Copy Number Variation on *ABCC2-DNMBP* Loci Affects the Diversity and Composition of the Fecal Microbiota in Pigs

Yuliaxis Ramayo-Caldas,[a] Daniel Crespo-Piazuelo,[a] Jordi Morata,[b] Olga González-Rodríguez,[a] Cristina Sebastià,[c,d] Anna Castello,[c,d] Antoni Dalmau,[e] Sebastian Ramos-Onsins,[c] Konstantinos G. Alexiou,[c] Josep M. Folch,[c,d] Raquel Quintanilla,[a] Maria Ballester[a]

[a]Animal Breeding and Genetics Program, Institute of Agrifood Research and Technology, Caldes de Montbui, Spain

[b]Centro Nacional de Análisis Genómico, Centre for Genomic Regulation, Barcelona Institute of Science and Technology, Barcelona, Spain

[c]Plant and Animal Genomics Program, Centre for Research in Agricultural Genomics, Consejo Superior de Investigaciones Científicas (CSIC)-Institute of Agrifood Research and Technology-Autonomous University of Barcelona-UB, Bellaterra, Spain

[d]Animal and Food Science Department, Autonomous University of Barcelona, Bellaterra, Spain

[e]Animal Welfare Program, Institute of Agrifood Research and Technology, Girona, Spain

Yuliaxis Ramayo-Caldas and Daniel Crespo-Piazuelo contributed equally to this work. Author order was determined in order of increasing seniority.

ABSTRACT Genetic variation in the pig genome partially modulates the composition of porcine gut microbial communities. Previous studies have been focused on the association between single nucleotide polymorphisms (SNPs) and the gut microbiota, but little is known about the relationship between structural variants and fecal microbial traits. The main goal of this study was to explore the association between porcine genome copy number variants (CNVs) and the diversity and composition of pig fecal microbiota. For this purpose, we used whole-genome sequencing data to undertake a comprehensive identification of CNVs followed by a genome-wide association analysis between the estimated CNV status and the fecal bacterial diversity in a commercial Duroc pig population. A CNV predicted as gain (DUP) partially harboring ABCC2-DNMBP loci was associated with richness ($P = 5.41 \times 10^{-5}$, false discovery rate [FDR] = 0.022) and Shannon α-diversity ($P = 1.42 \times 10^{-4}$, FDR = 0.057). The *in silico* predicted gain of copies was validated by real-time quantitative PCR (qPCR), and its segregation, and positive association with the richness and Shannon α-diversity of the porcine fecal bacterial ecosystem was confirmed in an unrelated F1 (Duroc $\times$ Iberian) cross. Our results advise the relevance of considering the role of host-genome structural variants as potential modulators of microbial ecosystems and suggest the ABCC2-DNMBP CNV as a host-genetic factor for the modulation of the diversity and composition of the fecal microbiota in pigs.

IMPORTANCE A better understanding of the environmental and host factors modulating gut microbiomes is a topic of greatest interest. Recent evidence suggests that genetic variation in the pig genome partially controls the composition of porcine gut microbiota. However, since previous studies have been focused on the association between single nucleotide polymorphisms and the fecal microbiota, little is known about the relationship between other sources of genetic variation, like the structural variants and microbial traits. Here, we identified, experimentally validated, and replicated in an independent population a positive link between the gain of copies of *ABCC2-DNMBP* loci and the diversity and composition of pig fecal microbiota. Our results advise the relevance of considering the role of host-genome structural variants as putative modulators of microbial ecosystems and open the possibility of implementing novel holobiont-based management strategies in breeding programs for the simultaneous improvement of microbial traits and host performance.

KEYWORDS diversity, microbiota, modulators, porcine, structural variants

Address correspondence to Yuliaxis Ramayo-Caldas, yuliaxis.ramayo@irta.cat, or Maria Ballester, maria.ballester@irta.cat.

Gut microbiomes have a profound impact on many aspects of pig health, such as the modulation of metabolic functions, physiological processes, and relevant porcine traits like growth (1), feed efficiency (2, 3), and immunocompetence (4). Pig gut microbiota composition is highly variable among individuals. Host-microbiome interactions are mediated by both environmental and host factors. Among them, genetic variation in the pig genome can modulate, in a taxa-specific manner, the composition and function of the pig gut eukaryotic and prokaryotic communities. Several studies have reported low to medium heritability values for pig gut microbiota composition (2, 5, 6) that varies according to specific taxon or taxonomic levels. In addition, quantitative trait loci (QTLs), genetic variants, and candidate genes linked to microbial traits have been identified in pigs (7–10). However, since previous studies were focused on the association between single nucleotide polymorphisms (SNPs) and microbial traits, little is known about the relationship between the fecal microbiota and structural variants in the porcine genome.

Copy number variants (CNVs) are structural variants that produce a change in the number of copies (gain or loss) of a genomic region. Compared to SNPs, CNVs involve large DNA segments that span a significant proportion of the genome and account for greater genomic variability than SNPs. Consequently, CNVs are a relevant source of genetic variation that contribute to evolutionary adaptations and variation in gene expression and phenotypic traits in human and domestic animals (11, 12). In humans, gain of copies of the salivary amylase (*AMY1*) gene was associated with oral and gut microbiome composition (13). In this seminal study, Poole et al. found that individuals with greater number of copies of *AMY1* showed greater levels of salivary *Porphyromonas*, followed by an increased abundance of resistant starch-degrading microbes in the gut (13). We hypothesized that as in humans, CNVs are likely to contribute to gut microbial variability in animals.

In pigs, CNVs have been extensively characterized (12, 14–16) and have been found to be associated with a variety of traits such as coat color (17, 18), fatty acid composition (19), and growth and reproductive traits (20–22). However, to the best of our knowledge, associations between CNVs and microbial traits have not been documented in livestock. Consequently, the putative modulatory role of CNVs in the diversity, composition, and function of livestock gastrointestinal microbiota remains to be elucidated. The main goal of this study was to assess the effect of porcine CNVs on the diversity and composition of pig fecal microbiota.

RESULTS

Detection of copy number variants and analysis. In this study, we used whole-genome sequencing data from 100 healthy 60-day-old Duroc pigs to undertake a comprehensive identification of CNVs. A total of 1,292 CNVs distributed across 531 copy number variant regions (CNVRs) on autosomal pig chromosomes were identified (Table S1). After quality control, 1,005 CNVs grouped into 291 CNVR, presented in at least the 5% of the samples, were used for the association analysis. Among them, a CNV predicted as gain located on CNVR454 (SSC14:111000000-111075999) that partially contain the ATP-binding cassette subfamily C member 2 (*ABCC2*) and the dynamin-binding protein (*DNMBP*) genes showed a significant association with richness ($P = 5.41 \times 10^{-5}$) and Shannon α-diversity ($P = 1.42 \times 10^{-4}$) (Fig. 1).

Validation by quantitative PCR. The *in silico* identification of CNVs may result in both false-positive and -negative results (14, 23). To confirm the presence of the CNVR454 in our animal material, we conducted a quantitative PCR (qPCR) assay with primers located on the *ABCC2* gene. The *in silico* predicted genotype was confirmed by qPCR in 46 of 48 Duroc samples, corresponding to a precision of 95.83%. To be noted, all the 24 samples *in silico* predicted as DUP presented the gain in number of copies. Thus, deviations from the diploid status were observed in 2 of 24 animals, in which variation in the number of copies was not predicted by Control-FREEC (24).

Since the presence of false-negative samples for CNV status could affect the genome-wide association studies (GWAS) results, we evaluated the correlation between
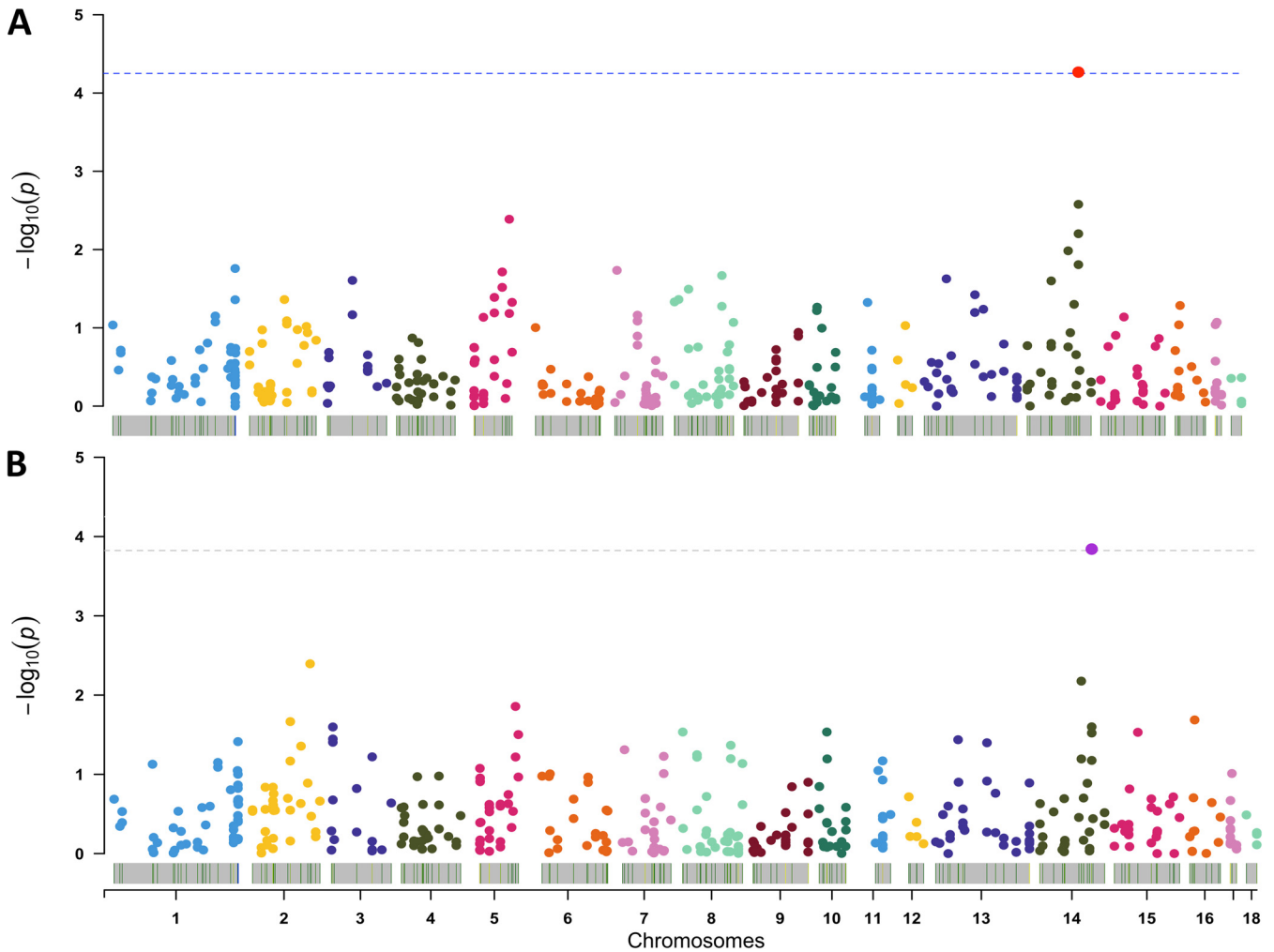
FIG 1 Results from the association analyses of copy number variants (CNVs) identified across the pig genome with gut bacterial richness (A) and Shannon α-diversity (B). The x axis represents the CNV position in the pig autosomal chromosomes (1 to 18), and the y axis reflects the significance level represented as the $-\log_{10}$ (P). The dashed lines correspond to the significance threshold after multiple test correction. The red and purple points on SSC14, Sus scrofa chromosome 14 represent the CNV comprising ABCC2-DNMBP genes.

the CNV relative quantification by qPCR (2N = 22 versus DUP = 26) and the diversity index. In agreement with the CNV-GWAS, qPCR results corroborated that DUP pigs had significantly greater richness ($P = 1.8 \times 10^{-3}$) and α-diversity values ($P = 3.8 \times 10^{-3}$) (Table 1). Furthermore, relative quantification (RQ) of the number of copies was positively correlated with the richness ($r = 0.474$, $P = 6.72 \times 10^{-4}$) and the α-diversity ($r = 0.401$, $P = 7.77 \times 10^{-3}$) (Fig. 2). We also observed that compared to their diploid counterparts, DUP samples had lower β-diversity (2N = 0.419 versus DUP = 0.375, $P = 6.36 \times 10^{-3}$), suggesting a more homogeneous fecal microbial ecosystem (i.e., more similar microbiota between samples). Furthermore, a positive relationship between the RQ of the number of

TABLE 1 Description of the diversity index of the Duroc and F1 (Duroc × Iberian) cross

Population	Data set (n)	Groups	Richness (SD)	α-Diversity (SD)
Duroc (discovery)	Complete (100)		604.92 (98.17)	6.05 (0.22)
	Validate qPCR (48)	Diploid	553.73[a] (92.71)	5.95[a] (0.21)
		DUP	657.85[b] (89.70)	6.16[b] (0.17)
Duroc × Iberic (validation)	Complete (285)		459.65 (104.33)	5.82 (0.25)
	Validate qPCR (24)	Diploid	292.36[c] (67.18)	5.36[c] (0.27)
		DUP	543.19[d] (158)	6.06[d] (0.33)

[a,b,c,d]Within a column of samples validated by quantitative PCR (qPCR), values without a common superscript differ ($P < 0.05$).
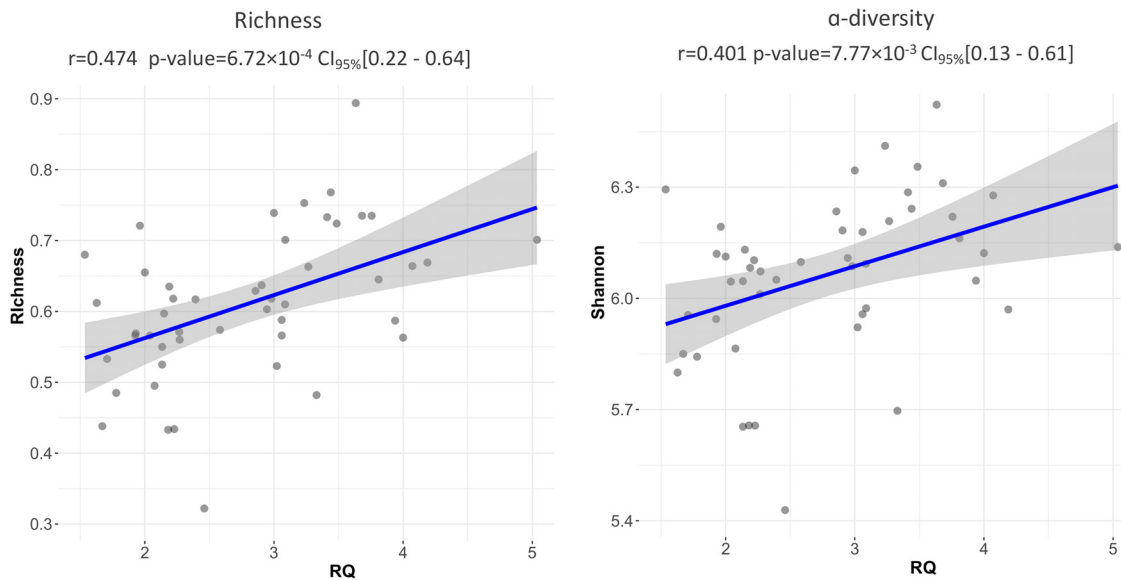
Richness
r=0.474 p-value=6.72×10⁻⁴ CI$_{95\%}$[0.22 - 0.64]

α-diversity
r=0.401 p-value=7.77×10⁻³ CI$_{95\%}$[0.13 - 0.61]

FIG 2 Relationship between the CNV relative quantification (RQ) of the number of copies with the richness and Shannon α-diversity index.

copies and the nucleotide variability of the CNV genomic interval estimators ATajima ($r = 0.428$, $P = 2.4 \times 10^{-3}$) and RTajima ($r = 0.71$, $P = 1.26 \times 10^{-8}$). The high correlation observed between RTajima and the RQ values (Fig. S1) can be explained by the characteristic of RTajima statistic, which gives more importance to intermediate frequencies observed in the whole population. Finally, a positive and significant correlation was observed with RTajima between the richness ($r = 0.27$, $P = 0.01$) and the Shannon α-diversity ($r = 0.38$, $P = 6.9 \times 10^{-3}$) (Fig. S1).

Remarkably, variation in the number of copies of the *ABCC2-DNMBP* loci was also segregating in an unrelated commercial F1 Duroc × Iberian crossbred pigs, with 13 of 24 pigs showing a gain of copies. Nonsignificant differences between males and females were observed ($P > 0.05$). The 12 samples with the highest diversity index (Table 1) presented the gain in the number of copies of the *ABCC2-DNMBP* loci. Meanwhile, the diploid status was observed in 11 of the 12 samples with the lowest diversity. Moreover, despite differences on genetic background, age, or other environmental factors such as diet of farm of origin, the association between the *ABCC2-DNMBP* loci and the fecal microbial diversity was replicated in the F1 Duroc × Iberian cross (Table 1; Fig. 3). Indeed, in both Duroc and F1 Duroc × Iberian cross data sets, the qPCR reaffirmed that a gain of copies of the *ABCC2-DNMBP* loci was positively associated with the richness and Shannon α-diversity of the pig fecal microbiota.

Microbial signatures linked to variation in the number of copies. The results from the supervised classification model showed that the relative abundance of 122 of 3,052 ASVs allowed the classification (mean accuracy of 0.67) between groups of samples (Table S2). The taxa-set enrichment analysis pointed out a higher overall discriminant importance of ASVs members of the *Desulfovibrio*, *Blautia*, *Phascolarctobacterium*, *Fibrobacter*, *Roseburia*, *Faecalibacterium*, *Megasphaera*, *Succinivibrio*, *Coprococcus*, *RFN20*, and *Anaerovibrio* genera (Fig. 4A). Furthermore, supporting their discriminative role, we observed that compared to their diploid counterparts, the fecal microbiota of DUP pigs exhibited a higher relative abundance (false discovery rate [FDR] < 0.05) of the *Desulfovibrio*, *Blautia*, *Phascolarctobacterium*, *Faecalibacterium*, *Megasphaera*, *Succinivibrio*, and *Anaerovibrio* genera but lower relative abundance of the *Fibrobacter* and *RFN20* genera (Fig. 4B). To be noted, the results obtained from the differential abundance analysis done on the unrelated F1 Duroc × Iberian crossbred population confirmed a higher relative abundance of the *Phascolarctobacterium*, *Roseburia*, and *Faecalibacterium* genera in the fecal microbiota of samples with a gain of copies of the *ABCC2-DNMBP* loci (Fig. 5).
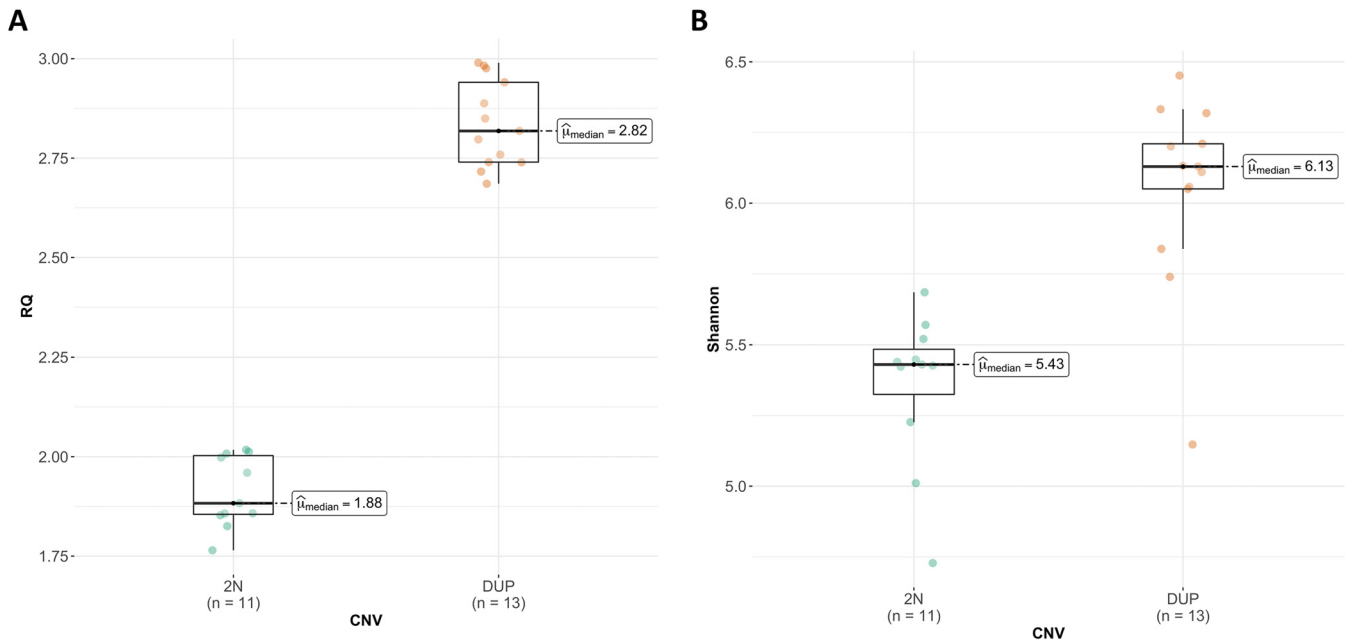
FIG 3 Results from the replication analysis comparing (A) mean RQ and (B) Shannon α-diversity of *ABCC2-DNMBP* loci in the F1 Duroc × Iberian cross. Green dots represents diploid (2N) samples, and orange dots represent DUP samples.

DISCUSSION

In this study, we report, for the first time in a livestock species, a CNV partially containing the *ABCC2* and *DNMBP* genes associated with the diversity and composition of the pig fecal microbiota. *ABCC2* encodes a multidrug resistance-associated protein 2 (MRP2) that plays a relevant role in preserving hepatic and intestinal homeostasis (25).
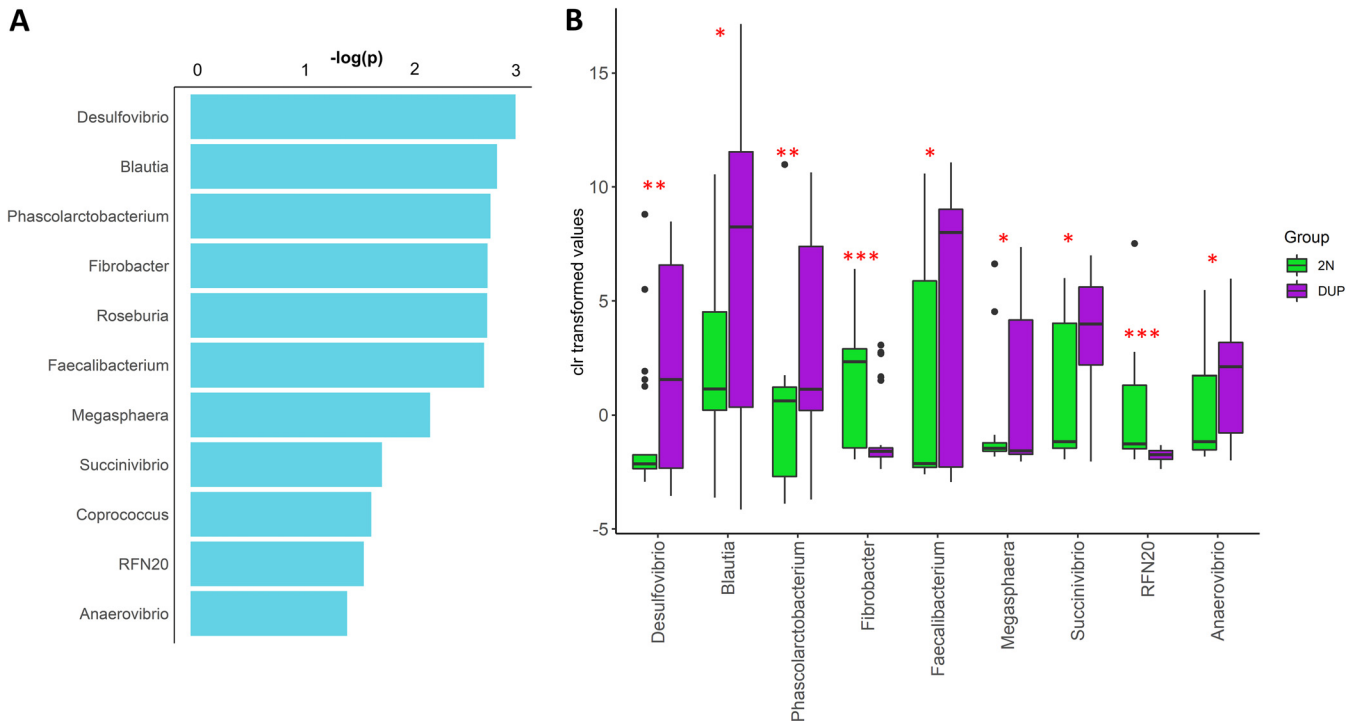

FIG 4 Results from microbial signature analyses at genus level. (A) Taxa-set enrichment. (B) Patterns of differential abundance analysis between DUP ($N = 26$) and 2N ($N = 22$) pigs in the purebred Duroc population. The mean values for the two groups are significantly different. *, $P < 0.05$; **, $P < 0.01$; **, $P < 0.001$.
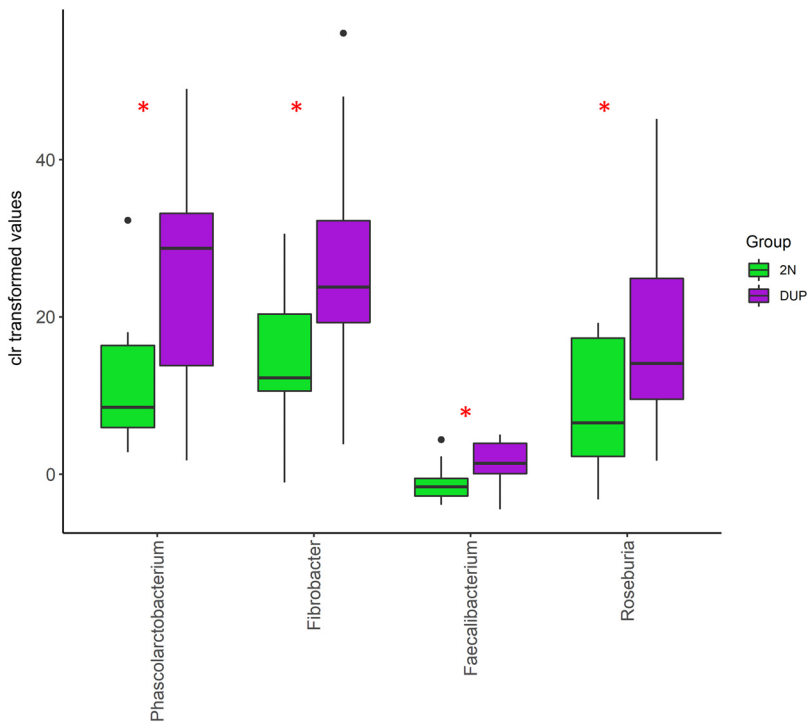
FIG 5 Differential abundance patterns at genus level between DUP ($N = 13$) and 2N ($N = 11$) samples in the F1 Duroc × Iberian crossbred population. The mean values for the two groups are significantly different. *, $P < 0.05$.

ABCC2 is involved in the excretion of conjugated bile acids (BAs), bilirubin, and xenobiotics and the transport of other organic anions (26, 27). In pigs, *ABCC2* has been reported as coassociated with the intramuscular profile of fatty acid composition in an Iberian × Landrace cross (28). In addition, the genomic interval harboring the *ABCC2-DNMBP*loci overlapped with QTLs associated with the muscle profile of palmitic (QTLId: 95385), stearic (QTLId: 95386), and palmitoleic (QTLId: 95387) fatty acids content in a Duroc × (Landrace × Yorkshire) cross (29). In other species such as mice, rats, or humans, mutations in *ABCC2* are related to hereditary liver diseases. *Mrp2*$^{-/-}$mice are viable (30, 31) but, like *ABCC2*-knockout rats, showed chronic hyperbilirubinemia followed by a reduction in biliary excretion of bilirubin glucuronides (30, 32). Meanwhile, mutations in the human *ABCC2* gene result in Dubin-Johnson syndrome, an autosomal recessive disorder characterized by a defect in the transport of endogenous and exogenous anionic conjugates from hepatocytes into the bile (33). It is noteworthy that a genomic duplication of 5,299 bp comprising exons 24 and 25 of human *ABCC2* gene was predicted to result in the insertion of a premature stop codon (34).

Based on the known role of *ABCC2* in the excretion of bilirubin and conjugated BAs, we propose that variations in the number of copies of *ABCC2* may influence gut levels of these substances in the gut. Previous research has also demonstrated bidirectional cross talk between the gut microbiota and the metabolism of bilirubin and conjugated BAs, which may be further influenced by variations in the number of copies in*ABCC2*. For example, bilirubin can regulate the composition of gut microbiota by being potentially toxic toward Gram-positive bacteria, while promoting the proliferation of Gram-negative species (35). In a similar way, a higher BA tolerance is evidenced by Gram-negative bacteria (36). In agreement with these studies, the fecal microbiota of DUP samples showed a higher relative abundance of Gram-negative bacteria, such as members of the *Desulfovibrio*, *Phascolarctobacterium*, *Faecalibacterium*, *Succinivibrio*, and *Anaerovibrio* genera. On another note, gut microbiota composition can regulate BA and bilirubin production and signaling. In addition, conjugated BAs can have a protective role on gut barrier integrity (37). The oral

administration of two major conjugated BAs, tauro-cholic acid and β-tauro-murocholic acid, increased the richness of neonatal small intestinal microbiota with a positive effect on the postnatal microbiota maturation (38). It is noteworthy that among the top discriminant ASVs, we observed butyrate-producer species with a potentially beneficial effect on the host, such as *Blautia obeum* (*ASV2433, ASV2171, ASV2278*), *Faecalibacterium prausnitzii* (*ASV2371, ASV2378, ASV2396*), and *Roseburia sp001940165* (*ASV1822*). Interestingly, the genome of all these species encodes bile salt hydrolases (BSHs, EC 3.5.1.24) (39, 40), enzymes that mediate the primary BA deconjugation and successive conversion to secondary BAs, thereby partly determining the amount of secondary BAs in the colonic epithelium, which in turn act as signaling molecules mediating different metabolic processes interconnected with health and diseases (41, 42).

The CNVR454 also included *DNMBP*, a gene that regulates the structure of apical junctions through F-actin organization in epithelial cells (43). *DNMBP* is also involved in luminal morphogenesis and enterocyte polarization (44, 45), thus potentially contributing to the function and homeostasis of the intestinal epithelial barrier (IEB). In fact, the cross talk between IEB and the gut microbiota is crucial for the maintenance of intestinal homeostasis. For example, enterocytes, which are the most abundant population among intestinal epithelial cells, express a range of pattern recognition receptors for sensing the microbe-associated molecular patterns. Further, enterocyte apex is covered by thousands of microvilli that are vital in colonic wound repair and the transport of molecules and nutrients such as bile salts, electrolytes, and vitamins (46–49). Interestingly, compared with conventional piglets, germfree (GM) pigs displayed aberrant intestinal morphology with longer villi and shorter crypts. Meanwhile, the oral administration of commensal bacteria increased crypt depth and induced enterocyte brush border microvilli enzyme activities on these GM piglets (50–53). Therefore, considering the functional roles of *DNMBP* in the IEB, we cannot rule out the contribution of *DNMBP* to the modulation of the diversity and composition of the pig gut microbiota.

Altogether, our results pinpointed a positive association of the variation in the number of copies of the *ABCC2-DNMBP* loci with the richness, α-diversity, and composition of the pig fecal microbial ecosystems. Such findings open the possibility of modulating the fecal microbiota, which has emerged as a promising breeding or therapeutic tool to optimize livestock production efficiency, animal health, and well-being. Greater gut microbial diversity is usually desired and generally accepted as an indicator of a resilient microbial ecosystem, gut, and host health. Indeed, a diverse and healthy gut has a positive effect on the absorption of dietary nutrients, feed efficiency, and animal well-being.

We are aware of some limitations of our study like the limited taxonomic resolution achieved by targeting the V3-V4 16S rRNA genomic region with short-read sequencing. We are also aware of the convenience of performing further analyses to confirm the raised hypotheses by assessing the metabolic profile of BAs, as well as evaluating the role of the CNV on gene expression (at both the microbial and host levels) of genes involved in BA metabolism. Despite these limitations, our findings contribute to the understanding of host-microbiome interactions. Moreover, our results open the possibility of breeding the holobiont via the incorporation of this source of variation on custom-made arrays that can be used in routine genotyping tasks applied to porcine breeding programs and, together with nutritional or management strategies, will favor the simultaneous improvement of microbial traits, gut health, and host performance.

Conclusions. Here, we report the first study exploring associations between porcine CNV and the diversity and composition of the pig fecal microbiota. In an unrelated population, we identified, functionally validated, and replicated a positive association between the gain of copies of *ABCC2-DNMBP* loci and the composition and diversity of the pig fecal microbiota. These results suggest a role for the host-genome structural variants in the modulation of microbial ecosystems and open the possibility of including CNVs in selection programs to simultaneously improve microbial traits, gut health, and host performance.

Ethics approval and consent to participate. The animal care and experimental procedures were carried out following the institutional guidelines for good experimental practices and the Spanish Policy for Animal Protection RD 53/2013, which meets

the European Union Directive 2010/63/EU for protection of animals used in experimentation, and were approved by the Institute of Agrifood Research and Technology Ethical Committee. Consent to participate is not applicable in this study.

MATERIALS AND METHODS

Animal samples. The samples employed in this study are a subset of pigs reported in references 9 and 54. In brief, a total of 100 healthy piglets (50 males and 50 females) aged 60 ± 8 days from a commercial Duroc pig line were used as a discovery data set (Table 1). All animals were raised on the same farm and had *ad libitum* access to the same commercial cereal-based diet. Furthermore, a subset of 24 unrelated F1 (Duroc × Iberian) crossbred pigs (300 days old, 12 males and 12 females) with phenotypically extreme fecal microbial diversity index (12 high and 12 low) from reference 7 were employed as an independent validation data set (Table 1).

Microbial DNA extraction, sequencing, and bioinformatics analysis. Fecal samples were collected from the Duroc piglets at 60 ± 8 days of age, and microbial DNA was extracted with the DNeasy PowerSoil kit (Qiagen, Hilden, Germany) following the manufacturer's recommendations. Extracted DNA was sent to the University of Illinois Keck Center for paired-end (2 × 250 bp) sequencing on an Illumina NovaSeq (Illumina, San Diego, CA, USA). The 16S rRNA gene fragment was amplified using the primers V3_F357_N (5′-CCTACGGGNGGCWGCAG-3′) and V4_R805 (5′-GACTACHVGGGTATCTAATCC-3′). The sequences were analyzed with QIIME2 (55); barcode sequences, primers, and low-quality reads (Phred score < 30) were removed. The quality control process also trimmed sequences based on expected amplicon length and removed chimeras. Afterwards, the sequences were clustered into amplicon sequence variants (ASVs) at 99% identity. ASVs were classified to the lowest possible taxonomic level based on a primer-specific trained version of GreenGenes2 Database (released October 2022) (56, 57). Before the estimation of the diversity indices, to correct for the sequencing depth, the samples were rarefied at 10,000 reads. The diversity metrics at ASV levels were estimated with the vegan R package version 2.6-2 (58). The α-diversity was evaluated with the Shannon index (59), and the β-diversity was assessed using the Whittaker index (60). The graphical representations of means comparison were done with the *ggstatsplot* R library (61).

Host-genome data analysis and CNV calling. Simultaneously with fecal sampling, blood was collected at 60 ± 8 days of age via the external jugular vein. Host genomic DNA was extracted from blood using the NucleoSpin blood protocol (Macherey-Nagel). The whole genome was paired-end sequenced (2 × 150 bp) in an Illumina NovaSeq 6000 platform (Illumina) at the Centro Nacional de Análisis Genómico (Centre for Genomic Regulation, Barcelona, Spain). Reads were mapped to the porcine reference assembly Sscrofa.11.1 with BWA-MEM 0.7.17 (62). Alignment files containing only properly paired, uniquely mapping reads without duplicates were processed using Picard (63) to add read groups and to remove duplicates. Variant calling was performed with the Haplotype Caller tool from the Genome Analysis Toolkit (GATK 4.1.8.0) (64). Applying GATK best practices, variants with minimum read depth of 5 on at least one sample were retained. Joint genotyping was conducted with combined genomic variant call format (gVCFs). Functional annotations were added using SnpEff v.5 (65) against the Sscrofa.11.1 reference database. CNV prediction was performed with Control-FREEC 11.5 (24), using a pool of samples as CNV baseline and using intervals of 20 kb. CNV calls from all samples that were less than 10 kb apart were merged with Survivor (66). Individual CNV calls were combined into copy number variant regions (CNVRs) following the reciprocal overlap approach (11) with CNVRange (67). Therefore, contiguous CNVs intervals with at least 50% of mutual overlap were merged into the same CNVR.

Nucleotide diversity pattern estimates per individual were calculated considering the SNPs present in each individual separately. Here, we tested two estimators: (i) Tajima's theta estimator (π or nucleotide diversity, called ATajima here), that is simply the number of variants present in the individual; and (ii) RTajima estimator, which considers the frequency observed in the entire sample but was calculated given the SNPs present in each individual. This estimate is calculated with the same principles that are in ATajima but need to be corrected by the probability that only a portion of the total SNPs from a sample is present in each of the samples (by using a hypergeometrical distribution). That is, for each individual, this is an estimate of the nucleotide diversity of the entire population if the individuals behave as if they belonged in a stationary neutral population. Specifically, for this new variability estimate expression, the calculation for a single diploid individual is as follows:

$$\hat{\theta} = \frac{1}{\sum_{i=1}^{n-1}(n-i)} \sum_{i=1}^{n-1}(n-i)i\xi_i \left(1 - \frac{\binom{n-i}{2}}{\binom{n}{2}}\right)^{-1}$$

θ is the estimate of population variability using a single individual, n is the number of samples in the population (2× number of individuals), and ξ_i is the number of SNPs observed in this individual that are at frequency i in the whole population. Note that this expression is constructed using an approach (68) to estimate nucleotide diversity (left-side expression after equality) and the hypergeometric correction (right-side expression).

The data set in VCF format was converted to FASTA and from FASTA to transposed FASTA (tFASTA). This tFASTA file was read with the software mstatspop (https://github.com/CRAGENOMICA/mstatspop) to obtain the frequencies of SNPs from each of the pigs at the desired region. Finally, we calculated

RTajima per fragment using self-made R scripts. All these estimates were finally divided by the effective length size of the studied region to obtain comparative estimates per nucleotide.

CNV-wide association analysis. A genome-wide association analysis between the estimated CNV status and fecal bacterial diversity index was done with GCTA (69) using the following mixed model:

$$y_{ijk} = \mathrm{sex}_j + b_k + u_i + \mathrm{cnv}_{li} + e_{ijk}$$

where y_{ijk} corresponds to the microbial index under scrutiny (richness or Shannon α-diversity) of the ith individual animal of sex j in the kth batch; sex_j and b_{ki} correspond to the systematic effects of jth sex (two levels) and kth batch (three levels), respectively; u_i is the random additive genetic effect of the ith individual, collectively distributed as $u \sim N (0, G \; \sigma_u^2)$ where σ_u^2 is the additive genetic variance, and G is the numerator of the genomic relationship matrix calculated using the autosomal SNPs; cnv_{li} is the genotype (recoded as 11 = loss, 12 = diploid, and 22 = gain) for the lth CNV of the ith individual, and e_{ijk} is the residual. To correct for multiple testing, the FDR was calculated with the *p.adjust* function of R.

Quantitative real-time PCR. Real-time qPCR was used to validate the CNV on the *ABCC2* gene in a total of 72 samples, including a subset of 48 Duroc samples (24 diploid and 24 *in silico* predicted as DUP), and 24 unrelated F1 Duroc × Iberian cross (Table 1). The CNV breakpoint was re-estimated with Manta version 1.6.0 (70). All primers were designed using the Primer Express 2.0 software (Applied Biosystems). The pair of primers ABCC2_CNV_F (5′-TGGCATCATTTATGTGGCTGTT-3′) and ABCC2_CNV_R (5′-AGGAAGGAGCTTGGGCTTTTA-3′) amplify a specific region of the *ABCC2* gene (exon 25-intron 25 of transcript ABCC2-201 with Ensembl ID ENSSSCT00000011534.5) containing the CNV, while the pair of primers ABCC2_F (5′-TGGACAAGAACCAGAGTCAAAGC-3′) and ABCC2_R (5′-ACATAGAGCGCATTTGAACGAA-3′) amplify a region outside the estimated CNV (exon 7-intron 7 of transcript ABCC2-201 with Ensembl ID ENSSSCT00000011534.5) breakpoint that was used as single copy control region. The $2 - \Delta\Delta Ct$ method for RQ of CNVs was used as previously described (14). qPCRs were carried out using SYBR green chemistry (SYBRTM Select Master Mix, Applied Biosystems) and the instruments ABI PRISM 7900HT and 7500 real-time PCR system (Applied Biosystems, Inc., Foster City, CA). The reactions were carried out in a 20-μL volume containing 10 ng of genomic DNA. All primers were used at 900 nM. The thermal cycle was 10 min at 95°C, 40 cycles of 15 s at 95°C, and 1 min at 60°C. Each sample was analyzed in quadruplicate. PCR efficiencies (<95%) were evaluated with standard curves, and dissociation curves were drawn for each primer pair to assess for the specificity of the PCRs. Three samples without CNV were used as reference. The results were analyzed with Thermo Fisher Cloud software 1.0 (Applied Biosystems) and qBase Plus v3.2 (Biogazelle). Pearson correlation was employed to evaluate the relationship between the RQ and the estimated nucleotide diversity patterns.

Identification of microbial signatures. The identification of ASVs that discriminate samples according to the number of copies of the *ABCC2-DNMBP* loci was performed based on the compositional kernel as implemented the function "classify" of the kernInt R package (71). The classify function runs a supervised classification model based on Support Vector Machine. For that purpose, the available data set was split at random into a training set (80% of data) and a validation set (20%). The *C* hyperparameter's optimal value was obtained by 10 × 10 cross-validation on the training set. To estimate the mean classification accuracy, the classify function was run 10 times using different training/test splits of the data set. Microbial signatures were obtained from the hyperplane vector$w$, and the importance of the ASV was computed by *kernInt*as(wk)² (72). Initially, the top 5% relevant taxa were retained, but a conservative approach was applied afterwards, keeping for subsequent analyses only the ASVs reported as relevant in at least 50% of the replicates. Finally, to identify over-representation at the genus level, the list of selected features was submitted to a taxa-set enrichment analysis (73).

Data availability. The raw whole-genome sequencing data have been submitted to the European Nucleotide Archive: ERA19192875. The raw 16S rRNA sequencing data employed in this article has been submitted to the NCBI sequence read archive (https://www.ncbi.nlm.nih.gov/sra) and BioProject: PRJNA608629, and Table S3 includes the corresponding individual biosample accession number under the PRJNA608629 BioProject.

SUPPLEMENTAL MATERIAL

Supplemental material is available online only.

SUPPLEMENTAL FILE 1, PDF file, 0.2 MB.

REFERENCES

1. Ramayo-Caldas Y, Mach N, Lepage P, Levenez F, Denis C, Lemonnier G, Leplat J-J, Billon Y, Berri M, Doré J, Rogel-Gaillard C, Estellé J. 2016. Phylogenetic network analysis applied to pig gut microbiota identifies an ecosystem structure linked with growth traits. ISME J 10:2973–2977. https://doi.org/10.1038/ismej.2016.77.

2. Camarinha-Silva A, Maushammer M, Wellmann R, Vital M, Preuss S, Bennewitz J. 2017. Host genome influence on gut microbial composition and microbial prediction of complex traits in pigs. Genetics 206:1637–1644. https://doi.org/10.1534/genetics.117.200782.

3. McCormack UM, Curião T, Metzler-Zebeli BU, Magowan E, Berry DP, Reyer H, Prieto ML, Buzoianu SG, Harrison M, Rebeiz N, Crispie F, Cotter PD, O'Sullivan O, Gardiner GE, Lawlor PG. 2019. Porcine feed efficiency-associated intestinal microbiota and physiological traits: finding consistent cross-locational biomarkers for residual feed intake. mSystems 4:e00324-18. https://doi.org/10.1128/mSystems.00324-18.

4. Ramayo-Caldas Y, Zingaretti LM, Pérez-Pascual D, Alexandre PA, Reverter A, Dalmau A, Quintanilla R, Ballester M. 2021. Leveraging host-genetics and gut microbiota to determine immunocompetence in pigs. Anim Microbiome 3:74. https://doi.org/10.1186/s42523-021-00138-9.

5. Bergamaschi M, Maltecca C, Schillebeeckx C, McNulty NP, Schwab C, Shull C, Fix J, Tiezzi F. 2020. Heritability and genome-wide association of swine gut microbiome features with growth and fatness parameters. Sci Rep 10:10134. https://doi.org/10.1038/s41598-020-66791-3.

6. Aliakbari A, Zemb O, Cauquil L, Barilly C, Billon Y, Gilbert H. 2022. Microbiability and microbiome-wide association analyses of feed efficiency and performance traits in pigs. Genet Sel Evol 54:29. https://doi.org/10.1186/s12711-022-00717-7.

7. Crespo-Piazuelo D, Migura-Garcia L, Estellé J, Criado-Mesas L, Revilla M, Castelló A, Muñoz M, García-Casco JM, Fernández AI, Ballester M, Folch JM. 2019. Association between the pig genome and its gut microbiota composition. Sci Rep 9:8791. https://doi.org/10.1038/s41598-019-45066-6.

8. Ramayo-Caldas Y, Prenafeta-Boldú F, Zingaretti LM, Gonzalez-Rodriguez O, Dalmau A, Quintanilla R, Ballester M. 2020. Gut eukaryotic communities in pigs: diversity, composition and host genetics contribution. Anim Microbiome 2:18. https://doi.org/10.1186/s42523-020-00038-4.

9. Reverter A, Ballester M, Alexandre PA, Mármol-Sánchez E, Dalmau A, Quintanilla R, Ramayo-Caldas Y. 2021. A gene co-association network regulating gut microbial communities in a Duroc pig population. Microbiome 9:52. https://doi.org/10.1186/s40168-020-00994-8.

10. Ryu EP, Davenport ER. 2022. Host genetic determinants of the microbiome across animals: from *Caenorhabditis elegans* to cattle. Annu Rev Anim Biosci 10:203–226. https://doi.org/10.1146/annurev-animal-020420-032054.

11. Conrad DF, Pinto D, Redon R, Feuk L, Gokcumen O, Zhang Y, Aerts J, Andrews TD, Barnes C, Campbell P, Fitzgerald T, Hu M, Ihm CH, Kristiansson K, Macarthur DG, Macdonald JR, Onyiah I, Pang AWC, Robson S, Stirrups K, Valsesia A, Walter K, Wei J, Tyler-Smith C, Carter NP, Lee C, Scherer SW, Hurles ME, Wellcome Trust Case Control Consortium. 2010. Origins and functional impact of copy number variation in the human genome. Nature 464:704–712. https://doi.org/10.1038/nature08516.

12. Clop A, Vidal O, Amills M. 2012. Copy number variation in the genomes of domestic animals. Anim Genet 43:503–517. https://doi.org/10.1111/j.1365-2052.2012.02317.x.

13. Poole AC, Goodrich JK, Youngblut ND, Luque GG, Ruaud A, Sutter JL, Waters JL, Shi Q, El-Hadidi M, Johnson LM, Bar HY, Huson DH, Booth JG, Ley RE. 2019. Human salivary amylase gene copy number impacts oral and gut microbiomes. Cell Host Microbe 25:553–564.e7. https://doi.org/10.1016/j.chom.2019.03.001.

14. Ramayo-Caldas Y, Castelló A, Pena RN, Alves E, Mercadé A, Souza CA, Fernández AI, Perez-Enciso M, Folch JM. 2010. Copy number variation in the porcine genome inferred from a 60 k SNP BeadChip. BMC Genomics 11:593. https://doi.org/10.1186/1471-2164-11-593.

15. Gong H, Liu W, Wu Z, Zhang M, Sun Y, Ling Z, Xiao S, Ai H, Xing Y, Yang B, Huang L. 2022. Evolutionary insights into porcine genomic structural variations based on a novel-constructed dataset from 24 worldwide diverse populations. Evol Appl 15:1264–1280. https://doi.org/10.1111/eva.13455.

16. Paudel Y, Madsen O, Megens H-J, Frantz LAF, Bosse M, Crooijmans RPMA, Groenen MAM. 2015. Copy number variation in the speciation of pigs: a possible prominent role for olfactory receptors. BMC Genomics 16:330. https://doi.org/10.1186/s12864-015-1449-9.

17. Giuffra E, Evans G, Törnsten A, Wales R, Day A, Looft H, Plastow G, Andersson L. 1999. The belt mutation in pigs is an allele at the dominant white (I/KIT) locus. Mamm Genome 10:1132–1136. https://doi.org/10.1007/s003359901178.

18. Giuffra E, Törnsten A, Marklund S, Bongcam-Rudloff E, Chardon P, Kijas JMH, Anderson SI, Archibald AL, Andersson L. 2002. A large duplication associated with dominant white color in pigs originated by homologous recombination between LINE elements flanking KIT. Mamm Genome 13:569–577. https://doi.org/10.1007/s00335-002-2184-5.

19. Revilla M, Puig-Oliveras A, Castelló A, Crespo-Piazuelo D, Paludo E, Fernández AI, Ballester M, Folch JM. 2017. A global analysis of CNVs in swine using whole genome sequence data and association analysis with fatty acid composition and growth traits. PLoS One 12:e0177014. https://doi.org/10.1371/journal.pone.0177014.

20. Zheng X, Zhao P, Yang K, Ning C, Wang H, Zhou L, Liu J. 2020. CNV analysis of Meishan pig by next-generation sequencing and effects of AHR gene CNV on pig reproductive traits. J Animal Sci Biotechnol 11:42. https://doi.org/10.1186/s40104-020-00442-5.

21. Ran X-Q, Pan H, Huang S-H, Liu C, Niu X, Li S, Wang J-F. 2018. Copy number variations of MTHFSD gene across pig breeds and its association with litter size traits in Chinese indigenous Xiang pig. J Anim Physiol Anim Nutr 102:1320–1327. https://doi.org/10.1111/jpn.12922.

22. Qiu Y, Ding R, Zhuang Z, Wu J, Yang M, Zhou S, Ye Y, Geng Q, Xu Z, Huang S, Cai G, Wu Z, Yang J. 2021. Genome-wide detection of CNV regions and their potential association with growth and fatness traits in Duroc pigs. BMC Genomics 22:332. https://doi.org/10.1186/s12864-021-07654-7.

23. Yalcin B, Wong K, Agam A, Goodson M, Keane TM, Gan X, Nellåker C, Goodstadt L, Nicod J, Bhomra A, Hernandez-Pliego P, Whitley H, Cleak J, Dutton R, Janowitz D, Mott R, Adams DJ, Flint J. 2011. Sequence-based

characterization of structural variation in the mouse genome. Nature 477: 326–329. https://doi.org/10.1038/nature10432.

24. Boeva V, Popova T, Bleakley K, Chiche P, Cappo J, Schleiermacher G, Janoueix-Lerosey I, Delattre O, Barillot E. 2012. Control-FREEC: a tool for assessing copy number and allelic content using next-generation sequencing data. Bioinformatics 28:423–425. https://doi.org/10.1093/bioinformatics/btr670.

25. Nies AT, Keppler D. 2007. The apical conjugate efflux pump ABCC2 (MRP2). Pflugers Arch 453:643–659. https://doi.org/10.1007/s00424-006-0109-y.

26. Jemnitz K, Heredi-Szabo K, Janossy J, Ioja E, Vereczkey L, Krajcsi P. 2010. ABCC2/Abcc2: a multispecific transporter with dominant excretory functions. Drug Metab Rev 42:402–436. https://doi.org/10.3109/03602530903491741.

27. Klaassen CD, Aleksunes LM. 2010. Xenobiotic, bile acid, and cholesterol transporters: function and regulation. Pharmacol Rev 62:1–96. https://doi.org/10.1124/pr.109.002014.

28. Ramayo-Caldas Y, Ballester M, Fortes MRS, Esteve-Codina A, Castelló A, Noguera JL, Fernández AI, Pérez-Enciso M, Reverter A, Folch JM. 2014. From SNP co-association to RNA co-expression: novel insights into gene networks for intramuscular fatty acid composition in porcine. BMC Genomics 15:232. https://doi.org/10.1186/1471-2164-15-232.

29. Zhang W, Zhang J, Cui L, Ma J, Chen C, Ai H, Xie X, Li L, Xiao S, Huang L, Ren J, Yang B. 2016. Genetic architecture of fatty acid composition in the longissimus dorsi muscle revealed by genome-wide association studies on diverse pig populations. Genet Sel Evol 48:5. https://doi.org/10.1186/s12711-016-0184-2.

30. Chu X-Y, Strauss JR, Mariano MA, Li J, Newton DJ, Cai X, Wang RW, Yabut J, Hartley DP, Evans DC, Evers R. 2006. Characterization of mice lacking the multidrug resistance protein MRP2 (ABCC2). J Pharmacol Exp Ther 317:579–589. https://doi.org/10.1124/jpet.105.098665.

31. Vlaming MLH, Mohrmann K, Wagenaar E, de Waart DR, Elferink RPJO, Lagas JS, van Tellingen O, Vainchtein LD, Rosing H, Beijnen JH, Schellens JHM, Schinkel AH. 2006. Carcinogen and anticancer drug transport by Mrp2 in vivo: studies using Mrp2 (Abcc2) knockout mice. J Pharmacol Exp Ther 318:319–327. https://doi.org/10.1124/jpet.106.101774.

32. Jansen PL, Peters WH, Lamers WH. 1985. Hereditary chronic conjugated hyperbilirubinemia in mutant rats caused by defective hepatic anion transport. Hepatology 5:573–579. https://doi.org/10.1002/hep.1840050408.

33. Dubin IN, Johnson FB. 1954. Chronic idiopathic jaundice with unidentified pigment in liver cells; a new clinicopathologic entity with a report of 12 cases. Medicine 33:155–197. https://doi.org/10.1097/00005792-195409000-00001.

34. Dixon PH, Sambrotta M, Chambers J, Taylor-Harris P, Syngelaki A, Nicolaides K, Knisely AS, Thompson RJ, Williamson C. 2017. An expanded role for heterozygous mutations of ABCB4, ABCB11, ATP8B1, ABCC2 and TJP2 in intrahepatic cholestasis of pregnancy. Sci Rep 7:11823. https://doi.org/10.1038/s41598-017-11626-x.

35. Nobles CL, Green SI, Maresso AW. 2013. A product of heme catabolism modulates bacterial function and survival. PLoS Pathog 9:e1003507. https://doi.org/10.1371/journal.ppat.1003507.

36. Ridlon JM, Harris SC, Bhowmik S, Kang D-J, Hylemon PB. 2016. Consequences of bile salt biotransformations by intestinal bacteria. Gut Microbes 7: 22–39. https://doi.org/10.1080/19490976.2015.1127483.

37. Li DK, Chaudhari SN, Lee Y, Sojoodi M, Adhikari AA, Zukerberg L, Shroff S, Barrett SC, Tanabe K, Chung RT, Devlin AS. 2022. Inhibition of microbial deconjugation of micellar bile acids protects against intestinal permeability and liver injury. Sci Adv 8:eabo2794. https://doi.org/10.1126/sciadv.abo2794.

38. van Best N, Rolle-Kampczyk U, Schaap FG, Basic M, Olde Damink SWM, Bleich A, Savelkoul PHM, von Bergen M, Penders J, Hornef MW. 2020. Bile acids drive the newborn's gut microbiota maturation. Nat Commun 11: 3692. https://doi.org/10.1038/s41467-020-17183-8.

39. Molinero N, Ruiz L, Sánchez B, Margolles A, Delgado S. 2019. Intestinal bacteria interplay with bile and cholesterol metabolism: implications on host physiology. Front Physiol 10:185. https://doi.org/10.3389/fphys.2019.00185.

40. Song Z, Cai Y, Lao X, Wang X, Lin X, Cui Y, Kalavagunta PK, Liao J, Jin L, Shang J, Li J. 2019. Taxonomic profiling and populational patterns of bacterial bile salt hydrolase (BSH) genes based on worldwide human gut microbiome. Microbiome 7:9. https://doi.org/10.1186/s40168-019-0628-3.

41. Shapiro H, Kolodziejczyk AA, Halstuch D, Elinav E. 2018. Bile acids in glucose metabolism in health and disease. J Exp Med 215:383–396. https://doi.org/10.1084/jem.20171965.

42. Cipriani S, Mencarelli A, Chini MG, Distrutti E, Renga B, Bifulco G, Baldelli F, Donini A, Fiorucci S. 2011. The bile acid receptor GPBAR-1 (TGR5) modulates integrity of intestinal barrier and immune response to experimental colitis. PLoS One 6:e25637. https://doi.org/10.1371/journal.pone.0025637.

43. Otani T, Ichii T, Aono S, Takeichi M. 2006. Cdc42 GEF Tuba regulates the junctional configuration of simple epithelial cells. J Cell Biol 175:135–146. https://doi.org/10.1083/jcb.200605012.

44. Qin Y, Meisen WH, Hao Y, Macara IG. 2010. Tuba, a Cdc42 GEF, is required for polarized spindle orientation during epithelial cyst formation. J Cell Biol 189:661–669. https://doi.org/10.1083/jcb.201002097.

45. Ubelmann F, Chamaillard M, El-Marjou F, Simon A, Netter J, Vignjevic D, Nichols BL, Quezada-Calvillo R, Grandjean T, Louvard D, Revenu C, Robine S. 2013. Enterocyte loss of polarity and gut wound healing rely upon the F-actin-severing function of villin. Proc Natl Acad Sci U S A 110:E1380–E1389. https://doi.org/10.1073/pnas.1218446110.

46. Bennet KM, Walker SL, Lo DD. 2014. Epithelial microvilli establish an electrostatic barrier to microbial adhesion. Infect Immun 82:2860–2871. https://doi.org/10.1128/IAI.01681-14.

47. Ferrary E, Cohen-Tannoudji M, Pehau-Arnaudet G, Lapillonne A, Athman R, Ruiz T, Boulouha L, El Marjou F, Doye A, Fontaine JJ, Antony C, Babinet C, Louvard D, Jaisser F, Robine S. 1999. In vivo, villin is required for Ca^{2+}-dependent F-actin disruption in intestinal brush borders. J Cell Biol 146: 819–830. https://doi.org/10.1083/jcb.146.4.819.

48. Wells JM, Rossi O, Meijerink M, van Baarlen P. 2011. Epithelial crosstalk at the microbiota-mucosal interface. Proc Natl Acad Sci U S A 108: 4607–4614. https://doi.org/10.1073/pnas.1000092107.

49. Gieryńska M, Szulc-Dąbrowska L, Struzik J, Mielcarska MB, Gregorczyk-Zboroch KP. 2022. Integrity of the intestinal barrier: the involvement of epithelial cells and microbiota—a mutual relationship. Animals 12:145. https://doi.org/10.3390/ani12020145.

50. Willing BP, van Kessel AG. 2007. Enterocyte proliferation and apoptosis in the caudal small intestine is influenced by the composition of colonizing commensal bacteria in the neonatal gnotobiotic pig. J Anim Sci 85: 3256–3266. https://doi.org/10.2527/jas.2007-0320.

51. Shirkey TW, Siggers RH, Goldade BG, Marshall JK, Drew MD, Laarveld B, Van Kessel AG. 2006. Effects of commensal bacteria on intestinal morphology and expression of proinflammatory cytokines in the gnotobiotic pig. Exp Biol Med 231:1333–1345. https://doi.org/10.1177/153537020623100807.

52. Danielsen M, Hornshøj H, Siggers RH, Jensen BB, van Kessel AG, Bendixen E. 2007. Effects of bacterial colonization on the porcine intestinal proteome. J Proteome Res 6:2596–2604. https://doi.org/10.1021/pr070038b.

53. Kozakova H, Kolinska J, Lojda Z, Rehakova Z, Sinkora J, Zakostelecka M, Splichal I, Tlaskalova-Hogenova H. 2006. Effect of bacterial monoassociation on brush-border enzyme activities in ex-germ-free piglets: comparison of commensal and pathogenic Escherichia coli strains. Microbes Infect 8:2629–2639. https://doi.org/10.1016/j.micinf.2006.07.008.

54. Ballester M, Ramayo-Caldas Y, González-Rodríguez O, Pascual M, Reixach J, Díaz M, Blanc F, López-Serrano S, Tibau J, Quintanilla R. 2020. Genetic parameters and associated genomic regions for global immunocompetence and other health-related traits in pigs. Sci Rep 10:18462. https://doi.org/10.1038/s41598-020-75417-7.

55. Bolyen E, Rideout JR, Dillon MR, Bokulich NA, Abnet CC, Al-Ghalith GA, Alexander H, Alm EJ, Arumugam M, Asnicar F, Bai Y, Bisanz JE, Bittinger K, Brejnrod A, Brislawn CJ, Brown CT, Callahan BJ, Caraballo-Rodríguez AM, Chase J, Cope EK, Da Silva R, Diener C, Dorrestein PC, Douglas GM, Durall DM, Duvallet C, Edwardson CF, Ernst M, Estaki M, Fouquier J, Gauglitz JM, Gibbons SM, Gibson DL, Gonzalez A, Gorlick K, Guo J, Hillmann B, Holmes S, Holste H, Huttenhower C, Huttley GA, Janssen S, Jarmusch AK, Jiang L, Kaehler BD, Kang KB, Keefe CR, Keim P, Kelley ST, Knights D, 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.

56. McDonald D, Jiang Y, Balaban M, Cantrell K, Zhu Q, Gonzalez A, Morton JT, Nicolaou G, Parks DH, Karst S, Albertsen M, Hugenholtz P, DeSantis T, Mirarab S, Knight R. 2022. Greengenes2 enables a shared data universe for microbiome studies. bioRxiv. https://doi.org/10.1101/2022.12.19.520774.

57. DeSantis TZ, Hugenholtz P, Larsen N, Rojas M, Brodie EL, Keller K, Huber T, Dalevi D, Hu P, Andersen GL. 2006. Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. Appl Environ Microbiol 72:5069–5072. https://doi.org/10.1128/AEM.03006-05.

58. Oksanen J, Blanchet G, Kindt R, Legendre P, Minchin P, O'Hara RB, Solymos P, Stevens MHH, Szoecs E, Wagner H, Barbour M, Bedward M, Bolker B, Borcard D, Carvalho G, Chirico M, FitzJohn R, Friendly M, Furneaux B, Hannigan G, Hill MO, Lahti L, McGlinn D, Ouellette M-H,

Cunha ER, Smith T, Stier A, Ter Braak CJF, Weedon J. 2009. Vegan: community ecology package. R package version 2.0-9. http://CRAN.R-project.org/package=vegan.

59. Shannon C. 1948. A mathematical theory of communication. Bell Syst Tech J 27:379–423, 623–656. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x.

60. Whittaker RH. 1972. Evolution and measurement of species diversity. Taxon 21:213–251. https://doi.org/10.2307/1218190.

61. Patil I. 2021. Visualizations with statistical details: the "ggstatsplot" approach. J Open Source Softw 6:3167. https://doi.org/10.21105/joss.03167.

62. Li H. 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv 1303.3997v2.

63. Broad Institute. Picard toolkit. 2018.

64. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA. 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res 20:1297–1303. https://doi.org/10.1101/gr.107524.110.

65. Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, Land SJ, Lu X, Ruden DM. 2012. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. Fly 6:80–92. https://doi.org/10.4161/fly.19695.

66. Jeffares DC, Jolly C, Hoti M, Speed D, Shaw L, Rallis C, Balloux F, Dessimoz C, Bähler J, Sedlazeck FJ. 2017. Transient structural variations have strong effects on quantitative traits and reproductive isolation in fission yeast. Nat Commun 8:14061. https://doi.org/10.1038/ncomms14061.

67. da Silva V, Ramos M, Groenen M, Crooijmans R, Johansson A, Regitano L, Coutinho L, Zimmer R, Waldron L, Geistlinger L. 2020. CNVRanger: association analysis of CNVs with gene expression and quantitative phenotypes. Bioinformatics 36:972–973. https://doi.org/10.1093/bioinformatics/btz632.

68. Achaz G. 2009. Frequency spectrum neutrality tests: one for all and all for one. Genetics 183:249–258. https://doi.org/10.1534/genetics.109.104042.

69. Yang J, Lee SH, Goddard ME, Visscher PM. 2011. GCTA: a tool for genome-wide complex trait analysis. Am J Hum Genet 88:76–82. https://doi.org/10.1016/j.ajhg.2010.11.011.

70. Chen X, Schulz-Trieglaff O, Shaw R, Barnes B, Schlesinger F, Källberg M, Cox AJ, Kruglyak S, Saunders CT. 2016. Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. Bioinformatics 32:1220–1222. https://doi.org/10.1093/bioinformatics/btv710.

71. Ramon E, Belanche-Muñoz L, Molist F, Quintanilla R, Perez-Enciso M, Ramayo-Caldas Y. 2021. A kernel framework for integrating supervised and unsupervised analyses in spatio-temporal metagenomic datasets. Front Microbiol 12:609048. https://doi.org/10.3389/fmicb.2021.609048.

72. Guyon I, Weston J, Barnhill S, Vapnik V. 2002. Gene selection for cancer classification using support vector machines. Mach Learn 46:389–422. https://doi.org/10.1023/A:1012487302797.

73. Zingaretti LM, Renand G, Morgavi DP, Ramayo-Caldas Y. 2020. Link-HD: a versatile framework to explore and integrate heterogeneous microbial communities. Bioinformatics 36:2298–2299. https://doi.org/10.1093/bioinformatics/btz862.