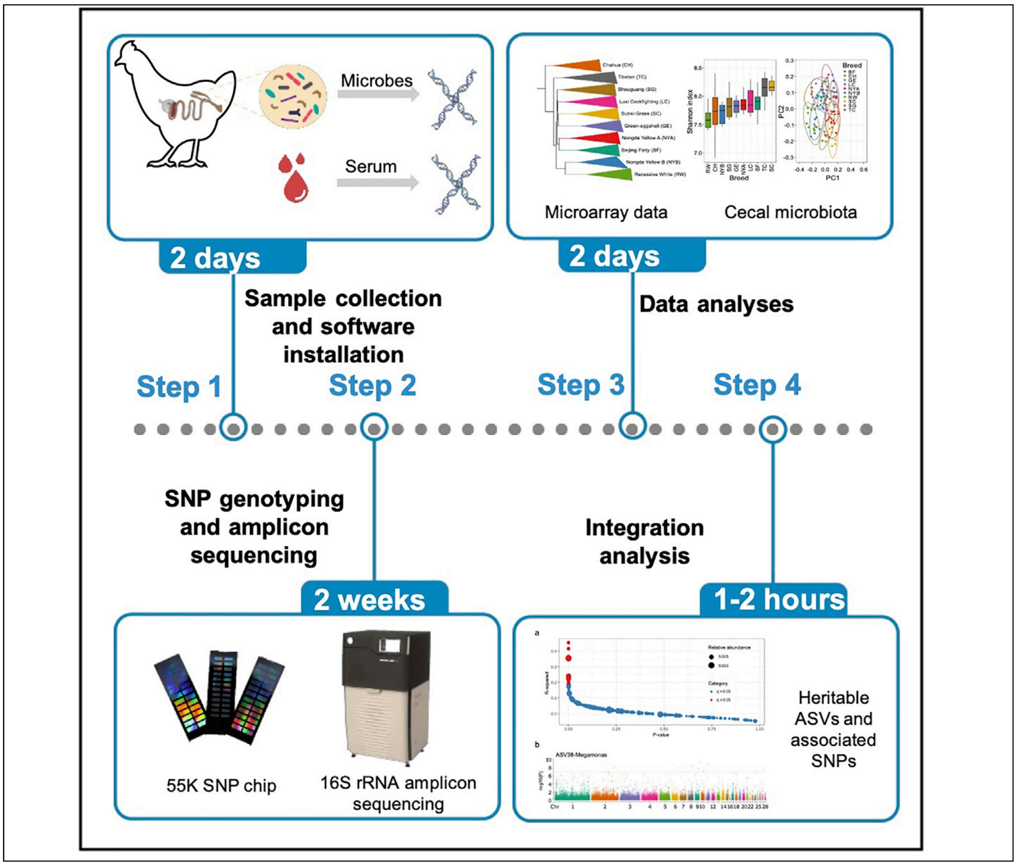
STAR Protocols

CellPress
OPEN ACCESS

Protocol

Integration of SNP genotyping and 16S rRNA amplicon sequencing to identify heritable gut microbes in chickens


Microbes
Serum
2 days
Sample collection and software installation
Microarray data
Cecal microbiota
2 days
Data analyses
Step 1
Step 2
Step 3
Step 4
SNP genotyping and amplicon sequencing
Integration analysis
2 weeks
1-2 hours
55K SNP chip
16S rRNA amplicon sequencing
Heritable ASVs and associated SNPs

Jinxin Zhang,
Yuqing Feng,
Yongfei Hu

fengyuqing2012@gmail.com (Y.F.)
huyongfei@cau.edu.cn (Y.H.)

Highlights

Step-by-step protocol for the analysis of SNP genotyping and 16S rRNA gene sequencing

Correlation analysis between chicken genetics and the gut microbiota

Identify the heritable taxa and related SNPs in the chicken genome

The effect of host genetics on the gut microbiota is not fully understood. Here, we introduce a protocol that describes the steps necessary to analyze the SNP genotyping and amplicon sequencing data to identify heritable microbes in chicken gut. We apply this protocol to infer the cecal heritable taxa and their associated SNPs in chicken genome sequence. This will be beneficial for the identification of gut microbes that are influenced by host genetics in both humans and animals.

Publisher's note: Undertaking any experimental protocol requires adherence to local institutional guidelines for laboratory safety and ethics.

Zhang et al., STAR Protocols 4, 102071
March 17, 2023 © 2023 The Author(s).
https://doi.org/10.1016/j.xpro.2023.102071

STAR Protocols

CellPress OPEN ACCESS

Protocol

Integration of SNP genotyping and 16S rRNA amplicon sequencing to identify heritable gut microbes in chickens

Jinxin Zhang,[1] Yuqing Feng,[1,2,*] and Yongfei Hu[1,3,*]

[1]State Key Laboratory of Animal Nutrition, College of Animal Science and Technology, China Agricultural University, Beijing 100193, PR China

[2]Technical contact

[3]Lead contact

*Correspondence: fengyuqing2012@gmail.com (Y.F.), huyongfei@cau.edu.cn (Y.H.)

https://doi.org/10.1016/j.xpro.2023.102071

SUMMARY

The effect of host genetics on the gut microbiota is not fully understood. Here, we introduce a protocol that describes the steps necessary to analyze the SNP genotyping and amplicon sequencing data to identify heritable microbes in chicken gut. We apply this protocol to infer the cecal heritable taxa and their associated SNPs in chicken genome sequence. This will be beneficial for the identification of gut microbes that are influenced by host genetics in both humans and animals.

For complete details on the use and execution of this protocol, please refer to Feng et al. (2022).[1]

BEFORE YOU BEGIN

Overview

The gut microbiota of animals can be affected by different factors including diet, environment, host genetics, etc. In recent years, the interactions between host genetics and gut microbiota earn wide attentions.[2–4] cumulated evidence has shown that host genetics can affect the colonization of gut microbes and thus influence gut microbiota composition and function. Meanwhile, the host-genotype-dependent gut microbes, i.e., heritable gut microbes, are involved in different processes of host physiology and metabolism, contributing to the development of complex host phenotypes.[1]

The gastrointestinal tract of chickens is densely colonized by a complex microbial community (bacteria, fungi, archaea, and viruses).[5] However, the study of the interactions between chicken genetics and the gut microbiota is still in its early stage. This protocol examines how the genetic makeup of the poultry host affects their cecal microbiota, mining the relationship between chicken genetic variation and gut microbiota composition and finding the heritable microbes. Our protocol also sheds light on investigating the interactions between host genetics and gut microbes in other animals and in humans.

Preparation: Prerequisite software installation

Institutional permissions

The animal study was reviewed and approved by Animal Welfare Committee of China Agricultural University (CAU).

Installing bioinformatic tools described in the following steps.

Illumina microarray data

1. Install the Plink v1.90 software for filtering the SNP data and downstream analyses. We can run the bash commands using the command prompt or terminal:

```
> wget http://s3.amazonaws.com/plink1-assets/plink_linux_x86_64_20200921.zip

> unzip plink_linux_x86_64_20190215.zip
```

   Note: Here, we installed the plink version 1.90 for Linux, the versions for Windows and mac OSX can be available from the plink homepage[6] (https://www.cog-genomics.org/plink/).

Installation QIIME2 using Miniconda

2. Before the installation of QIIME2,[7] we recommend installing Miniconda, a package management system, which provides applications for running commands in Linux OS, Windows as well as mac OSX. Miniconda can be downloaded from this website (https://docs.conda.io/en/latest/miniconda.html).

   Note: We also suggest installing the Miniconda in Windows Subsystem for Linux (WSL) for Windows users. On WSL, we can access the files stored in Windows system via drvFS (/mnt). For example, if there is a file named "data" under the directory of your username ("Administrator"), you can access this file via the directory /mnt/c/Users/Administrator/data/.

3. After the installation of Miniconda, we can run the bash commands in the command line to install QIIME2 in Linux OS:

```
> wget https://data.qiime2.org/distro/core/qiime2-2022.2-py38-linux-conda.yml

> conda env create -n qiime2-2022.2 --file qiime2-2022.2-py38-linux-conda.yml
```

Install the QIIME2 using conda by typing the following bash commands in mac OSX (QIIME2 usually updates four times per year, we suggest using the latest version, which has less bugs and more features).

   Note: 1) QIIME2 usually updates four times per year, we suggest installing the latest version, which has fewer bugs and more features; The bash commands for WSL are the same as the previous. 2) When dealing with a large project with over one hundred samples, we suggest doing the analyses in Linux OS. Before the installation of QIIME2 in Windows OS, we must follow these steps to install WSL on this website (https://learn.microsoft.com/en-us/windows/wsl/install).

4. Download the SILVA database used for training feature classifiers in QIIME2. We can download the SILVA database and train the classifier by running the following bash commands in Linux OS. Before the analysis we need to activate the conda environment in the terminal:

```
> conda activate qiime2-2022.2

> wget https://data.qiime2.org/2022.2/common/silva-138-99-seqs.qza
> wget https://data.qiime2.org/2022.2/common/silva-138-99-tax.qza
> qiime feature-classifier extract-reads \
  --i-sequences silva-138-99-seqs.qza \
  --p-f-primer CCTACGGGNBGCASCAG \
  --p-r-primer GACTACNVGGGTATCTAATCC \
  --o-reads ref-seqs.qza
> qiime feature-classifier fit-classifier-naive-bayes \
  --i-reference-reads ref-seqs.qza \
  --i-reference-taxonomy silva-138-99-tax.qza \
  --o-classifier classifier_V3V4.qza
```

As an alternative, we can also use the Greengenes database for training feature classifiers in QIIME2. We can download the Greengenes database by running the bash commands in Linux OS:

```
> wget ftp://greengenes.microbio.me/greengenes_release/gg_13_5/gg_13_8_otus.tar.gz
> tar zxvf gg_13_8_outs.tar.gz
> qiime tools import \
  --type ‘FeatureData[Sequence]’ \
  --input-path gg_13_8_outs/rep_set/99_outs.fasta \
  --outptu-path 99_outs.qza
> qiime tools import \
  --type ‘FeatureData[Taxonomy]’ \
  --input-format HeaderlessTSVTaxonomyFormat \
  --input-path gg_13_8_outs/taxonomy/99_out_taxonomy.txt \
  --output-path ref-taxonomy.qza
> qiime feature-classifier extract-reads \
  --i-reference-reads 99_outs.qza \
  --p-f-primer CCTACGGGNBGCASCAG \
  --p-r-primer GACTACNVGGGTATCTAATCC \
  --o-classifier greengenes_classifier_V3V4.qza
```

Note: The output files (SILVA_classifier_V3V4.qza and greengenes_classifier_V3V4.qza) are prepared for the taxonomic assignment of each sequence. In addition, these commands are suited for the 16S V3-V4 library with the specific primers. For another library, we need to change the parameters "--p-f-primer" and "--p-r-primer", which should be consistent with the library in the experiment.

5. Download and install R3.6.3 or 4.0+ and optionally RStudio software for further analyses. Detailed instructions for installation can be found at https://www.r-project.org/. The R packages involved in the study are DirichletMultinomial, ggplot2, dplyr, vegan, lme4. We recommend installing R and RStudio on your Personal Computer (such as Windows and mac OSX).

6. GEMMA software (Zhou and Stephens, 2012) is used to Genome-Wide Association Study (GWAS). Detailed instructions for installation can be found at https://github.com/genetics-statistics/GEMMA. We can download the static binary file and install them as the following bash commands:

```
> wget https://github.com/genetics-statistics/GEMMA/releases/download/v0.98.5/gemma-0.98.5-linux-static-AMD64.gz

> gunzip gemma-0.98.5-linux-static-AMD64.gz

> chmod +x gemma-0.98.5-linux-static-AMD64
```

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Biological samples		
	This study	NA
Chicken blood	This study	NA
Cecal content	This study	NA
Critical commercial assays		
TIANamp Blood DNA Kit	Tiangen	Cat#DP348-02
QIAamp DNA Stool Mini Kit	Qiagen	Cat#51604
Chicken 55K SNP genotyping array	Liu et al.[8]	NA
Deposited data		
Amplicon sequencing data	Feng et al.[1]	NCBI SRA: PRJNA693673
Illumina microarray data	Zenodo	https://doi.org/10.5281/zenodo.7501543
Data tables	Zenodo	https://doi.org/10.5281/zenodo.7501562
Software and algorithms		
PLINK Version 1.90	Purcell et al.[6]	https://www.cog-genomics.org/plink/
R v3.6.3	R Core Team	https://www.r-project.org/
RStudio v1.4.1717	RStudio Team	
QIIME2-2019.9	Bolyen et al.[7]	https://qiime2.org/
Cutadapt Version 2.4	Martin[9]	https://cutadapt.readthedocs.io/en/stable/
GEMMA	Zhou[10]	https://bioinformaticshome.com/tools/gwas/descriptions/GEMMA.html

MATERIALS AND EQUIPMENT

All the bioinformatics analyses were carried out on either the Linux OS server or Personal Computer (Windows or mac OSX). We suggest running QIIME2 on Linux OS server.

STEP-BY-STEP METHOD DETAILS

Part 1. Sample collection

⏱ Timing: 2 days

1. We randomly selected 100 60-week-old female chickens (ten birds per breed) from 10 breeds. And all birds were randomly selected from different cages in order to avoid cage effect bias.
2. Blood was collected from the wing veins of chickens into vessels containing EDTA anticoagulant and stored at −20°C. As soon as the cecal contents were collected and frozen in liquid nitrogen immediately, they were stored at −80°C for further use.

3. Chicken genomic DNA from blood was extracted using the Blood DNA Kit following the manufacturer's instructions. As an alternative, tissues can also be used to extract host genomic DNA.
4. Microbial genomic DNA from cecal contents or feces was extracted using the DNA Stool Mini Kit following the manufacturer's instructions. To increase the efficiency during the extraction, a bead-beating homogenization step was applied during the DNA extraction.[11]

Part 2. SNP genotyping and amplicon sequencing

Timing: 2 weeks

This section describes the experimental process of the SNP genotyping and 16S rRNA amplicon sequencing data of the gut microbiota.

5. SNP genotyping.

Samples were genotyped by the commercial service company with 55K high-density SNP chips. The 55K SNP chip contains 52,184 SNP probes across 28 autosomes and two sex chromosomes (chrZ and chrW). We should select the SNP chip type according to the host and aim of study accordingly.

6. Microbial genomic DNA of cecal contents or feces was used as templates to amplify regions of the 16S rRNA gene.

We use a two-stage PCR protocol as recommended by Illumina (https://support.illumina.com/documents/documentation/chemistry_documentation/16s/16s-metagenomic-library-prep-guide-15044223-b.pdf), using a KAPA HiFi HotStart ReadyMix (2×). Compared with the traditional protocol, it will decrease the ratio of chimeras and increase the accuracy. The primers we used here are CCTACGGGNBGCASCAG (forward) and GACTACNVGGGTATCTAATCC (reverse) to amplify V3-V4 variable regions of the 16S rRNA genes. The primers used in the study may be different from the protocol.

7. After converting the genome coordinates to the chicken reference genome (galGal5), 28 autosomes were extracted for further analyses.

```
>plink --file rawdata --recode --allow-extra-chr -autosome-num 28 --out outfile_autosome
```

Note: --autosome-num: The number of autosomes of the host. This parameter should adjust according to the host.

8. After PCR product purification, quantification and homogenization, the library was completed. The pooled library was sequenced on the Illumina HiSeq platform (2 × 250 bp). To avoid the bias caused by the sequencing depth, we recommend that it is better to have a depth of approximately 40,000 raw sequencing reads per sample, which is sufficient for comparative analysis.[12]

Note: Before sequencing, a dsDNA quantification is required. This procedure can be performed by Qubit™ dsDNA HS and BR Assay Kits or real-time PCR.

Part 3. Analyses of microarray and 16S rRNA amplicon sequencing data

Timing: 2 days

This section describes the bioinformatic analyses of the SNP genotyping and 16S rRNA amplicon sequencing data of the gut microbiota.

9. Before the analysis, we must have obtained the two files (rawdata.ped and rawdata.map) from the sequencing company. Generally, genetic variations in sex chromosomes would be irrelevant to the aim of our study. Thus, we need to extract the SNPs in autosomes and filter out some SNPs for the downstream analyses. We can run the bash commands in terminal:

```
>plink --file outfile_autosome -geno 0.05 -mind 0.05 -maf 0.05 --recode --out filterdata
```

Note: Here, the plink software would process the two input files (filterdata.ped and filterdata.map) and generate two output files (filterdata.ped and filterdata.map). There are some parameters can be adjusted in this procedure:

```
--geno: Removed all SNPs with a deletion rate higher than 0.05.

--mind: Remove all individuals with a miss rate over 0.05.

--maf: Remove all SNPs with minor allele frequencies (MAF) less than 0.05.
```

10. Statistical analysis of Illumina microarray data.
    a. The population structure was calculated by PLINK, and the bash commands are as follows. Principal component analysis (PCA):

```
>plink --file filterdata --pca 10 --out plink
```

Note: The eigenvectors are written to the file named plink.eigenvec; top eigenvalues are written to the file named plink.eigenval.

Identity by state (IBS):

```
>plink --file filterdata --distance --out filterdata_ibs
```

Note: output formats: –distance causes a matrix file to be written to plink.dist, and a list of corresponding sample IDs to plink.dist.id.

    b. The PCA result was visualized in R, the R commands are as follows.

```
> require(ggplot2)

> require(ggsci)

> pca0 <- read.table(‘plink.eigenvec’,header = T,sep = ‘\t’)

> pca1 <- pca0[,2:4]

> pca1$Breeds <- c(rep(‘<Group1>’,<N1>), rep(‘<Group2>,<N2>))

> p1 <- ggplot(pca1, aes(x = PC1, y = PC2, color = Breeds)) +

  geom_point(size = 1) +

  cale_color_npg() +
  theme_bw()
> p1
```

Note: We should modify "<Group1>", "<Group2>" with our own group names, and "<N1>" and "<N2>" with the number of samples of each group, if the order of plink.eigenvec was sorted by groups.

11. 16S rRNA gene amplicon sequencing data processing using QIIME2.
    a. Importing raw sequence files into QIIME2. Run the bash commands in terminal:

```
> conda activate qiime2-2022.2
> qiime tools import \
  --type ‘SampleData[PairedEndSequencesWithQuality]’ \
  --input-path manifest \
  --output-path demux.qza \
  --input-format PairedEndFastqManifestPhred64V2
```

⚠ CRITICAL: We need create a text file called a "manifest file", which maps sample identifiers to fastq.gz or fastq absolute filepaths that contain sequence and quality data for the sample. The file manifest contained three columns. The first column is the sample-id, and the second and the third columns are the absolute path of the forward and reverse read files (Figure 1). There may exist some changes in the format of the file, we should adjust it accordingly.

    b. Generate a summary of the amplicon sequencing results.

```
> qime demux summarize \
  --i-data demux.qza \
  --o-visualization demux.qzv
```

⚠ CRITICAL: we can visualize the output file demux.qzv by dragging the file into the website (view.qiime.org, Figure 2). It will help us to determine the position to be truncated in the next step. In this example, we have a quality score over 30 for the first 240 bases for forward and reverse reads. So, the number of "–p-trunc-len-f" and "–p-trunc-len-r" can be set to 240 in the next step. Good quality of the forward and reverse sequences could be merged after trimming off low-quality bases.

    c. Perform quality control and denoising sequences with DADA2.[13] Running these bash commands in the QIIME2 conda environment:

```
> qiime dada2 denoise-paired \
  --i-demultiplexed-seqs demux.qza \
  --p-trim-left-f 17 \
  --p-trim-left-r 21 \
  --p-trunc-len-f 240 \
  --p-trunc-len-r 240 \
  --o-table table.qza \
  --o-representative-sequences rep-seqs.qza \
  --o-denoising-stats denoising-stats.qza \
  --p-n-threads 16
```

Note: The "–p-trim-left-f" and "–p-trim-right-r" are the lengths of the forward and reverse primers applied during the library. In the downstream analyses, we used DADA2 to detect and correct Illumina amplicon sequence data. So, we need remove primer sequences from the sequences if they were present in your sequencing data. As an alternative, we can use the plugin "cutadapt" in the QIIME2 conda environment to trim the primer sequences before this process.

```
> qiime cutadapt trim-paired \
  --i-demultiplexed-sequences demux.qza \
  --p-cores 8 \
  --p-front-f CCTACGGGNBGCASCAG \
  --p-front-r GACTACNVGGGTATCTAATCC \
  --o-trimmed-sequences trimed_demux.qza
```

d. Assigning taxonomy of each amplicon sequence variant (ASV). We can assign the taxonomy of each ASV by running the bash commands:

```
> qiime feature-classifier classify-sklean \
  --i-classifier SILVA_classifier_V3V4.qza \
  --i-reads rep-seqs.qza \
  --o-classification taxonomy.qza
```

Note: We can use the feature classifiers (greengenes_classifier_V3V4.qza) generated by Greengenes database to assign taxonomy of each ASV.

e. Filter feature tables. In this step, we need to: i) Retain the features that contain a phylum-level annotation; ii) Remove all features that contain either mitochondria or chloroplast; iii) Filtering features that show up in only one sample. We can run the bash commands in the QIIME2 conda environment:

```
> qiime taxa filter-table \
  --i-table table.qza \
  --i-taxonomy taxonomy.qza \
  --p-include D_1__ \
  --p-exclude mitochondria,chloroplast \
  --o-filtered-table table_filt.qza
> qiime taxa filter-seqs \
  --i-sequences rep-seqs.qza \
  --i-taxonomy taxonomy.qza \
  --p-include D_1__ \
  --p-exclude mitochondria,chloroplast \
  --o-filtered-sequences rep-seqs_filt.qza
> qiime feature-table filter-features \
  --i-table table_filt.qza \
  --p-min-samples 2 \
  --o-filtered-table table_filt_2nd.qza
> qiime feature-table summarize \
  --i-table table_filt_2nd.qza \
  --o-visualization table_filt_2nd.qzv \
  --m-sample-metadata-file sample_metadata.tsv
```

Note: Here, we should create a file (“sample_metadata.tsv”) manually as the input file. The file contained two columns. The first column is the “sample-id”, and the second column is the “Group”.

f. View the taxonomic composition of our samples with interactive bar plots. We need to export the file to be recognized by the website (https://view.qiime2.org) by running the bash commands:

```
> qiime taxa barplot \
  --i-table table_filt_2nd.qza \
  --i-taxonomy taxonomy.qza \
  --m-metadata-file sample_metadata.tsv \
  --o-visualization taxa-bar-plots.qzv
```

g. Alpha and beta diversity of the gut microbiota. The diversity is an important aspect should be valued in the study of the gut microbiota, we can calculate the alpha and beta diversity by running the bash commands in the QIIME2 conda environment:

```
> qiime phylogeny align-to-tree-mafft-fasttree \
    --i-sequences rep-seqs_filt.qza \
    --o-alignment aligned-rep-seqs.qza \
    --o-masked-alignment masked-aligned-rep-seqs.qza \
    --o-tree unrooted-tree.qza \
   --o-rooted-tree rooted-tree.qza
> qiime diversity core-metrics-phylogenetic \
  --i-table table_filt_2nd.qza \
  --i-phylogeny rooted-tree.qza \
  --m-metadata-file sample_metadata.tsv \
  --p-sampling-depth 16786 \
  --output-dir core-metrics-results
```

Note: The parameter of "–p-sampling-depth" should be adjusted according to the minimum number of the file named "table_filt_2nd.qzv". This file can be visualized by dragging into the website (https://view.qiime2.org). We can also set to a certain number to filter out the samples of poor sequencing.

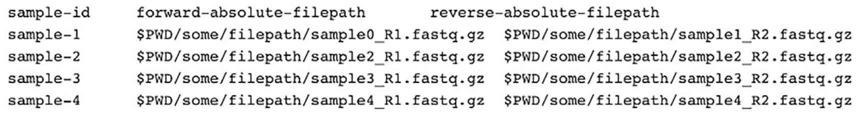
```
> qiime diversity alpha-group-significance \
  --i-alpha-diversity core-metrics-results/shannon_vector.qza \
  --m-metadata-file sample_metadata.tsv \
  --o-visualization core-metrics-results/shannon-group-significance.qzv
> qiime diversity beta-group-significance \
  --i-distance-matrix core-metrics-results/bray_curtis_distance_matrix.qza \
  --m-metadata-file sample_metadata.tsv \
  --m-metadata-column Group \
  --o-visualization core-metrics-results/bray_curtis-significance.qzv \
  --p-pairwise
```

Note: We can set multiple groups for our samples by adding columns to the file named sample_metadata.tsv.

h. Export the abundance of all ASVs and their taxonomic assignments. We can extract the data stored in the qza files by running the bash commands in the QIIME2 conda environment:

```
> qiime tools export \
  --input-path table_filt_2nd.qza \
  --output-path exported-feature-table
```

Note: The detail tutorials of this analysis using QIIME2 can be found in the following link (https://docs.qiime2.org/2022.2/tutorials/). After running these bash commands, we can get some useful files for the downstream analyses. 1) The file "taxa-bar-plots.qzv" can be visualized by dragging into the website (https://view.qiime2.org). Further, we can also explore the relative abundance on different taxonomy levels. 2) Under the folder "exported-feature-table", we can obtain the abundance of all filtered ASVs and their taxonomic assignments. It is recommended to filter out the ASVs with an average relative abundance less than a certain cut-off. In our study, we keep 200 ASVs with an average relative abundance less than 0.1%.



```
sample-id     forward-absolute-filepath         reverse-absolute-filepath
sample-1      $PWD/some/filepath/sample0_R1.fastq.gz  $PWD/some/filepath/sample1_R2.fastq.gz
sample-2      $PWD/some/filepath/sample2_R1.fastq.gz  $PWD/some/filepath/sample2_R2.fastq.gz
sample-3      $PWD/some/filepath/sample3_R1.fastq.gz  $PWD/some/filepath/sample3_R2.fastq.gz
sample-4      $PWD/some/filepath/sample4_R1.fastq.gz  $PWD/some/filepath/sample4_R2.fastq.gz
```

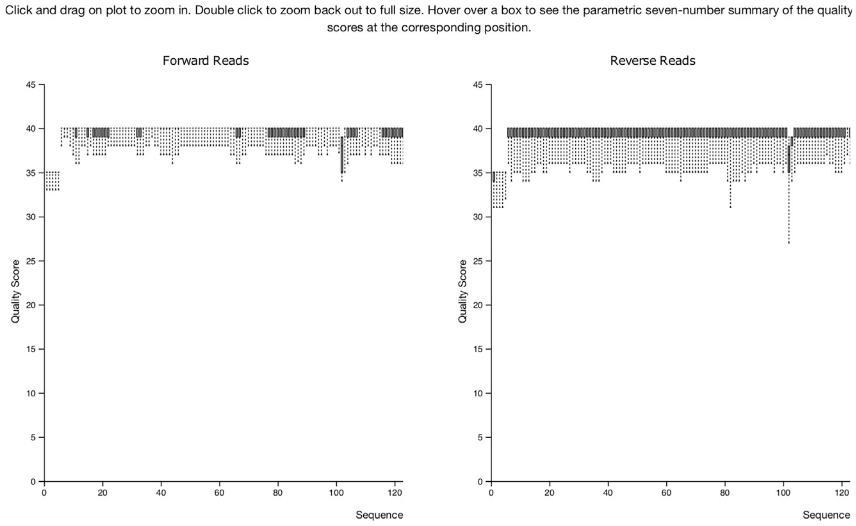
Figure 1. Format of the file used for the import of raw sequencing reads

Under the folder "core-metrics-results", we can get the results of alpha and beta-diversity, as well as the results of the statistical tests.

12. Dirichlet multinomial mixture (DMM) model.

Using DMM models, individuals were grouped according to the composition of their microbial communities. We can get the community type of our samples by running the R commands:

```
> require(DirichletMultinomial)
> require(dplyr)
> require(reshape2)
> require(ggplot2)
> L6 <- read.csv('level-6.csv', header = T, row.names = 1) %>%
  select(-c('Group')) %>%
  as.matrix()
> fit <- lapply(1:10, dmn, count = L6, verbose = TRUE)
> lplc <- sapply(fit, laplace)
> best <- fit[[which.min(unlist(lplc))]]
> ass <- apply(mixture(best), 1, which.max)
> imp_value <- melt(fitted(best))
> for (k in seq(ncol(fitted(best)))){
  d <- melt(fitted(best))
  colnames(d) <- c("Genus", "Cluster", "Value")
  d <- subset(d, Cluster == k) %>%
    arrange(Value) %>%
    mutate(Genus = factor(Genus, levels = unique(Genus))) %>%
    filter(abs(Value) > quantile(abs(Value), 0.8))
> p <- ggplot(d, aes(x = Genus, y = Value)) +
  geom_bar(stat = "identity") +
  coord_flip( ) +
  labs(title = paste("Top drivers: community type", k))
> print(p)
}
```



Click and drag on plot to zoom in. Double click to zoom back out to full size. Hover over a box to see the parametric seven-number summary of the quality scores at the corresponding position.

Figure 2. Summary of the quality scores at the corresponding position of the forward and reverse sequences, respectively

⚠ CRITICAL: The file "level-6" is an output file exported from "taxa-bar-plots.qzv" via the website (https://view.qiime2.org). It is an important file for the analysis in the gut microbiota. Here, we just show how to get the count abundance on the genus level in Figure 3.

Part 4. Investigation analysis to identify heritable ASVs and associated SNPs

⏲ Timing: 1–2 h

13. The correlation between the genomic relationship matrices and the Bray-Curtis dissimilarity of the cecal microbiota was then calculated by the Mantel test in R, the commands are as follows.

```
> require(ade4)

> array0 <- read.table('filterdata_ibs.mdist', header = FALSE)

> array <- as.dist(as(array0, "matrix"))

> micro <- read.csv ('level-6.csv', header = T, row.names = 1) %>%

select(-c('Group')) %>%

as.matrix()

> mantel.rtest (micro, array, nrepet = 9999)
```

Note: Before this analysis, it is necessary to convert the original data into distance matrices.

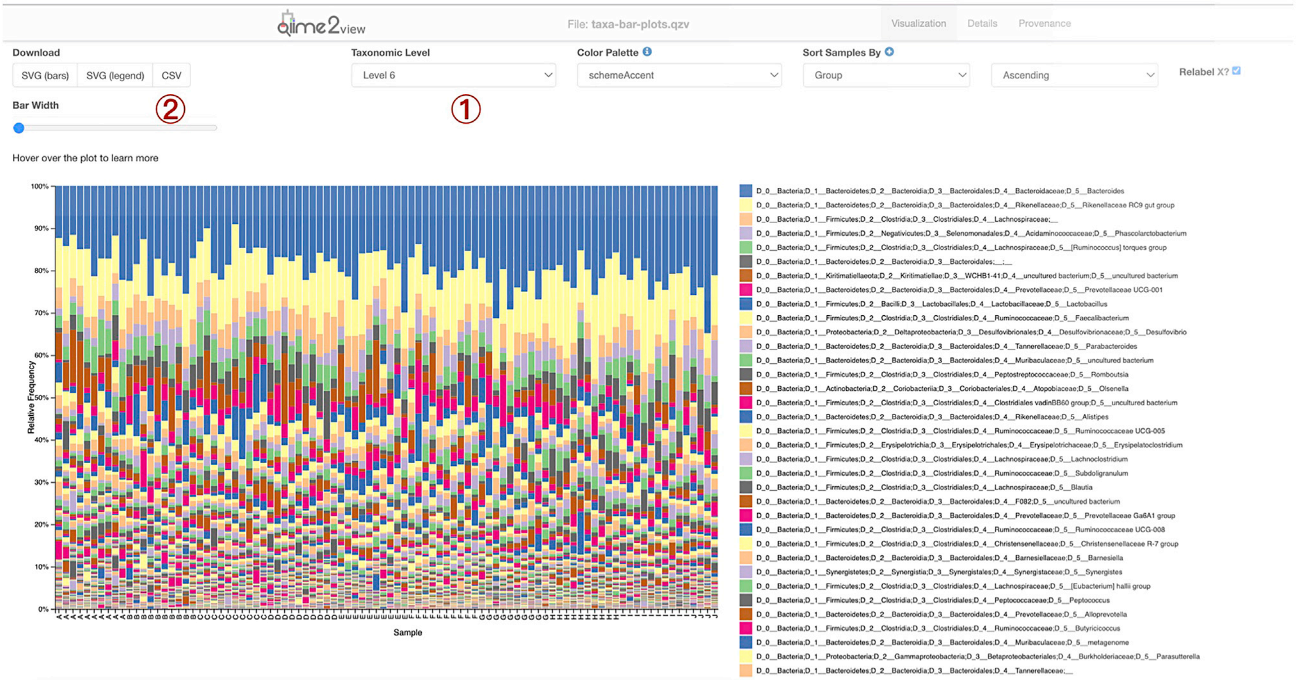
14. We performed a Procrustes analysis on the Bray-Curtis dissimilarity of the cecal microbiota and the genomic relationship matrices using the "procrustes" function from the R package Vegan, the R commands are as follows.

```
> require(vegan)

> require(dplyr)
> array0 <- read.csv(’genetic_distance.csv’, header = T, row.names = 1)

> array <- as.dist(as(array0, "matrix"))

> micro <- read.csv(’level-6.csv’, header = T, row.names = 1) %>%

  select(-c(’Group’)) %>%

  as.matrix()

> vare.proc <- procrustes(array, micro)

> summary(vare.proc)
```

15. To associate genomic PCA with the microbiota, we fit a linear model to each of the microbial features using the first five principal components, the R commands are as follows.

```
> ASV1 <- read.table (’ASV_filt.txt’, header = F, sep = ’\t’, row.names = 1)

> pca0 <- read.table (’plink.eigenvec’, header = T)

> ASV_pca <- merge (ASV1, pca0, by.x = ’ID’, by.y = ’ID’,all.y = TRUE)

> R_sq <- c()

> pvalue <- c()

> lmp <- function(modelobject){

  if(class(modelobject) != "lm") stop("Not an object of class ’lm’ ")

  f <- summary(modelobject)$fstatistic

  p <- pf(f[1], f[2], f[3], lower.tail = F)

  attributes(p) <- NULL

  return(p)

}

> for (i in 1: dim (ASV1)[1]){

  storage <- lm(ASV_pca[,i+1] ~ ASV_pca[,i+3] + ASV_pca[,i+4] + ASV_pca[,i+5] + ASV_pca
[,i+6] + ASV_pca[,i+7])

  R_sq[i] <- summary(storage)$adj.r.squared

  pvalue[i] <- lmp(storage)

}

> R2 <- data.frame(IDs, R_sq, pvalue)

> R2

> R2$qvalue <- p.adjust(R2$pvalue)

> R2$Col[R2$qvalue<=0.05] <- ’Significant’

> R2$Col[R2$qvalue>0.05] <- ’Not Significant’
```

Note: The number of the principal components is based on the proportion of explained variance of the first few principal components (PCs). Here, we used the five PCs to fit the model. The number of PCs applied here is determined by the accumulated explained variances of the PCs. The total proportions of explained variances of the selected PCs must reach the desired



File: taxa-bar-plots.qzv

Taxonomic Level: Level 6

Color Palette: schemeAccent

Sort Samples By: Group, Ascending

Relabel X? ☑

qiime2view
Visualization
Details
Provenance
Download
SVG (bars)
SVG (legend)
CSV
②
①
Bar Width
Hover over the plot to learn more
Relative Frequency
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
Sample
D_0__Bacteria;D_1__Bacteroidetes;D_2__Bacteroidia;D_3__Bacteroidales;D_4__Bacteroidaceae;D_5__Bacteroides
D_0__Bacteria;D_1__Bacteroidetes;D_2__Bacteroidia;D_3__Bacteroidales;D_4__Rikenellaceae;D_5__Rikenellaceae RC9 gut group
D_0__Bacteria;D_1__Firmicutes;D_2__Clostridia;D_3__Clostridiales;D_4__Lachnospiraceae;__
D_0__Bacteria;D_1__Firmicutes;D_2__Negativicutes;D_3__Selenomonadales;D_4__Acidaminococcaceae;D_5__Phascolarctobacterium
D_0__Bacteria;D_1__Firmicutes;D_2__Clostridia;D_3__Clostridiales;D_4__Lachnospiraceae;D_5__[Ruminococcus] torques group
D_0__Bacteria;D_1__Bacteroidetes;D_2__Bacteroidia;D_3__Bacteroidales;__;__
D_0__Bacteria;D_1__Kiritimatiellaeota;D_2__Kiritimatiellae;D_3__WCHB1-41;D_4__uncultured bacterium;D_5__uncultured bacterium
D_0__Bacteria;D_1__Bacteroidetes;D_2__Bacteroidia;D_3__Bacteroidales;D_4__Prevotellaceae;D_5__Prevotellaceae UCG-001
D_0__Bacteria;D_1__Firmicutes;D_2__Bacilli;D_3__Lactobacillales;D_4__Lactobacillaceae;D_5__Lactobacillus
D_0__Bacteria;D_1__Firmicutes;D_2__Clostridia;D_3__Clostridiales;D_4__Ruminococcaceae;D_5__Faecalibacterium
D_0__Bacteria;D_1__Proteobacteria;D_2__Deltaproteobacteria;D_3__Desulfovibrionales;D_4__Desulfovibrionaceae;D_5__Desulfovibrio
D_0__Bacteria;D_1__Bacteroidetes;D_2__Bacteroidia;D_3__Bacteroidales;D_4__Tannerellaceae;D_5__Parabacteroides
D_0__Bacteria;D_1__Bacteroidetes;D_2__Bacteroidia;D_3__Bacteroidales;D_4__Muribaculaceae;D_5__uncultured bacterium
D_0__Bacteria;D_1__Firmicutes;D_2__Clostridia;D_3__Clostridiales;D_4__Peptostreptococcaceae;D_5__Romboutsia
D_0__Bacteria;D_1__Actinobacteria;D_2__Coriobacteriia;D_3__Coriobacteriales;D_4__Atopobiaceae;D_5__Olsenella
D_0__Bacteria;D_1__Firmicutes;D_2__Clostridia;D_3__Clostridiales;D_4__Clostridiales vadinBB60 group;D_5__uncultured bacterium
D_0__Bacteria;D_1__Bacteroidetes;D_2__Bacteroidia;D_3__Bacteroidales;D_4__Rikenellaceae;D_5__Alistipes
D_0__Bacteria;D_1__Firmicutes;D_2__Clostridia;D_3__Clostridiales;D_4__Ruminococcaceae;D_5__Ruminococcaceae UCG-005
D_0__Bacteria;D_1__Firmicutes;D_2__Erysipelotrichia;D_3__Erysipelotrichales;D_4__Erysipelotrichaceae;D_5__Erysipelatoclostridium
D_0__Bacteria;D_1__Firmicutes;D_2__Clostridia;D_3__Clostridiales;D_4__Lachnospiraceae;D_5__Lachnoclostridium
D_0__Bacteria;D_1__Firmicutes;D_2__Clostridia;D_3__Clostridiales;D_4__Ruminococcaceae;D_5__Subdoligranulum
D_0__Bacteria;D_1__Firmicutes;D_2__Clostridia;D_3__Clostridiales;D_4__Lachnospiraceae;D_5__Blautia
D_0__Bacteria;D_1__Bacteroidetes;D_2__Bacteroidia;D_3__Bacteroidales;D_4__F082;D_5__uncultured bacterium
D_0__Bacteria;D_1__Bacteroidetes;D_2__Bacteroidia;D_3__Bacteroidales;D_4__Prevotellaceae;D_5__Prevotellaceae Ga6A1 group
D_0__Bacteria;D_1__Firmicutes;D_2__Clostridia;D_3__Clostridiales;D_4__Ruminococcaceae;D_5__Ruminococcaceae UCG-008
D_0__Bacteria;D_1__Firmicutes;D_2__Clostridia;D_3__Clostridiales;D_4__Christensenellaceae;D_5__Christensenellaceae R-7 group
D_0__Bacteria;D_1__Bacteroidetes;D_2__Bacteroidia;D_3__Bacteroidales;D_4__Barnesiellaceae;D_5__Barnesiella
D_0__Bacteria;D_1__Synergistetes;D_2__Synergistia;D_3__Synergistales;D_4__Synergistaceae;D_5__Synergistes
D_0__Bacteria;D_1__Firmicutes;D_2__Clostridia;D_3__Clostridiales;D_4__Lachnospiraceae;D_5__[Eubacterium] hallii group
D_0__Bacteria;D_1__Firmicutes;D_2__Clostridia;D_3__Clostridiales;D_4__Peptococcaceae;D_5__Peptococcus
D_0__Bacteria;D_1__Bacteroidetes;D_2__Bacteroidia;D_3__Bacteroidales;D_4__Prevotellaceae;D_5__Alloprevotella
D_0__Bacteria;D_1__Firmicutes;D_2__Clostridia;D_3__Clostridiales;D_4__Ruminococcaceae;D_5__Butyricicoccus
D_0__Bacteria;D_1__Bacteroidetes;D_2__Bacteroidia;D_3__Bacteroidales;D_4__Muribaculaceae;D_5__metagenome
D_0__Bacteria;D_1__Proteobacteria;D_2__Gammaproteobacteria;D_3__Betaproteobacteriales;D_4__Burkholderiaceae;D_5__Parasutterella
D_0__Bacteria;D_1__Bacteroidetes;D_2__Bacteroidia;D_3__Bacteroidales;D_4__Tannerellaceae;__

Figure 3. Visualization of the relative frequency of the gut microbiota at the genus level

We should change the taxonomic level to ''Level 6'', and then press the Download bottom ''csv'' to get the relative frequency of the gut microbiota at the genus level.

cut-off. It can be changed according to our real data. We suggest using no more than ten PCs to fit the model, although increasing the number of PCs could increase the accumulated explained variances. We can get the explained variances of each PC (''vairances2'') by running these R commands:

```
> variances <- read.table(''plink.eigenval'', header = FALSE)

> all_variances <- sum(variances[,1])

> variances2 <- variances
```

16. GWAS of the major ASVs (n = 200) were performed with GEMMA, a genome-wide efficient mixed model association algorithm. Before identifying the heritable ASVs, we need to prepare the input files by running the bash commands in terminal:

```
> plink --file plink --make-bed --out gemma_input

> gemma-0.98.1-linux-static --bfile gemma_input -gk 2 -o kinshipfile
```

Note: the parameter ''-gk'' has two options: 1: centered matrix; 2: standardized matrix. We choose the default parameter, getting a standardized matrix.

To fit a linear mixed model to get the associated SNPs, we can run the bash commands as following:

```
> gemma -bfile gemma_input -k ./output/kinshipfile.sXX.txt -lmm 4 -n ${i} -o final ${i}
```

Note: We set the parameter "-lmm" to 4, achieving all the three test results, including wald test, likelihood ration test and score test. We can also set to 1, 2 or 3, corresponding to the wald test, likelihood ration test and score test, respectively. The parameter "-n" is refer to phenotype.

17. The Manhattan plots were drawn by using the R package CMplot.

```
> d1 <- read.table('final_1.assoc.txt', header = T, sep = '\t')

> d2 <-data.frame(SNP=d1$rs,CHR=d1$chr,BP=d1$ps,P=d1$p_wald)

> require(CMplot)

> CMplot(d2, plot.type = "m", LOG10 = TRUE, threshold = 8.7e-08, file = "pdf", memo = "", file.output = TRUE, verbose = TRUE, width = 18, height = 7, signal.col = NULL)
```

Note: Here, we set the threshold to 8.7e-08, which is calculated according to the following formula:

$$\text{Threshold} = 0.05/N \quad (1)$$

N is the number of SNPs analyzed in the study.

EXPECTED OUTCOMES

Our protocol will guide people to perform integration analysis of SNP genotyping and 16S rRNA amplicon sequencing data. Using this protocol, it is possible to analyze the differences in the host genetics and the gut microbiota. Further, we can identify the heritable microbes and their associated SNPs.

The direct outcomes of the protocol contain the phylogenetic relationship of the birds from different breeds, as well as the composition and diversity of the cecal microbiota. This protocol will help us to infer the phylogenetic relationship among the ten breeds and explore the differences in the gut microbiota. In addition, we could infer the heritable ASVs and the SNPs associated with the heritable ASVs. In our study, we show here there were 15 heritable ASVs under the influence of host genetics, as well as the SNPs associated with them (n = 170, Figure 4). By exploring the function of the associated genes, it provides the candidate SNPs associated with the colonization of the gut microbiota. Further, we can explore the associations between the heritable taxa and the phenotypes (such as, meat quality, weight, and feed conversion rate).

LIMITATIONS

The sample size is limited. In general, 10 samples per group are sufficient for a basic analysis of chicken genetic variation[14] (Nazareno et al., 2017). However, as sample sizes increase, background differences among the birds can be reduced, leading to reliable results. Moreover, we omitted the effect of sex and only targeted the cecal microbial community, thus excluding the communities in the whole gastrointestinal tract. These limitations should be considered in future work.

Inadequacies in microarray genotyping. Due to the limited number and loci of SNPs on the 55K gene chip, there may be a small amount of missing SNP information in the measured chicken genome, leading to incomplete interpretations of the microbial information in the chicken genome, only improving detection techniques can fix this flaw.

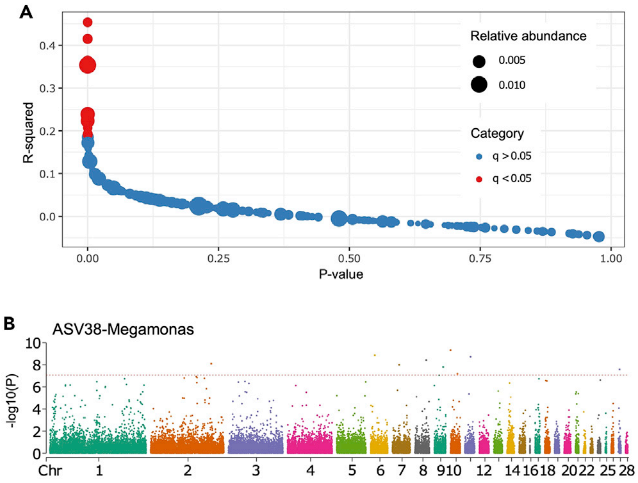


Figure 4. Associations between chicken genetics and the cecal microbiota

(A) Distribution of genetic principal component R^2 values for different ASVs in the cecal microbiota. The Y axis shows the variance explained, and the X axis shows p values for each ASV.

(B) Manhattan plots of genome-wide association p values for the relative abundance of ASV38-*Megamonas*.

The batch effect arising from different samples. Potential confounders that can affect the outcome of this kind of study are lack of sample replicates and the batch effect arising from different samples.

TROUBLESHOOTING

Problem 1

DNA yield does not meet the quality of microarray.

Potential solution

The tissue can also be used to extract the host genomic DNA as an alternative of the chicken serum.

Problem 2

Failed to install qiime2 with conda.

Potential solution

Try to type the bash commands as follows.

```
> conda clean --all

> conda update --all
```

Problem 3

The R Package installation failed.

Potential solution

Update the R language version and recommend using R 4.0 or later.

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Yongfei Hu (huyongfei@cau.edu.cn).

Materials availability

This study did not generate new unique reagents.

Data and code availability

Amplicon sequencing data have been deposited at NCBI SRA and are publicly available as of the data of publication. The accession number is listed in the key resources table. Microarray source data have been deposited at figshare and publicly available as the data of publication. The accession number is listed in the key resources table.

All original code has been deposited at figshare and is publicly available as of the data of publication. The DOI is listed in the key resources table. Any additional information required to reanalyze the data reported in this paper is available from the lead contact.

ACKNOWLEDGMENTS

This study was supported by the Beijing Natural Science Foundation (6222032), the Starting Grants Program for Young Talents at China Agricultural University, the 2115 Talent Development Program of China Agricultural University, and Chinese Universities Scientific Fund.

AUTHOR CONTRIBUTIONS

Conceptualization, Y.H.; methodology, J.Z., Y.F.; writing – original draft, J.Z., Y.F.; writing – review & editing, J.Z., Y.F.; funding acquisition, Y.H.; supervision, Y.H.

DECLARATION OF INTERESTS

The authors declare no competing interests.

REFERENCES

1. Feng, Y., Liu, D., Liu, Y., Yang, X., Zhang, M., Wei, F., Li, D., Hu, Y., and Guo, Y. (2022). Host-genotype-dependent cecal microbes are linked to breast muscle metabolites in Chinese chickens. iScience *25*, 104469. https://doi.org/10.1016/j.isci.2022.104469.

2. Benson, A.K., Kelly, S.A., Legge, R., Ma, F., Low, S.J., Kim, J., Zhang, M., Oh, P.L., Nehrenberg, D., Hua, K., et al. (2010). Individuality in gut microbiota composition is a complex polygenic trait shaped by multiple environmental and host genetic factors. Proc. Natl. Acad. Sci. USA *107*, 18933–18938. https://doi.org/10.1073/pnas.1007028107.

3. Chen, C., Zhou, Y., Fu, H., Xiong, X., Fang, S., Jiang, H., Wu, J., Yang, H., Gao, J., and Huang, L. (2021). Expanded catalog of microbial genes and metagenome-assembled genomes from the pig gut microbiome. Nat. Commun. *12*, 1106. https://doi.org/10.1038/s41467-021-21295-0.

4. Wen, C., Yan, W., Sun, C., Ji, C., Zhou, Q., Zhang, D., Zheng, J., and Yang, N. (2019). The gut microbiota is largely independent of host genetics in regulating fat deposition in chickens. ISME J. *13*, 1422–1436. https://doi.org/10.1038/s41396-019-0367-2.

5. Wang, Y., Sun, J., Zhong, H., Li, N., Xu, H., Zhu, Q., and Liu, Y. (2017). Effect of probiotics on the meat flavour and gut microbiota of chicken. Sci. Rep. *7*, 6400. https://doi.org/10.1038/s41598-017-06677-z.

6. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A.R., Bender, D., Maller, J., Sklar, P., de Bakker, P.I.W., Daly, M.J., and Sham, P.C. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. *81*, 559–575. https://doi.org/10.1086/519795.

7. Bolyen, E., Rideout, J.R., Dillon, M.R., Bokulich, N.A., Abnet, C.C., Al-Ghalith, G.A., Alexander, H., Alm, E.J., Arumugam, M., Asnicar, F., et al. (2019). Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. Nat. Biotechnol. *37*, 852–857. https://doi.org/10.1038/s41587-019-0209-9.

8. Liu, R., Xing, S., Wang, J., Zheng, M., Cui, H., Crooijmans, R.P.M.A., Li, Q., Zhao, G., and Wen, J. (2019). A new chicken 55K SNP genotyping array. BMC Genom. *20*, 410. https://doi.org/10.1186/s12864-019-5736-8.

9. Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet J. *17*, 3. https://doi.org/10.14806/ej.17.1.200.

10. Zhou, X., and Stephens, M. (2012). Genome-wide efficient mixed-model analysis for association studies. Nat. Genet. *44*, 821–824. https://doi.org/10.1038/ng.2310.

11. Sinha, R., Abu-Ali, G., Vogtmann, E., Fodor, A.A., Ren, B., Amir, A., Schwager, E., Crabtree, J., Ma, S., Microbiome Quality Control Project Consortium, et al.. (2017). Assessment of variation in microbial community amplicon sequencing by the Microbiome Quality Control (MBQC) project consortium. Nat. Biotechnol. *35*, 1077–1086. https://doi.org/10.1038/nbt.3981.

12. Myer, P.R., Kim, M., Freetly, H.C., and Smith, T.P.L. (2016). Evaluation of 16S rRNA amplicon sequencing using two next-generation sequencing technologies for phylogenetic analysis of the rumen bacterial community in steers. J. Microbiol. Methods *127*, 132–140. https://doi.org/10.1016/j.mimet.2016.06.004.

13. Callahan, B.J., McMurdie, P.J., Rosen, M.J., Han, A.W., Johnson, A.J.A., and Holmes, S.P. (2016). DADA2: high-resolution sample inference from Illumina amplicon data. Nat. Methods *13*, 581–583. https://doi.org/10.1038/nmeth.3869.

14. Nazareno, A.G., Bemmels, J.B., Dick, C.W., and Lohmann, L.G. (2017). Minimum sample sizes for population genomics: an empirical study from an Amazonian plant species. Mol. Ecol. Resour. *17*, 1136–1147. https://doi.org/10.1111/1755-0998.12654.