

RESEARCH

Open Access



Altitude-dependent agro-ecologies impact the microbiome diversity of scavenging indigenous chicken in Ethiopia

Laura Glendinning^{1*†}, Xinzheng Jia^{2*†}, Adebabay Kebede^{3,4}, Samuel O. Oyola⁵, Jong-Eun Park^{6,7}, Woncheoul Park⁷, Abdulwahab Assiri^{8,9}, Jacob Bak Holm^{10,11}, Karsten Kristiansen^{10,12}, Jianlin Han¹³ and Olivier Hanotte^{3,8*}

Abstract

Background Scavenging indigenous village chickens play a vital role in sub-Saharan Africa, sustaining the livelihood of millions of farmers. These chickens are exposed to vastly different environments and feeds compared to commercial chickens. In this study, we analysed the caecal microbiota of 243 Ethiopian village chickens living in different altitude-dependent agro-ecologies.

Results Differences in bacterial diversity were significantly correlated with differences in specific climate factors, top-soil characteristics, and supplemental diets provided by farmers. Microbiota clustered into three enterotypes, with one particularly enriched at high altitudes. We assembled 9977 taxonomically and functionally diverse metagenome-assembled genomes. The vast majority of these were not found in a dataset of previously published chicken microbes or in the Genome Taxonomy Database.

Conclusions The wide functional and taxonomic diversity of these microbes highlights their importance in the local adaptation of indigenous poultry, and the significant impacts of environmental factors on the microbiota argue for further discoveries in other agro-ecologies.

Keywords Chicken, Microbiota, Metagenomics, Ethiopia, Poultry, Agro-ecology

[†]Laura Glendinning and Xinzheng Jia contributed equally to this work.

*Correspondence:

Laura Glendinning

Laura.glendinning@roslin.ed.ac.uk

Xinzheng Jia

xinzhengjia@fosu.edu.cn

Olivier Hanotte

Olivier.Hanotte@nottingham.ac.uk; O.Hanotte@cgiar.org

¹ The Roslin Institute and Royal (Dick) School of Veterinary Studies, University of Edinburgh, Easter Bush, Midlothian, UK

² Guangdong Provincial Key Laboratory of Animal Molecular Design and Precise Breeding, School of Life Sciences and Engineering, Foshan University, Foshan, People's Republic of China

³ CTLGH - LiveGene, International Livestock Research Institute (ILRI), Addis Ababa, Ethiopia

⁴ Amhara Regional Agricultural Research Institute, Bahir Dar, Ethiopia

⁵ International Livestock Research Institute (ILRI), Nairobi, Kenya

⁶ Department of Animal Biotechnology, Faculty of Biotechnology, College of Applied Life Sciences, Jeju National University, 63243 Jeju, Republic of Korea

⁷ Animal Genomics and Bioinformatics Division, National Institute of Animal Science, 55365 Wanjū, Republic of Korea

⁸ School of Life Sciences, the University of Nottingham, University Park, Nottingham, UK

⁹ Department of Animal and Fisheries Production, King Faisal University, Al-Hofuf, Saudi Arabia

¹⁰ Laboratory of Genomics and Molecular Biomedicine, Department of Biology, University of Copenhagen, Copenhagen, Denmark

¹¹ Clinical Microbiomics, Copenhagen, Denmark

¹² Qingdao-Europe Advanced Institute for Life Sciences, BGI-Shenzhen, Qingdao, People's Republic of China

¹³ CAAS-ILRI Joint Laboratory On Livestock and Forage Genetic Resources, Institute of Animal Science, Chinese Academy of Agricultural Sciences (CAAS), Beijing, People's Republic of China



Background

Indigenous village chickens play a key role in sub-Saharan African and Asian countries, sustaining the livelihood of millions of farmers. They predominantly comprise indigenous chicken genotypes well adapted to the local environment [1] and acquire all or a substantial proportion of their diet from scavenging (scavenging or semi-scavenging). As these birds are exposed to high predation and disease challenges [2], survival traits rather than production traits are favoured by natural selection and, to some extent, human selection. These chickens also contribute to the spread of zoonotic diseases such as campylobacteriosis and salmonellosis, which have a high disease burden in African human populations [3, 4]. By understanding the microbiota of these chickens, we may be able to suggest ways of improving their nutrition and disease resistance, and henceforth productivity.

Over 95% of poultry products sold in Ethiopia originate from indigenous village chickens [5]. Ethiopia is an ecologically diverse country with distinct environmental zones, ranging from the hot and arid climate of the lowlands to the cold and humid climate of the highlands. These climatic zones, and a diverse geographical topography, form naturally varied environmental conditions for smallholder crop-livestock farming systems. Geographical location [6], temperature [7] and altitude [8] have previously been demonstrated to impact the composition of the chicken microbiota using 16S rRNA gene amplicon sequencing.

The gut microbiota plays important roles in chicken health and productivity, contributing to nutrition, immune development and pathogen resistance [9]. The highest concentration of microbes in chicken can be found in the caeca. Here, the microbial communities play a vital nutritional role, fermenting fibre into short-chain fatty acids (SCFAs) that can be used as an energy source by the bird, as well as contributing to colonisation resistance, and nitrogen recycling [10, 11].

Through the use of culturing techniques [12, 13], metabarcoding [14, 15] and metagenomics [13, 16, 17], there have been many recent advances in our understanding of the chicken caecal microbial ecosystems. However, most of these studies have examined grain-fed, commercial chicken breeds that were hatched and housed in biosecure facilities without maternal contact. These commercial-like conditions aim to enhance productivity by standardising/controlling host-environment interactions. Commercial chicken breeds, and by extension their microbiota, have also been shaped by human selection for high productivity traits. On the contrary, scavenging or semi-scavenging chicken populations are exposed to far greater predation and disease challenges [2], and therefore, survival traits have been selected for rather

than production. Not only has this affected the chicken genome [1], but it would also reasonably be expected to have an impact on the gut microbiota.

To characterise the microbiota of indigenous chickens in Ethiopia, we collected 243 chicken caecal content samples from 26 villages in 15 districts. Shotgun metagenomics was used to characterise the caecal microbiota. To profile all taxa, including low-abundance taxa, we constructed a catalogue of non-redundant genes. We identified three distinct enterotypes, one of which was particularly more abundant in the highest altitude samples. We also constructed metagenome-assembled genomes (MAGs). Previously, MAGs have been constructed from chicken breeds, including Ross 308s, Lohman Browns and Silkies [13, 16, 17]. We constructed 9977 high-quality, strain-level MAGs and 1790 species-level MAGs, representing diverse taxonomies. We found that the vast majority of the MAGs were not present in a dataset of microbial genomes from previous chicken microbiota studies. The MAGs generated in this study contained genes encoding a large diversity of carbohydrate-active enzymes (CAZymes) and metabolic functions.

Results

After the removal of three samples during quality control, we characterised the microbiota of the caecal contents of 240 indigenous Ethiopian scavenging chickens from 26 sampling sites, aiming to further understand the impact of agro-ecology on the gut microbiota composition. The sampling sites were highly diverse, representing different latitudes, longitudes and altitudes (Fig. 1B). Five distinct climate zones were defined based on climate variation analysis, with altitude and annual mean temperature as major predictors (Fig. 1C).

Construction of an Ethiopian indigenous chicken microbiota gene catalogue

By constructing a gene catalogue, we can detect more rare taxa than by using MAGs alone and also identify taxa from which it is difficult to construct high-quality MAGs due to their large genome sizes or complex genome structures. We constructed a reference gene catalogue of 33,629,587 non-redundant genes (Additional file 3: Fig. S1). Rarefaction analysis of sampling size revealed that 90% of genes were captured with a sampling size of 60 individuals (Additional file 3: Fig. S2). Genes identified across many individuals within a population can be defined as “core genes”. Only 420,891 (1.2%) genes were shared amongst 80% of the samples. This may be due to the high inter-individual diversity of microbiota-derived genes within the caecal samples or to insufficient sequencing depth to detect the total gene numbers.

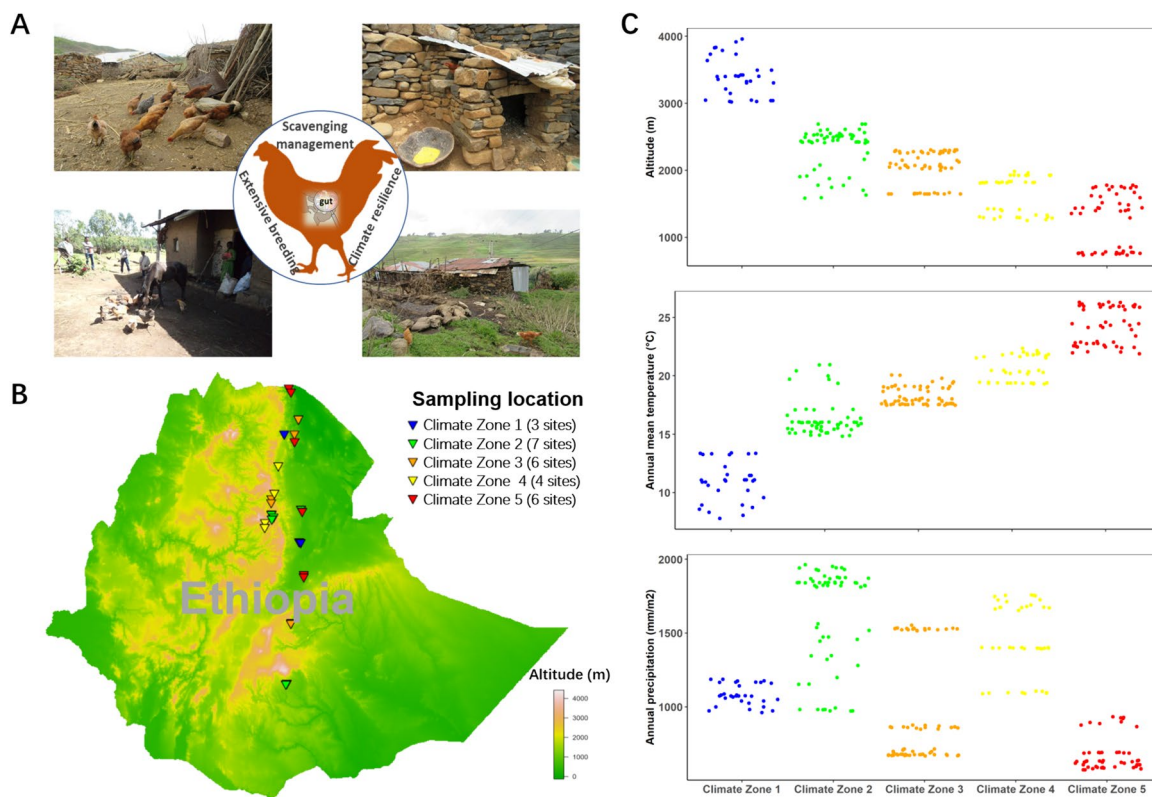


Fig. 1 Sample site information for Ethiopian indigenous chickens under scavenging production systems. **A** Examples of scavenging production systems in Ethiopia. **B** Geographic distribution of sampling sites within Ethiopia. **C** Five climate zones were defined based on climate factors. These included altitude, precipitation and temperature, as shown in the figure

We next characterised the taxonomic origin of the genes in the gene catalogue. Fifty-five percent of the genes were assigned a taxonomic label. Of these genes, 87.9%, 56.5% and 49.2% were assigned to a phylum, genus and species, respectively (Fig. 2A). After removing DNA that likely originated from the host diet (plants and insects), a total of 33,435,297 genes remained that were used for microbiota analysis.

Our analysis focused on taxa averaging at least one count per million sequences. Each sample had an average relative abundance of 98% bacteria ($\pm 1.22\%$), 0.86% archaea ($\pm 0.54\%$), 0.98% Eukaryota ($\pm 1.08\%$) and 0.16% viruses ($\pm 0.13\%$). Bacteroidota represented the most abundant phylum (48.0%), followed by Firmicutes (32.9%), Proteobacteria (7.3%), Spirochaetota (4.2%), Actinobacteriota (1.9%) and Deferribacterota (1.1%). Six of the 10 most abundant phyla showed significant differences in abundance between climate zones (Fig. 2C). The most abundant genera were *Bacteroides* (31.9%), *Alistipes* (7.9%) and *Prevotella* (7.6%), all members of the Bacteroidota phylum. *Prevotella*, in particular, showed a high level of variation between climate zones (Fig. 2D–E). Notably, chickens from climate zone 1 (high altitude)

had more than 2.5-fold higher abundance of *Prevotella* (17.6%) than chickens from other climate zones. Archaea, bacteria and Eukaryota showed significant differences in alpha-diversity between climate zones (Kruskal–Wallis p value < 0.01), while viruses showed no significant differences (Fig. 2B). For bacteria, diversity gradually increased from climate zone 1 (> 3000 m, high altitude) to climate zone 4 (around 1300 m, medium altitude) and was slightly decreased in climate zone 5 (around 1000 m).

Indigenous Ethiopian chickens contain three enterotypes related to environmental distribution

Using bacterial genera abundances as estimated from the gene catalogue, the Ethiopian chicken caecal microbiota clustered into three enterotypes (Fig. 3A). Samples belonging to enterotype 3 were particularly distinct from enterotypes 1 and 2. Climate zone 1 is clearly dominated by microbiota of enterotype 3, which accounts for 62% of samples from this zone (Fig. 3B). Linear discriminant analysis effect size (LEfSe) analysis was carried out to identify differential enrichment of genera between enterotypes. For enterotype 1, the top ten most discriminating genera were *Alistipes*, *Treponema*, *Brachyspira*,

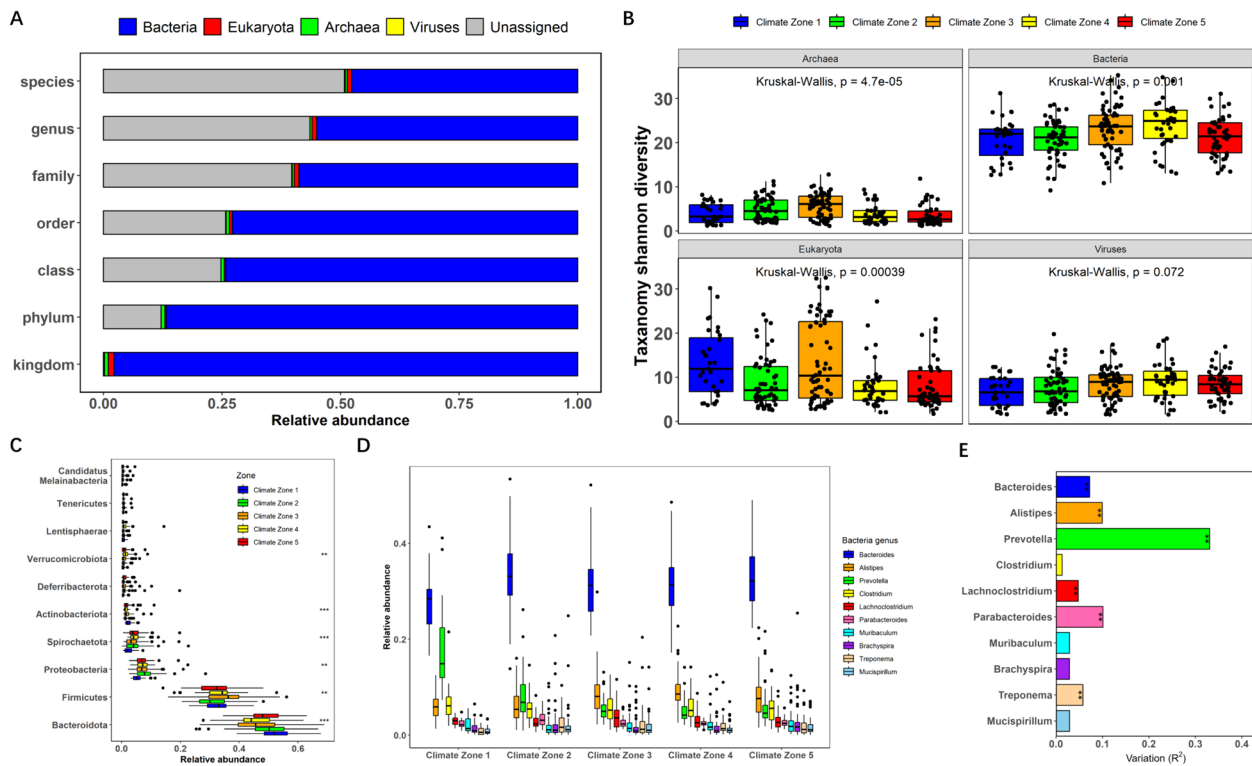


Fig. 2 Structure and phylogenetic diversity of the caecal microbiota from indigenous Ethiopian chickens. **A** Abundance distribution of taxa at each phylogenetic level. **B** Alpha-diversity (Shannon index) of the archaea, bacteria, viruses and Eukaryota in caecal samples. **C** Relative abundance of microbiota at phylum level among climate zones (** $p < 0.01$, *** $p < 0.001$). **D** Genus relative abundance among climate zones. **E** Comparison of genera abundance among populations from distinct climate zones

Mucispirillum, *Muribaculum*, *Parabacteroides*, *Sutterella*, *Acidaminococcus*, *Sphaerochaeta* and *Akkermansia*; for enterotype 2, the top genera were *Bacteroides*, *Lachnospirillum*, *Clostridium*, *Blautia*, *Pseudoflavonifractor*, *Flavonifractor*, *Erysipelatoclostridium*, *Azospirillum*, *Merdimonas* and *Gemmiger*; and for enterotype 3, *Prevotella*, *Megamonas*, *Faecalibacterium*, *Olsenella*, *Lactobacillus*, *Bifidobacterium*, *Mediterranea*, *Collinsella*, *Megasphaera* and *Dialister* (Fig. 3C, D). Co-occurrence networks based on correlation analysis revealed that the abundance of major discriminating genera had strong positive correlations with the abundance of other genera (Fig. 3E).

Agro-ecological factors contributing to microbiota composition

We next wanted to identify which agro-ecological factors were correlated with differences in the beta-diversity of the caecal microbiota, using the bacterial gene catalogue data. Beta-diversity was significantly associated with supplementary diets provided by farmers and the location’s topsoil characteristics. Significant differences in

beta-diversity (Bray–Curtis dissimilarity: genera, species and strains) were associated with temperature, altitude, precipitation and seasonal cycles, followed by supplementary diets, co-raising styles and local topsoil contents (adj- $p < 0.01$). Redundancy analysis was used to capture the driving factors contributing to bacterial microbiota diversity. For climate factors, altitude and seasonal precipitation (bio 15) were the major factors, explaining 10% of the total variation (Fig. 4A). Cation exchange capacity of topsoil (CECSOL) and silt percentage of topsoil (SLTPPT) were identified as the major contributors to diversity, explaining 11% of the variation (Fig. 4B). The provision of supplementary diets to chickens was also considered. We found that three common grains (maize, wheat and barley) significantly impacted the bacterial microbiota diversity, explaining 8% of the total variation. Diversity significantly decreased as altitude increased ($P < 0.01$, Fig. 4C). Amongst the main taxonomic contributors to enterotypes, *Prevotella* and *Megamonas* were significantly positively correlated to altitude (Fig. 4D, E), while *Coralloccoccus* was negatively correlated (Fig. 4F).

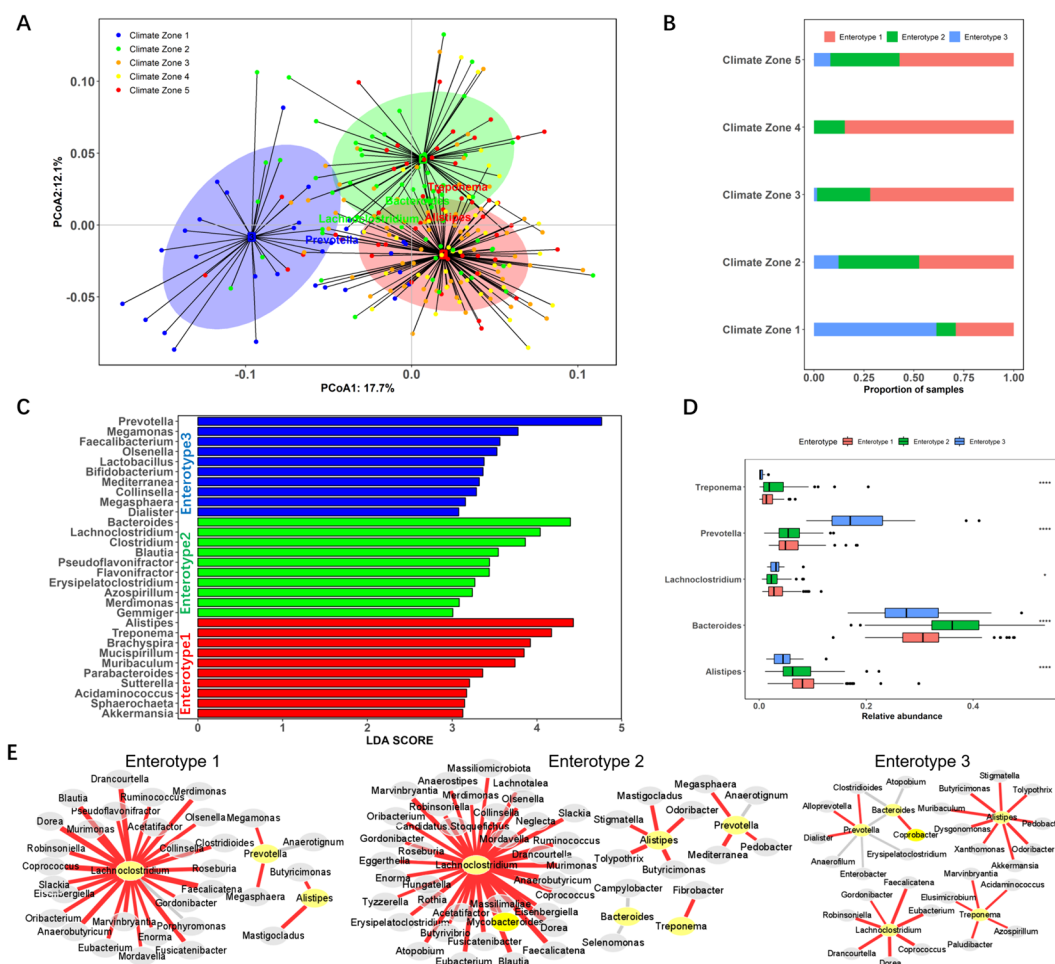


Fig. 3 Three enterotypes observed in the caecal microbiota of Ethiopian indigenous chickens. **A** Caecal microbiota enterotypes clustered by PCA, with discriminating genera highlighted. **B** Distribution of enterotypes as proportions of samples from different climate zones. **C** Top ten highest linear discriminant analysis (LDA) scores for genera contributing to the discrimination of each enterotype. **D** Abundance profiles of the main genera contributing to each enterotype, as defined by LefSe ($*p < 0.05$, $***p < 0.001$). **E** Co-occurrence networks based on correlations between genera abundance. Red and grey lines indicated positive and negative correlations ($p < 0.01$, $\rho > 0.5$ relative abundance $> 0.01\%$). Yellow nodes represented the main genus contributors (LDA score > 4)

Assembly of 9977 microbial genomes from diverse taxonomies

Whilst gene catalogues can provide information about the taxonomies present in a dataset, including those that are present only in low abundances, it is also possible to generate genomic-level information about components of the microbiota by constructing metagenome-assembled genomes (MAGs) from the more abundant taxa in the samples.

We constructed 9977 high-quality, non-redundant MAGs at strain-level, and 1790 at species-level (Additional file 4: Table S1). For strain-level MAGs, 9815 were identified as bacteria and 162 as Archaea (Additional file 4: Table S2). Archaea comprised three phyla, Halobacteriota ($n = 39$), Methanobacteriota ($n = 28$) and Thermoplasmata ($n = 95$). Bacteria belonged to

19 different phyla, with the most abundant being Bacteroidota ($n = 2846$), Firmicutes_A ($n = 2696$), Proteobacteria ($n = 985$), Firmicutes ($n = 970$) and Spirochaetota ($n = 842$) (Figs. 5A and 6A).

Various bacteria of interest as potential food-borne pathogens or poultry pathogens were identified. *Campylobacter* spp. are a common cause of food-borne diarrhoeal disease in humans. Several species of *Campylobacter* were identified amongst the MAGs, including *Campylobacter avium* (15 strains), *Campylobacter coli* (4 strains), *Campylobacter jejuni* (10 strains) and two novel species. Various strains of *Escherichia* and *Shigella* can cause disease in humans through contaminated poultry meat. Surprisingly, only one strain of *Escherichia coli* was identified among the MAGs, with most

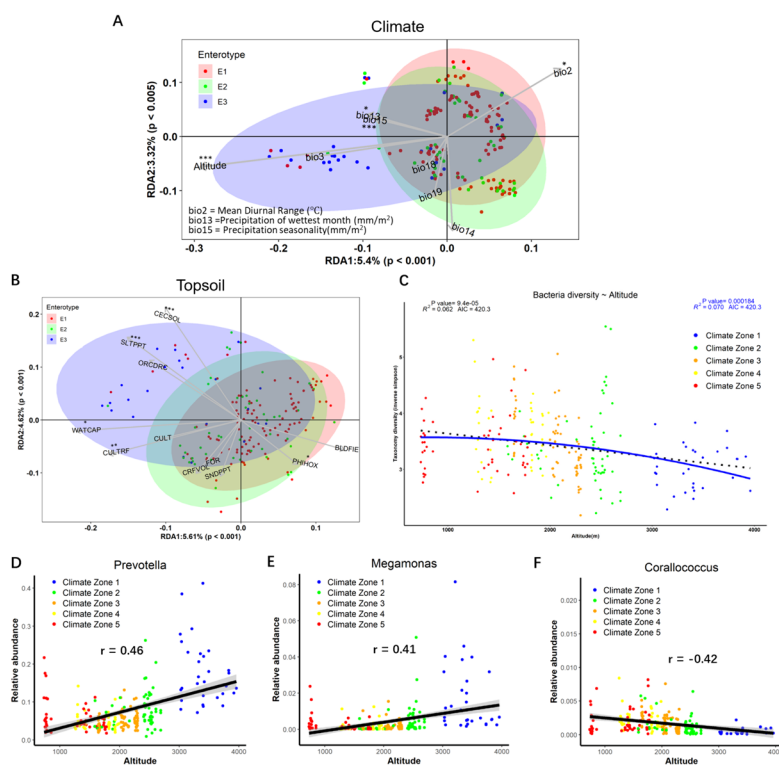


Fig. 4 Bacterial microbiota composition exhibited significant geographic and ecological diversity. **A** RDA analysis on the bacterial microbiota composition of samples (gene catalogue) and climate factors including altitude, temperature and precipitation. **B** RDA analysis on the bacteria composition and ecological factors including common topsoil characteristics. **C** Bacterial diversity (Inverse Simpsons index) decreases with altitude gradient. The best polynomial fit in blue was determined on the basis of the corrected Akaike information criterion (AIC) of the order polynomial models. The ANOVA test was used to test for significance. **D–F** Spearman correlation was performed to test the relationship between the abundance of genera and altitude

Escherichia strains identified as *Escherichia flexneri* ($n=17$, also known as *Shigella flexneri*), and one strain being identified as *Escherichia dysenteriae*. Members of the *Helicobacter* genus are common causes of gastroenteritis. Ninety-five strains of *Helicobacter* were identified amongst the MAGs; 40 of these MAGs were identified as *Helicobacter pullorum*, with the remaining MAGs defined as belonging to five separate “*Helicobacter*” genera. Only one MAG was identified as belonging to the Chlamydia, *Chlamydophila gallinacean*.

Compared to using the gene catalogue, after mapping raw sequencing reads to MAGs, the proportion of reads annotated as bacterial and archaeal taxa increased from 41.1 to 69.9% and from 0.4 to 1.3%, respectively (Additional file 3: Fig. S4). The relative abundance of MAGs in the samples was estimated, and the phylum abundance profiles were highly correlated between mapping using the gene catalogue or MAG database ($r=0.98$). Using species-level MAG abundances, the diversity and richness of the caecal microbiota of chickens from different climate zones were compared. Around 66% of species were identified as core species, i.e., present in all 26

sampling sites. Far fewer core species were found when comparing individual samples (present in at least 90% of samples), but despite their smaller numbers, these core species still accounted for around 50% of the average microbiota composition of samples (Fig. 5E). The richness (Fig. 5C, $p=1.8e-09$) and diversity (Fig. 5D, $p=0.038$) of the MAGs differed significantly between climate zones. No significant differences were observed in the alpha-diversity of the core species between climate zones (Fig. 5F). However, for non-core members of the microbiota, there were clear differences in alpha-diversity between climate zones ($p=8.7e-6$), with diversity tending to decrease with altitude, except in climate zone 5 (Fig. 5G).

We next wanted to identify whether any of our MAGs were differentially abundant in the three enterotypes that we had previously defined using our gene catalogue data. Of our 1790 species-level MAGs, 1404 were differentially abundant between enterotypes (Kruskal–Wallis adj- $p < 0.05$, Additional file 4: Table S3). In total, 114 of these MAGs were found at tenfold higher abundance in enterotype 1 compared to the other enterotypes. In

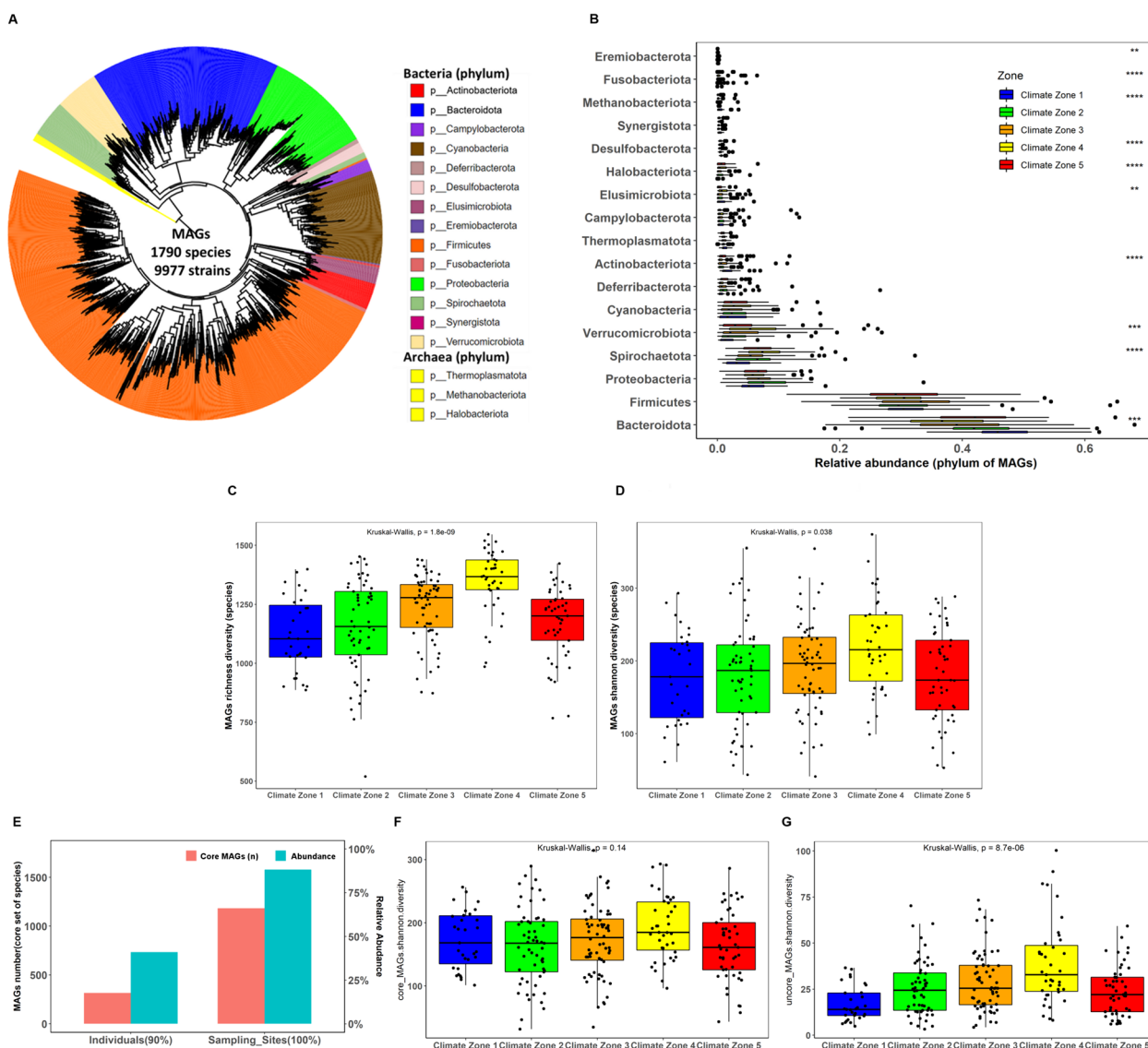


Fig. 5 Species-level MAG taxonomy and diversity between samples and climate zones. To aid readability, all phyla classed as Firmicutes (Firmicutes_A, Firmicutes_B, etc.) have been concatenated under the label “Firmicutes”. **A** Phylogenetic tree of species-level MAGs. **B** Abundance of phyla between climate zones (** $p < 0.01$, *** $p < 0.001$). **C** Richness of MAGs between 100% of sampling sites and at least 90% of individual samples. **D** Shannon diversity of MAGs between climate zones. **E** The prevalence and relative abundance of core MAGs shared between 100% of sampling sites and at least 90% of individual samples. **F** Shannon diversity of core MAGs between climate zones. **G** Shannon diversity of non-core MAGs between climate zones

contrast, only eight MAGs were found at tenfold higher abundance in enterotype 2. Thirty-six significantly differently abundant MAGs were found at tenfold higher abundance in enterotype 3 compared to the other enterotypes; these MAGs originated from a wide diversity of taxa (eight phyla, Table 1). Previously, we identified *Prevotella* as being highly variable in abundance between climate zones and enterotypes. Of the 14 *Prevotella* species that were differentially abundant between enterotypes, 12 were more abundant in enterotype 3 than in the other two enterotypes, with two MAGs at tenfold higher

abundance (*Prevotella* sp000431975 and *Prevotella* copri).

Comparing Ethiopian chicken MAGs to taxonomies found in non-scavenging chickens

The microbes identified in this study originate from indigenous scavenging chickens that are very distinct from commercial breeds in terms of genetics, diets and environments. As such, it may be expected that many of the microbes we have identified would be taxonomically and functionally distinct from those found in

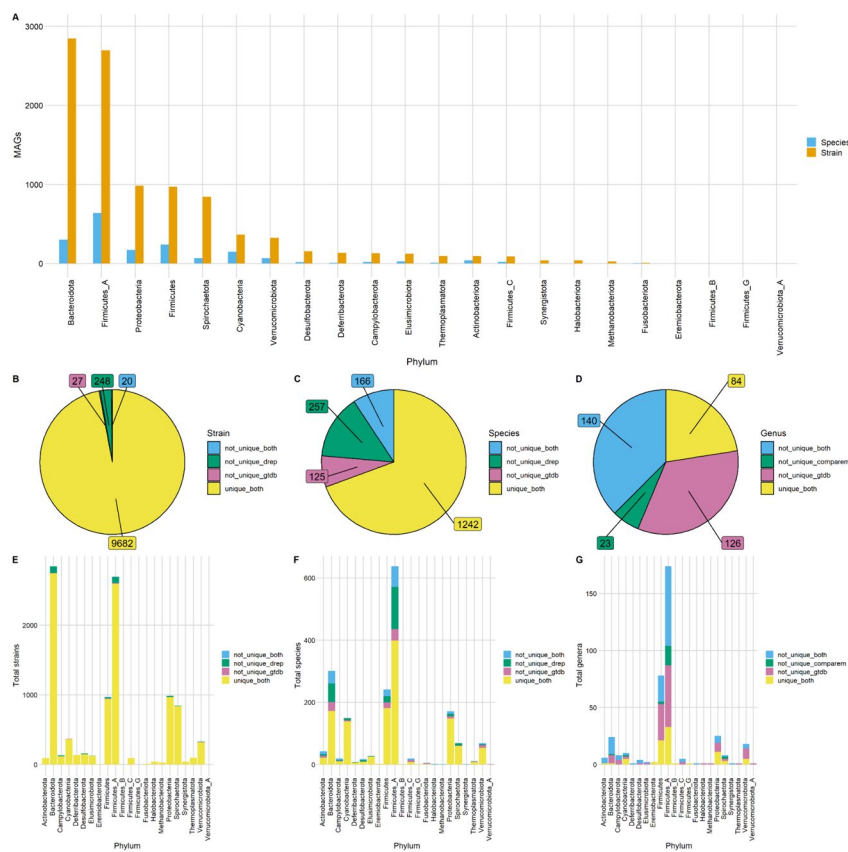


Fig. 6 Comparison of Ethiopian chicken caecal MAGs to microbial genomes from non-scavenging chickens and the GTDB. **A** Barplot showing the number of MAGs assigned to different phyla at both strain and species-level. The number of strain-level MAGs (**B** and **E**), species-level MAGs (**C** and **F**) and genus-level clusters (**D** and **G**) that were assigned as “unique” to our dataset according to the following criteria: Strain and species-level MAGs were not unique based on GTDB (“not_unique_gtodb”) if the average nucleotide identity (ANI) between the query and GTDB reference genome were > 99% or > 95%, respectively. Genus-level clusters were not unique based on GTDB if any MAGs within that cluster were assigned taxonomy at genus level. MAGs were defined as not unique when compared to previous chicken microbial datasets (“not_unique_drep”) if they clustered at 99% (strain) or 95% (species) ANI with any NSC microbial genome. Genera were defined as not unique when compared to previous chicken microbial datasets (not_unique_compare) if they clustered at 60% AAI with any NSC microbial genome

non-scavenging chickens. We therefore compared our MAGs to microbial genomes from a non-scavenging chickens (NSCs) dataset and to the Genome Taxonomy Database (GTDB).

The majority of our strain-level and species-level MAGs were absent in either the NSC dataset or GTDB (Fig. 6B, C, E, F, Additional file 4: Table S4). For the strain-level MAGs, only 268 were identified in the NSC dataset and 47 in the GTDB, leaving 9682 strains not identified in either. At the species-level, 423 of our MAGs were identified in the NSC dataset, and 291 were identified in the GTDB, leaving 1242 species that were not identified. We clustered MAGs into 373 genus-level clusters, according to their amino acid identities (AAI). Of these genera, 163 were found in the NSC dataset, while 266 were found in the GTDB, leaving 84 genera that were not identified in either (Fig. 6D, G). Genera found to be unique to our dataset originated from a wide variety of taxa.

The majority of species-level MAGs carried at least one antimicrobial resistance (AMR) gene ($n=783$), with tetracycline being the most common class of drugs targeted by the AMR genes. Tetracycline resistance was also the most commonly targeted drug class among the NSC genomes. One MAG in particular, identified as *Escherichia flexneri*, was noted as containing a large number of AMR genes ($n=54$) in comparison to other MAGs. Five NSC genomes also contained large numbers of AMR genes: *Pseudomonas aeruginosa* ($n=59$), *Escherichia* sp. *Cl-CZ-1* ($n=56$), *Escherichia whittamii* ($n=49$), *Enterobacter roggkampii* ($n=35$) and *Klebsiella pneumonia* ($n=34$).

Functional characterisation of MAGs isolated from Ethiopian indigenous chickens

The caecal microbiota plays an important role in the fermentation of carbohydrates that are not able to be

Table 1 Species-level MAGs that differed significantly in abundance between enterotypes

Phylum	Total differentially abundant MAGs (n)*	Enterotype 1: tenfold higher abundance (n) ^a	Enterotype 2: tenfold higher abundance (n) ^a	Enterotype 3: tenfold higher abundance (n) ^a
Actinobacteriota	35	1	0	4
Bacteroidota	218	0	1	4
Campylobacterota	4	0	0	0
Cyanobacteria	129	19	0	3
Deferribacterota	5	0	0	0
Desulfobacterota	14	0	1	0
Elusimicrobiota	23	4	1	0
Eremiobacterota	2	1	0	0
Firmicutes	180	16	1	6
Firmicutes_A	510	48	3	11
Firmicutes_B	1	0	0	0
Firmicutes_C	14	0	0	4
Firmicutes_G	1	0	0	0
Halobacteriota	1	0	0	0
Methanobacteriota	1	0	0	0
Proteobacteria	147	21	1	3
Spirochaetota	52	2	0	1
Synergistota	1	0	0	0
Thermoplasmata	9	1	0	0
Verrucomicrobiota	56	1	0	0
Verrucomicrobiota_A	1	0	0	0

* adj- $p < 0.05$, species-level MAGs^a Relative abundance in comparison to other enterotypes. Only includes MAGs that were significantly differentially abundant

digested and absorbed in the small intestine of the host. SCFAs are produced through the fermentation of these fibrous compounds. These SCFAs can be absorbed by the host and used as an energy source. Scavenging chickens likely consume different fibrous compounds and a greater diversity of fibre than non-scavenging chickens.

Our MAGs contained a large diversity of CAZymes (Additional file 3: Table S1, Supplementary table - MAGs metabolism strain ([dx.doi.org/10.6084/m9.figshare.22154597](https://doi.org/10.6084/m9.figshare.22154597)) and Supplementary table - MAGs metabolism species ([dx.doi.org/10.6084/m9.figshare.22154627](https://doi.org/10.6084/m9.figshare.22154627))). CAZymes are enzymes involved in the synthesis, metabolism and binding of carbohydrates. As such, it would be expected that a microbe that was rich in CAZymes would be able to thrive on a more diverse set of carbohydrates and would be more of a nutritional generalist than a microbe with a less rich CAZyme profile [18]. Phyla clustered significantly by CAZyme composition for all domains (ANOVA, $p = 1e - 05$, Fig. 7).

Overall, MAGs from the Bacteroidota and Verrucomicrobiota contained the highest numbers of CAZyme genes, both by total CAZyme gene count (Bacteroidota, 92 ± 39 ; Verrucomicrobiota, 127 ± 84) and the number of unique CAZyme families (Bacteroidota, 44 ± 18 ;

Verrucomicrobiota, 49 ± 15). For the Verrucomicrobiota, 41 MAGs contained over 250 CAZyme genes and therefore represent some of the most CAZyme-rich genomes in our dataset. These genomes belong to either of two families, Victivallaceae and UBA1829, and 16 were not identified in the GTDB or NSC dataset.

MAGs from the three Archaeal phyla had particularly low numbers of CAZymes: Thermoplasmata (total CAZyme genes: 6 ± 3 , unique CAZyme families: 4 ± 1), Halobacteria (total CAZyme genes: 9 ± 3 , unique CAZyme families: 7 ± 2) and Methanobacteriota (total CAZyme genes: 24 ± 8 , unique CAZyme families: 10 ± 2). The MAGs with the lowest CAZyme richness (< 1 CAZyme genes) were members of the Mycoplasmatales (phylum: Firmicutes).

The average numbers of CAZymes per genome were similar between our dataset (60.08 ± 40.99) and the NSC dataset (68.24 ± 45.97) (Additional file 3: Fig. S5). In total, 266 CAZymes were shared between the two datasets, while nineteen were unique to the NSC dataset. A further nineteen CAZymes were identified in our MAGs but not in the NSC genomes: four carbohydrate-binding molecules (CBM11, CBM65, CBM68 and CBM79),

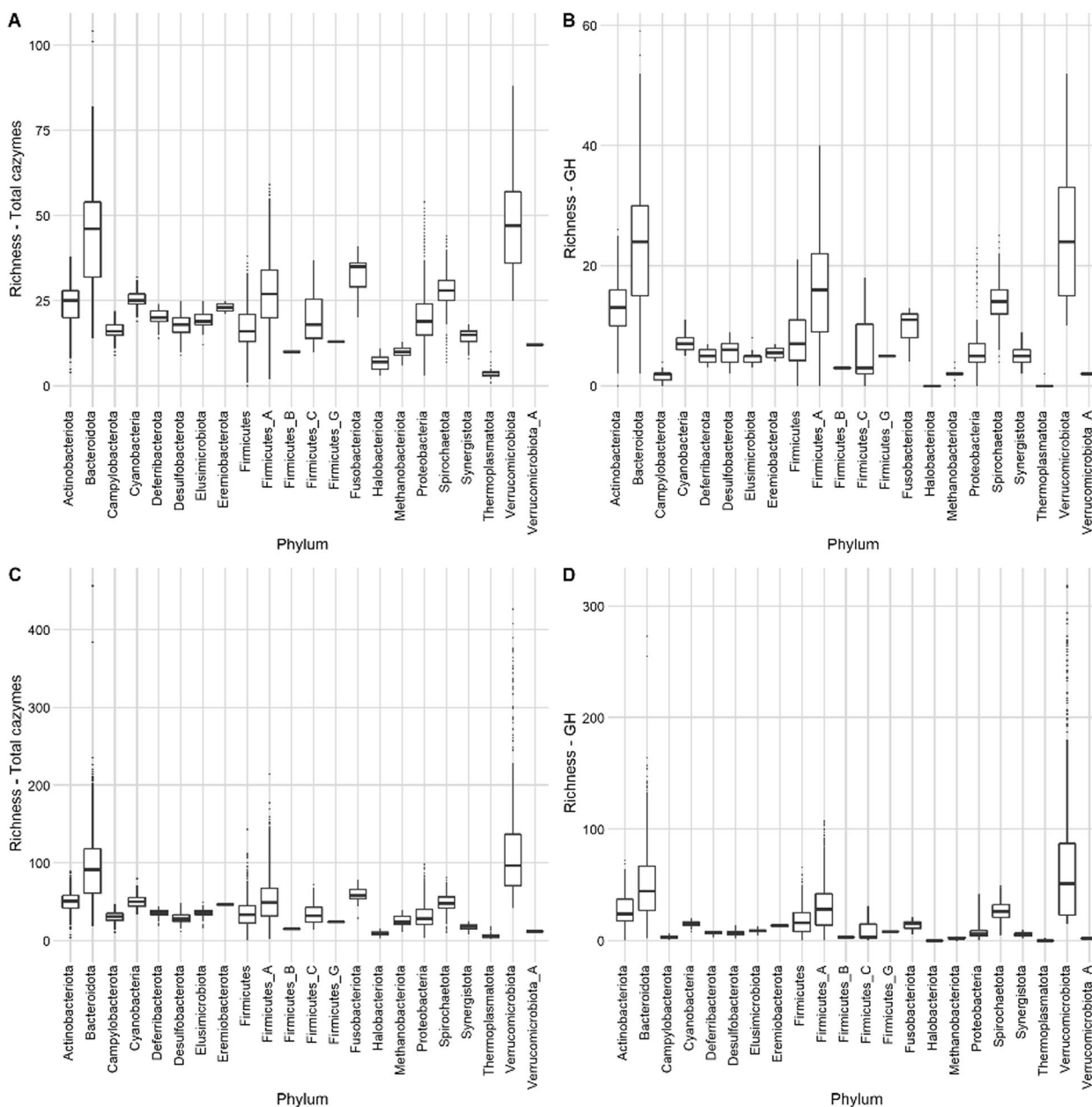


Fig. 7 Boxplots showing the number of CAZyme genes per strain-level MAG by phylum. **A** Total unique CAZyme families. **B** Total unique glycoside hydrolases (GH) families. **C** Total CAZyme genes. **D** Total GH genes

five glycoside hydrolases (GH47, GH86, GH107, GH119 and GH160), eight glycosyltransferases (GT13, GT15, GT40, GT44, GT60, GT74, GT75 and GT103) and two polysaccharide lyases (PL25 and PL32). These CAZyme genes originated from MAGs from a wide range of taxonomies. For example, for the glycoside hydrolases: GH47 (α -1,2-mannosidases) originated from two Bacteroidales strains, GH86 (β -agarase/ β -porphyranase) originated from 8 strains of the family UBA1829 (phylum

Verrucomicrobiota), GH107 (endo- α -1,4-L-fucanase) originated from one strain of the family UBA3636 (phylum Verrucomicrobiota), GH119 (α -amylase) originated from 2 Succinivibrionaceae strains, and GH160 originated from one *Parabacteroides* strain. Our MAGs also showed a wide diversity of predicted growth rates. These were found to significantly relate ($P < 0.05$) to CAZyme richness (Additional file 2).

As well as identifying individual CAZyme genes present in MAGs, we can also use CAZyme data to identify which forms of carbohydrate are likely to be able to be digested by these strains by reconstructing metabolic pathways. Our MAGs demonstrated the capacity to degrade various carbohydrates (Fig. 8, Additional file 3: Fig. S6 and Additional file 4: Table S5). The most commonly encoded carbohydrate degradation pathway was for chitin, a polysaccharide commonly found in fungi

and arthropods (6828 of 9977 strain-level MAGs), which reflects the findings of the distilled and refined annotation of metabolism (DRAM) developers (48). This is closely followed by arabinose cleavage, present in 6648 MAGs.

Pathways to digest common fibrous plant compounds that are indigestible by chickens were present in many of the MAGs. Pathways for digestion of hemi-celluloses were very common, with mixed linkage glucan degradation encoded by 4689 MAGs, xyloglucans by 4420,

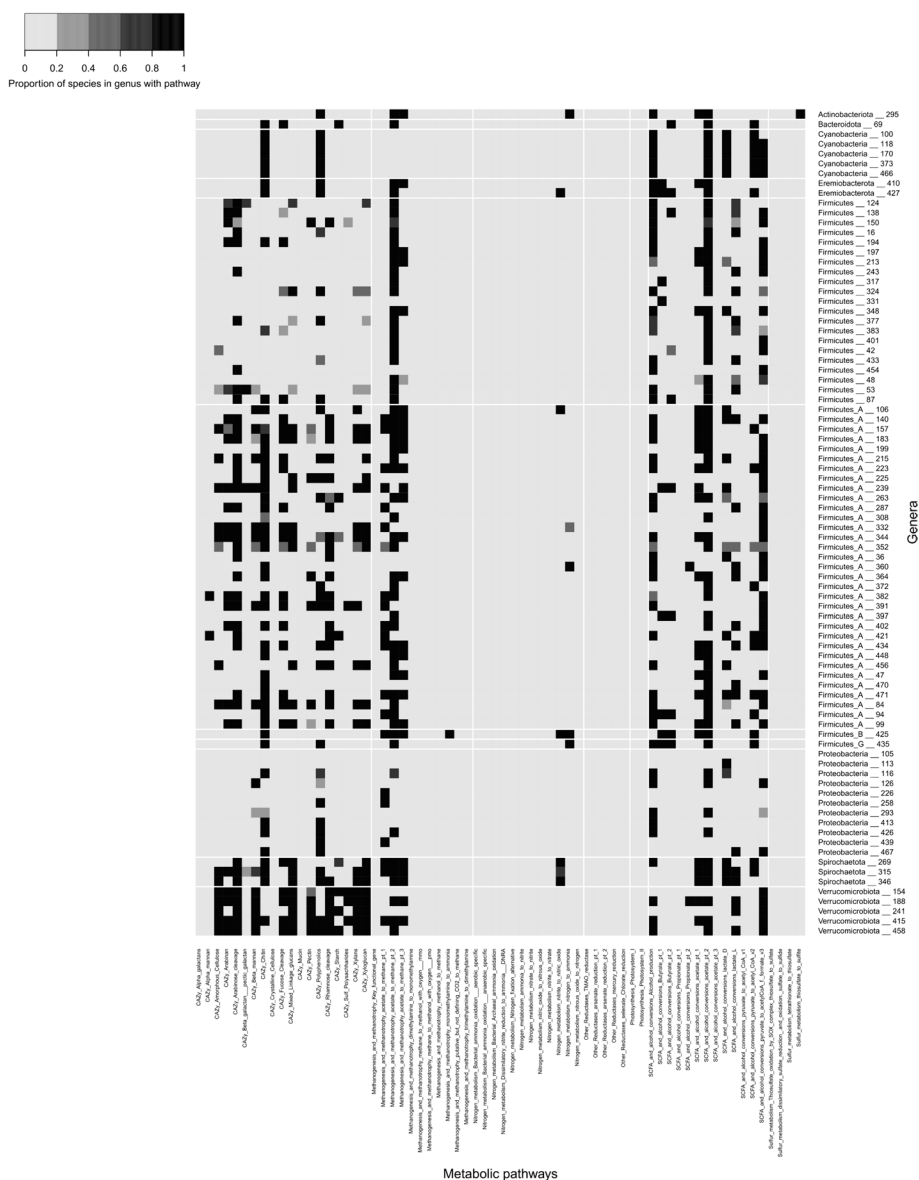


Fig. 8 Heatmap showing the percentage of species-level MAGs within each “unique” genera with particular metabolic pathways. Genera were clustered at 40% AAI using the output from CompareM. Genera were classified as “unique” if no MAGs within that genus cluster were assigned a taxonomy at genus-level by GTDB and if no NSC genomes clustered at >60% AAI with MAGs within that cluster

xylans by 4151, beta-mannan by 3056 and alpha-mannan by 621. As expected from the diversity of CAZymes, members of the Bacteroidota were found to be the most likely to digest hemi-cellulose. In comparison to hemi-cellulose, the capacity to degrade amorphous cellulose ($n=3567$) was less common. In general, those phyla that contained more strains that could degrade amorphous cellulose were also more likely to contain strains that degraded hemi-cellulose and pectin. As well as plant/diet-derived carbohydrates, microbes in the caeca have access to host-derived carbohydrates such as mucin. Only 1.3% of strains showed the potential to degrade mucin.

Pathways for nitrogen metabolism were far less abundant amongst the MAGs than those for fibre degradation. MAGs from phyla that frequently harboured nitrogen metabolism pathways were rarely able to degrade plant carbohydrates. For example, the majority of the members of the Desulfobacterota, Deferribacterota and Campylobacterota phyla harbour the dissimilatory nitrite reduction to ammonia (DNRA) pathway and/or were able to convert nitrate into nitrite, but less than 1% of these MAGs show any cellulose/hemi-cellulose degrading capacity. In contrast, members of the Spirochaetota were commonly able to metabolise nitrite to nitric oxide while also showing the capacity to degrade plant-derived carbohydrates.

While chickens produce little methane in comparison to other livestock, such as ruminants, members of their gut microbiota can carry out methanogenesis. Of the MAGs, 132 contained the key functional methanogenesis gene (methyl-coenzyme M reductase: *mcr*). Of those MAGs with the *mcr* gene, 22 were identified as having genes for all eight steps required for methanogenesis, with a further 25 having at least four of the required steps. Interestingly, none of the genomes from the NSC dataset had genes for all eight steps required for methanogenesis, and only five genomes had >50% of the required genes.

Gut microbes produce short-chain fatty acids, principally butyrate, acetate, and propionate, by the fermentation of indigestible polysaccharides. These SCFAs can then be used as an energy source by the host animal. While it is difficult to be certain using metagenomic data whether a particular strain produces SCFAs, we can predict the potential for SCFA production using DRAM. The potential to produce SCFAs was widely encoded across taxonomies (Additional file 4: Table S5). We visualised which SCFA/lactate encoding potentials occurred most together within the MAGs (Fig. 9). By far, the most common was the sole production of acetate, followed by the coproduction of acetate and lactate, then the coproduction of acetate and butyrate.

Discussion

In this study, we examined the caecal microbiota of 240 indigenous Ethiopian scavenging chickens originating from farms exposed to diverse climatic and geographic conditions. We constructed a gene catalogue containing 33 million genes and 9977 high-quality, strain-level MAGs originating from diverse taxonomies and with diverse functional capacities. We found that Ethiopian chicken caecal microbiota clustered into three distinct enterotypes, with one of these enterotypes being characterised by a high abundance of *Prevotella* and being particularly abundant in chickens living at high altitudes.

The chicken caecal microbiota is commonly dominated by bacteria, with low proportions of archaea and eukaryotes [19, 20], which was reflected in our findings. Bacteroidota was the most abundant bacterial phylum in our samples, followed by Firmicutes, Proteobacteria and Spirochaetota. This reflects what has previously been found in scavenging/feral birds and adult hens, but contrasts with taxa commonly found in intensively raised commercial broilers and young birds raised in biosecure poultry facilities. Commercial, intensively raised chicken breeds are reared without contact with older chickens in highly biosecure poultry facilities. They are therefore exposed to a lower diversity of microbial species than if they were raised by a maternal hen or exposed to an outdoor environment. As such, these birds usually have a low-diversity gut microbiota dominated by Firmicutes until around 2 months of age, before developing a microbiota composition more similar to our findings by 30–50 weeks of age [21, 22]. Commercial broilers are commonly slaughtered at 5–6 weeks of age and therefore do not have time to develop a “mature” microbiota. This is frequently also true for chickens that are part of microbiota studies, which often include only young birds. As such, the microbiota of these birds usually does not develop beyond a Firmicute-dominated composition [14, 23, 24]. This is in contrast to chicks raised by a hen or exposed to adult caecal/faecal contents, which have caecal microbiota more similar to our findings within a few days of hatch [25].

At an inter-country level, geography has been demonstrated to significantly impact the chicken caecal microbiota [6]. We found significant differences in the alpha-diversity of the caecal microbiota between climate zones, with bacterial diversity generally decreasing as altitude increased. Our samples were also found to cluster into three enterotypes. Enterotypes are generally defined as a stratification of the gut microbiota based on similarities in terms of taxonomic compositions between samples [26]. Interestingly, one of our three enterotypes, characterised by a high prevalence of *Prevotella*, was

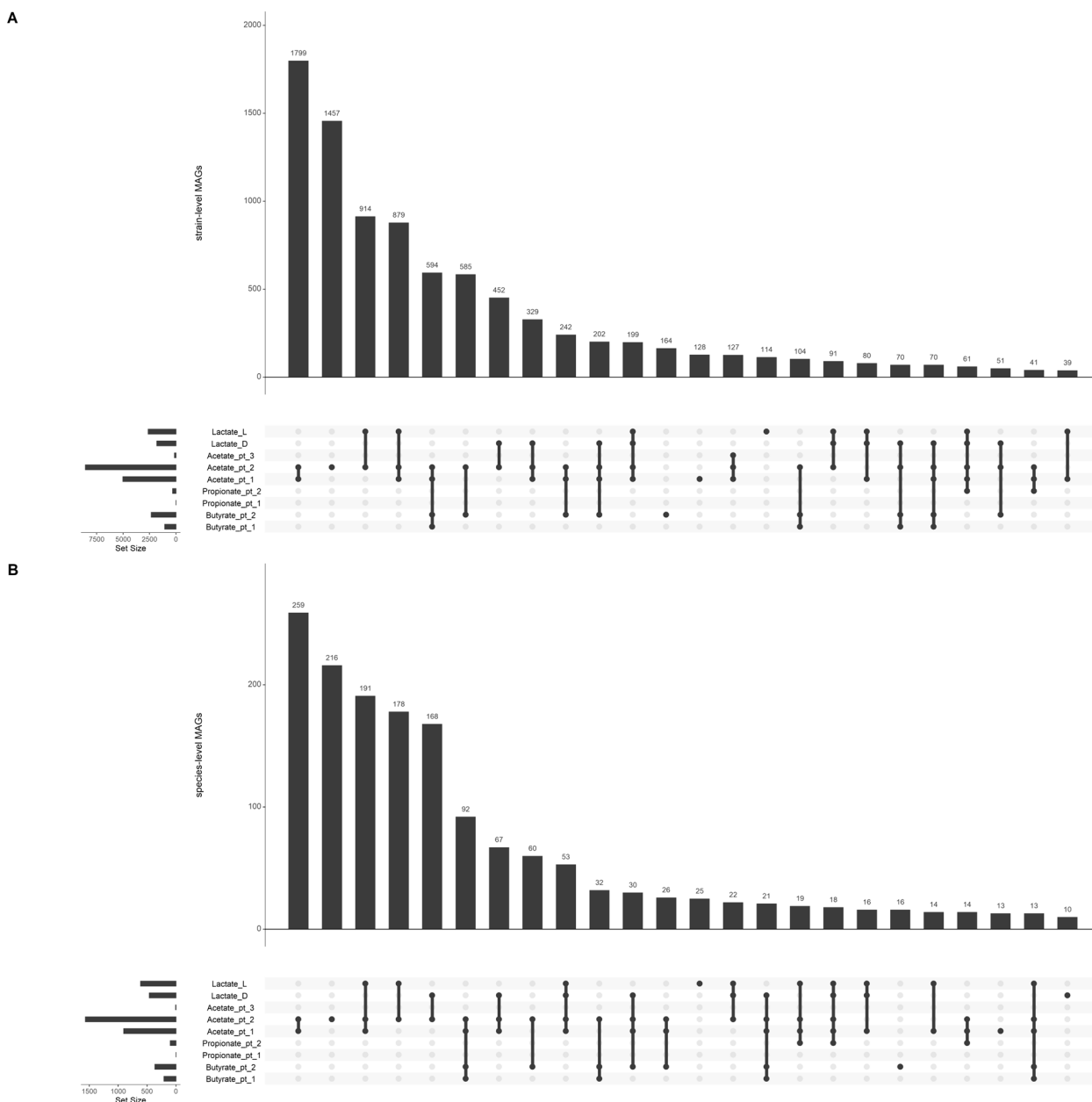


Fig. 9 UpSet plots showing the number of MAGs with production potential for SCFA and lactate production, as defined by DRAM. Only includes intersections with ≥ 10 MAGs. **A** Strain-level MAGs. **B** Species-level MAGs

predominantly associated with the highest altitude (climate zone 1).

Decreases in gut microbiota diversity have also been observed in a study of Tibetan chickens raised at different altitudes [8]. Tibetan chicken from high altitudes also had a greater abundance of Prevotellaceae in their caeca, in agreement with our findings. This increase in *Prevotella* at high altitudes has also been noted in house mice [27], and cranes adapted to high altitudes have lower gut

microbiota diversity than cranes that normally reside at low altitudes [28].

Gut microbiota samples from pigs and humans living at high altitudes were also found to be significantly lower in diversity than those from low altitudes [29]. In contrast, the diversity of the gut microbiota of rhesus macaques was increased in high altitudes, and members of the Prevotellaceae were found in lower abundance [30]. The microbiota of humans living in high-altitude areas of

China have also been associated with increased microbiota diversity [31]; however, as in our samples, *Prevotella* abundance was found to increase at high altitudes.

In chickens, while the reasons for this low diversity microbiota at high altitudes are currently unknown, it may be related to the effects of hypoxia [32], decreased temperature [33], decreased humidity, available crops (e.g., wheat and barley in high elevations and maize and finger millet at moderate elevations) or the relative lack of diversity in feed present at these elevations. Our study also found correlations between the gut microbiota composition and factors such as dietary components and soil characteristics. Further studies in these areas, alongside investigations of temporal dynamics within climate zones, would contribute to an even greater understanding of the factors influencing the gut microbiota in smallholder systems of different altitudes.

We compared our MAGs to a dataset made from microbial genomes previously constructed from the chicken gastrointestinal tract [13, 16, 34, 35] (NSC dataset) and to the Genome Taxonomy Database [36]. The number and diversity of taxa identified in our dataset but not in the GTDB or NSC dataset emphasise the importance of studying the microbiota of indigenous livestock as well as commercial breeds. Our results also demonstrate the need to include microbial genomes isolated from indigenous livestock in commonly used genetic databases such as GTDB, as many of our MAGs were unable to be identified by GTDB-tk at the level of taxonomic order ($n=39$), family ($n=204$) and genus ($n=2252$).

Despite the taxonomic novelty observed in our dataset, the functional novelty was less apparent. While there were significant differences between the types of CAZymes found in the NSC dataset vs our MAGs, the vast majority of CAZymes were found in both. This correlates with previous studies that found that differences in taxonomy tend to be greater than changes in the function of the microbiota [37]. Particularly high numbers of CAZymes were found in the phyla Bacteroidota and Verrucomicrobiota, these phyla having been previously shown to have a high percentage of their total proteins being identified as CAZymes in comparison to other taxa [17, 38]. Several CAZymes were found in our MAGs, but not in the NSC dataset. These included several enzymes linked to the breakdown of algae: PL25 (ulvan lyase: 3 *Sphaerochaetaceae* MAGs), PL32 (poly(β -mannuronate) lyase/M-specific alginate lyase: 2 *Paludibacteraceae* MAGs), GH86 (β -agarase/ β -porphyranase: 8 *UBA1829* MAGs) and GH107 (endo- α -1,4-L-fucanase: 1 *UBA3636* MAG) [39]. It is possible that this is due to the chickens consuming algae from local stagnant water sources.

As well as examining specific carbohydrate-degrading enzymes, we also characterised the metabolic pathways in our MAGs. The caeca is the main site of fibre fermentation in the chicken gastrointestinal tract. Both our MAGs and genomes from the NSC dataset contained pathways for the fermentation of a wide variety of fibrous compounds, including various forms of hemi-cellulose and amorphous cellulose. However, the ability to degrade crystalline cellulose was rare, being encoded by only 6 MAGs. The result of fibre fermentation is the production of SCFAs, which can be used by the host as an energy source and can also play a role in pathogen resistance [10, 11].

The SCFAs that had the most potential for production amongst our MAGs were acetate, then butyrate, then propionate, similar to the human gut [40] and the NSC dataset. Several MAGs also showed the capacity to degrade mucin. Mucin degradation by gut microbiota can be related to a lack of dietary fibre, leading bacteria to rely on host-derived glycans [41]. However, some mucin-degrading species, such as *Akkermansia muciniphila* in humans, are often also common members of healthy microbiota [42]. The caecal microbiota also plays an important role in the nitrogen nutrition of the chicken, particularly when the bird is consuming little protein [43], due to the reflux of urine into the caeca [11]. The capacity for nitrogen metabolism was present amongst our MAGs but far less so than plant fibre fermentation.

We have identified a wide functional and taxonomic diversity of microbes originating in the caecal contents of Ethiopian scavenging chickens under smallholder settings. The vast majority of these microbes were not identified in either the GTDB or a dataset of publicly available microbial genomes isolated from predominantly non-scavenging chickens. We detected differences in the alpha and beta diversity of the chicken caecal microbiota relating to various climate and geographical factors. These findings highlight the potential hidden microbial diversity amongst indigenous scavenging chickens that may be missed when examining only commercially reared animals. They also highlight the potential limitations of extrapolating results from microbiota studies in commercial animals to smallholder settings and the urgent need for more microbiota research in these smallholder systems.

Methods

Sample collection

Two hundred forty-three indigenous chickens were included in this study. They originated from 26 sites across 15 districts of Ethiopia that represent diverse agroclimatic conditions (Fig. 1A, B, Additional file 4: Table S6 and S7). Samples were collected from smallholder farms within 3 km² of the sampling site centre (within 0.03°

of longitude or latitude). In order to avoid bias associated with households, each bird collected from a single site belonged to a different smallholder. Whole genome sequences had previously been produced for the majority of chickens in our study except for samples from Arginjona and Dehina_Maria that were mapped following previously described methods to the chicken reference genome GRRcG6a [44]. Principle coordinate analysis was conducted using PLINK v1.9 [45] to show the clustering of chickens by autosomal single-nucleotide polymorphisms (SNPs). SNPs were pruned for LD in PLINK (using PLINK option `-indep-pairwise 50 10 0.1`), then principle component analysis (PCA) plots were constructed in R. The chickens in our study did not belong to specific breeds; indigenous Ethiopian chickens do not constitute distinct breeds, instead weak sub-structuring of populations is observed based on geographic closeness [1] (Additional file 3: Fig. S7).

Climate predictors were identified from the DIVA-GIS database based on a single geographic coordinate taken from the sampling site by GPS, as previously described [1]. Chickens were raised in a typical low-input system, acquiring most of their food from scavenging and occasional supplementary feeding. Data was collected on rearing conditions, supplementary feeding and local climate (Additional file 4: Table S7). Most chickens lived in simple poultry houses, often constructed with a stone wall and grass roof, and with limited cleaning frequency (Fig. 1A). All farmers practised supplementary feeding, which included items such as kitchen waste, grains or vegetables (for more details on the feeding conditions of Ethiopian smallholder chickens, see Additional file 1). Chickens acquired most of their nutrition through scavenging on food sources located near the household, including vegetation, insects, worms, wasted grains and animal faeces. Chickens frequently co-habited with other domestic animal species and are also likely to have come into contact with wild animals.

Clustering of climate zones

Climate predictions were obtained from WorldClim 2.0 Beta version 1 (June 2016). Clustering of climate zones was produced in R using the K-means cluster method based on annual temperature, annual precipitation and precipitation of the driest quarter of the sampling locations, during the years 1970 to 2000 (Fig. 1C and Additional file 3: Fig. S8). This resulted in the clustering of sampling areas into five climate zones. A similar method for climate regionalization has previously been used by Yang et al. [46].

DNA extraction and shotgun metagenomic sequencing

Caecal contents were collected at the farm from freshly slaughtered scavenging chickens, and stored in RNAlater solution (Ambio). Samples were kept on ice for a maximum of 24 h before being stored at -80°C prior to DNA extraction. DNA was extracted using the QIAamp Fast DNA Stool Mini Kit (Qiagen) following the manufacturer's instructions, with some adjustments as described previously [47], including the addition of a bead beating step and increasing the cell lysis temperature from 70 to 95°C to increase the likelihood of lysing Gram-positive bacteria. DNA sequencing libraries were prepared using the Nextera XT DNA Library Prep Kit (Illumina) and sequenced with an Illumina Novaseq (2×150 bp) (Berry Genomics Co.). All 243 samples were sequenced in a single run, yielding up to 10 Gb per sample. Adapter trimming and quality filtering were performed using Fastp [48] (v.0.1.24). Host reads were removed by mapping the *Gallus gallus* genome (GRCg6a) to the trimmed sample fastq files using BWA-MEM (v.0.7.15) [49], followed by SAMtools (v.1.3.1) [50] to select reads where both paired-end reads were unmapped. Three samples contained high host contamination and were thereby not taken forward for further metagenomic analysis (except for the construction of MAGs), leaving a total of 240 samples.

Gene catalogue construction and analysis

After quality control and host removal, a gene catalogue was constructed from non-redundant genes. Firstly, single-sample assembly was conducted using MEGAHIT (v.1.2.9) (contig length >500 bp) [51]. Then, MetaGeneMark-1 [52] (GeneMark.hmm v.3.38: `gmhmm -a -d -f G -m; MetaGeneMark_v1.mod`; gene length >200 nt) was used to predict open reading frames (ORFs). Non-redundant genes were clustered using CD-HIT [53] (v.4.6.6) at 95% identity over 90% of the shorter ORF length (`-c 0.95 -aS 0.9 -M 0 -T 0`), resulting in 33,629,587 genes. As many as 97.9% (95.2~98.5%) of the sequencing reads could be included in the non-redundant gene catalogue. Dual-BLAST least common ancestor strategies were used for microbiome taxonomic annotation. Diamond [54] (v.2.0.2; `diamond blastp -c 1 -k 5 -f 6`) was used to search for homologous genes in the Uniprot database (version 2019_03) using the protein sequences in the non-redundant gene catalogues. Then, the aligned Uniprot regions (E -value $<10^{-5}$) were used in a second alignment against the Uniprot database, resulting in the identification of homolog neighbourhoods of the initial query genes by reporting e -values that were equal to or less than the e -value from the first alignment. Query genes were then assigned the taxonomy of the least common ancestor of the neighbourhood. Sequencing reads were directly mapped to the non-redundant gene

catalogues by BWA-MEM (v.0.7.12—default options) [49]. The relative abundance of genes in the samples was calculated based on the count of aligned reads, normalized to the gene length and sum of the abundance of all genes. Based on the taxonomic annotation of the gene catalogue, aligned reads were used to quantify taxonomic abundance profiles within samples. Relative taxonomic abundance was normalized to sequencing depth and the length of the genes originating from the same level of taxonomic classification.

Enterotype clustering

Gut enterotype analysis was performed by multidimensional cluster analysis and PCA on genus abundance, according to previously described methods [55], using Jensen–Shannon divergence. All genera from the gene catalogue were used to define enterotypes in order to include taxa from all kingdoms. LEfSe was used to determine the microbiota features that most likely explained differences between enterotypes [56]. Spearman correlations were performed between the main genera contributing to enterotype clustering and all other genera. Genus networks were visualized using the Cytoscape [57] platform by transforming Spearman correlations ($p < 0.01$, $\rho > 0.5$) into links. Mantel tests (Spearman correlation) were performed between our previously identified non-redundant environmental factors and genera and phyla. Spearman pairwise correlations between continuous metadata variables were calculated (SparCC) [58]. *P* values were corrected for multiple tests using Benjamini–Hochberg false discovery rate (FDR) correction. Significant features were used as input for building linear models using stepwise regression based on the Akaike information criterion [59]. The Shannon diversity index and inverse Simpsons index were used to compare alpha-diversity between groups (Kruskal–Wallis). Bray–Curtis dissimilarity values were calculated to assess beta-diversity, and PERMANOVA was used to compare groups.

Metagenome-assembled genome assembly

Metagenomic bins were constructed using two different methods:

Method 1: Coassemblies and single-sample assemblies were performed using MEGAHIT (v.1.1.3) with a minimum contig length of 500 bp. Using MEGAHIT, five separate co-assemblies were performed on samples from each climate zone. BWA MEM was used to map reads from each sample to the assembly from the same sample. MetaBAT2 (option -m 1500) was applied for contig binning.

Method 2: All 243 samples were used for sequence assemblies. IDBA-UD (v.1.1.3) [60] was used for sin-

gle sample assembly (options: `-num_threads 16 -pre_correction -min_contig 300`). BWA MEM was used to map reads from each sample to the assembly from the same sample. SAMtools was used to create BAM files, and coverage for each assembly was calculated by running the command `jgi_summarize_bam_contig_depths` on these files. MEGAHIT was used for coassembly of all samples (v.1.1.1) (options: `-kmin-1pass -m 100e+10 -k-list 27,37,47,57,67,77,87 -min-contig-len 1000`), in six randomised batches of samples. Contigs were filtered to a minimum length of 2 kb and mapped as single assemblies. MetaBAT2 (v.2.11.1) [61] was used to construct metagenomic bins for both single sample assemblies and coassemblies (options: `-minContigLength 2000, -minContigDepth 2`).

The completeness and contamination of bins were assessed using CheckM [62] (v.1.1.3; options: `lineage_wf -t 30 -x fa -nt -tab_table`). Bins with completeness $\geq 80\%$ and contamination $\leq 5\%$ were concatenated and used as an input for DAS Tool [63] (v1.1.2; option: `-search_engine diamond -c merge.contigs.fa -threads 12 -write_bins 1`), as an additional quality control step. The bins output by the DAS tool were dereplicated using drep [64] at 99% average nucleotide identity (ANI), which is equivalent to a microbial strain, and at 95% ANI, equivalent to a microbial species. Dereplication is the process of reducing a set of genomes based on their sequence similarity. Mapping rates for our strain and species-level MAGs can be found in Additional file 4: Table S8. GTDB-Tk [65] (v.0.3.2, Database release 95) was used to assign taxonomy to MAGs. The taxonomic names of phyla used throughout the text of this manuscript are based on those used in this version of GTDB-Tk. Phylogenies were constructed using PhyloPhlan 3.0 [66] (v.3.0.60; options: `-d phylophlan -min_num_markers 60 -subsample phylophlan -f tol_config.cfg -diversity high -fast -genome_extension fa -nproc 2`). Phylogenetic trees were visualised using Graphlan [67] (v.1.1.3.1) and iTOL [68] (Interactive Tree Of Life; v.5). The Table2itol package (<https://github.com/mgoeker/table2itol>) was used for phylogenetic tree annotations.

MAG genes were identified using Prodigal [69] (v.2.6.3). CD-HIT [53] (v.4.6.8) was used to cluster all MAG proteins at 100% similarity and 90% similarity. The relative abundance of MAGs in the samples was estimated using the `quant_bins` module of MetaWRAP (v.1.3) with default parameters [70]. Kruskal–Wallis with Benjamini–Hochberg FDR correction was used to identify species-level MAGs that were differentially abundant between enterotypes.

Comparing MAGs to previous chicken microbiota datasets

We constructed a dataset of chicken-derived microbial genomes from previously published chicken microbiota studies (Additional file 4: Table S9): including 5595 MAGs and 41 genomes of representative cultured isolates of novel species from Gilroy et al. [13], 469 MAGs from Glendinning et al. [16], 133 genomes of cultured isolates from Medvecky et al. [34], and 16 genomes of cultured isolates from Zenner et al. [35]. These genomes were dereplicated using dRep (v.3.2.2) (options: -comp 80 -con 10 -str 100 -strW 0) at 99% ANI (strain-level) and 95% ANI (species-level) to produce two datasets of dereplicated, high-quality genomes. The vast majority of samples from which these genomes were isolated originated from non-scavenging chicken (except 22 (prior to dereplication) sampled from NCBI Bioproject PRJNA616250). We therefore labelled these datasets NSC (non-scavenging chickens) and compared our MAGs to these datasets to assess whether they were taxonomically and functionally distinct.

To calculate whether our MAGs were taxonomically distinct at strain and species-level, dRep was used on our MAGs and the NSC datasets at 99% and 95% ANI. Any MAG which did not cluster at 99% ANI with any genome from the NSC dataset was classed as distinct at strain level. Any MAG which did not cluster at 95% ANI with any genome from the NSC dataset was classed as distinct at species-level. Species-level genomes from both our dataset and the NSC dataset were compared using CompareM [71] (v.0.1.2) `aai_wf` to generate average AAI. Genera were clustered at >60% AAI; genera were defined as distinct if they contained no NSC genomes. Our MAGs were also compared to the GTDB [65]. MAGs were defined as distinct strains in comparison to genomes in the GTDB if the ANI output by GTDB-Tk was <99%, and distinct species if the ANI output by GTDB-tk was <95%. Genera clusters (>60% AAI) were defined as distinct from those in the GTDB if no MAG within that cluster was assigned a genus by GTDB-Tk.

Functional annotation of MAGs

Our MAGs and genomes from the NSC dataset were annotated in order to understand the potential metabolic function of these microbes. Genomes were annotated using DRAM [40] (v.1.2.2) with the “annotate” command; these annotations were then curated and summarised using the “distill” command. DRAM is a tool for annotating MAGs, using various databases, including UniRef90 [72], PFAM [73], dbCAN [74], RefSeq viral [75], VOGDB (<http://vogdb.org/>) and the MEROPS peptidase database [76], and a user-supplied version of the KEGG [77] database (downloaded Sep 15, 2018). This tool provides

information on the overall metabolic pathways encoded by the genomes, as well as specific information on the presence of CAZymes, rRNA genes and tRNA genes. Permutational multivariate analysis of variances (PERMANOVAs) was conducted using the `adonis` command from Vegan [78] (v.2.5.7), to compare the types of CAZymes present between groups. The Kruskal–Wallis test was used to compare the abundance of CAZyme genes between groups. Microbial growth rates were predicted using gRodon [79] (v.0.0.0.9). This tool estimates maximal microbial growth rates by comparing codon usage patterns in highly expressed genes versus other genes. We first used `prokka` [80] (v.1.14.6) (options: -centre X -compliant) to identify genes, including highly expressed genes (genes annotated as ribosomal proteins). Predicted genes and highly expressed genes were used as input for the gRodon “predictGrowth” command, which was run in partial mode to account for incomplete genomes. As suggested by the gRodon creators [79], copiotrophs were defined as having a <5 h doubling time, whereas oligotrophs were defined as having a >5 h doubling time. AMR genes were identified in species-level MAGs using the resistance gene identifier (v. 5.1.1), and comprehensive antibiotic resistance database reference sequences downloaded on March 30, 2021 [81].

Selection of environmental variables

The relationships between environmental variables and the abundance of microbial taxa were estimated. In order to reduce model complexity by limiting the amount of environmental variables included in our analyses, variance inflation factors (VIF) were calculated for a set of environmental variables (Additional file 4: Table S10), to identify collinearity among explanatory variables (VIF value less than 20). The highest contributing set of uncorrelated environmental variables was identified. This led to the selection of 24 variables that were included in our analyses: eight climate/geographical variables (altitude, bio2, bio3, bio13, bio14, bio15, bio18, bio19), ten soil/land cover factors (CULT, FOR, SNDPPT, SLTPPT, CRFVOL, BLDFIE, CECSOL, ORCDRC, PHIHOX, WATCAP) and the dominance of five feeding crops (Wheat, Maize, Barley, Millet, Teff), and Ingera (a food derivative from Teff). Redundancy analysis (RDA; Vegan 2.6.2 package) was used to identify environmental variables that contributed to microbiota variation.

Graphical analyses

Graphs were created in R. Plots were constructed using the packages `ggplot2` [82], `Cowplot` [83], `UpsetR` [84], `cluster`, `clusterSim`, `ggpurb` and `corrplot`. Heatmaps were constructed using `heatmap.2` from `gplots` [85] (v. 3.0.1).

Abbreviations

ANOVA	Analysis of variance
AMR	Antimicrobial resistance genes
AA	Auxiliary activities
AAI	Average amino acid identities
ANI	Average nucleotide identity
CAZymes	Carbohydrate active enzymes
CBM	Carbohydrate-binding modules
CE	Carbohydrate esterases
DNRA	Dissimilatory nitrite reduction to ammonia
dt	Doubling time
FDR	False discovery rate
GTDB	Genome Taxonomy Database
GPS	Global positioning system
GH	Glycoside hydrolases
GT	Glycosyltransferases
IACUC	Institutional Animal Care and Use Committee
ILRI	International Livestock Research Institute
LEfSe	Linear discriminant analysis Effect Size
LD	Linkage disequilibrium
MAGs	Metagenome assembled genomes
NSC	Non-scavenging chickens
ORF	Open reading frame
PERMANOVA	Permutational multivariate analysis of variance
PCA	Principal component analysis
PL	Polysaccharide lyases
RDA	Redundancy analysis
SCFAs	Short-chain fatty acids
SNPs	Single-nucleotide polymorphisms
VIF	Variance inflation factor

Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s40168-024-01847-4>.

Additional file 1. Feed of indigenous Ethiopian chickens.

Additional file 2: Supplementary Results 1. Detection of slow-growing microbes: Fig. S1. Boxplot showing the predicted doubling time of strain-level MAGs, by phylum. The red, dashed line indicates the cut-off point above which MAGs are considered oligotrophs. Fig. S2. Growth rate distributions for strain-level genomes from MAGs produced in this study (Current_MAGs: 9977) and dereplicated genomes from the NSC dataset, which originated from cultured isolates (NSC_cultured: 170) or were MAGs (NSC_MAGs: 2343). The red, dashed line indicates the cut-off point above which MAGs are considered oligotrophs. Relating CAZyme abundance to oligotrophy: Fig. S3. Total CAZyme families count vs maximal growth rate. A) Unique CAZyme families, B) Total CAZyme genes. The red, dashed line indicates the cut-off point above which MAGs are considered oligotrophs. Copiotrophs were found to have significantly more CAZyme genes and unique CAZyme families than oligotrophs (Kruskal-Wallis: $P < 2.2 \times 10^{-16}$).

Additional file 3: Fig. S1. Construction of a chicken caecal reference gene catalogue. The 3,629,587 non-redundant genes contained in the catalogue represent the metagenomes of 240 chicken caecal contents samples. Non-redundant genes were assigned to different taxon levels based on their last common ancestor in the Uniprot database (version 2019_03). Fig. S2. Description of the gene catalogue constructed from the caecal microbiota of Ethiopian indigenous chickens. A) Rarefaction analysis of the number of non-redundant genes vs sampling number. B-E) Break down of the taxa identified in the Ethiopian chicken caecal microbial gene catalogue, by Kingdom. Fig. S3. Genome statistics of high-quality, non-redundant strain-level (A-E) and species-level (F-J) metagenome-assembled genomes, as defined by CheckM. A and F: Completeness and contamination – dashed red lines indicate cutoffs for defining genomes as high-quality. B and G: Percentage GC content. C and H: log₁₀ number of contigs per genome. D and I: Genome size (mb). E and J: log₁₀ N50 of contigs. Fig. S4. The proportions of annotated read after mapping raw sequencing reads to the non-redundant gene catalogue (A) and MAGs

(B). Fig. S5. Violin plot showing the number of CAZyme genes per strain-level MAG by dataset. A) Total unique CAZyme families. B) Total unique Glycoside Hydrolases (GH) families. C) Total CAZyme genes. D) Total GH genes. Fig. S6. Heatmap showing the percentage of species-level MAGs within genera with particular metabolic pathways. Genera were clustered at 40% AAI using the output from comparem. The uniqueness of genera in comparison to previous datasets is indicated. Genus-level clusters were not unique based on GTDB if any MAGs within that cluster were assigned a taxonomy at the genus level. MAGs were defined as not unique when compared to previous chicken microbial datasets ("not_unique_drep") if they clustered at 99% (strain) or 95% (species) ANI with any non-scavenging chickens (NSC) microbial genome. Genera were defined as not unique when compared to previous chicken microbial datasets (not_unique_comparem) if they clustered at 60% AAI with any NSC microbial genome. Fig. S7. Principle coordinate analysis showing the clustering of samples by autosomal SNPs. Samples are labelled by the region in which the sample was collected. Fig. S8. Five climate zones, clustered using Kmeans according to annual temperature, annual precipitation and precipitation of the driest quarter of the sampling location between 1970 - 2000. Components 1 and 2 explain 85.1% of sampling site variation. Table S1: Diversity of CAZymes in strain level MAGs.

Additional file 4: Table S1. MAG statistics and metadata. Table S2. Taxonomic assignment of strain and special level MAGs. Table S3. Differentially abundant MAGs between enterotypes. Table S4. Comparison of MAGs and genera to the NSC dataset. Table S5. MAG metabolic potential. Table S6. Sampling site characteristics. Table S7. Sample metadata. Table S8. Mapping rates for strain and species level MAGs. Table S9. Components of the NSC dataset. Table S10. Environmental variables included in our analysis.

Acknowledgements

We thank ILRI staff Michael Tesmegen for field support and Mick Watson for advice on the analysis of our data.

Authors' contributions

LG and XJ contributed equally to the data analysis and writing of the paper. OH, JH and XJ designed the study. AK, JH, OH, JBH, JEP, and WP organised and/or collected the samples in the field. SO provided oversight and performed the DNA extraction of the samples. JBH, KK, HJ, and OH contributed to the writing of the paper. AA contributed to the data analysis. All authors read and approved the final manuscript.

Funding

The ILRI livestock genomics program is supported by the CGIAR Research Program on Livestock (CRP Livestock), which is supported by contributors to the CGIAR Trust Fund (<http://www.cgiar.org/about-us/our-funders/>). This research was funded in part by the Bill & Melinda Gates Foundation and with UK aid from the UK Foreign, Commonwealth and Development Office (Grant Agreement OPP1127286) under the auspices of the Centre for Tropical Livestock Genetics and Health (CTLGH), established jointly by the University of Edinburgh, SRUC (Scotland's Rural College) and the International Livestock Research Institute. This research work was also supported from the Cooperative Research Program for Agriculture Science and Technology Development under Africa chicken microbiome project (Project No. PJ0127562018, PJ0145202021), Rural Development Administration (RDA), Republic of Korea and International Livestock Research Institute (ILRI), Nairobi, Kenya. The findings and conclusions contained within are those of the authors and do not necessarily reflect positions or policies of the Bill & Melinda Gates Foundation or the UK Government. The microbiota genome sequencing was supported by The Chinese Government contribution to the CAAS-ILRI Joint Laboratory on Livestock and Forage Genetic Resources in Beijing (2018-GJHZ-01). The Roslin Institute forms part of the Royal (Dick) School of Veterinary Studies, University of Edinburgh. This project was supported by the Biotechnology and Biological Sciences Research Council, including institute strategic programme and national capability awards to The Roslin Institute (BBSRC: BB/P013759/1, BB/P013732/1, BB/J004235/1, BB/J004243/1). For the purpose of open access, the author has applied a Creative Commons Attribution (CC BY) licence to any Author Accepted Manuscript version arising from this submission.

Availability of data and materials

The paired-read fastq files and species-level MAG fasta files generated and analysed during the current study are available in the European Nucleotide Archive under project PRJEB57055. Strain-level MAG fasta files (<https://doi.org/10.6084/m9.figshare.22140284>), and the gene catalogue (<https://doi.org/10.6084/m9.figshare.23641128.v1>) are available through figshare. Metabolic data on species-level (<https://doi.org/10.6084/m9.figshare.22154627>) and strain-level (<https://doi.org/10.6084/m9.figshare.22154597>) MAGs are available through figshare.

Declarations

Ethics approval and consent to participate

This study was reviewed and approved by the Institutional Animal Care and Use Committee (IACUC), International Livestock Research Institute (ILRI) (Reference number: ILRI-IREC2015-08/1).

Consent for publication

NA.

Competing interests

The authors declare no competing interests.

Received: 4 July 2023 Accepted: 28 May 2024

Published online: 23 July 2024

References

- Gheyas AA, Vallejo-Trujillo A, Kebede A, Lozano-Jaramillo M, Dessie T, Smith J, et al. Integrated environmental and genomic analysis reveals the drivers of local adaptation in African indigenous chickens. *Mol Biol Evol.* 2021;38(10):4268–85. <https://doi.org/10.1093/molbev/msab156>.
- Bettridge JM, Lynch SE, Brena MC, Melese K, Dessie T, Terfa ZG, et al. Infection-interactions in Ethiopian village chickens. *Prev Vet Med.* 2014;117(2):358–66. <https://doi.org/10.1016/j.prevetmed.2014.07.002>.
- Thomas KM, de Glanville WA, Barker GC, Benschop J, Buza JJ, Cleaveland S, et al. Prevalence of *Campylobacter* and *Salmonella* in African food animals and meat: a systematic review and meta-analysis. *Int J Food Microbiol.* 2020;315: 108382. <https://doi.org/10.1016/j.jfoodmicro.2019.108382>.
- Asrat D, Hathaway A, Ekwall E. Studies on enteric campylobacteriosis in Tikur Anbessa and Ethio-Swedish Children's Hospital, Addis Ababa Ethiopia. *Ethiop Med J.* 1999;37(2):71–84.
- Alemneh T, Getabalew M. Exotic chicken production performance, status and challenges in Ethiopia. *Int J Vet Sci Res.* 2019;5(2):039–45.
- Pin Viso N, Redondo E, Diaz Carrasco JM, Redondo L, Garcia JSY, Fernández Miyakawa M, et al. Geography as non-genetic modulation factor of chicken caecal microbiota. *Plos One.* 2021;16(1):e0244724. <https://doi.org/10.1371/journal.pone.0244724>.
- Shi D, Bai L, Qu Q, Zhou S, Yang M, Guo S, et al. Impact of gut microbiota structure in heat-stressed broilers. *Poult Sci.* 2019;98(6):2405–13. <https://doi.org/10.3382/ps/pez026>.
- Du X, Li F, Kong F, Cui Z, Li D, Wang Y, et al. Altitude-adaptation of gut microbiota in Tibetan chicken. *Poult Sci.* 2022;101(9):101998. <https://doi.org/10.1016/j.psj.2022.101998>.
- Shang Y, Kumar S, Oakley B, Kim WK. Chicken gut microbiota: importance and detection technology. *Front Vet Sci.* 2018;5:254. <https://doi.org/10.3389/fvets.2018.00254>.
- Varmuzova K, Kubasova T, Davidova-Gerzova L, Sisak F, Havlickova H, Sebkova A, et al. Composition of gut microbiota influences resistance of newly hatched chickens to *Salmonella enteritidis* infection. *Front Microbiol.* 2016;7:957. <https://doi.org/10.3389/fmicb.2016.00957>.
- Svihus B, Choct M, Classen HL. Function and nutritional roles of the avian caeca: a review. *Poult Sci J.* 2013;69(2):249–64. <https://doi.org/10.1017/S0043933913000287>.
- Crhanova M, Karasova D, Juricova H, Matiasovicova J, Jahodarova E, Kubasova T, et al. Systematic culturomics shows that half of chicken caecal microbiota members can be grown *in vitro* except for two lineages of Clostridiales and a single lineage of Bacteroidetes. *Microorganisms.* 2019;7(11):496. <https://doi.org/10.3390/microorganisms7110496>.
- Gilroy R, Ravi A, Getino M, Pursley I, Horton DL, Alikhan N-F, et al. Extensive microbial diversity within the chicken gut microbiome revealed by metagenomics and culture. *PeerJ.* 2021;9:e10941. <https://doi.org/10.7717/peerj.10941>.
- Glendinning L, Watson KA, Watson M. Development of the duodenal, ileal, jejunal and caecal microbiota in chickens. *Anim Microbiome.* 2019;1(1):17. <https://doi.org/10.1186/s42523-019-0017-z>.
- Ferrario C, Alessandri G, Mancabelli L, Gering E, Mangifesta M, Milani C, et al. Untangling the caecal microbiota of feral chickens by culturomic and metagenomic analyses. *Environ Microbiol.* 2017;19(11):4771–83. <https://doi.org/10.1111/1462-2920.13943>.
- 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(1):34. <https://doi.org/10.1186/s13059-020-1947-1>.
- Feng Y, Wang Y, Zhu B, Gao GF, Guo Y, Hu Y. Metagenome-assembled genomes and gene catalog from the chicken gut microbiome aid in deciphering antibiotic resistomes. *Commun Biol.* 2021;4(1):1305. <https://doi.org/10.1038/s42003-021-02827-2>.
- Hamaker BR, Tuncil YE. A perspective on the complexity of dietary fiber structures and their potential effect on the gut microbiota. *J Mol Biol.* 2014;426(23):3838–50. <https://doi.org/10.1016/j.jmb.2014.07.028>.
- Robinson K, Yang Q, Stewart S, Whitmore MA, Zhang G. Biogeography, succession, and origin of the chicken intestinal mycobiome. *Microbiome.* 2022;10(1):55. <https://doi.org/10.1186/s40168-022-01252-9>.
- Saengkerdsud S, Anderson RC, Wilkinson HH, Kim W-K, Nisbet DJ, Ricke SC. Identification and quantification of methanogenic archaea in adult chicken caeca. *Appl Environ Microbiol.* 2007;73(1):353–6. <https://doi.org/10.1128/AEM.01931-06>.
- Joat N, Van TTH, Stanley D, Moore RJ, Chousalkar K. Temporal dynamics of gut microbiota in caged laying hens: a field observation from hatching to end of lay. *Appl Microbiol Biotechnol.* 2021;105(11):4719–30. <https://doi.org/10.1007/s00253-021-11333-8>.
- Videnska P, Sedlar K, Lukac M, Faldynova M, Gerzova L, Cejkova D, et al. Succession and replacement of bacterial populations in the caecum of egg laying hens over their whole life. *Plos One.* 2014;9(12):e115142. <https://doi.org/10.1371/journal.pone.0115142>.
- Richards P, Fothergill J, Bernardeau M, Wigley P. Development of the Caecal microbiota in three broiler breeds. *Front Vet Sci.* 2019;6:201. <https://doi.org/10.3389/fvets.2019.00201>.
- Di Marcantonio L, Marotta F, Vulpiani MP, Sonntag Q, Iannetti L, Janowicz A, et al. Investigating the caecal microbiota in broiler poultry farms and its potential relationships with animal welfare. *Res Vet Sci.* 2022;144:115–25. <https://doi.org/10.1016/j.rvsc.2022.01.020>.
- Kubasova T, Kollarcikova M, Crhanova M, Karasova D, Cejkova D, Sebkova A, et al. Contact with adult hen affects development of caecal microbiota in newly hatched chicks. *Plos One.* 2019;14(3):e0212446. <https://doi.org/10.1371/journal.pone.0212446>.
- Cheng M, Ning K. Stereotypess about enterotype: the old and new ideas. *Genom Proteom Bioinform.* 2019;17(1):4–12. <https://doi.org/10.1016/j.gpb.2018.02.004>.
- Suzuki TA, Martins FM, Nachman MW. Altitudinal variation of the gut microbiota in wild house mice. *Mol Ecol.* 2019;28(9):2378–90. <https://doi.org/10.1111/mec.14905>.
- Liu G, Li C, Liu Y, Zheng CM, Ning Y, Yang HG, et al. Highland adaptation of birds on the Qinghai-Tibet Plateau via gut microbiota. *Appl Microbiol Biotechnol.* 2022;106(19):6701–11. <https://doi.org/10.1007/s00253-022-12171-y>.
- Zeng B, Zhang S, Xu H, Kong F, Yu X, Wang P, et al. Gut microbiota of Tibetans and Tibetan pigs varies between high and low altitude environments. *Microbiol Res.* 2020;235: 126447. <https://doi.org/10.1016/j.micres.2020.126447>.
- Zhao J, Yao Y, Li D, Xu H, Wu J, Wen A, et al. Characterization of the gut microbiota in six geographical populations of Chinese rhesus macaques (*Macaca mulatta*), implying an adaptation to high-altitude environment. *Microb Ecol.* 2018;76(2):565–77. <https://doi.org/10.1007/s00248-018-1146-8>.
- Zuo H, Zheng T, Wu K, Yang T, Wang L, Nima Q, et al. High-altitude exposure decreases bone mineral density and its relationship with gut

- microbiota: results from the China multi-ethnic cohort (CMEC) study. *Environ Res.* 2022;215: 114206. <https://doi.org/10.1016/j.envres.2022.114206>.
32. Han N, Pan Z, Liu G, Yang R, Yujing B. Hypoxia: the "invisible pusher" of gut microbiota. *Front Microbiol.* 2021;12:690600. <https://doi.org/10.3389/fmicb.2021.690600>
 33. Sepulveda J, Moeller AH. The effects of temperature on animal gut microbiomes. *Front Microbiol.* 2020;11:384. <https://doi.org/10.3389/fmicb.2020.00384>.
 34. Medvecky M, Cejkova D, Polansky O, Karasova D, Kubasova T, Cizek A, et al. Whole genome sequencing and function prediction of 133 gut anaerobes isolated from chicken caecum in pure cultures. *BMC Genomics.* 2018;19(1):561. <https://doi.org/10.1186/s12864-018-4959-4>.
 35. Zenner C, Hitch Thomas CA, Riedel T, Wortmann E, Tiede S, Buhl Eva M, et al. Early-life immune system maturation in chickens using a synthetic community of cultured gut bacteria. *mSystems.* 2021;6(3):e01300–20. <https://doi.org/10.1128/mSystems.01300-20>.
 36. Parks DH, Chuvpochina M, Rinke C, Mussig AJ, Chaumeil P-A, Hugenholtz P. GTDB: an ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and complete genome-based taxonomy. *Nucleic Acids Res.* 2022;50(D1):D785–94. <https://doi.org/10.1093/nar/gkab776>.
 37. Tian L, Wang XW, Wu AK, Fan YH, Friedman J, Dahlin A, et al. Deciphering functional redundancy in the human microbiome. *Nat Commun.* 2020;11(1):6217. <https://doi.org/10.1038/s41467-020-19940-1>.
 38. Huang L, Zhang H, Wu PZ, Entwistle S, Li XQ, Yohe T, et al. dbCAN-seq: a database of carbohydrate-active enzyme (CAZyme) sequence and annotation. *Nucleic Acids Res.* 2018;46(D1):D516–21. <https://doi.org/10.1093/nar/gkx894>.
 39. Drula E, Garron M-L, Dogan S, Lombard V, Henrissat B, Terrapon N. The carbohydrate-active enzyme database: functions and literature. *Nucleic Acids Res.* 2022;50(D1):D571–7. <https://doi.org/10.1093/nar/gkab1045>.
 40. Shaffer M, Borton MA, McGivern BB, Zayed AA, La Rosa SL, Solden LM, et al. DRAM for distilling microbial metabolism to automate the curation of microbiome function. *Nucleic Acids Res.* 2020;48(16):8883–900. <https://doi.org/10.1093/nar/gkaa621>.
 41. Paone P, Cani PD. Mucus barrier, mucins and gut microbiota: the expected slimy partners? *Gut.* 2020;69(12):2232–43. <https://doi.org/10.1136/gutjnl-2020-322260>.
 42. Karcher N, Nigro E, Punčochář M, Blanco-Míguez A, Ciciani M, Manghi P, et al. Genomic diversity and ecology of human-associated *Akkermansia* species in the gut microbiome revealed by extensive metagenomic assembly. *Genome Biol.* 2021;22(1):209. <https://doi.org/10.1186/s13059-021-02427-7>.
 43. Karasawa Y, Maeda M. Role of ceca in the nitrogen nutrition of the chicken fed on a moderate protein-diet or a low-protein diet plus urea. *Br Poult Sci.* 1994;35(3):383–91. <https://doi.org/10.1080/00071669408417703>.
 44. Gheyas A, Vallejo-Trujillo A, Kebede A, Dessie T, Hanotte O, Smith J. Whole genome sequences of 234 indigenous African chickens from Ethiopia. *Scientific data.* 2022;9(1):53. <https://doi.org/10.1038/s41597-022-01129-4>.
 45. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet.* 2007;81(3):559–75. <https://doi.org/10.1086/519795>.
 46. Yang Y, Qian B, Xu Q, Yang Y. Climate regionalization of asphalt pavement based on the k-means clustering algorithm. *Adv Civ Eng.* 2020;2020:6917243. <https://doi.org/10.1155/2020/6917243>.
 47. Kumar H, Park W, Lim D, Srikanth K, Kim J-M, Jia X-Z, et al. Whole metagenome sequencing of cecum microbiomes in Ethiopian indigenous chickens from two different altitudes reveals antibiotic resistance genes. *Genomics.* 2020;112(2):1988–99. <https://doi.org/10.1016/j.ygeno.2019.11.011>.
 48. Chen S, Zhou Y, Chen Y, Gu J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics.* 2018;34(17):i884–90. <https://doi.org/10.1093/bioinformatics/bty560>.
 49. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. 2013 arXiv:1303.39972013.
 50. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The sequence alignment/map format and SAMtools. *Bioinformatics.* 2009;25(16):2078–9. <https://doi.org/10.1093/bioinformatics/btp352>.
 51. Li D, Liu C-M, Luo R, Sadakane K, Lam T-W. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics.* 2015;31(10):1674–6. <https://doi.org/10.1093/bioinformatics/btv033>.
 52. Zhu W, Lomsadze A, Borodovsky M. Ab initio gene identification in metagenomic sequences. *Nucleic Acids Res.* 2010;38(12):e132. <https://doi.org/10.1093/nar/gkq275>.
 53. Fu L, Niu B, Zhu Z, Wu S, Li W. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics.* 2012;28(23):3150–2. <https://doi.org/10.1093/bioinformatics/bts565>.
 54. Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. *Nat Methods.* 2015;12(1):59–60. <https://doi.org/10.1038/nmeth.3176>.
 55. Arumugam M, Raes J, Pelletier E, Le Paslier D, Yamada T, Mende DR, et al. Enterotypes of the human gut microbiome. *Nature.* 2011;473(7346):174–80. <https://doi.org/10.1038/nature09944>.
 56. Segata N, Izard J, Waldron L, Gevers D, Miropolsky L, Garrett WS, et al. Metagenomic biomarker discovery and explanation. *Genome Biol.* 2011;12(6):R60. <https://doi.org/10.1186/gb-2011-12-6-r60>.
 57. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* 2003;13(11):2498–504. <https://doi.org/10.1101/gr.1239303>.
 58. Friedman J, Alm EJ. Inferring correlation networks from genomic survey data. *Plos Comput Biol.* 2012;8(9): e1002687. <https://doi.org/10.1371/journal.pcbi.1002687>.
 59. Pérez-Jaramillo JE, Carrión VJ, Bosse M, Ferrão LFF, de Hollander M, Garcia AAF, et al. Linking rhizosphere microbiome composition of wild and domesticated *Phaseolus vulgaris* to genotypic and root phenotypic traits. *ISME J.* 2017;11(10):2244–57. <https://doi.org/10.1038/ismej.2017.85>.
 60. Peng Y, Leung HCM, Yiu SM, Chin FYL. IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics.* 2012;28(11):1420–8. <https://doi.org/10.1093/bioinformatics/bts174>.
 61. Kang DD, Li F, Kirtan 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. <https://doi.org/10.7717/peerj.7359>.
 62. 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(7):1043–55. <https://doi.org/10.1101/gr.186072.114>.
 63. Sieber CMK, Probst AJ, Sharrar A, Thomas BC, Hess M, Tringe SG, et al. Recovery of genomes from metagenomes via a dereplication, aggregation and scoring strategy. *Nat Microbiol.* 2018;3(7):836–43. <https://doi.org/10.1038/s41564-018-0171-1>.
 64. 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(12):2864–8. <https://doi.org/10.1038/ismej.2017.126>.
 65. Chaumeil P-A, Mussig AJ, Hugenholtz P, Parks DH. GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. *Bioinformatics.* 2019;36(6):1925–7. <https://doi.org/10.1093/bioinformatics/btz848>.
 66. Asnicar F, Thomas AM, Beghini F, Mengoni C, Manara S, Manghi P, et al. Precise phylogenetic analysis of microbial isolates and genomes from metagenomes using PhyloPhlAn 3.0. *Nat Commun.* 2020;11(1):2500. <https://doi.org/10.1038/s41467-020-16366-7>.
 67. Asnicar F, Weingart G, Tickle TL, Huttenhower C, Segata N. Compact graphical representation of phylogenetic data and metadata with GraPhlAn. *PeerJ.* 2015;3: e1029. <https://doi.org/10.7717/peerj.1029>.
 68. Letunic I, Bork P. Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. *Nucleic Acids Res.* 2021;49(W1):W293–6. <https://doi.org/10.1093/nar/gkab301>.
 69. Hyatt D, Chen G-L, LoCascio PF, Land ML, Larimer FW, Hauser LJ. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics.* 2010;11(1):119. <https://doi.org/10.1186/1471-2105-11-119>.
 70. Uritskiy GV, DiRuggiero J, Taylor J. MetaWRAP—a flexible pipeline for genome-resolved metagenomic data analysis. *Microbiome.* 2018;6(1):158. <https://doi.org/10.1186/s40168-018-0541-1>.
 71. Parks D. <https://github.com/dparks1134/CompareM>.
 72. Suzek BE, Huang H, McGarvey P, Mazumder R, Wu CH. UniRef: comprehensive and non-redundant UniProt reference clusters. *Bioinformatics.* 2007;23(10):1282–8. <https://doi.org/10.1093/bioinformatics/btm098>.

73. Finn RD, Tate J, Mistry J, Coghill PC, Sammut SJ, Hotz H-R, et al. The Pfam protein families database. *Nucleic Acids Res.* 2008;36(suppl_1):D281–8. <https://doi.org/10.1093/nar/gkm960>.
74. Yin Y, Mao X, Yang J, Chen X, Mao F, Xu Y. dbCAN: a web resource for automated carbohydrate-active enzyme annotation. *Nucleic Acids Res.* 2012;40(W1):W445–51. <https://doi.org/10.1093/nar/gks479>.
75. O'Leary NA, Wright MW, Brister JR, Ciufu 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(D1):D733–45. <https://doi.org/10.1093/nar/gkv1189>.
76. Rawlings ND, Barrett AJ, Bateman A. MEROPS: the peptidase database. *Nucleic Acids Res.* 2010;38(suppl_1):D227–33. <https://doi.org/10.1093/nar/gkp971>.
77. Kanehisa M, Goto S. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.* 2000;28(1):27–30. <https://doi.org/10.1093/nar/28.1.27>.
78. Oksanen J, Simpson GL, Blanchet GF, Kindt R, Legendre P, Minchin PR, O'Hara B, Solymos P, Stevens MHH, Szoecs E, Wagner H, Barbour M, Bedward M, Bolker B, Borcard D, Carvalho G, Chirico M, De Caceres M, Durand S, Antoniazzi Evangelista HB, FitzJohn R, Friendly M, Furneaux B, Hannigan G, Hill MO, Lahti L, McGlenn D, Ouellette MH, Ribeiro Cunha E, Smith T, Stier A, Cajo JF, Braak T, Weedon J, Oksanen MJ, et al. The vegan package. *Community ecology package*; 2022. R package version 2.6-4. <https://CRAN.R-project.org/package=vegan>.
79. Weissman JL, Hou S, Fuhrman JA. Estimating maximal microbial growth rates from cultures, metagenomes, and single cells via codon usage patterns. *PNAS.* 2021;118(12): e2016810118. <https://doi.org/10.1073/pnas.2016810118>.
80. Seemann T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics.* 2014;30(14):2068–9. <https://doi.org/10.1093/bioinformatics/btu153>.
81. Alcock BP, Raphenya AR, Lau TTY, Tsang KK, Bouchard M, Edalatmand A, et al. CARD 2020: antibiotic resistome surveillance with the comprehensive antibiotic resistance database. *Nucleic Acids Res.* 2019;48(D1):D517–25. <https://doi.org/10.1093/nar/gkz935>.
82. Wickham H. *Ggplot2: elegant graphics for data analysis.* New York: Springer-Verlag; 2016.
83. Wilke CO. Cowplot: streamlined plot theme and plot annotations for "ggplot2". <https://CRAN.R-project.org/package=cowplot>; 2019
84. Conway JR, Lex A, Gehlenborg N. UpSetR: an R package for the visualization of intersecting sets and their properties. *Bioinformatics.* 2017;33(18):2938–40. <https://doi.org/10.1093/bioinformatics/btx364>.
85. Warnes MGR, Bolker B, Bonebakker L, Gentleman R, Huber W, Liaw A, Lumley T, Maechler M, Magnusson A, Moeller S, Schwartz M, Venables B. *gplots: Various R Programming Tools for Plotting Data*; 2022. R package version 3.1.3. <https://CRAN.R-project.org/package=gplots>.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.