as a species by analysis of its genome sequence. The Type Strain is Sa1CUA4, which has been submitted for deposition in NCTC and DSMZ. The GC content of the Type Strain is 70.3% and the genome size is 3,345,699 base pairs. Further information can be found in the Methods and in Additional File 5.

DESCRIPTION OF *MICROBACTERIUM GALLINARUM* SP. NOV.

(gal.li.na'rum. L. pl. gen. n. *gallinarum* of hens)

A bacterial species cultured from chicken faeces and assigned to this genus and delineated as a species by analysis of its genome sequence. GTDB [88] has given this species the alphanumerical designation sp001878835. The Type Strain is Re1, which has been submitted for deposition in NCTC and DSMZ. The GC content of the Type Strain is 69.4% and the genome size is 2,791,888 base pairs. Further information can be found in the Methods and in Additional File 5.

DESCRIPTION OF *MICROBACTERIUM PULLORUM* SP. NOV.

(pul.lor'um. L. gen. pl. n. *pullorum* of chickens)

A bacterial species cultured from chicken faeces and assigned to this genus and delineated as a species by analysis of its genome sequence. The Type Strain is Sa4CUA7, which has been submitted for deposition in NCTC and DSMZ. The GC content of the Type Strain is 70.1% and the genome size is 3,113,556 base pairs. Further information can be found in the Methods and in Additional File 5.

DESCRIPTION OF *OCEANITALEA STEVENSII* SP. NOV.

(ste.ven'si.i. N.L. gen. n. *stevensii* named in honour of British microbiologist Mark Stevens)

A bacterial species cultured from chicken faeces and assigned to this genus and delineated as a species by analysis of its genome sequence. The Type Strain is Sa1BUA1, which has been submitted for deposition in NCTC and DSMZ. The GC content of the Type Strain is 73.4% and the genome size is 3,531,784 base pairs. Further information can be found in the Methods and in Additional File 5.

DESCRIPTION OF *OCHROBACTRUM GALLINARUM* SP. NOV.

(gal.li.na'rum. L. pl. gen. n. *gallinarum* of hens)

A bacterial species cultured from chicken faeces and assigned to this genus and delineated as a species by analysis of its genome sequence. GTDB [88] has given this species the alphanumerical designation sp002278035. The Type Strain is Sa2BUA5, which has been submitted for deposition in NCTC and DSMZ. The GC content of the Type Strain is 53.5% and the genome size is 4,986,285 base pairs. Further information can be found in the Methods and in Additional File 5.

DESCRIPTION OF *OERSKOVIA DOUGANII* SP. NOV.

(dou.gan'i.i. N.L. gen. n. *douganii* named in honour of British microbiologist Gordon Dougan)

A bacterial species cultured from chicken faeces and assigned to this genus and delineated as a species by analysis of its genome sequence. The Type Strain is Sa1BUA8, which has been submitted for deposition in NCTC and DSMZ. The GC content of the Type Strain is 72.5% and the genome size is 4,251,564 base pairs. Further information can be found in the Methods and in Additional File 5.

DESCRIPTION OF *OERSKOVIA GALLYI* SP. NOV.

(gall.y'i. N.L. gen. n. *gallyi* named in honour of British microbiologist David Gally)

A bacterial species cultured from chicken faeces and assigned to this genus and delineated as a species by analysis of its genome sequence. The Type Strain is Sa2CUA8, which has been submitted for deposition in NCTC and DSMZ. The GC content of the Type Strain is 72.5% and the genome size is 4,251,268 base pairs. Further information can be found in the Methods and in Additional File 5.

DESCRIPTION OF *OERSKOVIA MERDAVIUM* SP. NOV.

(merd.a'vi.um. L. fem. n. *merda* faeces; L. fem. n. *avis* bird; N.L. gen. n. *merdavium* of bird faeces)

A bacterial species cultured from chicken faeces and assigned to this genus and delineated as a species by analysis of its genome sequence. The Type Strain is Sa2CUA9, which has been submitted for deposition in NCTC and DSMZ. The GC content of the Type Strain is 72.1% and the genome size is 4,486,858 base pairs. Further information can be found in the Methods and in Additional File 5.

DESCRIPTION OF *OERSKOVIA RUSTICA* SP. NOV.

(rus.tic.a L fem adj. *rustica* of the countryside, as isolates obtained from soil and chickens)

A bacterial species cultured from chicken faeces and assigned to this genus and delineated as a species by analysis of its genome sequence. GTDB [88] has given this species the alphanumerical designation sp005937995. The Type Strain is Sa4CUA1, which has been submitted for deposition in NCTC and

DSMZ. The GC content of the Type Strain is 72.5% and the genome size is 4,351,592 base pairs. Further information can be found in the Methods and in Additional File 5.

DESCRIPTION OF *PAENIBACILLUS GALLINARUM* SP. NOV.

(gal.li.na'rum. L. pl. gen. n. *gallinarum* of hens)

A bacterial species cultured from chicken faeces and assigned to this genus and delineated as a species by analysis of its genome sequence. The Type Strain is Sa2CUA10, which has been submitted for deposition in NCTC and DSMZ. The GC content of the Type Strain is 41.2% and the genome size is 5,411,474 base pairs. Further information can be found in the Methods and in Additional File 5.

DESCRIPTION OF *PHOCAEICOLA CAECIGALLINARUM* SP. NOV.

(cae.ci.gal.li.na'rum. L. neut. n. caecum the caecum L. fem. n. gallina a hen N.L. gen. n. caecigallinarum of the caecum of hens)

A bacterial species cultured from chicken faeces and assigned to this genus and delineated as a species by analysis of its genome sequence. GTDB [88] has given this species the alphanumerical designation sp002161565 The Type Strain is Sa1CVN1, which has been submitted for deposition in NCTC and DSMZ. The GC content of the Type Strain is 45.7% and the genome size is 4,070,481 base pairs. Further information can be found in the Methods and in Additional File 5.

DESCRIPTION OF *PHOCAEICOLA GALLINARUM* SP. NOV.

(gal.li.na'rum. L. pl. gen. n. *gallinarum* of hens)

A bacterial species cultured from chicken faeces and assigned to this genus and delineated as a species by analysis of its genome sequence. GTDB [88] has given this species the alphanumerical designation sp900540105. The Type Strain is Sa1YUN3, which has been submitted for deposition in NCTC and

DSMZ. The GC content of the Type Strain is 45.6% and the genome size is 3,486,467 base pairs. Further information can be found in the Methods and in Additional File 5.

DESCRIPTION OF *PLANOCOCCUS WIGLEYI* SP. NOV.

(wig'ley.i. N.L. masc. gen. n. *wigleyi* named in honour of British microbiologist Paul Wigley)

A bacterial species cultured from chicken faeces and assigned to this genus and delineated as a species by analysis of its genome sequence. GTDB [88] has assigned a genus name with an alphabetic suffix *Planococcus_A* to this genus. The Type Strain is Sa1BUA2, which has been submitted for deposition in NCTC and DSMZ. The GC content of the Type Strain is 45.0% and the genome size is 3,752,086 base pairs. Further information can be found in the Methods and in Additional File 5.

DESCRIPTION OF *PSYCHROBACILLUS FAECIGALLINARUM* SP. NOV.

(fae.ci.gal.li.na'rum. L. fem. n. *faex, faecis* faeces; L. fem. n. *gallina* a hen; N.L. gen. n. *faecigallinarum* of chicken faeces)

A bacterial species cultured from chicken faeces and assigned to this genus and delineated as a species by analysis of its genome sequence. GTDB [88] has given this species the alphanumerical designation sp007115085. The type strain is Sa2BVA5, which has been submitted for deposition in NCTC and DSMZ. The GC content of the Type Strain is 36.5% and the genome size is 4,007,948 base pairs. Further information can be found in the Methods and in Additional File 5.

DESCRIPTION OF *PSYCHROBACTER COMMUNIS* SP. NOV.

(com.mun'is L. masc. adj. *communis* common, referring to diverse habitats from which this species has been isolated, including chickens and soil)

A bacterial species cultured from chicken faeces and assigned to this genus and delineated as a species by analysis of its genome sequence. GTDB [88] has given this species the alphanumerical designation

sp001652315. Further information can be found in the Methods and in Additional File 5. The type strain is Sa4CVA2, which has been submitted for deposition in NCTC and DSMZ. The GC content of the Type Strain is 43.7% and the genome size is 2,956,826 base pairs. Further information can be found in the Methods and in Additional File 5.

DESCRIPTION OF *SERPENS GALLINARUM* SP. NOV.

(gal.li.na'rum. L. pl. gen. n. *gallinarum* of hens)

A bacterial species cultured from chicken faeces and assigned to this genus and delineated as a species by analysis of its genome sequence. GTDB {Parks et al., 2020, Nat Biotechnol} has assigned a genus name with an alphabetic suffix *Planococcus_A*, which cannot be incorporated into a well-formed binomial. However, GDTB genus *Pseudomonas_H* includes *Pseudomonas flexibilis*, where the basonym is *Serpens* [99], so we have used this genus name. GTDB [88] has given this species the alphanumerical designation sp001660175. The Type Strain is Sa2CUA2, which has been submitted for deposition in NCTC and DSMZ. The GC content of the Type Strain is 61.0% and the genome size is 3,905,357 base pairs. Further information can be found in the Methods and in Additional File 5.

DESCRIPTION OF *SOLIBACILLUS FAECAVIUM* SP. NOV.

(faec.a'vi.um. L. fem. n. *faex, faecis* faeces; L. fem. n. avis bird; N.L. gen. n. faecavium of bird faeces)

A bacterial species cultured from chicken faeces and assigned to this genus and delineated as a species by analysis of its genome sequence. The Type Strain is A46, which has been submitted for deposition in NCTC and DSMZ. The GC content of the Type Strain is 37.1% and the genome size is 3,824,479 base pairs. Further information can be found in the Methods and in Additional File 5.

DESCRIPTION OF *SOLIBACILLUS MERDAVIUM* SP. NOV.

(merd.a'vi.um. L. fem. n. *merda* faeces; L. fem. n. *avis* bird; N.L. gen. n. *merdavium* of bird faeces)

A bacterial species cultured from chicken faeces and assigned to this genus and delineated as a species by analysis of its genome sequence. The Type Strain is Sa1YVA6, which has been submitted for deposition in NCTC and DSMZ. The GC content of the Type Strain is 37.0% and the genome size is 3,783,750 base pairs. Further information can be found in the Methods and in Additional File 5.

DESCRIPTION OF *SPOROSARCINA GALLISTERCORIS* SP. NOV.

(gal.li.ster'co.ris. L. masc. n. *gallus* a cock; L. neut. n. *stercus* dung; N.L. gen. n. *gallistercoris* of faeces of a cock)

A bacterial species cultured from chicken faeces and assigned to this genus and delineated as a species by analysis of its genome sequence. GTDB [88] has assigned a genus name with an alphabetic suffix *Sporosarcina_A* to this genus. The Type Strain is Sa2YVA2, which has been submitted for deposition in NCTC and DSMZ. The GC content of the Type Strain is 44.1% and the genome size is 3,135,381 base pairs. Further information can be found in the Methods and in Additional File 5.

DESCRIPTION OF *SPOROSARCINA QUADRAMI* SP. NOV.

(qua.dra'mi. N.L. gen. n. *quadrami* of the Quadram Institute, where the species was first cultured.)

A bacterial species cultured from chicken faeces and assigned to this genus and delineated as a species by analysis of its genome sequence. GTDB [88] has assigned a genus name with an alphabetic suffix *Sporosarcina_B* to this genus. The Type Strain is Sa2BUA9, which has been submitted for deposition in NCTC and DSMZ. The GC content of the Type Strain is 41.4% and the genome size is 3,576,489 base pairs. Further information can be found in the Methods and in Additional File 5.

DESCRIPTION OF *STENOTROPHOMONAS PENNII* SP. NOV.

(pen'ni.i. N.L. gen. n. *pennii* named in honour of British microbiologist Charles W. Penn)

A bacterial species cultured from chicken faeces and assigned to this genus and delineated as a species by analysis of its genome sequence. GTDB [88] has given this species the alphanumerical designation sp002836635. The Type Strain is Sa3BUA13, which has been submitted for deposition in NCTC and

DSMZ. The GC content of the Type Strain is 66.4% and the genome size is 3,928,648 base pairs. Further information can be found in the Methods and in Additional File 5.

DESCRIPTION OF *UREIBACILLUS GALLI* SP. NOV.

(gal'li. L. masc. gen. n. *galli* of a chicken)

A bacterial species cultured from chicken faeces and assigned to this genus and delineated as a species by analysis of its genome sequence. *Ureibacillus* is a synonym of GTDB [88] *Lysinibacillus_C* according to doi 10.3389/fmicb.2019.02821. The Type Strain is Re31, which has been submitted for deposition in NCTC and DSMZ. The GC content of the Type Strain is 35.2% and the genome size is 3,726,905 base pairs. Further information can be found in the Methods and in Additional File 5.

DESCRIPTION OF *XANTHOMONAS SURREYENSIS* SP. NOV.

(sur.rey.en'sis. N.L. fem. adj. *surreyensis* pertaining to the English county of Surrey where the samples were obtained)

A bacterial species cultured from chicken faeces and assigned to this genus and delineated as a species by analysis of its genome sequence. GTDB [88] has assigned a genus name with an alphabetic suffix *Xanthomonas_A* to this genus. The Type Strain is Sa2BVA3, which has been submitted for deposition in NCTC and DSMZ. The GC content of the Type Strain is 68.8% and the genome size is 5,377,401 base pairs. Further information can be found in the Methods and in Additional File 5.4,422

Figures

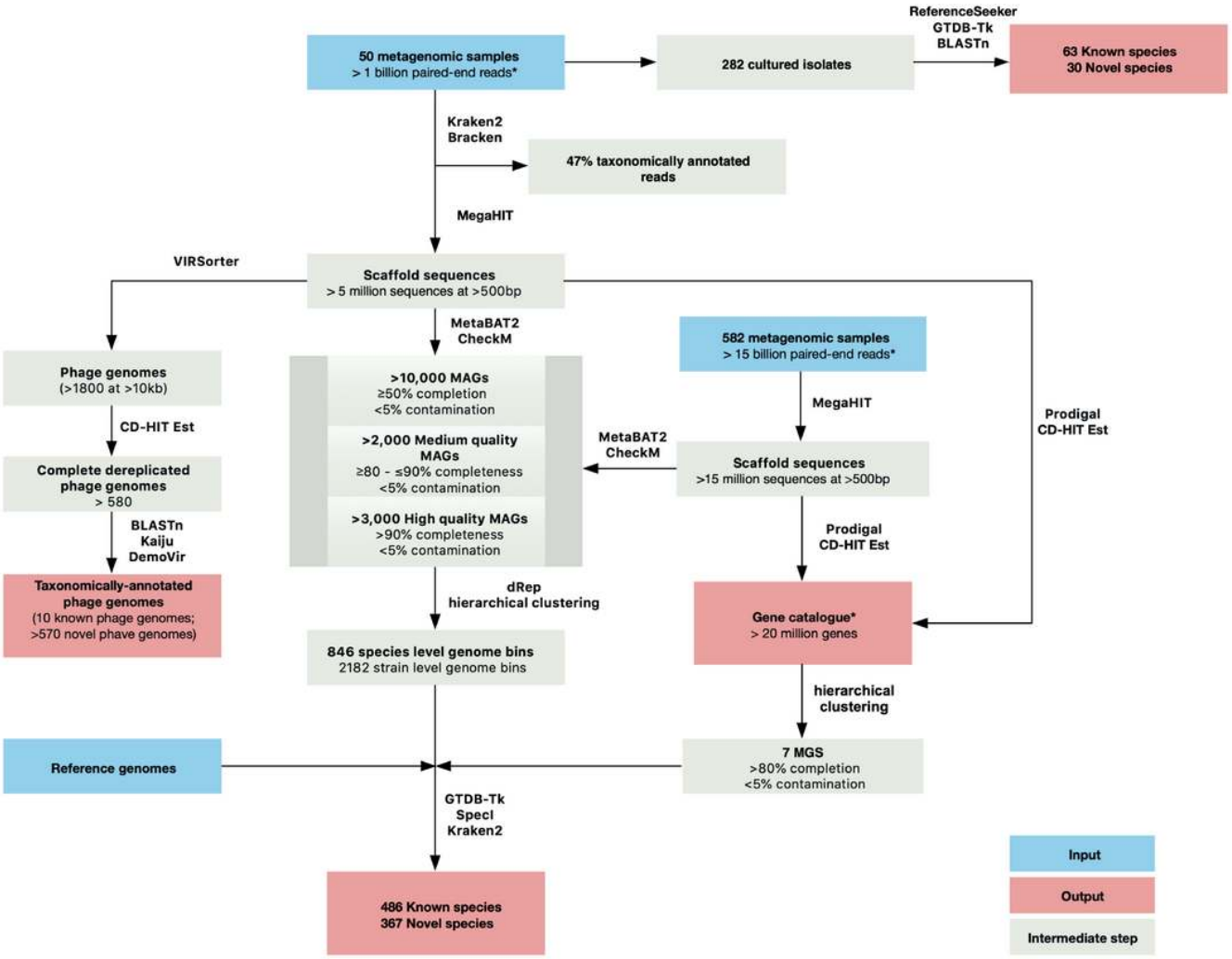

Figure 1

Analytical Workflow. * indicates read numbers are detailed post-filtering of diet and host associated reads.

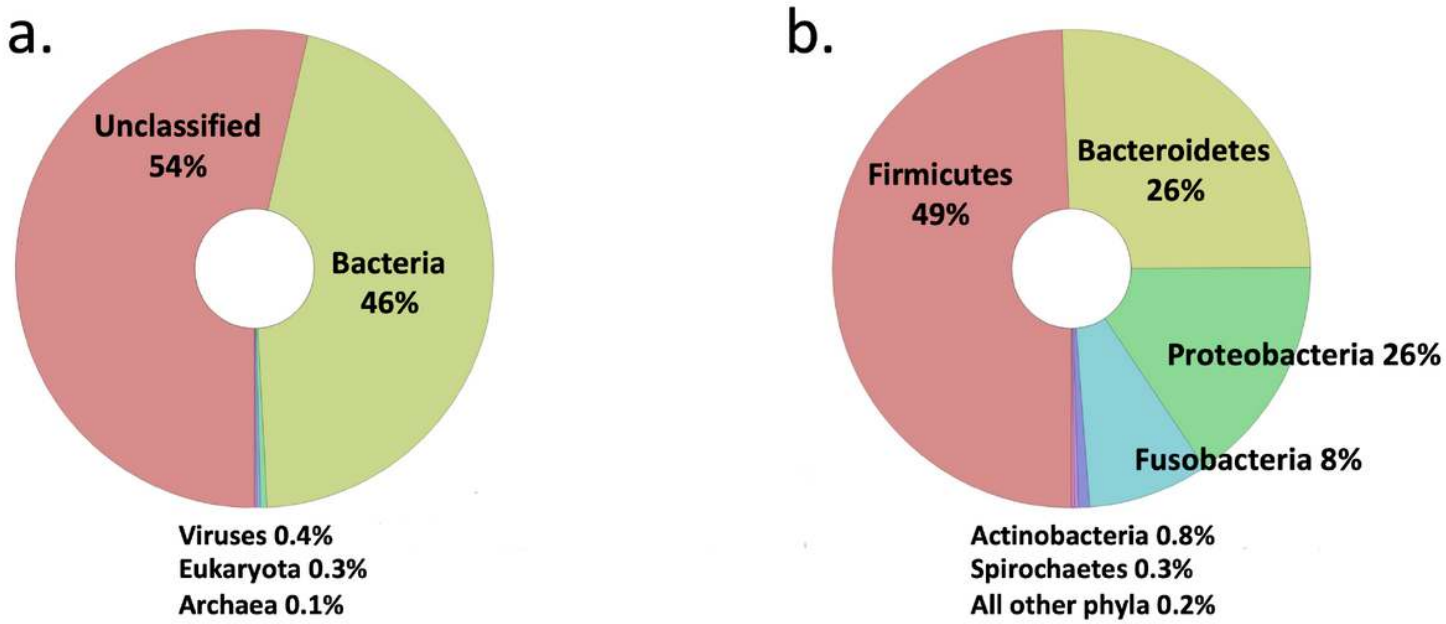

Figure 2

Krona plots of metagenomic reads from 50 chicken faecal samples. a all reads classified by domain. b bacterial reads classified by phylum.

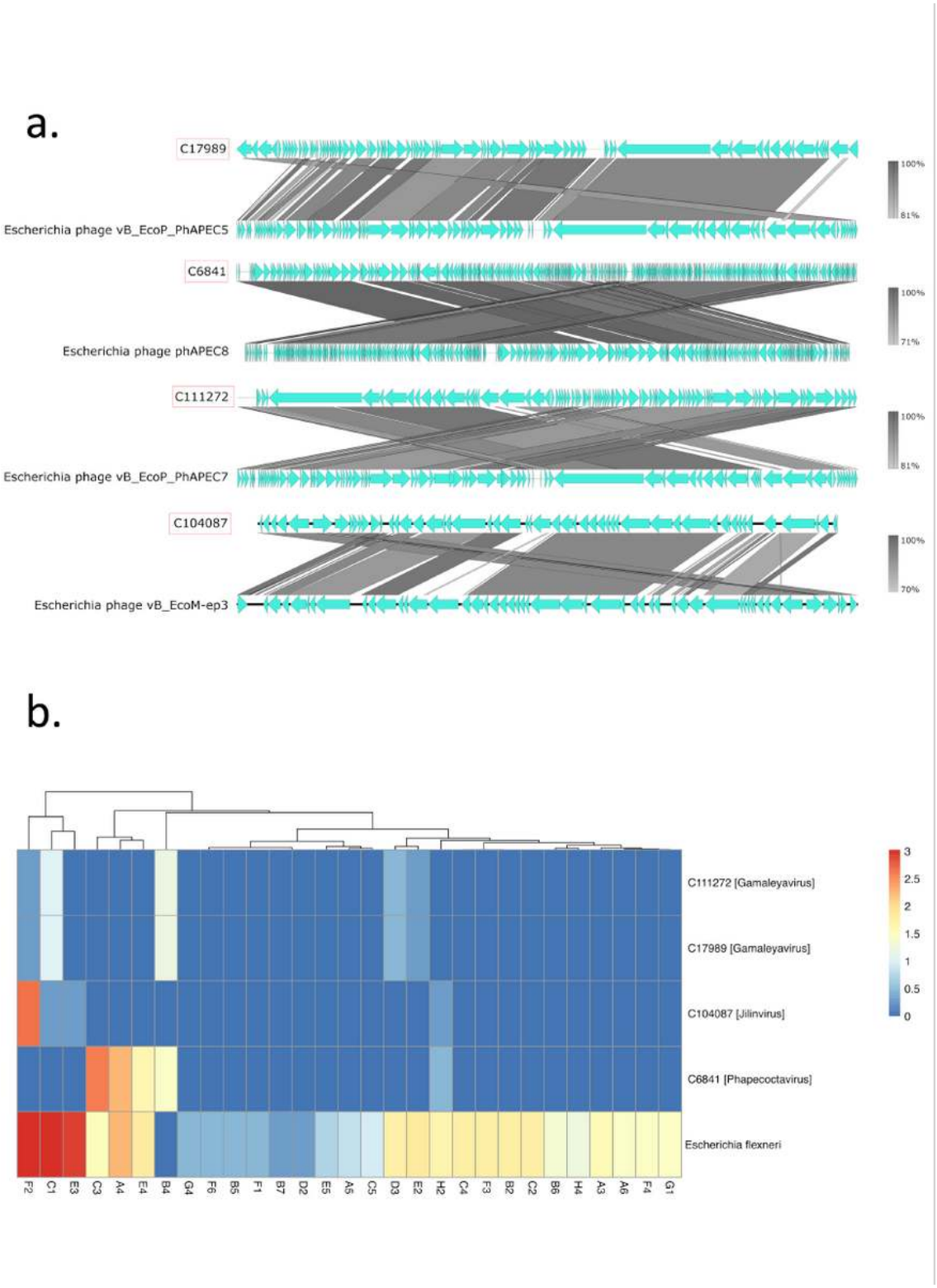

Figure 3

Sequence based analysis of four coliphage genomes recovered from the chicken faecal metagenomes a. Synteny plots comparing four novel coliphage genomes (in red) to closest reference genomes. b. Coverage of four coliphages and of the host bacterial species. Only samples in which at least one genome had ≥1x coverage are shown (n=29). All coverage values have been Log10 transformed with blue depicting low abundance and red high abundance.

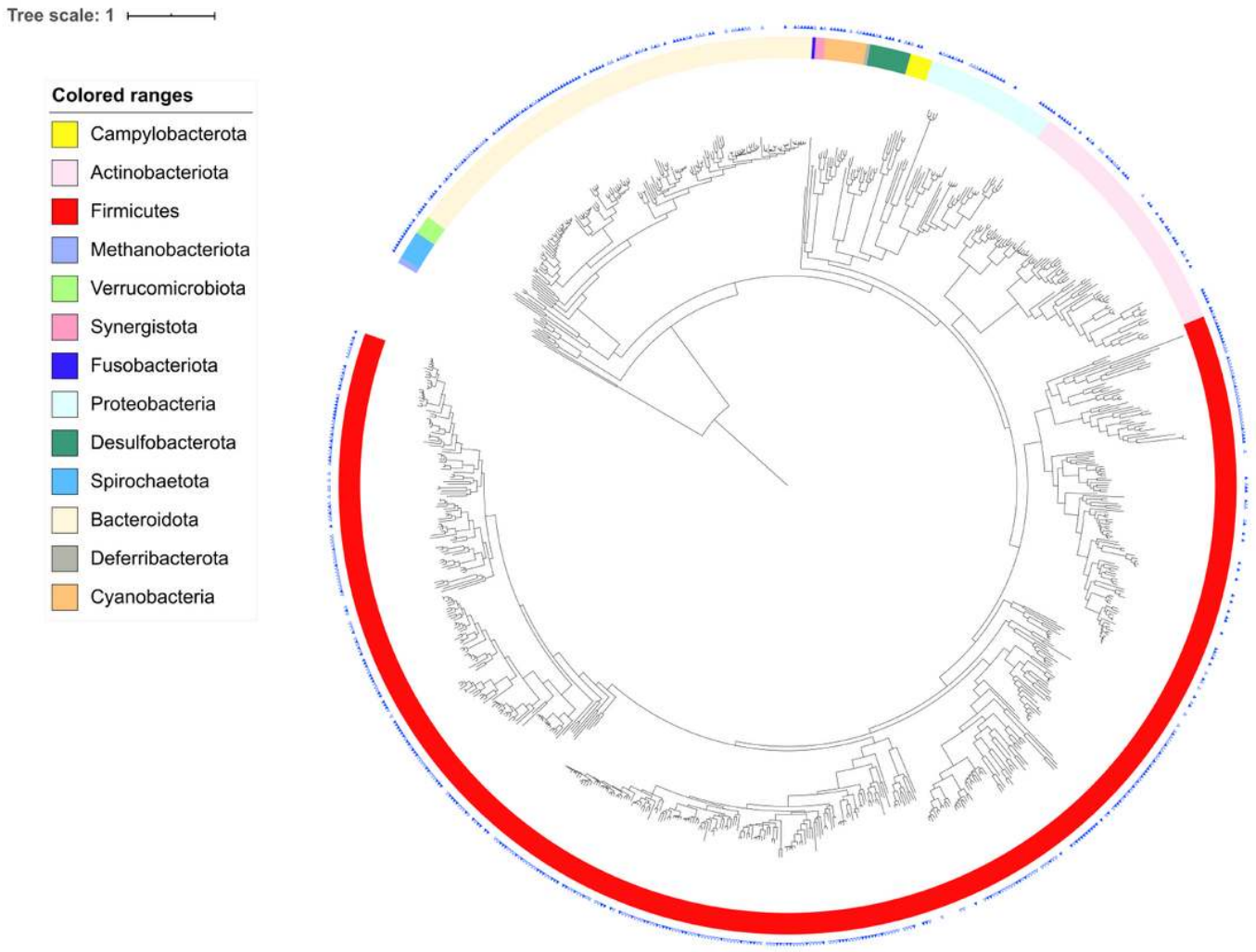

Figure 4

Sequence based analysis of four coliphage genomes recovered from the chicken faecal metagenomes a. Synteny plots comparing four novel coliphage genomes (in red) to closest reference genomes. b. Coverage of four coliphages and of the host bacterial species. Only samples in which at least one genome had ≥1x coverage are shown (n=29). All coverage values have been Log10 transformed with blue depicting low abundance and red high abundance.

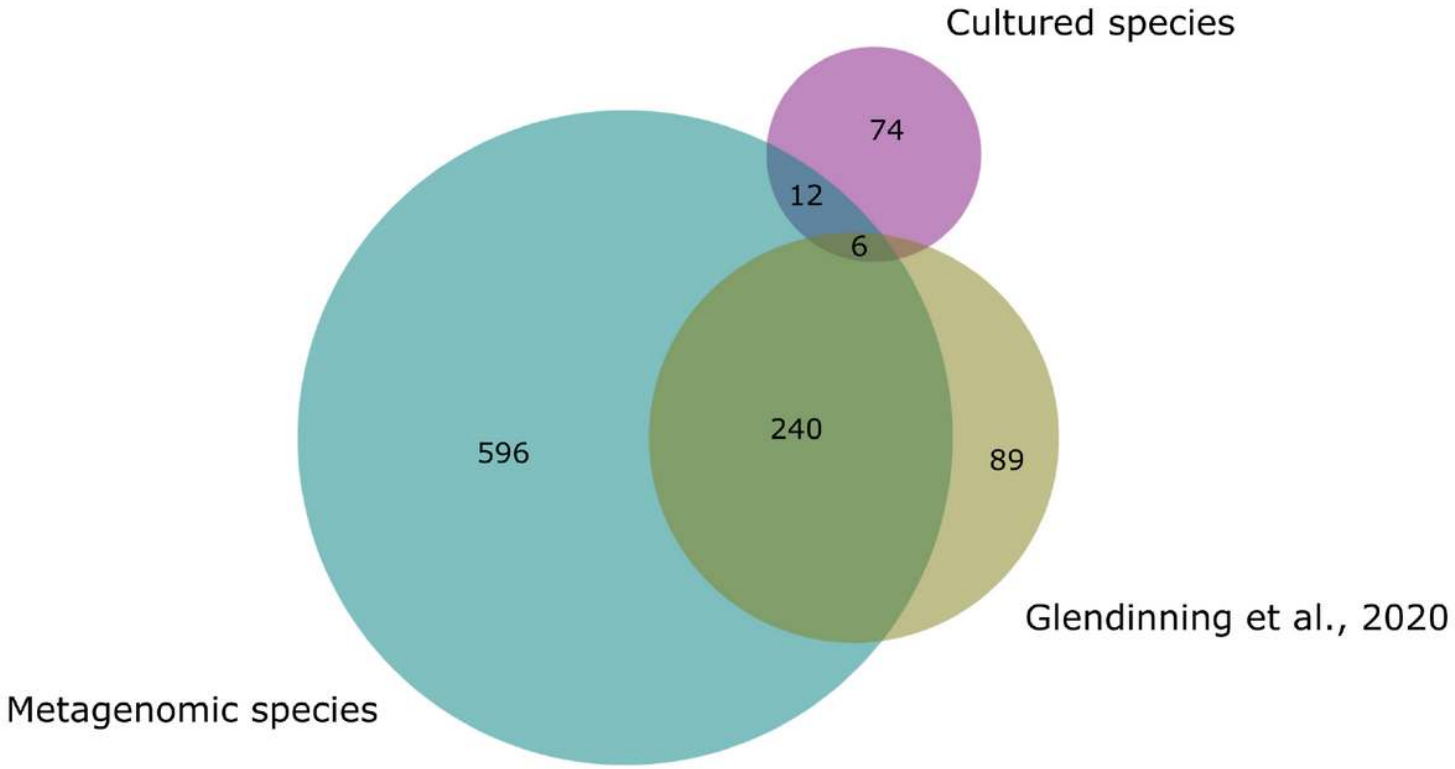

Figure 5

A Venn diagram showing shared and unique taxonomic species among three data sources; genomes of cultured isolates of 6 chicken faecal samples (Cultured species), metagenomic species identified from a combined dataset of >630 chicken gastrointestinal metagenome samples (Metagenomic species); MAGs also found by Glendinning et al., 2020 [15].

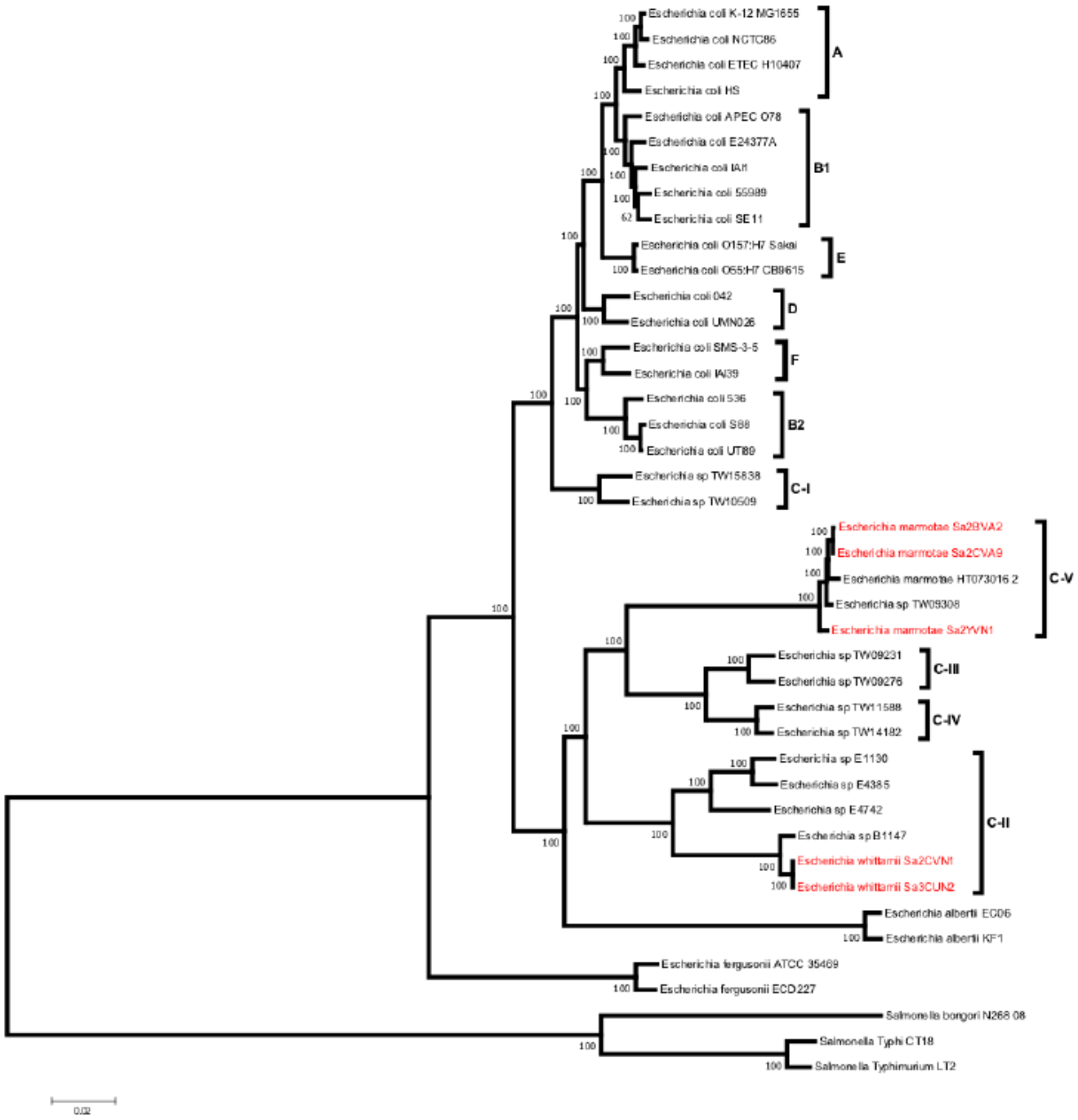

Figure 6

Phylogenetic tree showing the relationships between Escherichia marmotae, Escherichia whittamii and the other Escherichia species and cryptic clades. The tree was constructed by RAxML maximum likelihood analysis of a core genome alignment generated using Mugsy. The scale bar indicates the number of substitutions per site represented by the branch length shown. Numbers on branches indicate the percentage bootstrap support out of 100 replicates. Strains sequenced as part of this study are highlighted in red.

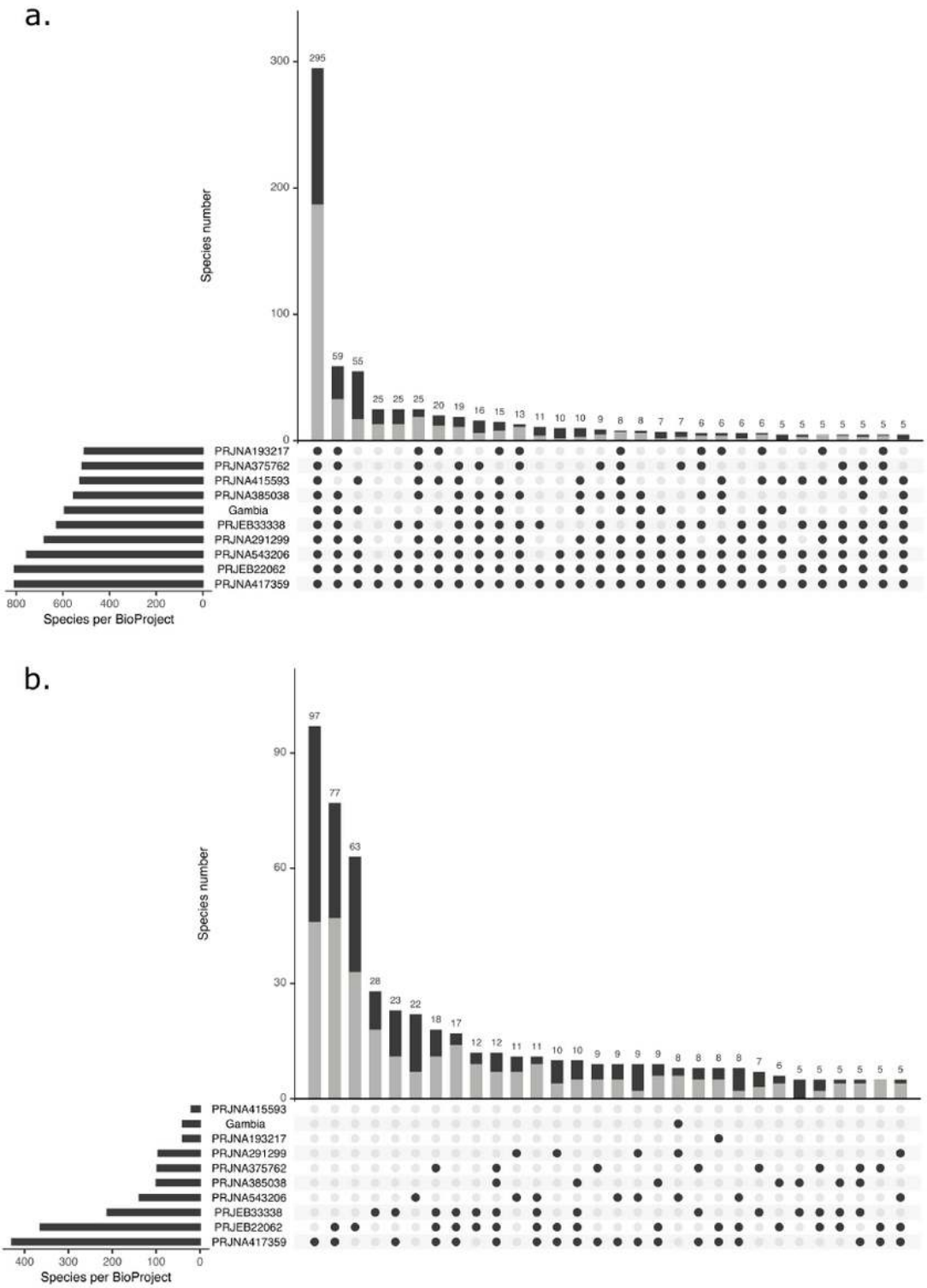

Figure 7

UpSet plots depicting presence of 820 metagenomic species across all BioProjects included within this study at a. 1x coverage and b. 10x coverage. Bars are stacked according to taxonomic species novelty, with black stacked bars depicting novel species and grey depicting species previously described in public databases or published studies. Only intersections with 5 or more species are shown.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- AdditionalFile1.xlsx
- AdditionalFile2.xlsx
- AdditionalFile3.xlsx
- AdditionalFile4.xlsx
- AdditionalFile5.xlsx
- SupplementaryFigure1.png
- SupplementaryFigure2.png
- SupplementaryFigure3.png
- SupplementaryFigure4.png
- SupplementaryFigure5.png