Article

ABO genotype alters the gut microbiota by regulating GalNAc levels in pigs

https://doi.org/10.1038/s41586-022-04769-z

Received: 29 July 2020

Accepted: 19 April 2022

Published online: 27 April 2022

Check for updates

Hui Yang[1,3], Jinyuan Wu[1,3], Xiaochang Huang[1], Yunyan Zhou[1], Yifeng Zhang[1], Min Liu[1], Qin Liu[1], Shanlin Ke[1], Maozhang He[1], Hao Fu[1], Shaoming Fang[1], Xinwei Xiong[1], Hui Jiang[1], Zhe Chen[1], Zhongzi Wu[1], Huanfa Gong[1], Xinkai Tong[1], Yizhong Huang[1], Junwu Ma[1], Jun Gao[1], Carole Charlier[1,2], Wouter Coppieters[2], Lev Shagam[2], Zhiyan Zhang[1], Huashui Ai[1], Bin Yang[1], Michel Georges[1,2,4]✉, Congying Chen[1,4]✉ & Lusheng Huang[1,4]✉

The composition of the intestinal microbiome varies considerably between individuals and is correlated with health[1]. Understanding the extent to which, and how, host genetics contributes to this variation is essential yet has proved to be difficult, as few associations have been replicated, particularly in humans[2]. Here we study the effect of host genotype on the composition of the intestinal microbiota in a large mosaic pig population. We show that, under conditions of exacerbated genetic diversity and environmental uniformity, microbiota composition and the abundance of specific taxa are heritable. We map a quantitative trait locus affecting the abundance of Erysipelotrichaceae species and show that it is caused by a 2.3 kb deletion in the gene encoding *N*-acetyl-galactosaminyl-transferase that underpins the ABO blood group in humans. We show that this deletion is a ≥3.5-million-year-old trans-species polymorphism under balancing selection. We demonstrate that it decreases the concentrations of *N*-acetyl-galactosamine in the gut, and thereby reduces the abundance of Erysipelotrichaceae that can import and catabolize *N*-acetyl-galactosamine. Our results provide very strong evidence for an effect of the host genotype on the abundance of specific bacteria in the intestine combined with insights into the molecular mechanisms that underpin this association. Our data pave the way towards identifying the same effect in rural human populations.

It is increasingly recognized that a comprehensive understanding of the physiology and pathology of organisms requires integrated analysis of the host and its multiple microbiota[1]. In humans, the composition of the intestinal microbiota is associated with physiological and pathological parameters, including HDL cholesterol, fasting glucose levels and body mass index[2]. In livestock, ruminal microbiome composition is associated with methane production and feed efficiency[3]. These correlations reflect a complex interplay between host and microbiota that may include direct (causal) effects of the microbiome on the host's physiology[4]. Several phenotypes correlated with microbiota composition are heritable[5,6]. This leads to the hypothesis that the genotype of the host may in part determine the composition of the microbiota, which may in turn affect the host's phenotype[4]. It implies that the composition of the microbiota is partially heritable. Although studies in rodents support this[7], evidence is less convincing in humans. Initial reports did not reveal a higher microbiota resemblance between monozygotic twins compared with dizygotic twins, suggesting a limited effect of the host genotype[8]. Better-powered studies provided evidence for a significant effect of host genetics on the abundance of taxa, particularly Christensenellaceae[9]. Loci that underpin microbiota heritability have remained difficult to identify in humans. Apart from variants that cause persistent expression of lactase (*LCT*) and are associated with decreased *Bifidobacterium* abundance, other GWAS loci have proven to be difficult to replicate[2,10–14]. The analysis of larger human cohorts is needed to gain a better understanding of the genetic architecture of microbiota composition.

To decipher the genetic architecture of intestinal microbiota composition in a large monogastric omnivore, we report the generation of a mosaic pig population and the longitudinal characterization of its intestinal microbiota. We observed a strong effect of the host genotype on microbiota composition and identified a locus with large effect on the abundance of specific taxa by controlling the concentration of *N*-acetyl-galactosamine in the gut and thereby affecting some of the species that use this metabolite as a carbon source.

A mosaic pig population to study complex traits

We generated a large (>7,500) mosaic population by intercrossing the offspring of 61 F_0 founders from four Chinese and four western breeds for more than 10 generations (Supplementary Table 1 and Extended Data Fig. 1). Animals were reared in uniform housing and feeding conditions. We analysed more than 200 phenotypes (pertaining to body

[1]National Key Laboratory for Swine Genetic Improvement and Production Technology, Ministry of Science and Technology of China, Jiangxi Agricultural University, Nanchang, PR China. [2]Unit of Animal Genomics, GIGA-Institute and Faculty of Veterinary Medicine, University of Liege, Liege, Belgium. [3]These authors contributed equally: Hui Yang, Jinyuan Wu. [4]These authors jointly supervised this work: Michel Georges, Congying Chen, Lusheng Huang. ✉e-mail: michel.georges@uliege.be; chcy75@hotmail.com; lushenghuang@hotmail.com

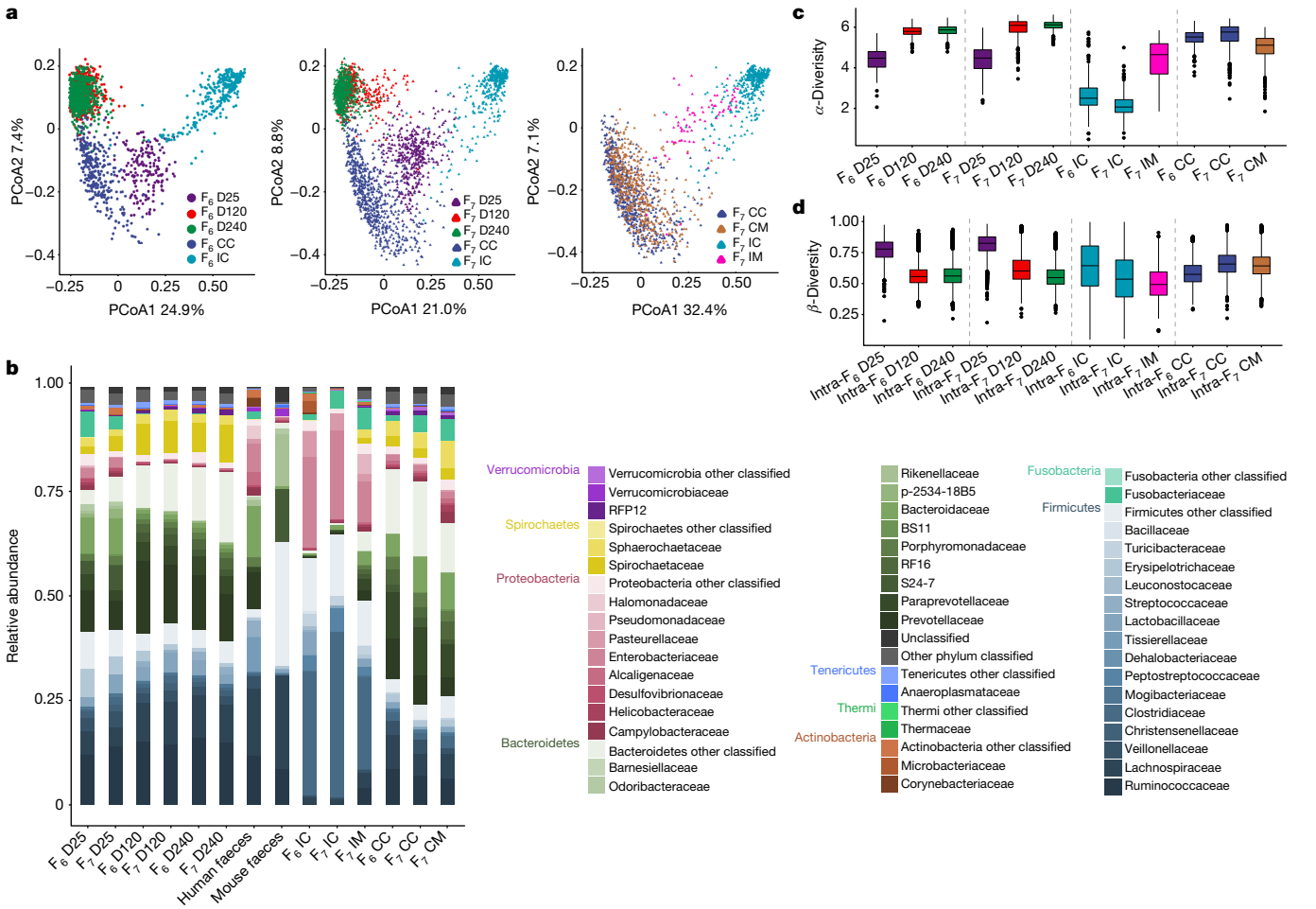


Fig. 1 | Intestinal microbiota of the healthy pig. a, Joint principal coordinate analysis (PCoA) of 5,110 16S rRNA profiles. F_6 day 25 faeces (D25, mauve), day 120 faeces (D120, red), day 240 faeces (D240, green), ileal content (IC, light blue), caecum content (CC, dark blue) (left). Middle, as described for the left plot but for F_7. Right, F_7 ileal content (IC, light blue), caecum content (CC, dark blue), ileal mucosa (IM, pink), caecal mucosa (CM, brown). **b**, The average microbiota composition of the 12 data series. Taxa are coloured by phylum and family within phylum, highlighting 43 families among the top 15 in at least one data series. The names of the corresponding phyla and families are shown in the key. The average composition of 106 human faeces and 6 mouse faeces (C57BL/6) samples is shown. **c**, α-Diversity values (Shannon's index) for the 12 data series coloured as in **a**. Sample sizes are provided in Supplementary Table 2.1. The box plots show the median (centre line), interquartile range (box limits), 1.5× the interquartile range span (whiskers) and outliers (dots). **d**, β-Diversity values (pair-wise Bray–Curtis distances) for the 12 data series coloured as in **a**. Distances were computed for all sample pairs. Sample numbers and box plots are as described in **c**.

composition, physiology, disease resistance and behaviour), obtained transcriptome, epigenome and chromatin interaction data from multiple tissues, and collected plasma metabolome and microbiome data in up to 954 F_6 and 892 F_7 animals. The F_0 animals were whole-genome sequenced at an average depth of 28.4-fold, and the F_6 and F_7 animals were sequence at an average depth of 8.0-fold. We called genotypes at 23.8 million single-nucleotide polymorphisms (SNPs) and 6.4 million insertion–deletions (indels) with a minor allele frequency (MAF) of ≥0.03 (>1/100 bp). The nucleotide diversity (π) (that is, the proportion of nucleotide sites that differ between homologous sequences in two breeds) between two Chinese breeds and between two European breeds was similar to that between *Homo sapiens* and *Homo neanderthalensis* ($\sim 3 \times 10^{-3}$)[15], whereas the π between a Chinese and a European breed approached half of that between human and chimpanzee ($\sim 4.3 \times 10^{-3}$)[16]. The proportion of the eight founder genomes in F_6 and F_7 ranged from 11.2% to 14.7% at the genome level, and from 4.9% to 22.1% at the chromosome level. The median number of variants in high linkage disequilibrium (LD) ($r^2 \geq 0.9$) with an index variant was 30, and the median maximal distance with a variant in high LD ($r^2 \geq 0.9$) was 54 kb (Extended Data Fig. 1).

The intestinal microbiota of the healthy pig

We collected faeces at days 25 (suckling period), 120 (growing period) and 240 (slaughter day), as well as caecal and ileal content (F_6 and F_7) and caecal and ileal mucosal scrapings (F_7 only) at day 240 (7 sample types, 12 data series). We generated 16S rRNA tags (V3–V4) for an average of 426 animals per data series (5,110 samples) (Supplementary Table 2.1). Tags were rarefied and clustered in 32,032 operational taxonomic units (OTUs). A total of 12,054 OTUs amounting to 98.7% of reads were retained and annotated to 41 phyla, 87 classes, 149 orders, 207 families, 360 genera and 150 species. The average microbiota composition of the 12 data series indicated high consistency across F_6 and F_7, yet marked differences between sample types (Fig. 1a, b and Supplementary Table 2.2). Even at the family level, some taxa were found to be nearly sample-type specific (Extended Data Fig. 2). The proteobacteria Enterobacteriaceae, Pseudomonadaceae and Pasteurellaceae, the firmicutes Clostridiaceae, Peptostreptococcaceae, Bacillaceae and Leuconostocaceae, and the actinobacteria Microbacteriaceae were at least ten times more abundant in ileal samples than in any other sample type. Among those, Leuconostocaceae were nearly digesta

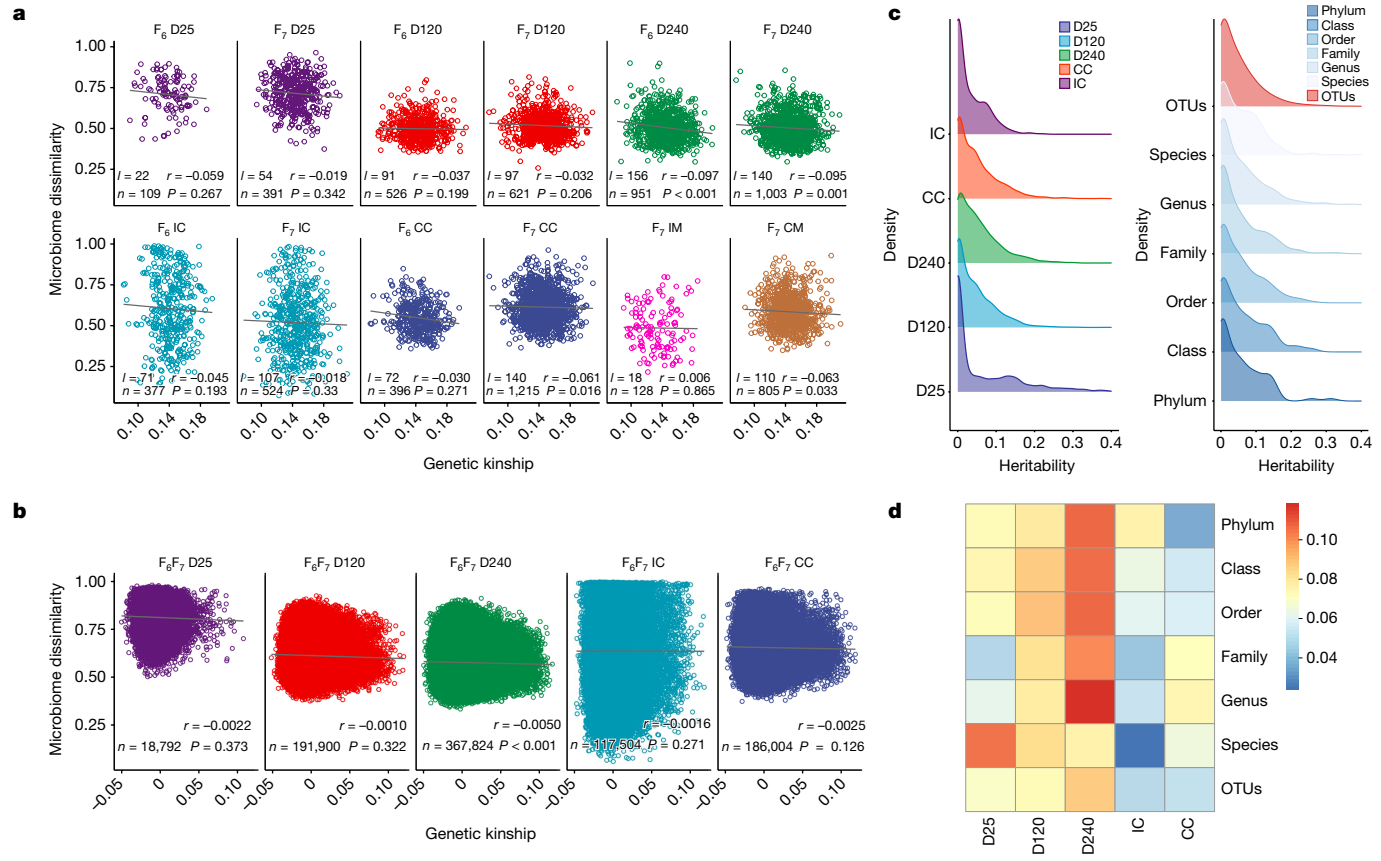


Fig. 2 | Heritability of microbiota composition in mosaic pigs. a, Correlation between genome-wide kinship (Θ) and microbiome dissimilarity (Bray–Curtis distance) within litter. The correlation (Spearman's r) was measured separately for the 12 data series. P values (one-sided) were computed by permutation. Adjusted r values were below the 50th percentile of the permutation values for 11 out of 12 ($P = 0.0029$). The empirical P value (one-sided) of r was $\leq 0.05/12 = 0.004$ (Bonferroni corrected) for two data series. P values were combined across the 12 data series yielding an overall P value of 3×10^{-4}. The number of litters (l) and animal pairs (n) used are given for each data series. **b**, The correlation between genome-wide kinship and microbiome dissimilarity (Bray–Curtis distance) across generations. We considered all possible pairs of F_6 and F_7 animals (not including sow–offspring pairs). Analyses were conducted for the five traits measured in both F_6 and F_7. r, p and n are as described in **a**. r values were below the 50th percentile of the permutation values for the five analysed sample types ($P = 0.03$). The empirical P value (one-sided) of r was $\leq 0.05/5 = 0.01$ (Bonferroni corrected) for one sample type. P values for the five sample types were combined yielding an overall P value of 0.013. **c**, The frequency distribution of heritabilities of individual taxa sorted by sample type (left) or taxonomic level (right). The values obtained by joint analysis of F_6 and F_7. **d**, Total heritabilities[2] computed by sample type and taxonomic level.

specific, whereas Pseudomonadaceae were nearly mucosa specific. The Bacteroidetes Odoribacteraceae and Rikenellaceae were at least ten times more abundant in day 25 faeces samples than in any other sample type. The firmicutes Christensenellaceae were nearly ten times more abundant in faeces samples than in any other sample type. This confirms that limiting the analysis of the intestinal microbiota to adult faeces can provide only a partial view of its complexity and the factors that determine it[17]. α-Diversity of faeces was lower at day 25 compared with at days 120–240, reminiscent of the enrichment of the intestinal flora between childhood and adulthood in humans[8,18]. α-Diversity was lower for ileal content than for caecal content and mucosa (Fig. 1c). β-Diversity tended to be inversely proportional to α-diversity, being higher for day 25 than for day 120 and 240 faeces. Variation in pairwise Bray–Curtis dissimilarities was highest for ileal content, which had the lowest α-diversity (Fig. 1d and Supplementary Table 2.3). The microbiota composition of pig day 240 (D240) faeces was more similar to that of human than to that of mouse faeces (Extended Data Fig. 2).

Microbiota heritability in mosaic pigs

We first examined the relationship between genetic relatedness (genome-wide SNP identity by state) and microbiota dissimilarity (Bray–Curtis distance, all taxa combined)[2]. We used two approaches to mitigate confounding of genetics and environment: (1) we restricted the analyses to full-sibling littermates raised in the same environment, and (2) we confronted genetic similarity and microbiota dissimilarity across generations (F_6 and F_7). Both approaches supported an effect of genetics on microbiota composition manifested by significant negative correlations between genetic similarity and microbiota dissimilarity (Fig. 2a, b).

We then evaluated the heritability (h^2) of the abundances of individual taxa/OTUs using F_6 and F_7 jointly (Fig. 2c and Supplementary Table 3.1). We computed total heritabilities (average heritability weighted by abundance)[2] separately by sample type and taxonomic level. Total heritabilities were generally low (≤11.8%), yet higher for faecal (average D240, 10.2%) than for content samples (average ileal content, 5.7%) and lower for OTUs (average, 6.6%) than for higher taxonomic levels (average, 7.5%) (Fig. 2d). The correlation between $\hat{h}^2$ estimates obtained separately in F_6 and F_7 was positive and significant for D120 ($P = 4.2 \times 10^{-5}$), D240 ($P = 2.2 \times 10^{-16}$) and caecum content ($P = 2.3 \times 10^{-3}$), supporting genuine genetic effects (Extended Data Fig. 3). We established a list of the 55 most likely heritable taxa/OTUs on the basis of congruent $\hat{h}^2$ estimates in F_6 (≥0.15), F_7 (≥0.15), and across F_6 and F_7 (≥0.10) (Supplementary Table 3.1). It included the order Campylobacterales in D240 faeces, the species *Bacteroides coprophilus* in D25 faeces, and 53 OTUs of which two (D240 faeces, caecum

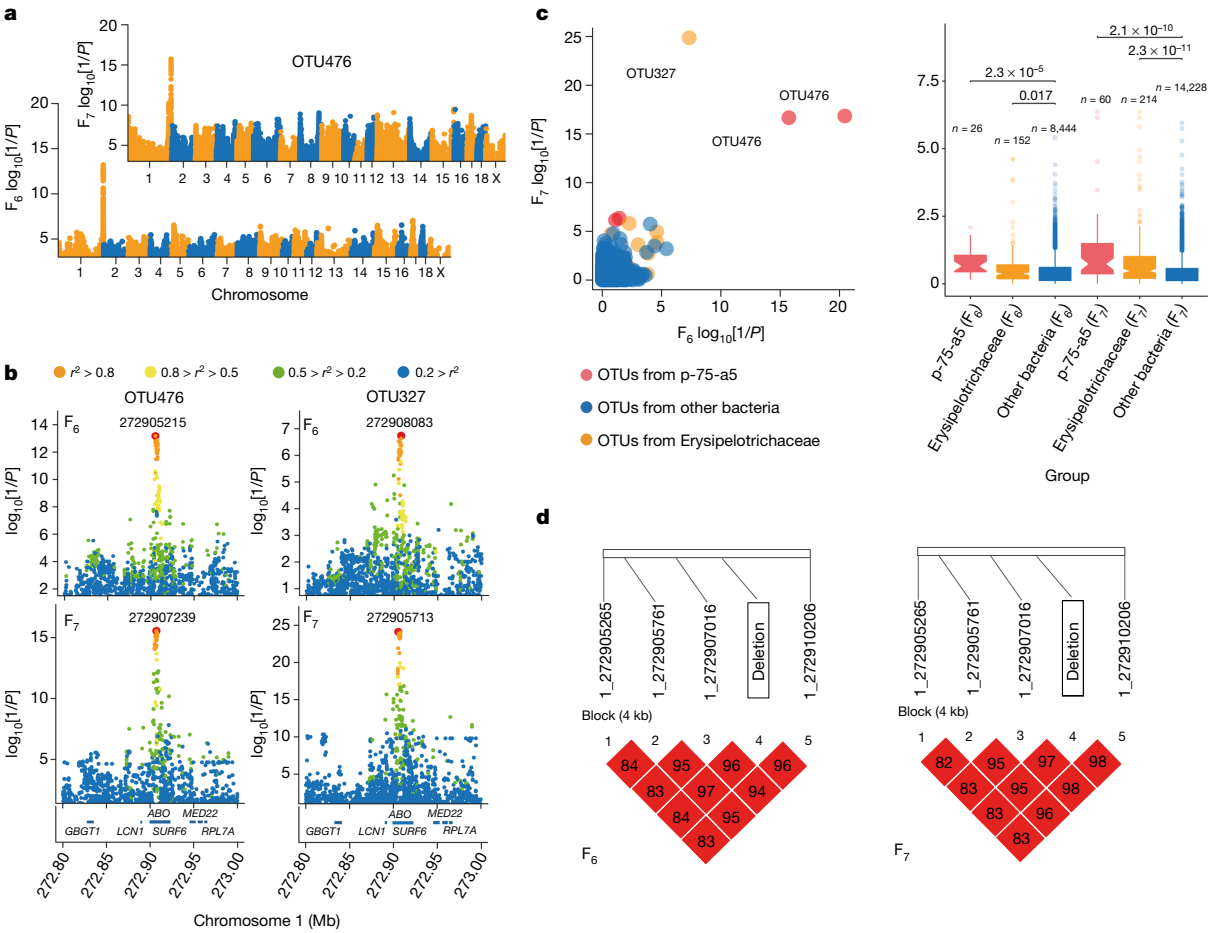


Fig. 3 | A miQTL affecting Erysipelotrichaceae species. a, The result of genome-wide meta-analysis (across sample types) in F_6 and F_7 for OTU476. Reported log-transformed *P* values are nominal (that is, not corrected for multiple testing). **b**, Local magnified views of chromosome 1 (272.8–273 Mb) of OTU476 and OTU327 in F_6 and F_7. log-transformed nominal *P* values as in **a**. **c**, log[1/*P*] values in F_6 (*x*-axis) and F_7 (*y*-axis) for the association between SNP 1_272907239 and the abundance of 8,490 OTUs for all of the sample types and two analyses methods (abundance and presence/absence, explaining the two OTU476 values) (left). OTUs belonging to p-75-a5 and Erysipelotrichaceae are shown in red and yellow, respectively. Right, comparison of the distribution of association (1_272907239) *P* values for p-75-a5 and Erysipelotrichaceae OTUs with other OTUs in F_6 and F_7. Box plots are as described in Fig. 1c. The notches in the boxes correspond to 95% confidence intervals of the median values. Distributions were compared using Wilcoxon's rank-sum test. *P* values (nominal) of the comparisons are given above the horizontal lines. **c**, LD (r^2) between the four top SNPs and the 2.3 kb ABO deletion in F_6 and F_7.

content) were assigned to Christensenellaceae, five to Ruminococcaceae (four in D240 faeces, one in caecum content), one (D25) to *Ruminococcus*, one (D25) to *Phascolarctobacterium*, and one (D240) to RF32; that is, taxons characterized by $h^2 \geq 0.10$ in human studies[19]. With the exception of RF32, all of these belong to the order Clostridiales.

A miQTL affecting Erysipelotrichaceae species

Heritability quantifies the proportion of additive genetic variance but does not foretell genetic architecture: phenotypes with low heritability may be affected by variants with major effects, whereas highly heritable traits may be very polygenic. To gain insights into the genetic architecture of the gut microbiota composition in this population, we performed a genome-wide association study (GWAS). We applied two statistical models testing SNP effects on taxa abundance and taxa presence/absence, respectively[11]. We ran GWAS separately by sample type, taxon and generation for a total of 57,557 GWAS (involving 8,490 taxa) (Supplementary Table 4.1). This yielded 1,527 genome-wide significant signals ($P \leq 5 \times 10^{-8}$). For these, we performed meta-analyses across sample types, separately in F_6 and F_7 (Extended Data Fig. 4). We identified six signals exceeding the experiment-wide discovery threshold ($P = 1.5 \times 10^{-12}$) in one cohort and the experiment-wide replication threshold ($P = 0.007$) in the other (Fig. 3a and Supplementary Table 4.2–4.3). All lead SNPs mapped within 3,037 bp from each other (chromosome 1) and were in high LD (Fig. 3b). They affected two OTUs (OTU476 and OTU327) as well as the genus p-75-a5, to which OTU476 is assigned. All three are Erysipelotrichaceae.

To determine whether this microbiota QTL (miQTL) affects other taxa, we plotted the F_6 and F_7 association log[1/*P*] for lead SNP 1_272907239 and the 8,490 studied OTUs. OTU476 and OTU327 stood out as highly significant in F_6 and F_7 (Fig. 3c). Yet the *P* values for the 31 other p-75-a5 OTUs and the 83 other Erysipelotrichaceae OTUs were significantly shifted towards lower *P* values in both cohorts (Fig. 3c), and the sign was consistent with that for OTU476, OTU327 and p-75-a5 (Extended Data Fig. 4). This suggests that the chromosome 1 miQTL affects other species in this family. Notably, the h^2 for p-75-a5 and OTU476 deviated significantly from 0 in D240 (0.10 and 0.14) and D120 (0.13 and 0.13) faeces, and h^2 estimates for OTU327 in D120 faeces (0.13) (Supplementary Table 4.4).

QQ plots obtained after removing chromosome 1 variants (272.8–273.1 Mb interval) did not show convincing evidence for residual inflation of log[1/*P*] (Extended Data Fig. 4). Thus, the residual h^2 is most likely highly polygenic.

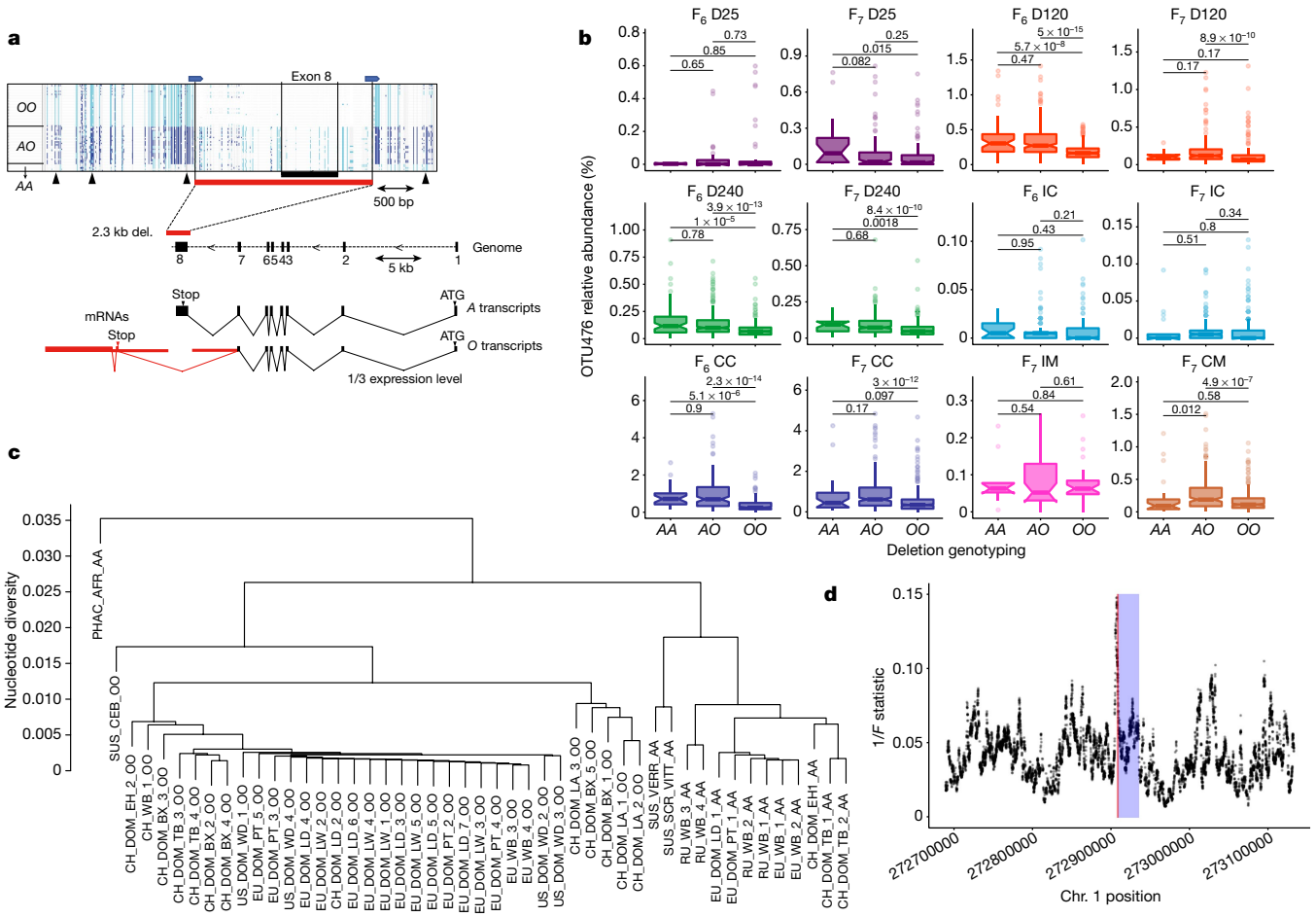


Fig. 4 | A 3.5-million-year-old deletion in the pig *ABO* orthologue causes the miQTL. a, The structure of the porcine *AO* blood group gene. IGV (Integrated Genome Viewer) view of the genotypes of the 61 F_0 animals showing 145 variants in a ~5 kb interval spanning the 2.3 kb deletion. The top two red rectangles show the 2.3 kb deletion. Homozygous alternative (light blue), heterozygous (dark blue) and homozygous reference (grey) genotypes are shown. The horizontal blue arrows mark SINEs that have mediated intrachromosomal recombination. The vertical black arrows mark the top variants from Fig. 3b. The effect of the 2.3 kb deletion on acetyl-galactosaminyl transferase transcripts is shown, including the creation of alternative exons 8 and 9, and the reduction of transcript levels to around one-third of normal. **b**, The effect of the *AO* genotype (*AA*, *AO* or *OO*) on the abundance of OTU476. The effect of the *A* allele is dominant over that of the *O* allele, and the miQTL effect is detected in the caecum (content and mucosa) and in day 120 and 240 faeces samples. Sample sizes are provided in Supplementary Table 4.1. The box plots are as described in Fig. 3d. **c**, Unweighted pair group method with arithmetic mean dendrogram based on the sequence similarity between 14 *AA* and 34 *OO* animals in a 5 kb window centred on the 2.3 kb deletion. CH, Chinese; EU, European; RU, Russian; DOM, domestic pigs; WB, wild boars; PHAC_AFR, common warthog; SUS_VERR, Javan warty pig; SUS_CEB, Visayan warty pig; SUS_SCR_VII, Sumatran wild boar. Breeds: BX, Bamaxiang; EH, Erhualian; LA, Laiwu; LD, Landrace; LW, Large White; PT, Piétrain; TB, Tibetan; WD, White Duroc. **d**, The peak of reduced population differentiation coinciding with the 2.3 kb deletion (red) in the porcine *AO* gene (blue). The position on chromosome 1 is shown on the *x*-axis; and 1/(mean *F* statistic) for all variants in a 2 kb sliding window is shown on the *y*-axis. The *F* statistic was computed as the ratio of the between-breed mean squares and the within-breed mean squares for the dosage of *O* allele. Chr., chromosome.

A 3.5-million-year-old deletion in the pig *ABO* orthologue

All lead SNPs map to the 3′ end of the porcine *N*-acetyl-galactosaminyl transferase gene underlying the ABO blood group in humans (Fig. 3b). This is a strong candidate that modulates interactions with several pathogens[20], and was recently suggested to affect the intestinal microbiota in healthy humans[21–23]. A 2.3 kb deletion encompassing the last exon has been reported in pigs[24]. We showed that it was in near-perfect LD ($r^2 \geq 0.94$) with the lead SNPs in our population (Fig. 3d). We confirmed its boundaries and showed that it results from an intrachromosomal recombination between short interspersed nuclear elements (SINEs). We showed using RNA-sequencing (RNA-seq) and expression quantitative trait loci (eQTL) analysis in 300 F_7 caecum samples that the deleted allele produces reduced amounts (around 1/3 of normal) of a 7.4 kb transcript (with alternative eighth and ninth exons) encoding a protein that lost 62% of its amino-acids, including 7 out of 8 active sites[25] (Fig. 4a). Our results indicate that the 2.3 kb deletion creates a null allele (hereafter, the *O* allele) and that is a very strong candidate for the causative miQTL mutation (Extended Data Figs. 5 and 6).

We examined the effect of the *AO* genotype on the abundance of OTU476, OTU327 and p-75-a5. The results showed that (1) the effect of the *A* allele is dominant over the *O* allele (Supplementary Table 5.1), and (2) that the effect manifests in D120 and D240 faeces, caecal content as well as mucosa, but not in D25 faeces and D240 ileal content and mucosa (Fig. 4b and Supplementary Table 5.1). In these samples (D120, D240, caecum content, caecal mucosa), the *AO* genotype explained 7.9%, 3.2% and 6.6% of the variation in the abundance of OTU476, OTU327 and genus p-75-a5 (Supplementary Table 5.2). Notably, the abundance of OTU476 and OTU327 was highest in caecal content (around 0.92% and 0.02% of reads in *AA/AO* animals, and around 0.47% and 0.003% of reads in *OO* animals, respectively; Extended Data Fig. 5).

In primates, the *ABO* locus is under strong balancing selection that has perpetuated identical-by-descent alleles segregating in humans,

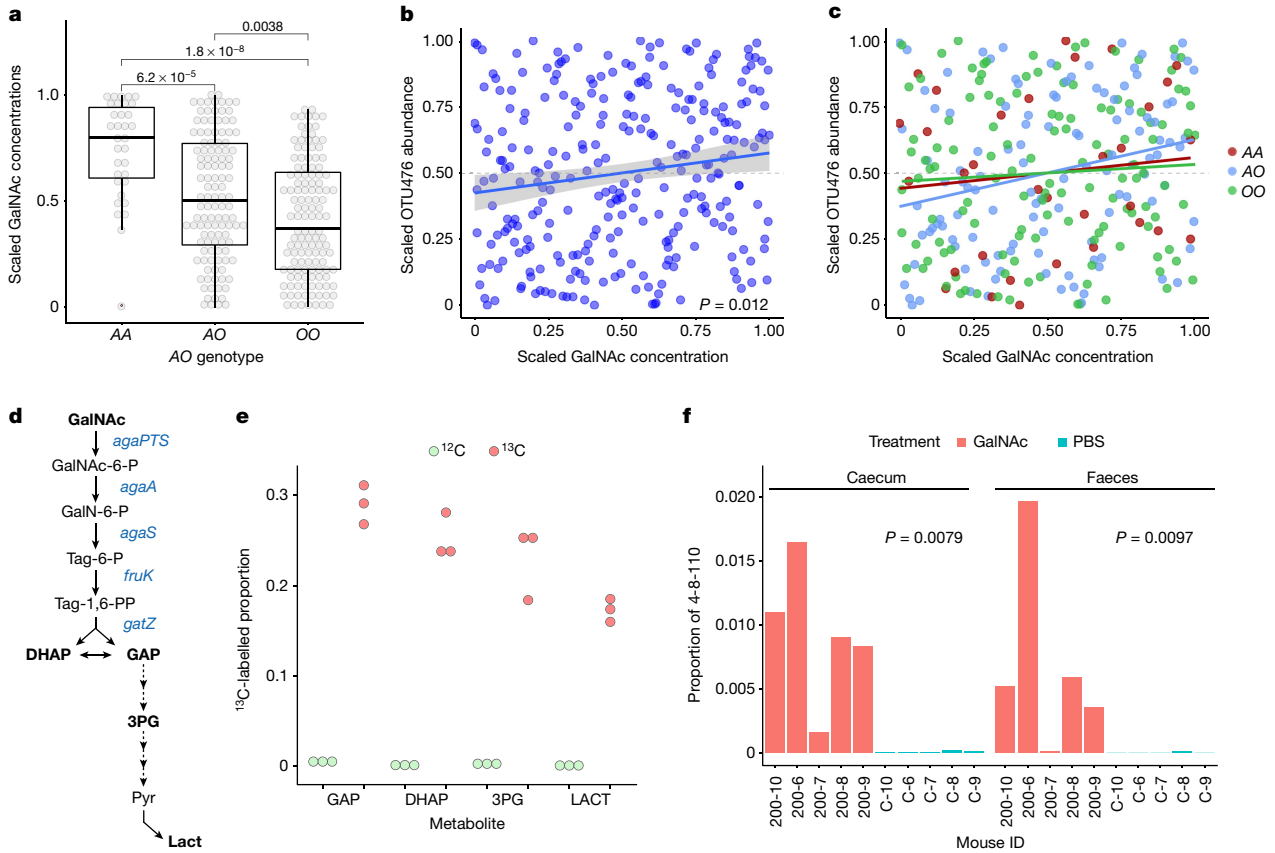


Fig. 5 | The miQTL acts by increasing GalNAc concentrations and affects GalNAc-using bacteria. a, The effect of *AO* genotype on GalNAc concentrations in caecal content ($n_{AA} = 33$, $n_{AO} = 118$, $n_{OO} = 127$). Concentrations were corrected for batch effect and scaled between 0 and 1 to equalize residual variance. *P* values (two-sided and nominal) for genotype contrasts were computed using Wilcoxon's tests. The box plots are as described in Fig. 1c. **b**, The correlation between GalNAc concentration and OTU476 abundance within the *AO* genotype. Area under the curve (AUC) values for GalNAc corrected for batch effect and *AO* genotype and scaled as described above. The *P* value (nominal, two-sided) of Spearman's correlation is given ($P = 0.012$). The shaded area corresponds to the 95% confidence region for the regression fit. **c**, Same as in **b**, with animals coloured by *AO* genotype. **d**, The GalNAc transport and catabolic pathway in OTU476-like strains. GalNAc-6-P, *N*-acetylgalactosamine-6-phosphate; GalN-6-P, galactosamine-6-P; Tag_6-P, tagatose-6-phosphate; Tag-1,6-PP, tagatose-1,6-biphosphate; GAP, glyceraldehyde-3-phosphate; DHAP, dihydroxyacetone-phosphate; 3PG, 3-phosphoglycerate; Pyr, pyruvate; Lact, lactate. Enzymes encoded in the GalNAc operon are shown in blue. Metabolites considered in the metabolic flux analysis are shown in bold. **e**, The proportion of ^{13}C-labelled metabolites determined by GC–MS in the OTU476-like strain (4-8-110) fed ^{13}C-labelled (red) versus regular GalNAc (green). **f**, In vivo (germ-free mice) *E. coli* versus OTU476-like strain competition with and without GalNAc. The proportion of 16S rRNA reads mapping to the 4-8-110 reference rRNA sequence versus *E. coli* rRNA sequence (that is, 1 minus the proportions shown in the figure correspond to reads mapping to the *E. coli* 16S rRNA) in the caecum content and faeces of 10 germ-free mice (Kunming line) inoculated by gavage with a pure culture of 4-8-110 and *E. coli* and force-fed with GalNAc (red bars) versus PBS (green bars). *P* values (nominal, two-sided, uncorrected) comparing the difference in abundance were determined using Wilcoxon tests.

gibbons and Old World monkeys for tens of millions of years[26]. We analysed the sequences of the 61 F_0 (*Sus scrofa domestica*), 18 wild boars (9 Asian, 9 European) (*Sus scrofa*), 1 Indonesian wild boar from Sumatra (*Sus scrofa vittatus*), 1 Visayan warty pig from the Philippines (*Sus cebifrons*), 1 Javan warty pig from Indonesia (*Sus verrucosus*) and one common warthog from Africa (*Phacochoerus africanus*) in a 50 kb window spanning the *AO* gene. Asian and European wild boars diverged from a common *S. scrofa* ancestor around 1 million years ago (Ma), *S. scrofa* and *S. s. vittatus* around 1.5 Ma, *S. scrofa* and *S. cebifrons/verrucosus* around 3.5 Ma, and *S. scrofa* and *P. africanus* around 10 Ma[27]. The 2.3 kb deletion segregated in all eight F_0 breeds, in all Asian and European/American wild-boar populations, and in *S. cebifrons* (Supplementary Table 5.3). The sequence of the deletion breakpoint was identical in all of the samples, confirming the identical-by-descent nature of the porcine *O* allele (Extended Data Fig. 7). Consistent with the hypothesis of an ancestral trans-species polymorphism (rather than hybridization), the *O* allele of *S. cebifrons* lay outside the cluster of *S. scrofa O* alleles (Fig. 4c and Extended Data Fig. 7). Further supporting balancing selection, the *AO* gene showed a marked decrease in population differentiation maximizing at the 2.3 kb deletion (Fig. 4d). Thus, although largely unknown, the underlying selective forces have operated in at least two mammalian branches (primates and suidae), over substantially long periods and broad geographical ranges, pointing towards their pervasive nature (Extended Data Fig. 7).

The miQTL affects caecal GalNAc concentrations

As in humans, the porcine *AO* gene is broadly expressed, yet is particularly expressed in the intestines (Extended Data Fig. 7). The *N*-acetyl-galactosaminyl-transferase encoded by the *A* allele adds *N*-acetyl-D-galactosamine (GalNAc, α1-3 linkage) to H and Lewis antigens present on glycan substrates, including the heavily glycosylated mucins, which can be used as carbon source by intestinal bacteria[20,28,29] (Extended Data Fig. 8). We reasoned that, owing to the abundance of intestinal mucus, the *O* allele might reduce the concentration of GalNAc in the intestine, thereby reducing the growth of bacterial species that

are dependent on this sugar. We measured the concentration of free GalNAc, *N*-acetylglucosamine (GlcNAc) and *N*-acetylmannosamine (ManNAc) isomers (hereafter, HexNAc) in the caecum content of 278 D240 animals (124 F_7 and 154 Duroc × Landrace × Large White) using liquid chromatography coupled with tandem mass spectrometry (LC–MS/MS). Intestinal HexNAc in pigs comprises approximately equal proportions of the GlcNAc and GalNAc isomers[30]. Dosage of the *A* allele (additive model) increased ($P = 5.7 \times 10^{-9}$) caecal HexNAc concentrations (Fig. 5a). Assuming that the increase is due only to GalNAc (given the enzymatic activity of the *A* allele), *AA* animals have at least twice the amount of free GalNAc compared with *OO* animals (Fig. 5a and Extended Data Fig. 9). Non-zero values in *OO* animals are primarily due to GlcNAc/ManNAc, and are probably also due to non-A antigen host and dietary GalNAc (GalNAc is ubiquitous at the core of all *O*-glycans, including intestinal mucins, and is a component of glycosaminoglycans such as chondroitin sulfate). Note that the effect of the *A* allele appears to be dominant on bacterial abundance, yet additive on GalNAc concentrations (*P* value of likelihood ratio test against the additive model = 0.12). This indicates that the additional increase in GalNAc availability in *AA* animals (versus *AO* animals) does not further favour the growth of OTU476, OTU327 and p-75-a5.

Showing that the *O* deletion (1) reduces caecal GalNAc concentrations, and (2) reduces the abundance of OTU476, OTU327 and p-75-a5 does not prove that one causes the other. The effect on bacterial abundance could be mediated by an unidentified mechanism independent of GalNAc. However, the hypothesis that it is the change in GalNAc concentrations that causes the change in bacterial abundance makes a testable prediction. In this case, we may expect to observe a 'residual' correlation between GalNAc concentrations and bacterial abundance 'within' the *AO* genotype, that is, by blocking the effect of the *AO* genotype and exploiting other sources of variation of GalNAc and bacterial abundance (Extended Data Fig. 8). We performed these analyses and observed the predicted positive correlation for OTU476 ($P = 0.012$) and p-75-a5 ($P = 0.010$), consistently within the three genotypes (*AA*, *AO* and *OO*). The correlation was positive (and consistent within the three genotypes) albeit not significant ($P = 0.3$) for the less abundant OTU327 (Fig. 5b, c and Extended Data Fig. 9).

Taken together, these results make a strong case that the *AO* genotype affects intestinal GalNAc concentrations, and that this affects the abundance of some Erysipelotrichaceae species.

miQTL-responsive bacteria use GalNAc

To be used as carbon source by intestinal bacteria, GalNAc needs (1) to be released from the glycan structures by secreted glycoside hydrolases (GHs), (2) to be imported across the bacterial membranes by dedicated transport systems (TR), and (3) to be converted into intermediates of central metabolism by a specific catabolic pathway (CP). Although some bacteria have both GHs and TR–CP for specific monosaccharides, other bacteria have only the GHs (donors) or the TR–CP (acceptors)[28]. We reasoned that the bacteria that would respond to the miQTL would have a complete GalNAc TR–CP system (with or without GHs). To test this, we (1) isolated two bacterial strains (4-8-110 and 4-15-1) with a V3–V4 sequence similarity of 100% and 99.8% with OTU476, respectively, and sequenced their genome using the Oxford Nanopore Promethion system, and (2) built 3,111 metagenomic assembled genomes (MAGs) from shotgun data of 92 porcine intestinal samples, including 248 Erysipelotrichaceae. We compiled a list of 24 genes implicated in GalNAc use (TR–CP)[28,31–36]. These encode (1) 11 components of three GalNAc transporter systems (AgaPTS: *agaE*, *agaF*, *agaV*, *agaW*; TonB dependent transporter: *omp*, *agaP*, *agaK*; GnbPTS: *gnbA*, *gnbB*, *gnbC*, *gnbD*), (2) two GalNAc-6P deacetylases (*agaA*, *nagA*), (3) two galactosamine-6P (*GalN-6P*) isomerase and/or deaminases (*agaI*, *agaS*), (4) three tagatose-6P kinases (*pfka*, *lacC*, *fruK*), (5) four tagatose-1,6-PP aldolases or aldolase subunits (*gatY-kbaY*, *gatZ-kbaZ*, *lacD*, *fba*) and (6) two regulon repressors (*agaR*, *gntR*), for a total of six essential pathway constituents (Fig. 5d and Supplementary Table 6). Genes involved in the use of specific sugars tend to cluster and form operons of potentially coregulated genes (regulons) that support all or most of the essential TR–CP steps. The steps that are not encoded by the operon may be complemented in *trans* by genes encoding enzymes that may be less substrate-specific[37,38]. We searched for orthologues of the 24 genes in the two OTU476-like genomes and 3,111 MAGs. We generated two scores to evaluate the ability of bacterial species to use GalNAc. The first (pathway score) counted the number of essential steps in GalNAc use (out of six) that could be accomplished by the orthologues detected in the genome, irrespective of map position. The second (regulon score) counted the number of essential GalNAc-use steps that could be fulfilled by orthologues that were clustered in the genome, that is, forming a potential operon. We used *agaS* as an anchor gene to establish the regulon score[28]. The first notable observation was that at least one orthologue of *agaS* was found in the two (that is, all) OTU476-like strains, in 31% of Erysipelotrichaceae MAGs ($n = 248$), yet in only 2.9% of other MAGs ($n = 2{,}863$). The second was that both scores were significantly higher for Erysipelotrichaceae compared with the other MAGs ($P_{pathway} = 2.0 \times 10^{-16}$ and $P_{regulon} = 2.0 \times 10^{-16}$), and for the two OTU476-like strains compared with Erysipelotrichaceae and non-Erysipelotrichaceae MAGs combined ($P_{pathway} = 2.2 \times 10^{-3}$ and $P_{regulon} = 1.2 \times 10^{-5}$) (Extended Data Fig. 10). The two OTU476-like strains were characterized by a cluster with eight GalNAc genes, including orthologues of the four components of the AgaPTS transporter system (*agaE*, *agaF*, *agaV*, *agaW*), of nagA deacetylase, of agaS deaminase/isomerase, of fruK kinase and of the gatZ-kbaZ aldolase subunit (Extended Data Fig. 10). This amounted to a score of five for both pathway and regulon, corresponding, respectively, to the top 4.7% and 0.35% out of 3,113 MAGs, demonstrating the uncommon status of OTU476-like strains with regard to GalNAc use. Neither the 4-15-1 nor 4-8-110 genome encode GHs that are known to have *α*-*N*-acetylgalactosaminidase activity specific for the A antigen[39–41], suggesting that these strains are acceptors only.

To confirm that the OTU476-like strains are able to import and catabolize GalNAc, we grew the 4-8-110 strain in the presence of ^{13}C-labelled GalNAc and checked the appearance of ^{13}C-labelled catabolites in the bacterial pellet using gas chromatography coupled with MS (GC–MS). After 13 h of culture, around 29% of glyceraldehyde-3-P and around 25% of dihydoxyacetone phosphate (the products of the conversion of tagatose-1,6-bisphosphate by gatZ aldolase) were labelled with ^{13}C, demonstrating the use of GalNAc by strain 4-8-110. Approximately 23% of 3-phosphoglycerate and around 17% of lactate molecules but none of TCA metabolites were ^{13}C-labelled, underscoring the predominance of glycolysis over oxidative phosphorylation in this anaerobe (Fig. 5d, e).

We further performed a gavage experiment in germ free mice to test the effect of the addition of GalNAc on the relative growth of 4-8-110 and *Escherichia coli* in vivo. We gavaged mice either with GalNAc (200 mg kg^{-1} live weight, yielding caecal GalNAc concentrations comparable to *AA* pigs; Extended Data Fig. 9) or phosphate-buffered saline (PBS), as well as a mixture of *E. coli* and the 4-8-110 OTU476-like strain. We euthanized the mice at day 12 and measured the relative abundance of *E. coli* and 4-8-110 using 16S rRNA sequencing in the caecum and faeces. Strain 4-8-110 was in essence not detectable in PBS-gavaged mice, whereas it accounted for an average of 0.9% of reads in the caecum content ($P = 0.0079$) and 0.7% of reads in the faeces ($P = 0.0097$) of GalNAc-gavaged mice (Fig. 5f).

GalNAc operon of miQTL-sensitive bacteria

At least 65 MAGs (including *E. coli* and other non-*Erysipelotrichacea* species) contain orthologues for the five key GalNAc TR–CP steps as do 4-15-1 and 4-8-110. We wondered why the chromosome 1 miQTL does not affect these species. The organization of their GalNAc operons

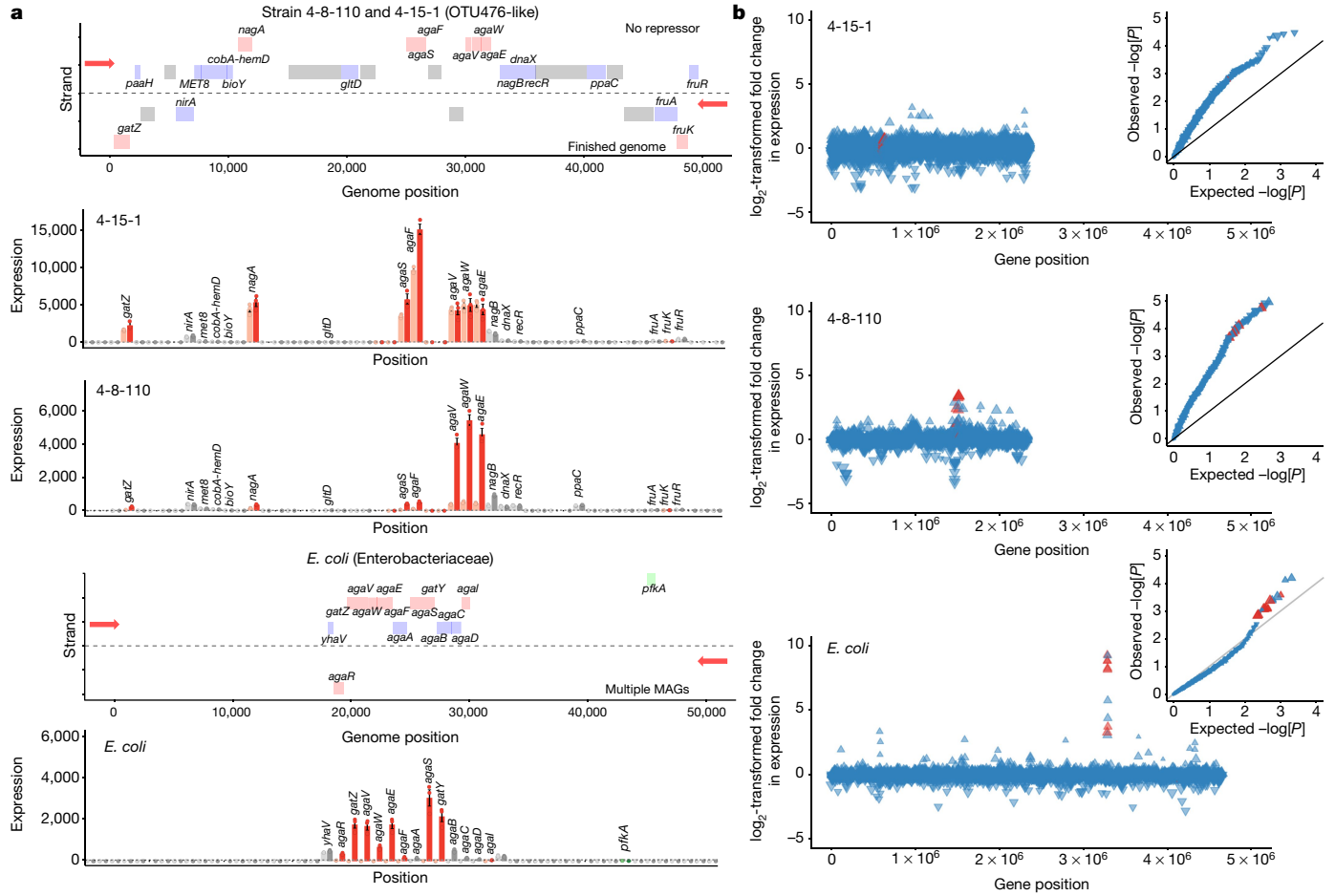


Fig. 6 | The GalNAc operon organization and transcriptome response of miQTL-responsive bacteria. a, The GalNAc operon organization and local transcriptome response to GalNAc addition in OTU476-like strains and *E. coli*. Top, GalNAc operon organization. Identified ORFs are represented as coloured boxes. Genes implicated in GalNAc import and catabolism are shown in red if they are part of the cluster and in green if located elsewhere in the genome. Genes with a known function unrelated to GalNAc are shown in blue. ORFs with an uncharacterized gene product are shown in grey. Gene acronyms are given next to the corresponding boxes. ORFs transcribed from the top and bottom strand are shown above and below the dotted line, respectively. The respective transcriptional directions are marked by arrows. Bottom, local gene expression levels (fragments per kb of exon model per million mapped reads (FPKM)) with (dark colour) and without (light colour) addition of GalNAc in the growth medium. The colours for the ORFs with a known function (GalNAc gene (red), other gene (blue)) are the same as in the top panels. The error bars are the s.e.m. from three replicates (individual values are shown as dots). **b**, The global transcriptome response to GalNAc addition in OTU476-like strains and *E. coli*. The $\log_2$-transformed fold change in expression for all genes in the respective genomes (4,419 in *E. coli*, 1,119 in OTU476-like strains, ranked according to genomic position) after addition of GalNAc in the medium is shown. GalNAc genes are shown in red and other genes are shown in blue. Insets: corresponding *QQ* plots showing the near absence of effects on genes other than the GalNAc regulon in *E. coli* versus the widespread response in OTU476-like strains. The *P* values used to generate the *QQ* plots are nominal, and were determined using DISEQ2.

may provide a hint. The GalNAc clusters of non-Erysipelotrichaceae species have the features of genuine regulons. The relevant ORFs tend to be adjacent to each other (spanning around 10 kb) and on the same strand, compatible with polycistronic mRNAs enabling coregulated expression. By contrast, the open reading frames (ORFs) of the GalNAc clusters of the two OTU476-like strains and at least one studied Erysipelotrichaceae span around 50 kb and 30 kb, respectively, and are distributed on both strands. Neither genome contained orthologues of *agaR* or *gntR*, which are negative regulators of GalNAc regulons and were observed in all other GalNAc-rich MAGs (Fig. 6a and Extended Data Fig. 11). This suggests that, in contrast to *E. coli* and other species, the OTU476-like strains and some Erysipelotrichaceae cannot sense GalNAc concentrations and induce expression of the genes necessary for GalNAc use only when needed[33,35,36], but rather may express these constitutively. To test this hypothesis, we grew the two OTU476-like strains and *E. coli* with and without GalNAc in the medium and analysed their transcriptome using RNA-seq. The GalNAc regulon of *E. coli* was tightly regulated, with a near complete absence of transcription in the absence of GalNAc, and around 300-fold upregulation in its presence. By contrast, in the OTU476-like 4-15-1 strain, GalNAc TR–CP genes were expressed at nearly equal levels with and without GalNAc, as predicted. Strain 4-15-1 expression levels in the absence of GalNAc were higher than *E. coli* expression levels in the presence of GalNAc. The expression pattern in the 4-8-110 strain was intermediate between *E. coli* and 4-15-1: GalNAc genes were expressed in the absence of GalNAc (albeit at low levels) yet were upregulated (around five to tenfold) in its presence (Fig. 6a). The difference in the response of the transcriptome to GalNAc addition between OTU476-like strains and *E. coli* response was not limited to the GalNAc operon. Although addition of GalNAc did not have noticeable effects on genes outside of the GalNAc regulon in *E. coli*, it altered the expression of ≥605 and 225 genes ($q \leq 0.05$) in 4-15-1 and 4-8-110, respectively (Fig. 6b). Six KEGG pathways were consequently perturbed in both strains (FDR ≤ 0.05): metabolic pathways, pyrimidine metabolism, alanine, aspartate and glutamate metabolism,

biosynthesis of antibiotics, ABC transporters and biosynthesis of secondary metabolites (Supplementary Table 7).

Discussion

In humans, the *ABO* genotype affects susceptibility to various viral (including SARS-CoV-1 and 2), bacterial and protozoan pathogens[20,42,43], and possibly the composition of the intestinal microbiota[21–23]. Invoked mechanisms are usually immune related, including pathogen adhesion, toxin binding, soluble decoys and natural antibodies[26,44]. We provide strong evidence that the miQTL reported in this research functions by affecting intestinal GalNAc concentrations, and thereby the growth of GalNAc-using bacteria. A puzzling finding is that not all species with the ability to use GalNAc as a carbon source appear to be affected. We suggest that this could be connected to the observation that the GalNAc operon is inducible in species like *E. coli*, while being constitutively expressed at high levels in at least some of the OTU476-like strains. The GalNAc gene cluster as seen in the OTU476-like strains is a possible evolutionary intermediate towards the formation of a genuine regulon as seen in *E. coli*, facilitating horizontal transmission of a 'selfish' functional gene ensemble even if not yet adaptively coregulated[37]. The distinct behaviour of the GalNAc operon in the 4-15-1 and 4-8-110 strains in response to GalNAc may be indicative of an evolving regulatory mechanism distinct from the canonical repressor-based regulon system. These findings suggest an alternative modus operandi of the miQTL. Bacteria affected by it may not be at an advantage when GalNAc is present at a relatively higher concentration in the intestinal content (as in *AA* and *AO* animals), but rather at a disadvantage when GalNAc is present at low concentrations (as in *OO* animals) due to wasting energy transcribing and translating useless genes. By tightly regulating the expression of their operon in response to ambient GalNAc availability, species like *E. coli* may fair equally well in the gut of *AA/AO* pigs as in that of *OO* pigs and, therefore, not be affected by the miQTL. Note that, in addition to the distinct organization of their GalNAc operon, the transcriptome of the two studied OTU476-like strain responds more to GalNAc addition than that of *E. coli*. The precise significance of this observation remains to be established but may indicate alternative or additional factors that underpin differential sensitivity to the miQTL.

FUT2-null alleles segregate in humans and in homozygotic individuals; this precludes synthesis of the type I H antigen and therefore epistatically of the type I A and/or B antigens in non-*O* individuals[20,21] (Extended Data Fig. 8). Although the $FUT2^{H117R}$ variant segregated in our population, there was no evidence for an effect of that or any other *FUT2* variant on the abundance of OTU476, OTU327 or p-75-a5 with or without conditioning on *AO* genotype (data not shown).

The effect of the *ABO* genotype on intestinal microbiota composition in humans remains somewhat controversial. Despite suggestive evidence in a small ($n = 71$) cohort of separate microbiota-based clustering of *AB* and *B* versus *A* and *O* individuals[45], a subsequent study in a larger cohort ($n = 1{,}503$) could not detect experiment-wide significant effects of the *ABO* genotype on gut microbiota composition[46]. More recently, an effect of *ABO* genotype was reported on the abundance of OTUs assigned respectively to *Faecalibacterium* and *Bacteroides* in a cohort of around 9,000 German individuals[21], but this was not confirmed in a subsequent study (~18,000 individuals)[47]. We examined the effect of the ABO blood group on the abundance of around 75 OTUs assigned to Erysipelotrichaceae in human gut samples (Extended Data Fig. 12). None of the OTUs detected in human were as closely related to the pig OTU476, OTU327 or p-75-a5 as these were to each other. We found no evidence for an effect of ABO blood group on the abundance of any of these. What underlies the difference between pigs and humans is unclear. Either Erysipelotrichaceae strains susceptible to the *ABO* genotype are not present at sufficient levels in humans, or the carbohydrate composition of human intestinal content makes these strains less sensitive to variations in GalNAc concentrations. The abundance of p-75-a5 was found to differ significantly between African subsistence categories and to be highest in pastoralists (as compared to hunter-gatherers and agro-pastoralists) possibly as a result of interaction with livestock[48,49]. Repeating the experiments in pastoralists may reveal the same miQTL effect detected in this study.

While this paper was in the last phase of the publication process two papers came out reporting an effect of *ABO* genotype on intestinal microbiota composition in human, albeit on distinct taxa[22,23].

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-022-04769-z.

1. Kundu, P., Blacher, E., Elinav, E. & Pettersson, S. Our gut microbiome: the evolving inner self. *Cell* **171**, 1481–1493 (2017).
2. Rothschild, D. et al. Environment dominates over host genetics in shaping human gut microbiota. *Nature* **555**, 210–215 (2018).
3. O'Hara, E., Neves, A. L. A., Song, Y. & Guan, L. L. The role of the gut microbiome in cattle production and health: driver or passenger? *Annu. Rev. Anim. Biosci.* **8**, 199–220 (2020).
4. Schmidt, T. S. B., Raes, J. & Bork, P. The human gut microbiome: from association to modulation. *Cell* **172**, 1198–1215 (2018).
5. Polderman, T. J. C. et al. Meta-analysis of the heritability of human traits based on 50 years of twin studies. *Nat. Genet.* **47**, 702–709 (2015).
6. Polubriaginof, F. C. G. et al. Disease heritability inferred from familial relationships reported in medical records. *Cell* **173**, 1692–1704 (2018).
7. Benson, A. K. et al. 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 (2010).
8. Yatsunenko, T. et al. Human gut microbiome viewed across age and geography. *Nature* **486**, 222–227 (2012).
9. Goodrich, J. K. et al. Human genetics shape the gut microbiome. *Cell* **159**, 789–799 (2014).
10. Blekhman, R. et al. Host genetic variation impacts microbiome compoistion across human body sites. *Genome Biol.* **16**, 191 (2015).
11. Turpin, W. et al. Association of host genome with intestinal microbial composition in a large healthy cohort. *Nat. Genet.* **48**, 1413–1417 (2016).
12. Bonder, M. J. et al. The effect of host genetics on the gut microbiome. *Nat. Genet.* **48**, 1407–1412 (2016).
13. Wang, J. et al. Genome-wide association analysis identifies variation in vitamin D receptor and other host factors influencing the gut microbiota. *Nat. Genet.* **48**, 1396–1406 (2016).
14. Hughes, D. A. et al. Genome-wide associations of human gut microbiome variation and implications for causal inference analyses. *Nat. Microbiol.* **5**, 1079–1087 (2020).
15. Sankararaman, S. et al. The genomic landscape of Neanderthal ancestry in present-day humans. *Nature* **507**, 354–357 (2014).
16. Patterson, N. et al. Genetic evidence for complex speciation of humans and chimpanzees. *Nature* **441**, 1103–1108 (2006).
17. Donaldson, G. P., Lee, S. M. & Mazmanian, S. K. Gut biogeography of the bacterial microbiota. *Nat. Rev. Microbiol.* **14**, 20–32 (2016).
18. Radjabzadeh, D. et al. Diversity, compositional and functional differences between gut microbiota of children and adults. *Sci. Rep.* **10**, 1040 (2020).
19. Goodrich, J. K. et al. Genetic determinants of the gut microbiome in UK twins. *Cell Host Microbe* **19**, 731–743 (2016).
20. Cooling, L. Blood groups in infection and host susceptibility. *Clin. Microbiol. Rev.* **28**, 801–870 (2015).
21. Rühlemann, M. C. et al. Genome-wide association study in 8,956 German individuals identifies influence of ABO histo-blood groups on gut microbiome. *Nat. Genet.* **53**, 147–155 (2021).
22. Lopera-Maya, E. E. et al. Effect of host genetics on the gut microbiome in 7,738 participants of the Dutch microbiome project. *Nat. Genet.* **54**, 143–151 (2022).
23. Qin, Y. et al. Combined effects of host genetics and diet on human gut microbiota and incident disease in a single population cohort. *Nat. Genet.* **54**, 134–142 (2022).
24. Choi, M. K. et al. Determination of complete sequence information of the human ABO blood group orthologous gene in pigs and breed differences in blood type frequencies. *Gene* **640**, 1–5 (2018).
25. Wang, S. et al. Design of glycosyl transferase inhibitors: serine analogues as pyrophosphate surrogates? *ChemPlusChem* **80**, 1525–1532 (2015).
26. Ségurel, L. et al. The ABO blood group is a trans-species poilymorphism in primates. *Proc. Natl Acad. Sci. USA* **109**, 18493–18498 (2012).
27. Groenen, M. A. M. A decade of pig genome sequencing: windo on pig domestication and evolution. *Genet. Sel. Evol.* **48**, 23–32 (2016).
28. Ravcheev, D. A. & Thiele, I. Comparative genomic analysis of the human gut microbiome reveals a broad distribution of metabolic pathways for the degradation of host-synthesized mucin glycans and utilization of mucin-derived monosaccharides. *Front. Genet.* **8**, 111 (2017).
29. Tailford, L. A. et al. Mucin glycan foraging in the human gut microbiome. *Front. Genet.* **6**, 81 (2015).

30. Lien, K. A., Sauer, W. C. & He, J. M. Dietary influences on the secretion into and degradation of mucin in the digestive tract of monogastric animals and humans. *J. Anim. Feed Sci.* **10**, 223–245 (2001).
31. Brinkkötter, A. B., Klöss, H., Alpert, C.-A. & Lengeler, J. W. Pathways for the utilization of *N*-acetyl-galactosamine and galactosamine in *Escherichia coli*. *Mol. Microbiol.* **37**, 125–135 (2000).
32. Rodionov, D. A. et al. Genomic encyclopedia of sugar utilization pathways in the *Shewanella* genus. *BMC Genom.* **11**, 494 (2010).
33. Leyn, S. A., Gao, F., Yang, C. & Rodionov, D. A. *N*-acetylgalactosamine utilization pathway and regulon in proteobacteria. *J. Biol. Chem.* **287**, 28047–28056 (2012).
34. Hu, Z., Patel, I. R. & Mukherjee, A. Genetic analysis of the roles of agaA, agaI, and agaS genes in the *N*-acetyl-D-galactosamine and D-galactosamine catabolic pathways in *Escherichia coli* strains O157:H7 and C. *BMC Microbiol.* **13**, 94 (2013).
35. Bidart, G. N., Rodriguez-Diaz, J., Monedoro, V. & Yebra, M. J. A unique gene cluster for the utilization of the mucosal and human milk-associated glycans galacto-*N*-biose and lacto-*N*-biose in *Lactobacillus casei*. *Mol. Microbiol.* **93**, 521–538 (2014).
36. Zhang, H. et al. Two novel regulators of *N*-acetyl-galactosamine utilization pathway and distinct roles in bacterial infections. *Microbiol. Open* **4**, 983–1000 (2015).
37. Lawrence, J. Selfish operons: the evolutionary impact of gene clustering in prokrayotes and eukaryotes. *Curr. Opin. Genet. Dev.* **9**, 642–648 (1999).
38. Koonin, E. V. Evolution of genome architecture. *Int. J. Biochem. Cell Biol.* **41**, 298–306 (2009).
39. Lombard, V. et al. The carbohydrate-active enzymes database (CAZy) in 2013. *Nucleic Acids Res.* **42**, D490–D495 (2014).
40. Rahfeld, P. et al. An enzymatic pathway in the human gut microbiome that converts A to universal O type blood. *Nat. Microbiol.* **4**, 1475–1585 (2019).
41. Rahfled, P. et al. Prospecting for microbial α-*N*-acetylgalactosaminidases yields a new class of GH31 O-glycanase. J. Biol. Chem. 294, 16400–16415.
42. Chen, Y. et al. ABO blood group and susceptibility to severe acute respiratory syndrome. *JAMA* **293**, 1450–1451 (2005).
43. Ellinghaus, D. et al. The ABO blood group locus and a chromosome 3 gene cluster associate with SARS-CoV-2 respiratory failure in an Italian-Spanish genome-wide association analysis. Preprint at *medRxiv* https://doi.org/10.1101/2020.05.31.20114991 (2020).
44. Blancher, A. Evolution of the *ABO* supergene family. *ISBT Sci. Ser.* **8**, 201–206 (2013).
45. Makivuokko, H. et al. Association between the ABO blood group and the human intestinal microbiota composition. *BMC Microbiol.* **12**, 94 (2012).
46. Davenport, E. R. et al. ABO antigen and secretor statuses are not associated with gut microbiota composition in 1,500 twins. *BMC Genom.* **17**, 941–955 (2016).
47. Kurilshikov, A. et al. Large-scale association analyses identify host factors influencing human gut microbiome composition. *Nat. Genet.* **53**, 156–165 (2021).
48. Malmuthuge, N., Griebel, P. J. & Guan, L. L. Taxonomic identification of commensal bacteria associated with the mucosa and digesta throughout the gastrointestinal tracts of preweaned calves. *Appl. Environ. Microbiol.* **80**, 2021–2028 (2014).
49. Hanson, M. E. B. et al. Population structure of human gut bacteria in a diverse cohort from rural Tanzania and Botswana. *Genome Biol.* **20**, 16 (2019).

Methods

Animal rearing and sample collection

This study focused on the sixth (F_6) and seventh (F_7) generation of a mosaic population generated as follows. An average of 3.6 boars (range, 3–4) and 4 sows (range, 2–5) from four indigenous Chinese pig breeds (Erhualian (EH), Bamaxiang (BX), Tibetan (TB), Laiwu (LA)) and four commercial European/American pig breeds (Landrace (LD), Large White (LW), Duroc (WD) and Piétrain (PT)) were successfully mated, constituting the F_0 generation. For each Chinese breed, the boars were mated with the sows of one European breed, and the sows with the boars of another European breed to produce the F_1 generation. Thus, every Chinese and every European breed is parent breed of two distinct F_1 hybrid combinations each, for a total of eight F_1 combinations (BX–LW, BX–PT, LA–PT, LA–LD, TB–LD, TB–WD, EH–WD, EH–LW). The F_2 generation was obtained by mating each F_1 hybrid combination with two others that did not share parental breeds for a total of eight F_2 combinations (BX–LW × LA–PT, BX–PT × LA–LD, LA–PT × TB–LD, LA–LD × TB–WD, TB–LD × EH–WD, TB–WD × EH–LW, EH–WD × BX–LW, EH–LW × BX–PT). Every F_2 combination was obtained by reciprocally crossing an average of 4 boars from one F_1 combination with an average of 7.25 sows from the other. The F_3 generation was obtained by mating each of the eight F_2 hybrid combinations with the only complementary F_2 combination that did not share any parental breeds for a total of four F_3 combinations (BX–LW–LA–PT × TB–LD–EH–WD, BX–PT–LA–LD × TB–WD–EH–LW, LA–PT–TB–LD × EH–WD–BX–LW, LA–LD–TW–WD × EH–LW–BX–PT) expected to each have around 12.5% of their genome from each of the founder breeds. Every F_3 combination was obtained by reciprocally crossing an average of 7 boars from one F_2 combination with an average of 10.8 sows from the complementary one. The F_4, F_5, F_6 and F_7 generations were obtained by intercrossing 57 boars × 75 sows (F_3 to F_4), 62 boars × 97 sows (F_4 to F_5), 85 boars × 170 sows (F_5 to F_6) and 82 boars × 111 sows (F_6 to F_7) (Supplementary Table 1).

All F_6 and F_7 animals were born and reared at the experimental farm of the National Key Laboratory for swine Genetic Improvement and Production Technology, Jiangxi Agricultural University (Nanchang, Jiangxi) under standard and uniform housing and feeding conditions. Piglets remained with their mother during the suckling period and were weaned at about 46 days of age. Litters were transferred to 12-pig fattening pens with automatic feeders (Osborne Industries), minimizing splitting and merging of litters. All pigs were fed twice per day with formula diets containing 16% crude protein, 3,100 kJ digestible energy, 0.78% lysine, 0.6% calcium and 0.5% phosphorus. Water was available ad libitum from nipple drinkers. Males were castrated at 80 days. Faecal samples were manually collected from the rectum of experimental pigs at the ages of 25, 120 and 240 days, dispensed in 2 ml tubes, flash-frozen in liquid nitrogen and stored at −80 °C. The animals were slaughtered at day 240. The ileum and caecum were sealed at both ends with a sterile rope and extracted from the carcass. Within 30 min after slaughter, ileal and caecal luminal content were collected (F_6 and F_7 animals), the ileum and caecum were rinsed with sterile saline solution, and samples of ileal and caecal mucosa were scraped with a sterile microscopic slide (F_7 animals only). Approximately 1 g of content or scrapings was packed in 2 ml sterile freezer tubes, flash-frozen in liquid nitrogen and stored at −80 °C. The number of samples of the different types available for further analysis are provided in Supplementary Table 2.1. All of the animals included in the analyses were healthy and did not receive any antibiotic treatment within one month of sample collection. All of the procedures involving animals were approved by the Ethics Committee of the Jiangxi Agricultural University (no. JXAU2011-006).

Genotyping by sequencing the F_0, F_6 and F_7 generations

Genomic DNA was extracted from ear punches using a standard phenol–chloroform-based DNA-extraction protocol. DNA concentrations were measured using the Nanodrop-1000 instrument (Thermo Fisher Scientific), and the DNA quality of all of the samples was assessed by agarose (0.8%) gel electrophoresis. Genomic DNA was sheared to 300–400 bp fragment size. The 3′-ends were adenylated and indexed primers were ligated. The libraries were amplified by PCR using Phusion High-Fidelity DNA polymerase (NEB) according to the recommendations of the manufacturer (Illumina). The libraries were loaded onto the Illumina X-10 instrument (Illumina) for 2 × 150 bp paired-end sequencing by Novogene. We removed reads with a quality score of ≤20 for ≥50% of bases or ≥10% missing (*N*) bases. Read quality was checked using Fastqc (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Clean reads were aligned to the *S. scrofa* reference genome assembly 11.1[50] using BWA (v.0.7.17)[51]. BAM files of mapped reads were sorted by chromosome position using SAMTools (v.1.6)[52]. Indel realignment and marking of duplicates were performed using Picard (v.2.21.4, http://broadinstitute.github.io/picard). Individual genotypes were called from BAM files using Platypus (v.0.8.1)[53]. Individual genotypes were merged into a single VCF file using PLINK (v.1.9)[54] encompassing a total of 39.3 million variants including 31,094,663 SNPs and 8,266,390 INDELs. Missing genotypes were imputed with Beagle (v.40)[55]. Genomic variants with MAF < 0.03 were removed.

Computing nucleotide diversity

Nucleotide diversity between pairs of breeds was computed from variant frequencies as follows:

$$\pi_i = \left(\sum_{j=1}^{n_i} 1 - \left(f_{ij}^{A} \times f_{ij}^{B} \right) - \left(\left(1 - f_{ij}^{A} \right) \times \left(1 - f_{ij}^{B} \right) \right) \right) / w$$

where π_i is the nucleotide diversity in window i, n_i is the number of variants in window i, f_{ij}^{A} is the frequency of variant j of window i in breed A, f_{ij}^{B} is the frequency of variant j of window i in breed B, and w is the size of the windows in base pairs. The overall nucleotide diversity for a pair of breeds A and B was computed as the average of π_i across all of the windows. The numbers reported are averages of overall nucleotide diversities for multiple pairs of breeds (within European, within Chinese, between European, between Chinese, between European and Chinese), computed for a window size of 1 million base pairs.

Estimating the contribution of the eight founder breeds in the F_6 and F_7 generation at the genome and chromosome level

We estimated the proportion of the genome of the eight founder breeds in the F_6 and F_7 generation according to ref. [56]. Assume that the total number of variants segregating in the mosaic population is n_T. Each of these variants has a frequency in each one of the founder breeds which we denote $f_1^{0.1} \rightarrow f_{n_T}^{0.1}$ for breed 1, $f_1^{0.2} \rightarrow f_{n_T}^{0.2}$ for breed 2, and so on, as well as a frequency in the F_6 (or F_7) generation, which we refer to as $f_1^6 \rightarrow f_{n_T}^6$. We assume that there is a total of B breeds. We denote the proportion of the genome of breed 1 in generation F_6 (or F_7) as P_1, of breed 2 in generation F_6 (or F_7) as P_2, and so on. We estimated the values of P_1, P_2, and so on, using a set of linear equations:

$$f_1^6 = \sum_{j=1}^{B} \left(P_j \times f_1^{0,j} \right)$$

$$\vdots$$

$$f_i^6 = \sum_{j=1}^{B} \left(P_j \times f_i^{0,j} \right)$$

$$\vdots$$

$$f_{n_T}^6 = \sum_{j=1}^{B} \left(P_j \times f_{n_T}^{0,j} \right).$$

We used standard least square methods (lm function in R) to find the solutions of P_j that minimize the residual sum of squares. This was done for the entire genome, as well as by autosome.

16S rRNA data collection and processing

Microbial DNA was extracted from faeces, luminal content and mucosal scrapings using the QIAamp Fast DNA stool Mini Kit according to the manufacturer's recommendations (Qiagen). DNA concentrations were measured using the Nanodrop-1000 instrument (Thermo Fisher Scientific), and DNA quality was assessed by agarose (0.8%) gel electrophoresis. The V3–V4 hypervariable region of the 16S rRNA gene was amplified with the barcode fusion primers (338F: 5-ACTCCTACGGGAGGCAGCAG-3, 806R: 5-GGACTACHVGGGTWTCTAAT-3) with 56 °C annealing temperature. After purification, PCR products were used for constructing libraries and sequenced on the Illumina MiSeq platform (Illumina) at Major Bio. The 16S rRNA sequencing data were submitted to the CNGB database and have accession number CNP0001069. The raw 16S rRNA gene sequencing reads were demultiplexed and primer and barcode sequences were trimmed using Trimmomatic (v.0.39)[57]. Reads with ≥10 consecutive same or ambiguous bases were eliminated. Clean paired-end reads were merged (minimum 10 bp overlap) into tags using FLASH (v.1.2.11)[58]. The average number of tags per sample was around 40,888 (Supplementary Table 2.1). Chimeric reads were removed using USEARCH (v.7.0.1090)[59]. Sequencing data were rarefied to 19,632 tags, that is, the lowest number of tags per sample. Tags were clustered in OTUs with VSEARCH (v.2.8.1)[60] using 97% as the similarity threshold. OTUs that would not have ≥3 reads in at least two samples or were detected in ≤0.2% of the samples were ignored. In the end, 12,054 OTUs accounting for an average of 98.7% of total reads per sample were used for further analysis. The mean number of tags for the 12,054 OTUs retained for further analyses was 1.6 (range, 0.01– 702.2). OTUs were matched to taxa using the Greengenes (v.13.5) database and the RDP classifier (v.2.2)[61]. PCoA was performed with the ape and vegan R packages (v.3.5.3) using Bray–Curtis dissimilarities. Shannon's index was used as α-diversity metric and computed using mothur (v.1.43.0)[62]. Bray–Curtis dissimilarity was used as β-diversity metric and computed using vegdist of the vegan package in R (v.3.5.3). The mouse faecal microbiome data were from ref. [63]. The human faecal microbiome data were 16S rRNA data from 106 healthy individuals (L.S. et al., manuscript in preparation).

Measuring the heritability of microbiota composition (all taxa combined)

We first estimated the effect of host genetics on the composition of the intestinal microbiome by measuring the correlation between genome-wide kinship and microbiome dissimilarity. We computed genome-wide kinship (Θ) for all pairs of relevant individuals using the SNP genotypes at the above-mentioned 30.2 million DNA variants using either GEMMA (v.0.97)[64] or GCTA (v.1.26)[65]. Both programs yielded estimates of Θ with the same distribution after standardization, albeit different raw values. We report results obtained using GEMMA (v.0.97). Microbiome dissimilarity was measured using the Bray–Curtis dissimilarity computed using the vegan R function[66] and abundances of all OTUs. We computed Spearman's (rank-based) correlations using the corrtest function in R (v.3.5.3). We first performed this analysis for each sample type and generation separately within litter, that is, only considering pairs of full-siblings born within the same litter, therefore in essence following ref. [67]. We then performed the analysis across the F_6 and F_7 generations. The pairs of individuals considered were all F_6–F_7 animal pairs, which included no parent–offspring pairs. To account for dependencies characterizing the data, *P* values were determined empirically by permutation testing. We performed 1,000 permutations of kinship coefficients and Bray–Curtis dissimilarities within litter. Vectors of OTU abundances were permuted within litters, Bray–Curtis distances were recomputed and correlated with the unpermuted kinships. The empirical *P* value was determined as the proportion of permutations that yielded a Spearman's correlation that was as low or lower than that obtained with the real data.

Measuring the heritability of the abundance of individual taxa

Heritabilities of log-transformed abundances of specific taxa/OTUs were estimated using a linear mixed model implemented with the lme4QTL R-package (v.3.5.3)[68]. *P* values for the heritability estimates were computed using the associated lrt_h2 function[68]. We first estimated heritabilities using information from the F_6 and F_7 generation jointly (F_6 plus F_7). The model included individual animal (that is, polygenic effect), dam, litter, pen and batch as random effects and generation as a fixed effect. The additive genetic relationship matrix determining covariances between individual animal effects was computed using GCTA (v.1.26)[65]. This F_6 plus F_7 analysis extracts information from the correlation between kinship and phenotypic resemblance (1) between F_6 animals (within F_6), (2) between F_7 animals (within F_7) and (3) between F_6 and F_7 animals (between F_6 and F_7). To evaluate whether these three sources of information were consistent, we also estimated—for each taxon or OTU—the three heritabilities separately. The within F_6 and within F7 heritabilities were estimated using the same mixed model (except for the absence of the fixed generation effect) and lme4QTL package as for the F_6 plus F_7 analysis. The between F_6 and F_7 heritabilities were estimated by regressing the squared difference in abundance on additive relationship according to ref. [69]—narrow sense heritability was estimated as $-\hat{\beta}/(2\hat{\sigma}_P^2)$ where $\hat{\beta}$ is the least-square regression coefficient and $\hat{\sigma}_P^2$ an estimate of the phenotypic variance (Extended Data Fig. 3). The total heritability of the intestinal microbiome was further computed from the heritabilities and abundance of individual taxa/OTU according to ref. [2].

Mapping miQTL

miQTL were mapped using the GenABEL R package (v.3.5.3)[70], applying two models according to ref. [11]. The first fitted a linear regression between allelic dosage and $\log_{10}$-transformed taxa abundance (additive model). It was applied to all SNPs with MAF ≥ 0.05 (in the corresponding data series) and taxa (that is, at all taxonomic levels including OTU) with non-null abundance in at least 20% of samples (in the corresponding data series), ignoring samples with null abundance if those represented more than 5% of samples. The second fitted a logistic regression model between allelic dosage and taxon presence/absence in the corresponding sample (binary model). It was applied only to taxa present in ≥20% and ≤95% of individuals and SNPs with MAF ≥ 10% (as the test statistic was inflated under the null when using this model with MAF < 10%; Extended Data Fig. 4). Both models included sex, slaughter batch (21 for F_6, 23 for F_7) and the three first genomic principal components as fixed covariates. GWAS were conducted separately for each taxon × data series combination and *P* values were concomitantly adjusted for residual stratification by genomic control. *P* values were combined across traits and/or taxa using a *Z*-score. *P* values were converted to signed *Z*-values using the inverse of the standard normal distribution and summed to give a *Z*-score. *Z*-scores were initially calculated using METAL (v.3.0)[71]. To compute the *P* value of the corresponding *Z*-score while accounting for the correlation that exists between the phenotypic values of a given cohort across traits, we also computed the genome-wide (that is, across all of the tested SNPs) average ($\overline{Z}$) and standard deviation (σ_Z) of the *Z* score. The *P* value of *Z* scores was (conservatively) computed by assuming that $(Z-\overline{Z})/\sigma_Z$ is distributed as $N(0,1)$ under the null hypothesis. Both approaches yielded similar results.

De novo assembly of the *A* allele of the porcine *AO* acetyl-galactosaminyl transferase gene

We extracted high-quality genomic DNA from longissimus dorsi of a Bamaxiang female using a phenol–chloroform-based extraction method (Novogene Biotech). A 40 kb SMRTbell DNA library (Pacific

Biosciences) was prepared using BluePippin for DNA size selection (Sage Science) and then sequenced on the PacBio Sequel platform (Pacific Biosciences) with P6-C4 chemistry at Novogene Biotech. We obtained a total of 18,148,470 subreads with an N_{50} length of 17,273 bp. Moreover, a paired-end library with insert size of 350 bp was constructed and sequenced on the Illumina Novaseq 6000 PE150 platform (2 × 150bp reads) at Novogene Biotech. PacBio reads were self-corrected using Canu (v.1.7.1) before assembly with Flye (v.2.4.2)[72]. Errors in the primary assembly were first corrected using PacBio subreads using racon (v.1.4.10)[73], and Illumina paired-end reads were then mapped to the contigs using bwa-mem[74] to polish the contigs using Pilon (v.1.23, Broad Institute)[75]. Lastz (v.1.02.00)[76] and Minimap2 (v.2.17-r941)[77] were used to compare the Bamaxiang contig and the 40 kb sequence spanning the *ABO* gene of the *S. scrofa* build 11.1 reference genome.

Developing a PCR assay to distinguish between *AA*, *AO* and *OO* pigs

We designed two pairs of primers to genotype the deletion in the F_6 and F_7 populations. The first pair of primers was located respectively within intron 7 of the ABO gene and downstream of the deletion (FP: 5′-GAGTTCCCCTTGTGGCTCAGT-3′, RP: 5′- TTGCCTAAGT CTACCCCTGTGC-3′). The second pair of primers was located in exon 8 (FP2: 5′-CGCCAGTCCTTCACCTACGAAC-3′, RP2: 5′-CGGTTC CGAATCTCTGCGTG-3′). PCR amplification was performed in a 25 µl reaction containing 50 ng genomic DNA and 1.5 U of LA Taq DNA polymerase (Takara) under thermocycle conditions of 94 °C for 4 min; 35 cycles of 94 °C for 1 min, 1 min at specific annealing temperature for each set of primers and 72 °C for 2 min; and 72 °C for 10 min on a PE 9700 thermal cycler (Applied Biosystem).

RNA-seq and eQTL analysis

A total of 300 caecum tissue samples from F_7 pigs that also had microbiota and genotype data were used to extract total RNA with TRIzol (Invitrogen) according to the manufacturer's manual. Total RNA was electrophoresed on 1% agarose gels. RNA purity and integrity were assessed using an eNanoPhotometer spectrophotometer (Implen) and a Bioanalyzer 2100 system (Agilent Technologies). A Qubit3.0 Fluorometer (Life Technologies) was used to measure RNA concentrations. Total RNA (2 µg) of each sample was used to constructed RNA-seq libraries, using the NEBNext UltraTMR NA Library Prep Kit for Illumina (NEB) according to the manufacturer's protocol. In brief, oligo(dT) magnetic beads (Invitrogen) were used to enrich mRNA, which was then fragmented using a fragmentation buffer (Ambion, USA). cDNA was synthesized using 6 bp random primers and reverse transcriptase (Invitrogen). After purification, cDNA was end-repaired, and index codes and sequencing adaptors were ligated. After PCR amplification, purification and quantification, the libraries were sequenced on the Novaseq-6000 platform using 2 × 150 bp paired-end sequencing. Clean data were obtained by removing adapter reads, poly-N and low-quality reads from raw data. Cleaned reads from each sample were mapped to the complete *ABO* sequence from the Bamaxiang reference genome with the *A* allele at the *ABO* locus constructed by the authors using STAR (v.020201)[78]. Samtools (v.1.6)[52] was used to convert SAM format to BAM format. The read counts mapping to *ABO* (exon 1 to 7) were quantified for each sample using featureCounts (v.1.6.4)[79]. The expression abundance of the *ABO* gene was normalized to FPKM. Gender and batch were treated as covariates to correct for gene expression levels, and the corrected residuals were used for subsequent analyses. GEMMA (v.0.97)[64] was used to analyse the association of *ABO* expression level with genome-wide variants using a linear mixed model.

Whole-genome sequencing and bioinformatic analysis for wild boars, *S. verrucosus* and *S. cebifrons*

The genomes of six Russian wild boars, one Sumatran wild boar and one African warty hog were sequenced on the Illumina HiSeq X Ten platform at Novogene Biotech. Furthermore, six Chinese wild boars were sequenced in a previous study[80], and we downloaded the genome sequencing data for eight other pigs from NCBI. Finally, we used a total of 22 genomes to call SNPs in the porcine *ABO* gene using GATK (v.4.2)[81]. We replaced the *ABO* gene of the *S. scrofa* build 11.1 genome with the 50 kb Bamaxiang contig sequence containing the *A* allele of *ABO* gene. The cleaned reads of the 22 individuals were aligned to the modified *S. scrofa* reference genome (build 11.1) using BWA (v.0.7.17)[51].

Phylogenetic analysis of the *O* alleles in the *Sus* genus

We applied GATK (v4.2) to perform indel realignment, and proceeded to SNP and INDEL discovery and genotyping with UnifiedGenotyper across all 83 samples simultaneously using standard hard filtering parameters according to GATK best practices recommendations[81,82]. We restricted the analysis to the 14 *AA* and 34 *OO* animals, therefore circumventing the need to phase the corresponding genotypes. We defined windows of varying size (0.5 to 50 kb) centred around the 2.3 kb deletion. For all pairs of individuals, we computed a running sum over all of the variants in the window adding 0 when both animals had genotype *AA* (alternate) or *RR* (reference), 1 when one animal was *AA* and the other *RR*, and 0.5 in all other cases. The nucleotide diversity for the corresponding animal pair was then computed as the running sum divided by the window size in bp. We ignored the variants located in the 2.3 kb deletion in this computation. The ensuing matrix of pairwise nucleotide diversities was then used for hierarchical clustering and dendrogram construction using the hclust (method="average") R function corresponding to the unweighted pair group method with arithmetic mean (UPGMA).

Analysis of population differentiation

We quantified the degree of population differentiation by computing the effect of breed on the variance of allelic dosage using a standard one-way ANOVA fixed-effect model and a *F*-statistic computed as the ratio of the between breed mean squares (BMS) and within breed mean squares (WMS)[83]. BMS and WMS were computed as:

$$\mathrm{BMS} = \left(\sum_{i=1}^{B} \sum_{j=1}^{n_i} \left(\overline{y_i} - \overline{y_\mathrm{T}} \right)^2 \right) / (B-1)$$

$$\mathrm{WMS} = \left(\sum_{i=1}^{B} \sum_{j=1}^{n_i} \left(y_{ji} - \overline{y_i} \right)^2 \right) / (N_\mathrm{T} - B)$$

where y_{ij} is the allelic dosage of the alternate allele in individual j (of n_i) of breed i (of B), $\overline{y_i}$ is the average allelic dosage in breed i (of B), and $\overline{y_\mathrm{T}}$ is the average allelic dosage in the entire dataset. We computed the average of the corresponding *F*-statistic for all variants within a sliding window of fixed physical size (2 kb in Fig. 5e), and took the inverse of this mean as measure of population similarity. The corresponding profiles were nearly identical to those obtained by computing average F_{ST} values (fixation index) across variants (and taking the inverse) following ref. [84].

Profiling *ABO* gene expression level at various adult and embryo tissues

Total RNA was extracted using Trizol from 15 tissues (lung, hypophysis, skin, spinal cord, liver, spleen, muscle, hypothalamus, heart, blood, brain, caecum, stomach, duodenum and kidney) collected from an adult Bamaxiang sow and a Duroc pig embryo (day 75). RNA quality was monitored by agarose (1%) gel electrophoresis, and using the RNA Nano 6000 Assay Kit of the Bioanalyzer 2100 system (Agilent Technologies). RNA concentration was measured using Qubit RNA Assay Kit in Qubit 2.0 Flurometer (Life Technologies). Total RNA (1 µg) of each sample was used to constructed RNA-seq libraries. Sequencing libraries were generated using the TruSeq RNA Library Preparation Kit (Illumina) according to the manufacturer's recommendations, and index codes

were added to attribute sequences to each sample. In brief, mRNA was purified from total RNA using poly-T oligo-attached magnetic beads. First-strand cDNA was synthesized using random hexamer primer and M-MuLV Reverse Transcriptase (RNase H^-). Second-strand cDNA synthesis was subsequently performed using DNA polymerase I and RNase H. The remaining overhangs were converted into blunt ends by exonuclease/polymerase activities. After adenylation of the 3′ ends of the DNA fragments, Illumina adaptors were ligated. To select cDNA fragments of preferentially around 350–400 bp in length, the library fragments were purified with AMPure XP system (Beckman Coulter). PCR was performed using Phusion High-Fidelity DNA polymerase, Universal PCR primers and Index (X) Primer. PCR products were purified (AMPure XP system) and the library quality was assessed on the Agilent Bioanalyzer 2100 system. Clustering of the index-coded samples was performed on the cBot Cluster Generation System using the TruSeq PE Cluster Kit v3-cBot-HS (Illumia) according to the manufacturer's instructions. Sequencing was performed on the Illumina NovaSeq platform and 150 bp paired-end reads were generated. Filtered reads were obtained by removing adapter sequences, poly-N and low-quality reads. Cleaned reads were mapped to the complete *ABO* sequence from the Bamaxiang reference genome sequence using HISAT2 (v.2.2.1)[85]. Samtools (v.1.6)[52] was used to convert SAM format to BAM format. The read counts mapping to ABO (exon 1 to 7) were quantified for each sample using featureCounts (v.1.6.4)[79]. To adjust for the effect of sequencing depth, the expression abundance of *ABO* gene was normalized to Transcripts Per Million (TPM). Expression abundance of *ABO* was used to cluster and visualize the expression level of the 15 tissues from an adult Bamaxiang sow and a Duroc pig embryo using the functions dist(), hclust(), as.dendrogram() and set() implemented in R package stats and dendextend.

Testing the effect of the *AO* genotype on other phenotypes

The associations between the 2.3 kb *ABO* deletion and 150 traits were calculated in the F_6 and F_7 populations based on a meta-analysis combining the effects. The observed *P* value for a trait was calculated by testing a weighted mean of *Z* scores from F_6 and F_7 generations as follows: $Z = (Z_1W_1 + Z_2W_2)/(W_1 + W_2)$, where $Z_1 = b_1/SE_1$ and $Z_2 = b_2/SE_2$, $W_1 = 1/(SE_1)^2$ and $W_2 = 1/(SE_2)^2$, where the subscripts 1 and 2 denote F_6 and F_7 generations, respectively; b_1, b_2, SE_1 and SE_2 were additive effects and standard errors of *ABO* locus on a given trait estimated from a linear mixed model, which accounted for population structure using a genomic relationship matrix derived from whole genome marker genotypes. A total of 250 and 254 traits were tested in the F_6 and F_7 generations, and 150 traits that were shared in the F_6 and F_7 generations were used for meta-analysis.

Determination of the concentration of *N*-acetyl-galactosamine in caecal lumen

Targeted LC–MS/MS analysis was performed to determine the concentration of the mixture of GalNAc, GlcNAc and ManNAc isomers (referred to as HexNAc) in caecal lumen samples. Samples of caecal content were collected at D240 for 124 F_7 and 154 animals from a Duroc × Landrace × Large White commercial population. The samples were lyophilized, grounded into powder by a Mixer Mill MM 400 (30 Hz, 1 min) (Retsch) and stored at −80 °C until use. Approximately 15 mg of powder was weighed and extracted with 500 µl of 70% methanol/water. The mixture was vortexed for 10 min, and centrifuged at around 16,000*g* and 4 °C for 10 min. Supernatant (300 µl) was transferred into a new centrifuge tube, placed at −20 °C for 30 min, and centrifuged again at 16,128*g* at 4 °C for 3 min. Finally, 150 µl of the supernatant was collected for further LC–MS analysis. Separation was performed in a 50 mm × 2.1 mm × 1.8 µm ACQUITY UPLC BEH C18 Column (Waters) in an ExionLC AD System (AB Sciex). Linear ion trap and triple quadrupole scans were performed on the Applied Biosystems 6500 Triple Quadrupole (QTRAP 6500) system equipped with an ESI Turbo Ion-Spray interface. The operation was performed under positive ion mode and controlled by Analyst software (v.1.6.3) (Sciex). The ESI source parameters were set as follows: ion source: ESI+, source temperature: 550 °C, ion spray voltage (IS): 5,500 V (Positive), curtain gas (CUR): 35 psi. The GalNAc was analysed using multiple reaction monitoring. We established a standard curve using increasing and known amounts of GalNAc standard and used it to determine GalNAc concentrations per gram of dried caecal content.

Measuring the correlation between GalNAc concentrations and bacterial abundance within the *AO* genotype

Raw AUC (GalNAc) or bacterial abundances were showing considerable batch-to-batch variation both with regards to means and variances. We therefore transformed them to ranks within batch and projected these uniformly between 0 and 1, yielding equalized means of 0.5 and standard deviations ≈ 0.3 (scaled batch-corrected ranks). We then applied the same approach within *AO* genotype to the scaled batch-corrected ranks yielding scaled, batch- and genotype-corrected ranks. We computed Spearman's correlations between the GalNAc and abundance scaled ranks.

Isolating 4-8-110 and 4-15-1

Faecal samples were collected from the rectum of healthy pigs at about 120 days and transferred immediately to anaerobic conditions. Fresh samples were homogenized with sterile 1× PBS (pH 7.0) in an anaerobic glovebox (Electrotek), which contained 10% hydrogen, 10% carbon dioxide and 80% nitrogen. The faecal suspension was diluted 10^{-6}, 10^{-7}, 10^{-8} and 10^{-9}-fold, and plated on GAM medium (Nissui Pharmaceutical). Plates were incubated at 37 °C for 3 days in an anaerobic glovebox. Single clones were picked and streaked until pure colonies were obtained on GAM medium. Full-length 16S rRNA gene sequencing was performed after amplification using the following primers: 27 forward: 5′-AGAGTTTGATCCTGGCCTCAG-3′; and 1492 reverse, 5′-GGTTACCTTGTTACGACTT-3′). The isolates were stored at −80 °C in GAM broth containing 16% of glycerol until further use.

Oxford Nanopore sequencing

The strains 4-8-110 and 4-15-1 were recovered on GAM medium. Cells were collected at the period of logarithmic growth. Genomic DNA was extracted using the Blood & Cell Culture DNA Midi Kit (Qiagen) according to the manufacturer's protocol. Libraries for whole-genome sequencing of the strains were constructed and sequenced on an ONT PromethION (Oxford Nanopore Technology) at NextOmics. To correct sequencing errors, a library for second-generation sequencing was also constructed for each of the two strains and sequenced (2 × 100 bp) on the BGISEQ platform (BGI). Bioinformatic analyses of sequencing data were performed following refs. [72,86]. In brief, after quality control, the sequencing data were assembled using flye (v.2.6)[72] with the parameter: --nano-raw, and the assembled genomes were corrected by combining the Oxford Nanopore data with the second-generation sequencing data using pilon with the default parameters. The encoded genes were predicted using prodigal (v.2.6.3) (parameter: -p none-g 11)[87].

MAG assembly

A total of 92 faecal samples from eight pig populations, four intestinal locations and different ages were used for metagenomic sequencing and construction of MAGs. Microbial DNA was extracted as described above. The libraries for metagenomic sequencing were constructed according to the manufacturer's instructions (Illumina), with an insert size of 350 bp for each sample, and 2 × 150 bp paired-ends sequenced on the NovaSeq 6000 platform. Raw sequencing data were filtered to remove adapter sequences and low-quality reads using fastp (v.0.19.41)[88]. Host genomic DNA sequences were filtered out using BWA (v.0.7.17)[51]. The clean reads of each sample were assembled into contigs using MEGAHIT (v.1.1.3) with the option '--min-count 2 --k-min 27 --k-max 87 --k-step 10 --min-contig-len 500'[89]. Single-sample metagenomic

binning was performed using two different binning algorithms '--metabat2 --maxbin2' using the metaWRAP package (v.1.1.1)[90]. The bins (MAGs) generated by the two binning algorithms were evaluated for quality and combined to form a MAG set using the bin_refinement module in metaWRAP (v.1.1.1). Metagenomic sequences were further assembled to optimize MAGs using the reassemble_bins module of metaSPAdes (v.3.15.3) in the metaWRAP pipeline. CheckM (v.1.0.12) was used to estimate the completeness and contamination of each MAG[91]. The MAGs with completeness <50% and contamination >5% were filtered out. Non-redundant MAGs were generated by dRep (v.2.3.2) at threshold of 99% average nucleotide identity (ANI)[92].

Bioinformatic analyses

Gene prediction in MAGs was carried out using the annotate_bins module in metaWRAP (v.1.1.1). The FASTA file of amino acid sequences translated from coding genes was used to perform KEGG annotation using Ghost KOALA tool (v.2.2)[93] on the KEGG website (https://www.kegg.jp/ghostkoala/). Taxonomic classification of MAGs was performed using PhyloPhlAn (v.0.99)[94]. The graphs in Fig. 5b, c were generated using custom Perl (v.5.10.1) and R scripts (v.3.5.3). Pathway and regulon scores were computed using a custom Perl script. Both scores included (1) one point for import (having orthologues of either the four components of AgaPTS (*agaE*, *agaF*, *agaV* and *agaW*) and/or the three components of the TonB dependent transporter (*omp*, *agaP* and *agaK*) and/or the four components of the GnbPTS transporter (*gnbA*, *gnbB*, *gnbC* and *gnbD*), (2) one point for GalNAc deacetylase activity (having an orthologue of *agaA* and/or *nagA*), (3) one point for GalN deaminase/isomerase (having on orthologue of *agaS*), (4) one point for tagatose-6-P kinase (having an orthologue of *pfkA* and/or *lacC* and/or *fruK*), and (5) one point for tagatose-1,6-PP aldolase (having an orthologue of *gatY* and/or *gatZ* and/or *lacD* and/or *fba*). For the pathway score, the orthologues could be located anywhere in the MAG; for the regulon score, they had to be located on the same sequence contig and in close proximity (2.5% of genome size) to the anchor gene *agaS*[28]. For the top hits, we manually checked whether proximity was confirmed either by the replication of the order in more than one MAG and/or by the colocalization of the genes on one and the same sequence contig. The effect of the MAG type (OTU476-like, Erysipelotrichaceae and others), completion, contig and genome size on pathway and regulon scores was estimated using the R lm function, and was highly significant. The *P* values for the Erysipelotrichaceae versus other contrast were directly obtained from the lm function. To conservatively estimate the *P* value of OTU476-like versus Erysipelotrichaceae + others we generated score residuals corrected for completion, contig number and genome size, and determined how many MAGs had scores as high or higher than the OTU476-like strains. The *P* values reported in Extended Data Fig. 10 correspond to the square of these proportions as there are two OTU476-like strains with same GalNAc cluster organization.

Determining the metabolic flux ratios from ^{13}C experiments by GC–MS

^{13}C-labelled GalNAc (*N*-acetyl-D-[UL-$^{13}C_6$]galactosamine) (99% isotopic purity of ^{13}C) was bought from ZZBIO (Shanghai). The 4-8-110 strain was cultured in 6 ml of GAM medium (Nissui Pharmaceuticals) adding 1 g l^{-1} of ^{13}C-labelled GalNAc or regular GalNAc (as control) in an anaerobic glovebox (Electrotek) in triplicate as described above. After 13 h of culture, the cultures were centrifuged for 5 min at 1,500*g*. Bacterial cells were resuspended in 0.6 ml of cold (−40 °C) 50% aqueous methanol containing 100 µM of norvaline as internal standard, plunged in dry ice for 30 min and then thawed on ice. We added 0.4 ml of chloroform, vortexed for 30 s, centrifuged for 10 min at 17,530*g* (4 °C), transferred the supernatant to new 1.5 ml tubes and dried the sample at −105 °C in a FreeZone Freeze Dryers (Labconco). Metabolites were derivatized for GC–MS analysis as follows. First, 70 µl of pyridine was added to the dried pellet and incubated for 20 min at 80 °C. After cooling, 30 µl of *N-tert*-butyldimethylsilyl-*N*-methyltrifluoroacetamide (Sigma-Aldrich) was added and the samples were reincubated for 60 min at 80 °C before centrifugation for 10 min at 17,530*g* (4 °C). The supernatant was transferred to an autosampler vial for GC/MS analysis. A Shimadzu QP-2010 Ultra GC-MS system was programmed for an injection temperature of 250 °C and injected with 1 µl of sample. The GC oven temperature started at 110 °C for 4 min, rising to 230 °C at 3 °C min^{-1} and then to 280 °C at 20 °C min^{-1} with a final hold at this temperature for 2 min. The GC flow rate with helium carrier gas was 50 cm s^{-1}. The GC column used was a 20 m × 0.25 mm × 0.25 mm Rxi-5 ms. The GC–MS interface temperature was 300 °C and the ion source temperature was set at 200 °C, with 70 V ionization voltage. The mass spectrometer was set to scan *m*/*z* range 50–800, with a 1 kV detector. GC–MS data were analysed to determine isotope labelling. To determine ^{13}C labelling, the mass distribution for known fragments of metabolites was extracted from the mass spectra. For each fragment, the retrieved data comprised mass intensities for the lightest isotopomer (without any heavy isotopes, M0), and isotopomers with increasing unit mass relative to M0. The mass distributions were normalized by dividing by the sum over all isoptomers and corrected for the natural abundance of heavy isotopes of the elements H, O and C using 7 × 7 matrix-based probabilistic methods as described previously[95–97] and implemented in MATLAB (release R2021a). Labelling results are expressed as the average fraction of the particular compound that contains isotopic label from the particular precursor.

Gavage experiment

4-8-110 and an *E. coli* strain isolated from caecal content (with 4-15-1 and 4-8-110) were cultured in GAM medium (Nissui Pharmaceuticals) for 13 h and 4 h, respectively, in an anaerobic glovebox (Electrotek) as described above. Optical density values were adjusted to 0.2 by addition of GAM medium, glycerol added to 1/6 of the final volume, and aliquots were stored at −80 °C. Germ-free mice (Kunming line) were provided by Huazhong Agricultural University. All of the experimental procedures involving mice were approved by the Ethics Committee in Huazhong Agricultural University (HZAUMO-2021-0077). Two groups (A and B) of five female mice each were housed in two separate cages (temperature, 25 ± 2 °C; humidity, 45–60%; light cycle, 12 h–12 h light–dark; light hours, 06:30–18:30). Mice were gavaged with 100 µl of PBS (group A) or 100 µl of PBS + 80 g l^{-1} GalNAc (corresponding to 8 mg per 100 µl or ~200 mg kg^{-1} of live weight) (group B) directly in the mouth with a sterile syringe twice per day (09:00 and 16:00) for 10 consecutive days (days 2 to 11). All of the mice were further gavaged with 150 µl of the 4-8-110 glycerol stocks (see above) once per day (at 16:00) for five consecutive days (days 3 to 7), and with 150 µl of the *E. coli* glycerol stocks (see above) once per day (at 16:00) for three consecutive days (days 5 to 7). The mice were euthanized 5 days after the last bacterial inoculation (on day 12, based on ref. [98]), and caecal content and faeces were collected for each mouse. Bacterial DNA was extracted using the QIAamp fast DNA stool mini kit (Qiagen). The V3–V4 region of the 16S rRNA gene was amplified and sequenced using the methods described above. Reads were mapped to the V3–V4 sequence of the previously determined 4-8-110 and *E. coli* strains. An average of 99.74% of the reads mapped either to the 4-8-110 or *E. coli* reference sequences.

Testing the inducibility of the GalNAc operon and transcriptome response upon GalNAc addition in *E. coli* and the OTU476-like strains

Bacterial growth. Bacteria were grown in an anaerobic glovebox (Electrotek) containing 10% hydrogen, 10% carbon dioxide and 80% nitrogen at 37 °C. Samples (4 ml) were pipetted after gentle shaking to measure optical density at 600 nm (OD_{600}) using an ultraviolet spectrophotometer (Yoke Instrument). Experiments were conducted in triplicate.

RNA extraction. The cultured bacterial cells of the two OTU476 like strains (4-8-110 and 4-15-1) and *E. coli* were collected at the end of the

exponential growth phase (11 h and 4 h, respectively). Total RNA was extracted from bacterial cells using TRIzol Reagent according to the manufacturer's manuals (Thermo Fisher Scientific), and genomic DNA was removed using DNase I (TaKara). The quantity and quality of total RNA were evaluated using a 2100 Bioanalyser (Agilent) and NanoDrop-2000 (Thermo Fisher Scientific). The RNA samples with $OD_{260/280}$ = 1.8–2.0, $OD_{260/230} \geq 2.0$, RIN ≥ 6.5, 28S:18S ≥ 1.0, ≥100 ng μl^{-1} and ≥2 μg were used to construct libraries for RNA-seq.

Construction of libraries and sequencing. The libraries for RNA-seq were constructed according to the TruSeq RNA sample preparation Kit (Illumina) using 2 μg of total RNA. In brief, rRNA was removed from total RNA using the Ribo-Zero Magnetic kit (Epicenter). All mRNAs were broken into short fragments (200 nucleotides) by adding fragmentation buffer. Double-stranded cDNA was synthesized using the SuperScript double-stranded cDNA synthesis kit (Invitrogen) with random hexamer primers (Illumina). cDNA was processed for end-repair, phosphorylation and A base addition according to library construction protocol (Illumina). The constructed libraries were sequenced on the NovaSeq 6000 platform with a 2 × 150 bp paired-end strategy. Base calling and quality value calculations were performed using the Illumina GA Pipeline (v.1.6) (Illumina).

Bioinformatics analysis. To obtain clean reads, low-quality sequences, reads with more than 5% of N bases (unknown bases) and reads containing adaptor sequences were removed from the raw data with a Perl script. The cleaned sequences were mapped to the reference genomic sequences of two OTU476 like strains (generated by ourselves as described above) and *E. coli* (GenBank: NC_000913.3) using Bowtie2 (v.2.4.2)[99]. The read counts mapping to the reference genomes were quantified for each sample using featureCounts (v.1.6.4). The fragments per kilobase of exon model per million mapped reads method was used to calculate the gene expression level. We used the DESeq2 package[100] in R (v.g3.5.3) to test for differential expression.

Analysing the effect of the *ABO* genotype on the abundance of Erysipelotrichaceae in humans

Samples. The data used correspond to the previously described CEDAR cohort[101], which included 300 healthy individuals of European descent who were visiting the University Hospital (CHU) from the University of Liège as part of a national screening campaign for colon cancer. Blood samples and intestinal biopsies (ileum, colon and rectum) were collected with full consent. The experimental protocol was approved by the ethics committee of the University of Liège Academic Hospital. Informed consent was obtained before donation in agreement with the recommendations of the declaration of Helsinki for experiments involving human participants.

Sequencing. For microbiota analysis, DNA was extracted from biopsies using the QIAamp DNA Stool Mini Kit (QIAgen). Three 16S rRNA amplicons corresponding to the V1–V2, V3–V4 and V5–V6 variable regions were generated in separate PCR reactions and processed for paired-end (2 × 300 bp) NGS sequencing on the MiSeq instrument (Illumina) following the standard protocol at the GIGA genomics core facility.

Data processing. Reads were QV20 trimmed from the 3′ end, demultiplexed, primer sequences were removed using the bbduk tool (version 38.82)[102]. Reads mapping to the human genome were eliminated using the BBTools suite (v.38.82). The corresponding pipeline was constructed using Snakemake (v.7.0.1)[103]. Further analyses were performed using QIIME 2 (v.2018.11)[104]. The paired-end reads were denoised and joined using the DADA2 plugin (v.1.16)[105] using batch-specific trimming length parameters yielding 9.1± 2.0 kb amplicon sequence variants (ASVs) per run for V1–V2, 4.5 ± 1.6 kb for V3–V4 and 6.8 ± 0.67 kb for V5–V6 amplicons. ASVs mapping to known contaminant taxa as well as ASVs with abundance negatively correlated with coverage depth were removed. Samples that contained more than 20% contaminant ASVs were eliminated from further analyses. ASVs were then clustered to 97% identity level OTUs using the DNACLUST program (v.r3)[106]. After OTU assignment, read counts were rarefied to 10,000 (V1–V2 and V5–V6) and 5,000 (V3–V4). As intestinal location only explored a minor proportion of the variance in OTU abundance (L.S. et al., manuscript in preparation), OTU abundances were averaged across locations. Local alignment identity of the detected ASVs with the OTU476 and OTU327 from the pig microbiome were measured using blastn[107].

Association analysis with ABO blood group. The effect of ABO blood group on standardized abundances of individual OTUs was performed using a linear model (lm R function) including (1) ABO blood group (A, B, AB or O), (2) secretor status, (3) sex, (4) smoking status, (5) age and (6) body mass index. Analyses were conducted separately for the different amplicons.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

All the 16S rRNA sequencing data, the metagenomics sequence data and the RNA-seq data were submitted to the GSA database under accession numbers CRA006230, CRA006239, CRA006240 and CRA006216. The genotype data were deposited at the GVM (http://bigd.big.ac.cn/gvm/getProjectDetail?project=GVM000310) under the GSA database under accession number GVM000310. The GWAS summary statistics are available at Figshare (https://doi.org/10.6084/m9.figshare.19313960). The whole-genome sequencing data of experimental pigs have been deposited in the GSA database (https://ngdc.cncb.ac.cn/gsa/browse/CRA006383) under accession number CRA006383. The source data are available at GitHub (https://github.com/yanghuijxau/Manuscript-microbiota-ABO).

Code availability

Codes to replicate the findings and the source data are available at GitHub (https://github.com/yanghuijxau/Manuscript-microbiota-ABO).

50. Warr, A. et al. An improved pig reference genome sequence to enable pig genetics and genomics research. Gigascience 9 (2019).
51. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26**, 589–595 (2010).
52. Li, H. et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
53. Rimmer, A. et al. Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications. *Nat. Genet.* **46**, 912–918 (2014).
54. Chang, C. C. et al. Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4**, 7 (2015).
55. Browning, S. R. & Browning, B. L. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am. J. Hum. Genet.* **81**, 1084–1097 (2007).
56. Coppieters, W., Karim, L. & Georges, M. SNP-based quantitative deconvolution of biological mixtures: application to the detection of cows with subclinical mastitis by whole genome sequencing of tank milk. *Genome Res.* **30**, 1201–1207 (2020).
57. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
58. Magoč, T. & Salzberg, S. L. FLASH: fast length adjustment of short reads to improve genome assemblies. *Bioinformatics* **27**, 2957–2963 (2011).
59. Edgar, R. C. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26**, 2460–2461 (2010).
60. Rognes, T., Flouri, T., Nichols, B., Quince, C. & Mahe, F. VSEARCH: a versatile open source tool for metagenomics. *PeerJ* **4**, e2584 (2016).
61. Wang, Q., Garrity, G. M., Tiedje, J. M. & Cole, J. R. Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl. Environ. Microbiol.* **73**, 5261–5267 (2007).
62. Schloss, P. D. et al. Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl. Environ. Microbiol.* **75**, 7537–7541 (2009).

63. Cheema, M. U. & Pluznick, J. L. Gut microbiota plays a central role to modulate the plasma and fecal metabolomes in response to angiotensin II. *Hypertension* **74**, 184–193 (2019).
64. Zhou, X. & Stephens, M. Genome-wide efficient mixed-model analysis for association studies. *Nat. Genet.* **44**, 821–824 (2012).
65. Yang, J., Lee, S. H., Goddard, M. E. & Visscher, P. M. GCTA: a tool for genome-wide complex trait analysis. *Am. J. Hum. Genet.* **88**, 76–82 (2011).
66. Dixon, P. VEGAN, a package of R functions for community ecology. *J. Veg. Sci.* **14**, 927–930 (2003).
67. Visscher, P. M. et al. Assumption-free estimation of heritability from genome-wide identity-by-descent sharing between full siblings. *PLoS Genet.* **2**, e41 (2006).
68. Ziyatdinov, A. et al. lme4QTL: linear mixed models with flexible covariance structure for genetic studies of related individuals. *BMC Bioinform.* **19**, 68 (2018).
69. Haseman, J. K. & Elston, R. C. The investigation of linkage between a quantitative trait and a marker locus. *Behav. Genet.* **2**, 3–19 (1972).
70. Aulchenko, Y. S., Ripke, S., Isaacs, A. & van Duijn, C. M. GenABEL: an R library for genome-wide association analysis. *Bioinformatics* **23**, 1294–1296 (2007).
71. Willer, C. J., Li, Y. & Abecasis, G. R. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* **26**, 2190–2191 (2010).
72. Kolmogorov, M., Yuan, J., Lin, Y. & Pevzner, P. A. Assembly of long, error-prone reads using repeat graphs. *Nat. Biotechnol.* **37**, 540–546 (2019).
73. Vaser, R., Sovic, I., Nagarajan, N. & Sikic, M. Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Res.* **27**, 737–746 (2017).
74. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. Preprint at *arXiv* https://arxiv.org/abs/1303.3997 (2013).
75. Walker, B. J. et al. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS ONE* **9**, e112963 (2014).
76. Harris, R. S. *Improved Pairwise Alignment of Genomic DNA*. PhD thesis, Pennsylvania State Univ. (2007).
77. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
78. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
79. Liao, Y., Smyth, G. K. & Shi, W. FeatureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923–930 (2014).
80. Ai, H. et al. Adaptation and possible ancient interspecies introgression in pigs identified by whole-genome sequencing. *Nat. Genet.* **47**, 217–225 (2015).
81. Van der Auwera, G. A. et al. From FastQ data to high confidence variant calls: the genome analysis Toolkit best practices pipeline. *Curr. Protoc. Bioinform.* **43**, 11.10.11–11.10.33 (2013).
82. DePristo, M. A. et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43**, 491–498 (2011).
83. Weir, B. S. & Cockerham, C. C. Estimating *F*-statistics for the analysis of population structure. *Evolution* **38**, 1358–1370 (1984).
84. Nei, M. *F*-statistics and analysis of gene diversity in subdivided populations. *Ann. Hum. Genet.* **41**, 225–233 (1977).
85. Kim, D., Paggi, J. M., Park, C., Bennett, C. & Salzberg, S. L. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat. Biotechnol.* **37**, 907–915 (2019).
86. Hunt, M. et al. Circlator: automated circularization of genome assemblies using long sequencing reads. *Genome Biol.* **16**, 294 (2015).
87. Hyatt, D. et al. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinform.* **11**, 119 (2010).
88. Chen, S., Zhou, Y., Chen, Y. & Jia, G. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**, i884–i890 (2018).
89. Li, D. et al. MEGAHIT v1.0: a fast and scalable metagenome assembler driven by advanced methodologies and community practices. *Methods* **102**, 3–11 (2016).
90. Uritskiy, G. V., DiRuggiero, J. & Taylor, J. MetaWRAP—a flexible pipeline for genome-resolved metagenomic data analysis. *Microbiome* **6**, 158 (2018).
91. Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P. & Tyson, G. W. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res.* **25**, 1043–1055 (2015).
92. Olm, M. R., Brown, C. T., Brooks, B. & Banfield, J. F. dRep: a tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication. *ISME J.* **11**, 2864–2868 (2017).
93. Kanehisa, M., Sato, Y. & Morishima, K. BlastKOALA and GhostKOALA: KEGG tools for functional characterization of genome and metagenome sequences. *J. Mol. Biol.* **428**, 726–731 (2016).
94. Segata, N., Bornigen, D., Morgan, X. C. & Huttenhower, C. PhyloPhlAn is a new method for improved phylogenetic and taxonomic placement of microbes. *Nat. Commun.* **4**, 2304 (2013).
95. Li, M. et al. Aldolase B suppresses hepatocellular carcinogenesis by inhibiting G6PD and pentose phosphate pathways. *Nat. Cancer* **1**, 737–747 (2020).
96. Nanchen, A., Fuhrer, T. & Sauer, U. Determination of metabolic flux partios from ^{13}C-experiments and gas chromatography-mass spectrometry data: protocol and principles. *Methods Mol. Biol.* **358**, 177–197 (2007).
97. van Winden, W. A. et al. Correcting mass isoptopomer distributions for naturally occurring isotopes. *Biotechnol. Bioeng.* **80**, 477–479 (2002).
98. Staley et al. Stable engraftment of human microbiota into mice with a single oral gavage following antibiotic conditioning. *Microbiome* **5**, 87 (2017).
99. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
100. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
101. Momozawa, Y. et al. IBD risk loci are enriched in multigenic regulatory modules encompassing putative causative genes. *Nat. Commun.* **9**, 2427 (2018).
102. Bushnell, B. *BBMap: A Fast, Accurate, Splice-Aware Aligner* (version 38.82) https://sourceforge.net/projects/bbmap/ (2014).
103. Köster, J. & Rahmann, S. Snakemake: a scalable bioinformatics workflow engine. *Bioinformatics* **28**, 2520–2522 (2012).
104. Bolyen, E., Rideout, J. R., Dillon, M. R., Bokulich, N. A. & Caporaso, J. G. Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nat. Biotechnol.* **37**, 852–857 (2019).
105. Callahan, B. J. et al. DADA2: high-resolution sample inference from Illumina amplicon data. *Nat. Methods* **13**, 581–583 (2016).
106. Ghodsi, M., Liu, B. & Pop, M. DNACLUST: accurate and efficient clustering of phylogenetic marker genes. *BMC Bioinform.* **12**, 271 (2011).
107. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
108. Srivastava, A. et al. Genomes of the mouse collaborative cross. *Genetics* **206**, 537–556 (2017).
109. Yu, N. et al. Global patterns of human DNA sequence variation in a 10-kb region on chromosome 1. *Mol. Biol. Evol.* **18**, 214–222 (2001).
110. The 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061–1073 (2010).
111. Frantz, L. A. F. et al. Evidence of long-term gene flow and selection during domestication from analyses of Eurasian wild and domestic pig genomes. *Nat. Genet.* **47**, 1141–1148 (2015).
112. Charlier, C. et al. NGS-based reverse genetic screen for common embryonic lethal mutations compromising fertility in livestock. *Genome Res.* **26**, 1333–1341 (2016).
113. Georges, M., Charlier, C. & Hayes, B. Harnessing genomic information for livestock improvement. *Nat. Rev. Genet.* **20**, 135–156 (2019).
114. Geraldes, A. et al. Inferring the history of speciation in house mice from autosomal, X-linked, Y-linked and mitochondrial genes. *Mol. Ecol.* **17**, 5349–5363 (2008).
115. Huang, H. et al. Fine-mapping inflammatory bowel disease loci to single-variant resolution. *Nature* **547**, 173–178 (2017).
116. Suzuki, T. A. & Nachman, M. W. Spatial heterogeneity of gut microbioal composition along the gastrointestinal tract in natural populatiions of house mice. *PLoS ONE* **11**, e0163720 (2016).
117. Vuik, F. E. R. et al. Composition of the mucosa-associated microbiota along the entire gastrointestinal tract of human individuals. *UEG J.* **7**, 897–907 (2019).
118. Rowe, J. A. et al. Blood group O protects against severe *Plasmodium falciparum* malaria through the mechanism of reduced rosetting. *Proc. Natl Acad. Sci. USA* **104**, 17471–17476 (2007).
119. Robinson, M. G., Tolchin, D. & Halpern, C. Enteric bacterial agents and the ABO blood groups. *Am. J. Hum. Genet.* **23**, 135–145 (1971).
120. Camus, D., Bina, J. C., Carlier, Y. & Santoro, F. ABO blood groups and clinical forms of schistosomiasis mansoni. *Trans. R. Soc. Trop. Med. Hyg.* **71**, 182 (1977).
121. Pereira, F. E. L., Bortolini, E. R., Carneiro, J. L. A., da Silva, C. R. M. & Neves, R. C. A, B, O blood groups and hepatosplenic form of schistosomiasis mansoni (Symmer's fibrosis). *Trans. R. Soc. Trop. Med. Hyg.* **73**, 238 (1977).
122. Ndamba, J., Gomo, E., Nyazema, N., Makaza, N. & Kaondera, K. C. Schistosomiasis infection in relation to the ABO blood groups among school children in Zimbabwe. *Acta Trop.* **65**, 181–190 (1997).
123. Chaudhuri, A. & De, S. Cholera and blood groups. *Lancet* **2**, 404 (1977).
124. Boren, T. et al. Attachment of *Helicobacter pylori* to human gastric epithelium mediated by blood group antigens. *Science* **262**, 1892–1895 (1993).
125. Lindesmith, L. et al. Human susceptibility and resistance to Norwalk virus infection. *Nat. Med.* **9**, 548–553 (2003).
126. Galili, U. in *α-Gal and Anti-Gal* (eds Galili, U. & Avila, J. L.) Vol. 32, 1–23 (Springer, 1999).
127. Prather, R. S., Shen, M. & Dai, Y. Genetically modified pigs for medicine and agriculture. *Biotechnol. Genetic Eng. Rev.* **25**, 245–266 (2008).

Acknowledgements We thank Y. He, S. Xiao, W. Li, Y. Guo and Y. Xing for assistance in the construction of the experimental mosaic pig populations; Y. Su and J. Li for preparation of reagents and management of samples; Y. Momozawa, R. Mariman, M. Mni, L. Karim and M. Dekkers for generating the CEDAR-1 16S rRNA data; the staff at the Jiangxi Department for Education, the Ministry of Science and Technology of P. R. China, the Ministry of Agriculture and Rural Affairs of P. R. China, and Jiangxi department of Science and Technology for their long-term support of the swine heterogeneous stock project; and the members of the MIQUANT consortium for comments and discussions. L.H. is supported by The National Natural Science Foundation of China (31790410) and National pig industry technology system (CARS-35); C. Chen by the National Natural Science Foundation of China (31772579); H.Y. by the National Postdoctoral Program for Innovative Talent (no. BX201700102); L.S. by the FNRS IBD-GI-Seq project; M.G. by the Chinese Thousand Talents Program, the Belgian EOS 'Miquant' project and the FNRS (CDR 'GEM' project). C. Charlier is a senior research associate at the FNRS.

Author contributions H.Y. analysed the 16S rRNA sequence data, performed GWAS, meta-analyses and local association analyses, computed heritabilities of individual taxa, contributed to ABO genotyping and analysed the effect of the 2.3 kb deletion on taxa abundance. J.W. analysed the composition of the microbiome, including PCoA analyses, *β*- and *α*-diversity, correlations between kinship and microbiome dissimilarities, isolated the OTU476-like strains, performed the GalNAc feeding experiments, measured the concentrations of GalNAc in the caecal lumen, analysed the GalNAc import and use pathway in the MAGs, and contributed to ABO genotyping. X.H. participated in 16S rRNA sequencing (F_6) and GWAS (F_6). Y. Zhou performed metagenome sequencing analysis, analysed the GalNAc import and use pathway in MAGs, analysed the RNA-seq data from caecum samples and contributed to ABO genotyping. Y. Zhang participated in the preparation of the genotype data from whole-genome sequence information, participated in the computation of the genomic contribution of the different breeds in the F_6 and F_7 generation and the definition of expected mapping resolution, performed LD analyses, performed eQTL analysis for the *ABO* gene, participated in the characterization and sequence analysis of the *ABO* gene, including definition of the 2.3 kb deletion, and in the balancing selection and trans-species

polymorphism analyses. M.L. assisted with the isolation of the OTU476-like strains, the GalNAc feeding experiments and genotyping of the *ABO* gene. Q.L. assisted with measuring the concentrations of GalNAc in caecal lumen. S.K., M.H., H.F., S.F., X.X., H.J., Z.C. and J.G. assisted with the experiments. Z.Z., X.T., Z.W., H.G. and Y.H. assisted with the preparation of genotype data from whole-genome sequencing data and conducted the analysis of the Nanopore data of the *ABO* region. J.M. assisted with the construction of the mosaic population. H.A. assisted with the bioinformatic analysis of the *ABO* region, de novo assembly of the *A* allele, and evolutionary analysis of the *ABO* alleles. L.S. analysed the effect of *ABO* genotype on intestinal microbiota composition in humans. W.C. assisted in the analysis of the sequencing data for the trans-species polymorphisms. C. Charlier supervised the characterization of the *ABO* gene and the 2.3 kb deletion and the corresponding haplotype structure in the F_0, F_6 and F_7 population and for the trans-species polymorphism. B.Y. prepared the genotype data of whole-genome variants, assisted with raising the heterogeneous stock, and participated in the computation of the genomic contribution of the different breeds in the F_6 and F_7 generation and the definition of expected mapping resolution. M.G. supervised the bioinformatic and statistical analyses, performed bioinformatic and statistical analyses, and wrote the paper. C. Chen codesigned the study, supervised experiments, supervised bioinformatic and statistical analyses of gut microbiome, and wrote the paper. L.H. created the swine heterogeneous stock, designed the study, directed the project, supervised the experiments and analyses, and wrote the paper.

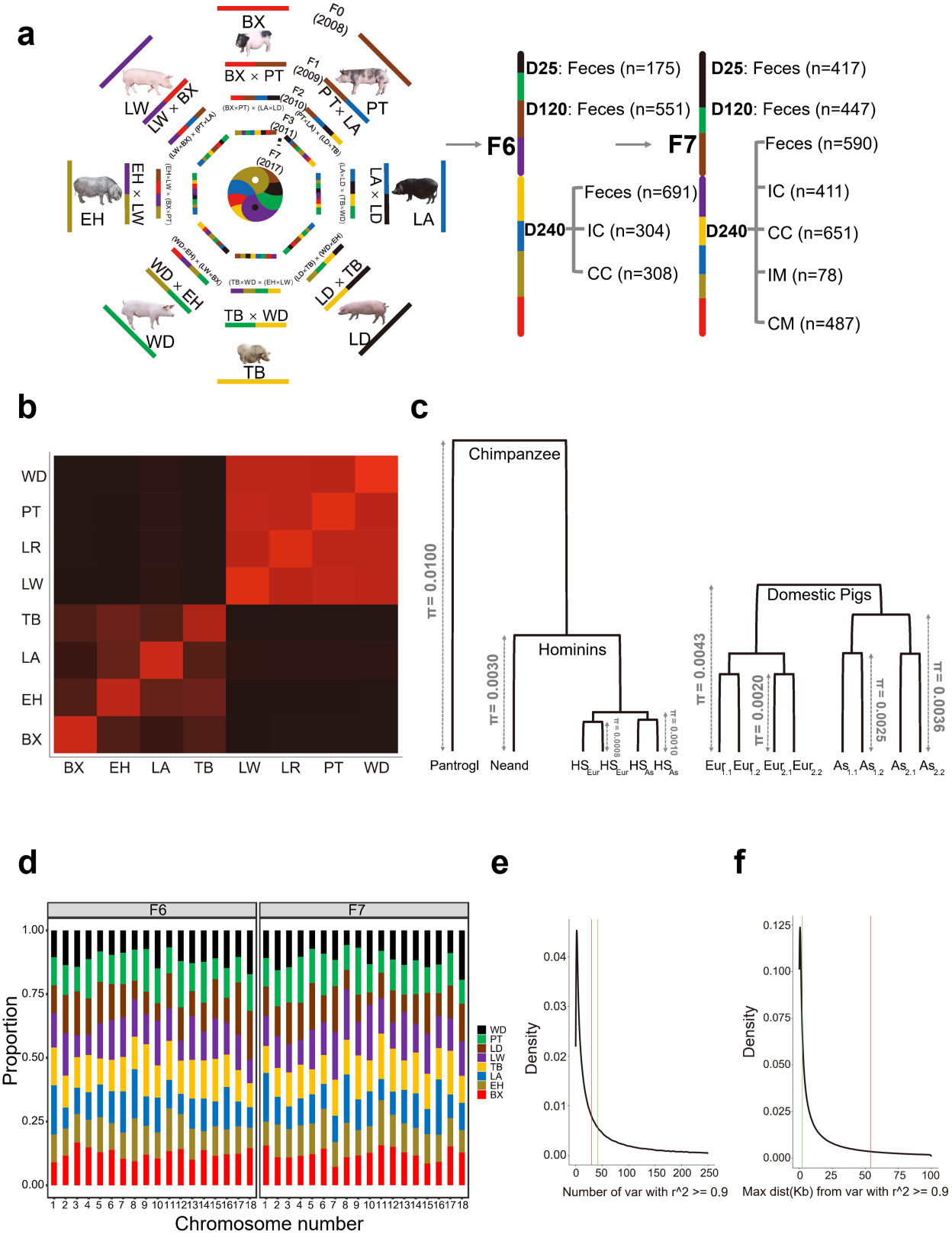
Competing interests The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41586-022-04769-z.

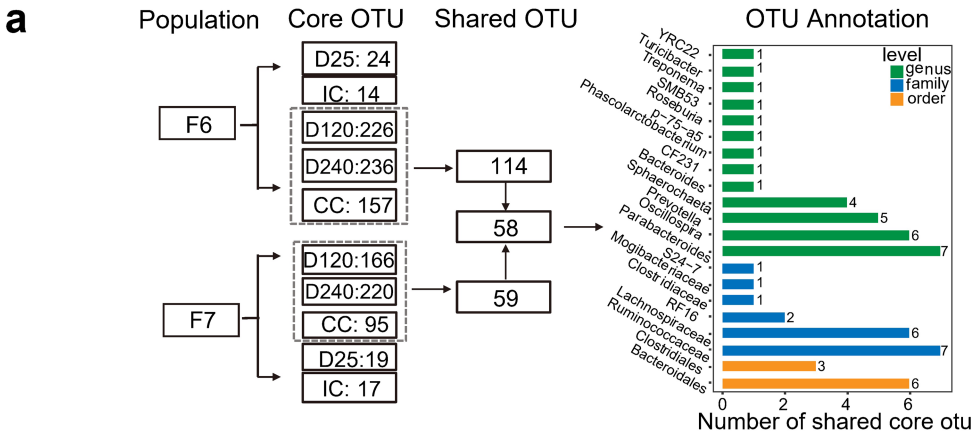
Correspondence and requests for materials should be addressed to Michel Georges, Congying Chen or Lusheng Huang.

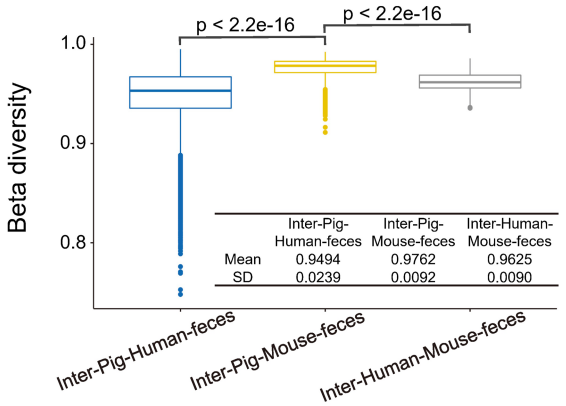
Peer review information *Nature* thanks Catherine Lozupone, Vincent Plagnol and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.



Extended Data Fig. 1 | See next page for caption.

Extended Data Fig. 1 | Generating a large mosaic pig population for genetic analysis of complex phenotypes. (a) Rotational breeding design used for the generation of a large mosaic pig population for the genetic analysis of complex phenotypes, with sampling scheme for faeces (D25, D120, D240), luminal content of the ileum (IC) and caecum (CC), and mucosal scrapings in the ileum (IM) and caecum (CM). BX: Bamaxiang, EH: Erhualian, LA: Laiwu, TB: Tibetan, LW: Large White, LD: Landrace, PT: Piétrain, WD: White Duroc. (b) Average similarity ($1 - \boldsymbol{\pi}$) between allelic sequences sampled within and between the eight founder breeds. The colour intensity ranges from black (breeds with lowest allelic similarity: BX vs WD, $1 - 4.3 \times 10^{-3}$) to bright red (breed with highest allelic similarity: WD, $1 - 1.8 \times 10^{-3}$). The acronyms for the breeds are as in (a). More than 30 million variants with MAF ≥ 3% segregate in this population, i.e. more than one variant every 100 base pairs. This is slightly lower than the 40 million high quality variants segregating in the mouse collaborative cross[108]. (c) Comparison of the average nucleotide diversity ($\boldsymbol{\pi}$, i.e. the proportion of sites that differ between two chromosomes sampled at random in the population(s)) within and between European (Eur) and Asian (As) domestic pigs, and between modern European (HS_{Eur}), Asian humans (HS_{As}), Neanderthal (Neand) and Chimpanzee (Pan Trogl). The average nucleotide diversity within the four Chinese founder breeds was ~2.5×10^{-3} and within the four European founder breeds ~2.0×10^{-3}. By comparison, π-values within African and within Asian/European human populations are ~9×10^{-4} and ~8×10^{-4}, respectively[109,110]. Thus, against intuition (as domestication is often assumed to have severely reduced effective population size) the within population diversity is >2-fold higher in domestic pigs than in human populations, as previously reported[111–113]. Nucleotide diversities between Chinese founder breeds and between European founder breeds were ~3.6×10^{-3} and ~2.5×10^{-3}, respectively, i.e. 1.44-fold and 1.25-fold higher than the respective within-breed π-values. These π-values are of the same order of magnitude as the sequence divergence between *Homo sapiens* and Neanderthals/Denosivans (~3×10^{-3}, ref. [15]). By comparison, π-values between Africans, Asians and Europeans are typically ≤ ~1×10^{-3} (ref. [109]). The nucleotide diversity between Chinese and European breeds averaged ~4.3×10^{-3}. This π-value is similar to the divergence between *M. domesticus* and *M. castaneus*[114], and close to halve the ~1% difference between chimpanzee and human[16]. Note that Chinese and European pig breeds are derived from Chinese and European wild boars, respectively, which are thought to have diverged ~1 million years ago[27], while *M. domesticus* and *M. castaneus* are thought to have diverged ≤ 500,000 years ago[114]. (d) Autosome-specific estimates of the genomic contributions of the eight founder breeds in the F6 and F7 generation. We used a linear model incorporating all variants to estimate the average contribution of the eight founder breeds in the F6 and F7 generation at genome and chromosome level[56]. At genome-wide level, the proportion of the eight founder breed genomes ranged from 11.2% (respectively 11.5%) to 14.1% (14.7%) in the F6 (F7) generations. At chromosome-specific level, the proportion of the eight founder breeds ranged from 6.7% (respectively 4.9%) to 20.7% (22.1%) in the F6 (F7) generations. The genomic contribution of the eight founder breeds in the F6 and F7 generation is remarkably uniform and close to expectations (i.e. 12.5%) both at genome-wide and chromosome-wide level, suggesting comparable levels of genetic diversity across the entire genome. This does not preclude that more granular examination may reveal local departures from expectations, or under-representation of incompatible allelic combinations at non-syntenic loci. (e-f) Indicators of achievable mapping resolution in the F6 generation: (e) Frequency distribution (density) of the number of variants in high LD ($r^2 \geq 0.9$) with an "index" variant (was computed separately for all variants considered sequentially as the "index"), corresponding to the expected size of "credible sets" in GWAS[115]. The red vertical line corresponds to the genome-wide median. The green vertical line corresponds to the mapping resolution achieved in this study for the ABO locus (see hereafter). (f) Frequency distribution (density) of the maximum distance between an index variant and a variant in high LD ($r^2 \geq 0.9$) with it, defining the spread of credible sets. Red and green vertical lines are as in (D).



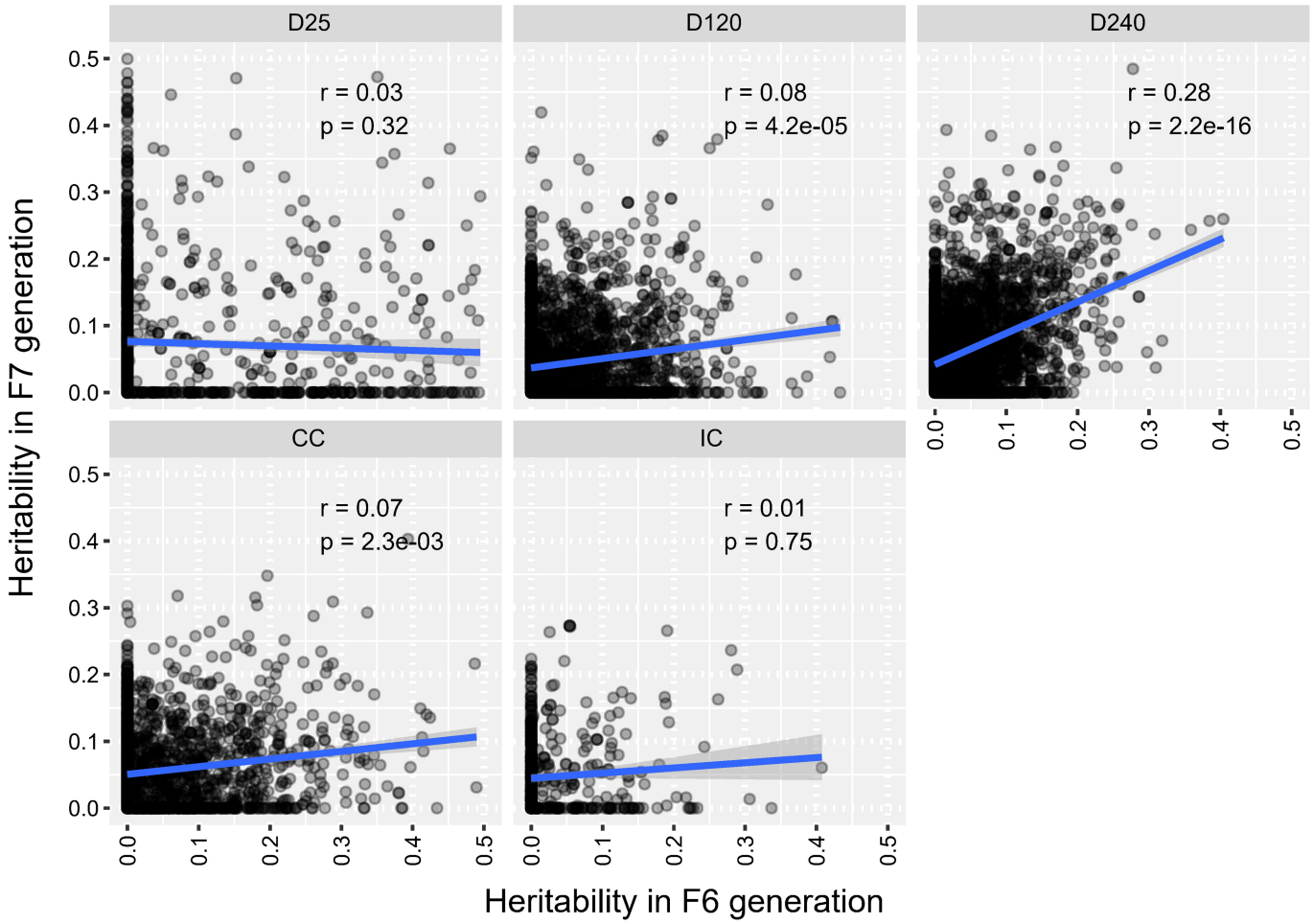


c

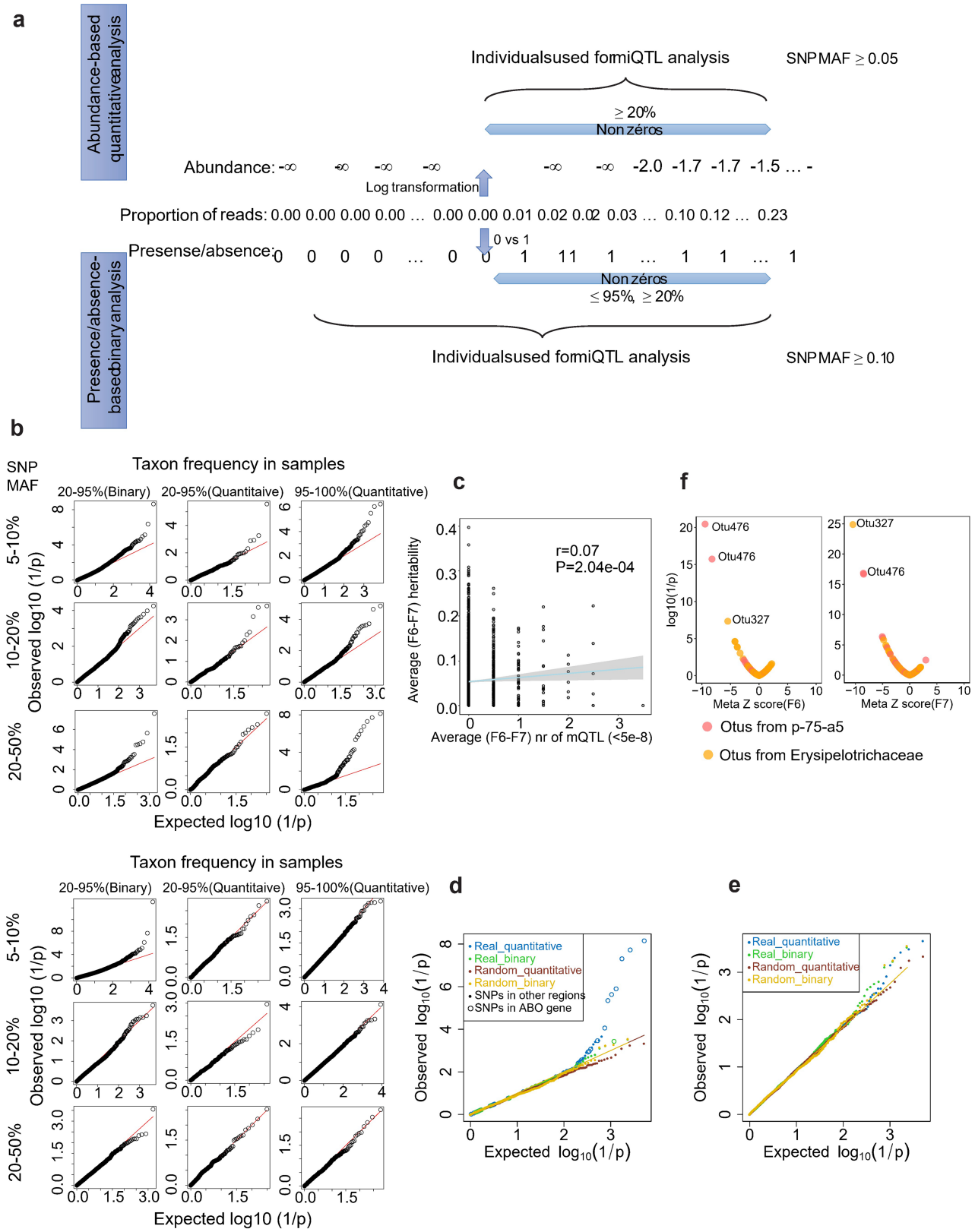
PHYLUM	FAMILY	ILEUM IC	ILEUM IM	CAECUM CC	CAECUM CM	FAECES 25	FAECES 120	FAECES 240	MOUSE P_I	MOUSE D-I	MOUSE C	MOUSE CO	MOUSE RE	MOUSE F	HUMAN P_I	HUMAN D_I	HUMAN CO	HUMAN RE
Firmicutes	Leuconostocaceae	1	0.1	0	0	0	0	0										
Actinobacteria	Microbacteriaceae	1	0.1	0	0	0	0	0										
Firmicutes	Bacillaceae	1	0.2	0	0	0.1	0	0										
Fusobacteria_other_classfied	Fusobacteria_other_classfied	1	0.3	0	0	0.1	0	0										
Firmicutes	Tissierellaceae	1	0.3	0.1	0.1	0.3	0.1	0.1							1	1	1	1
Proteobacteria	Pasteurellaceae	1	0.3	0.1	0.1	0	0	0	1	0.3	0	0	0	0	1	0.3	0.3	0.3
Actinobacteria	Corynebacteriaceae	1	0.4	0.1	0.2	0.3	0.2	0.1							1	0	0	0
Proteobacteria	Enterobacteriaceae	1	0.5	0.1	0.1	0.1	0	0							1	1	0.8	0.5
Firmicutes	Peptostreptococcaceae	1	0.5	0	0.1	0	0	0										
Firmicutes	Clostridiaceae	1	0.6	0.1	0.1	0	0	0							0.3	1	0.5	0.8
Firmicutes	Turicibacteraceae	1	0.9	0.1	0.2	0	0	0										
Thermi	Thermaceae	1	0.8	0	0	0.4	0	0										
Firmicutes_other_classfied	Firmicutes_other_classfied	1	0.8	0.3	0.4	0.6	0.3	0.3										
Proteobacteria	Pseudomonadaceae	0	1	0	0	0	0	0							1	0.1	0.2	0.2
Thermi_other_classfied	Thermi_other_classfied	0.3	1	0	0.2	0.3	0	0										
Fusobacteria	Fusobacteriaceae	0.6	1	0.5	1	0.9	0	0										
Proteobacteria_other_classfied	Proteobacteria_other_classfied	0.7	1	0.6	0.9	0.8	0.6	0.7										
Verrucomicrobia	Verrucomicrobiaceae	0	0.2	1	0.4	0.3	0.1	0.5							0	0.3	0.3	1
Bacteroidetes	Paraprevotellaceae	0	0.2	1	0.7	0.3	0.4	0.4							0.4	1	0.4	0.4
Bacteroidetes	Porphyromonadaceae	0	0.3	1	0.7	0.5	0.4	0.5	0	0	1	1	1	0.5	0.3	1	1	1
Bacteroidetes_other_classfied	Bacteroidetes_other_classfied	0	0.3	1	0.7	0.3	0.6	0.9										
Verrucomicrobia_other_classfied	Verrucomicrobia_other_classfied	0	0.2	0.8	1	0.2	0.1	0.2										
Proteobacteria	Alcaligenaceae	0.1	0.4	0.7	1	0.3	0.2	0.3							0	1	0.7	0.8
Tenericutes	Anaeroplasmataceae	0	0.4	0.7	1	0	0.5	0.8										
Proteobacteria	Campylobacteraceae	0.1	0.4	0.3	1	0.4	0.1	0.1										
Proteobacteria	Helicobacteraceae	0.2	0.3	0.2	1	0.3	0	0	0.1	1	0.3	0.3	0.2	0.2	1	0	0	0
Spirochaetes	Sphaerochaetaceae	0	0.3	0.6	1	0.3	0.4	0.3										
Bacteroidetes	RF16	0	0.3	0.7	1	0	0.4	0.6										
Unclassified	Unclassified	0	0.5	0.7	1	0.6	0.5	0.7										
Other_classfied	Other_classfied	0.2	0.7	0.9	1	1	0.8	0.8										
Bacteroidetes	Rikenellaceae	0	0.1	0.1	0.1	1	0	0.1	0	0	0.7	1	1	0.5	0	0.5	0.8	1
Bacteroidetes	Odoribacteraceae	0	0	0	0	1	0	0	0	0	1	0.5	0.5	0				
Firmicutes	Erysipelotrichaceae	0	0.1	0.3	0.3	1	0.3	0.2										
Bacteroidetes	p-2534-18B5	0	0.1	0.1	0.1	1	0.4	0.3										
Bacteroidetes	S24-7	0	0.2	0.5	0.3	1	0.5	0.3	0	0	0.8	0.8	0.8	1				
Bacteroidetes	Barnesiellaceae	0	0.5	0	0	1	0	0.1										
Actinobacteria_other_classfied	Actinobacteria_other_classfied	0.7	0.4	0.1	0.1	1	0.2	0.3										
Bacteroidetes	Bacteroidaceae	0.1	0.3	0.8	1	1	0.1	0.1	0	0.1	0.5	0.8	1	0.5	0.1	0.8	1	0.9
Proteobacteria	Desulfovibrionaceae	0.4	0.3	0.6	0.6	1	0.4	0.3	0	0	0.7	1	1	0.3				
Firmicutes	Mogibacteriaceae	0	0.4	0.9	0.7	1	0.8	0.9										
Firmicutes	Veillonellaceae	0.1	0.2	0.9	0.4	1	0.7	0.6							1	0.7	0.5	0.5
Bacteroidetes	BS11	0	0.3	0.2	0.6	0.4	1	1										
Firmicutes	Christensenellaceae	0	0.2	0.1	0.1	1	1	0.7										
Firmicutes	Dehalobacteriaceae	0.1	0.4	0.3	0.4	0.6	1	0.9										
Firmicutes	Lachnospiraceae	0.1	0.4	0.7	0.6	0.5	1	1	0	0	0.6	1	1	0.5	0.1	0.9	1	0.8
Firmicutes	Lactobacillaceae	0.8	0.4	0.2	0.3	0.8	1	0.7	0.4	1	0	0	0.1	0.1	1	0	0	0
Bacteroidetes	Prevotellaceae	0	0.2	0.5	0.3	0.5	1	0.8							0.4	1	0.9	0.7
Tenericutes_other_classfied	Tenericutes_other_classfied	0	0.3	0.4	0.5	0.5	1	0.6										
Proteobacteria	Halomonadaceae	0.7	0.1	0.1	0	0	0	1										
Verrucomicrobia	RFP12	0	0.3	0.9	0.7	0.2	0.7	1										
Firmicutes	Ruminococcaceae	0	0.3	0.5	0.4	0.8	1	1	0	0	1	0.7	0.7	1	0	0.1	0.2	1
Spirochaetes	Spirochaetaceae	0	0.1	0.3	0.3	0.3	0.9	1										
Spirochaetes_other_classfied	Spirochaetes_other_classfied	0.1	0.2	0.7	0.4	0.8	1	1										
Firmicutes	Streptococcaceae	0.2	0.2	0.1	0.2	0.1	0.6	1	1	1	0	0	0	0	1	0.2	0.1	0.1

Extended Data Fig. 2 | See next page for caption.

Extended Data Fig. 2 | Characterizing the age- and location-specific composition of the intestinal microbiome of the healthy pig. (a) Definition of a core intestinal microbiome of the pig. A total of 58 OTUs that were annotated to 21 taxa were identified in >95% of day 120 and 240 faeces and caecum content samples of both F6 and F7 generations, hence defined as core bacterial taxa. (b) The compositions of the porcine and human intestinal microbiota are closer to each other than either is to that of the mouse. Boxplots are as is Fig. 1c. The number of samples available for analysis were 1281 pigs, 106 humans and 6 mice. (c) Abundances (F6-F7 averages when available) of the 43 families represented in Fig. 1b in the seven sample types relative to the sample type in which they are the most abundant (red – blue scale). The families are ordered according to the sample type in which they are the most abundant. The colour-code for phyla is as in Fig. 1b. Columns are added for comparison with mouse and human. Mouse data are from Fig. 1 in Suzuki & Nachman[116], and human data from Fig. 6 in Vuik et al[117]. P_I: proximal ileum, D_IL: distal ileum, C: caecum, CO: colon, RE: rectum, F: faeces. The families differing the most with regards to location-specific distribution between species include *Helicobacteriaceae*, *Veillonellaceae*, *Lactobacillaceae* and *Streptococcaceae*.

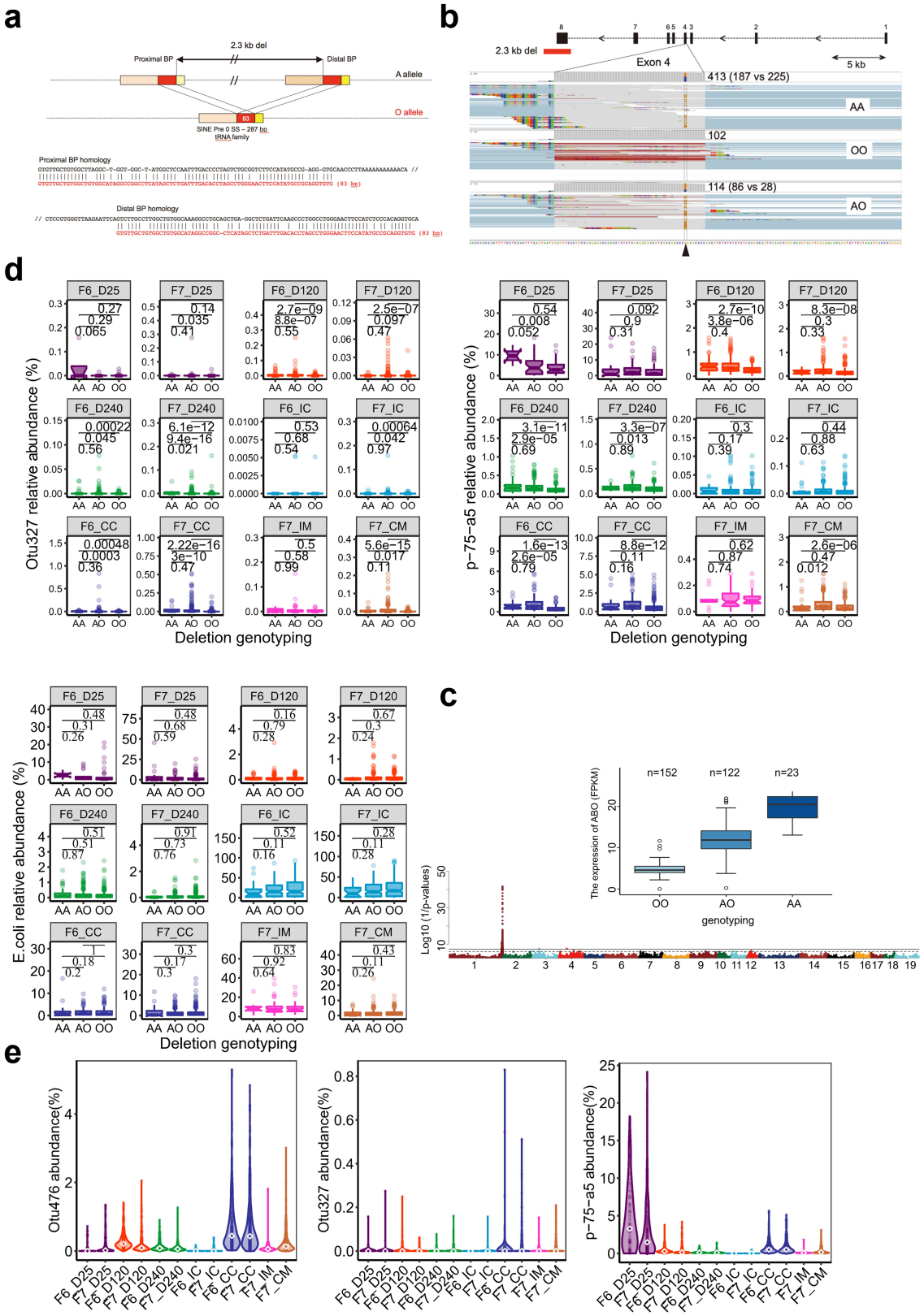


Extended Data Fig. 3 | Evaluating the heritability of intestinal microbiota composition in the mosaic pig population. Correlation between heritability estimates of taxa/OTUs in F6 and F7 generation by sample type (D25, D120, D240, CC and IC). Correlation coefficients (r) and associated p-values (p) were computed using heritability estimates that were pre-corrected for bacterial abundance (residuals of linear model). Heritability estimates indeed tend to slightly increase with taxa abundance. Yet, results show that this effect cannot account for the observed correlations between F6 and F7 estimates in D120, D240 and CC, hence pointing towards genuine genetic effects. The shaded areas correspond to the 95% confidence region for the regression fit. Correlation coefficients and two-sided p-values were computed using Spearman's rank-based method. Reported p-values are nominal (i.e. uncorrected for multiple testing).



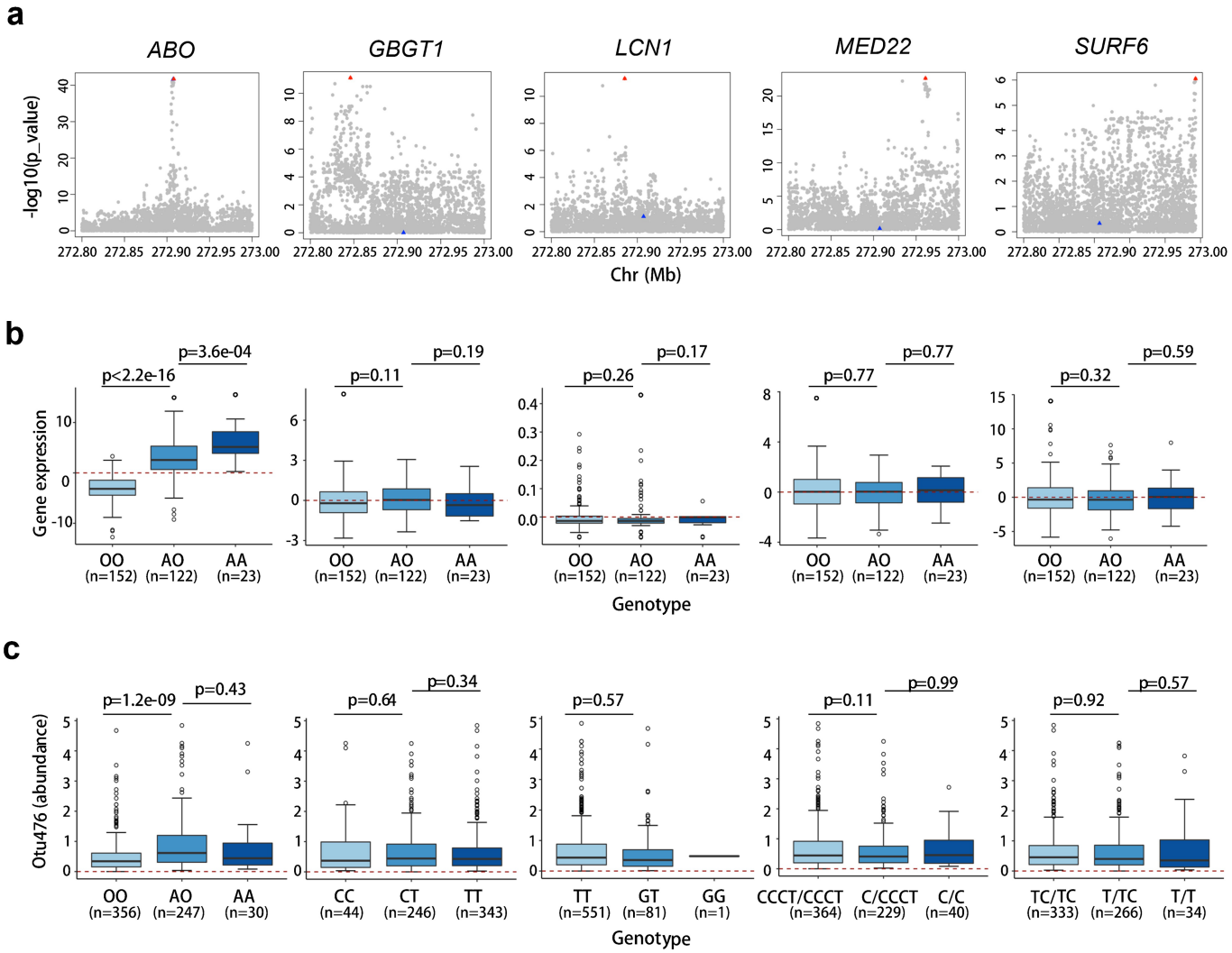
Extended Data Fig. 4 | See next page for caption.

Extended Data Fig. 4 | Identifying a microbiota QTL (miQTL) with major effect on the abundance of Erysipelotrichaceae species by whole genome sequence based GWAS. (a) Schematic illustration of the samples and SNPs used for the two types of analyses (abundance and presence/absence) performed for miQTL mapping. (b) (Upper) Distribution of log(1/p) values for 1,527 sets of 11 p-values obtained in 11 data-series for a SNP x taxon x analysis model combination that yielded a genome-wide significant signal ($p < 5 \times 10^{-8}$) in the 12th data-series. (Lower) Distribution of log(1/p) values for 1,527 sets of 11 p-values obtained in the same data-series and with the same analysis model as in (upper) but with randomly selected SNP x taxon combinations matching the ones in (upper) for MAF and taxa abundance. Log(1/p) values were computed using GenABEL as described in Methods. Corresponding p-values are nominal and two-sided. (c) Correlation between the average (F6 and F7) taxon heritability, and the average (F6 and F7) number of genome-wide significant ($p \leq 5 \times 10^{-8}$) miQTL for D240 faecal samples. The shaded area corresponds to the 95% confidence region for the regression fit. Correlation coefficient and associated p-values are Spearman's. (d) *QQ* plot for 1,527 (number of signals (SNP x taxon x model x one data series in one cohort) exceeding the genome-wide log(1/p) threshold value of 7.3) sets of ≤ 5-7 p-values (same SNP x taxon x model, all data series in the other cohort) for real SNPs (Blue: quantitative model; Green: binary model), and matched sets of ≤ 5-7 p-values corresponding to randomly selected SNP x taxon combinations matched for MAF and abundance or presence/absence rate (Brown: quantitative model; Yellow: binary model). Log (1/p) values were computed using GenABLE as described in Methods. Corresponding p-values are nominal and two-sided. (e) Same *QQ* plot as in (C) after removal of all SNPs in the chromosome 1: 272.8-273.1Mb interval. Log(1/p) values were computed using GenABLE as described in Methods. Corresponding p-values are nominal and two-sided. (f) Distribution of the association log(1/p) values and corresponding signed z-scores for SNP 1_272907239 and 31 p-75-a5 OTUs (red) and 83 Erysipelotrichaceae (yellow) OTUs, showing an enrichment of effects with same sign as for OTU476 and OTU327. Log(1/p) values were computed using Metal (v3.0) as described in Methods. Corresponding p-values are nominal and two-sided. See also Supplemental discussion 1.

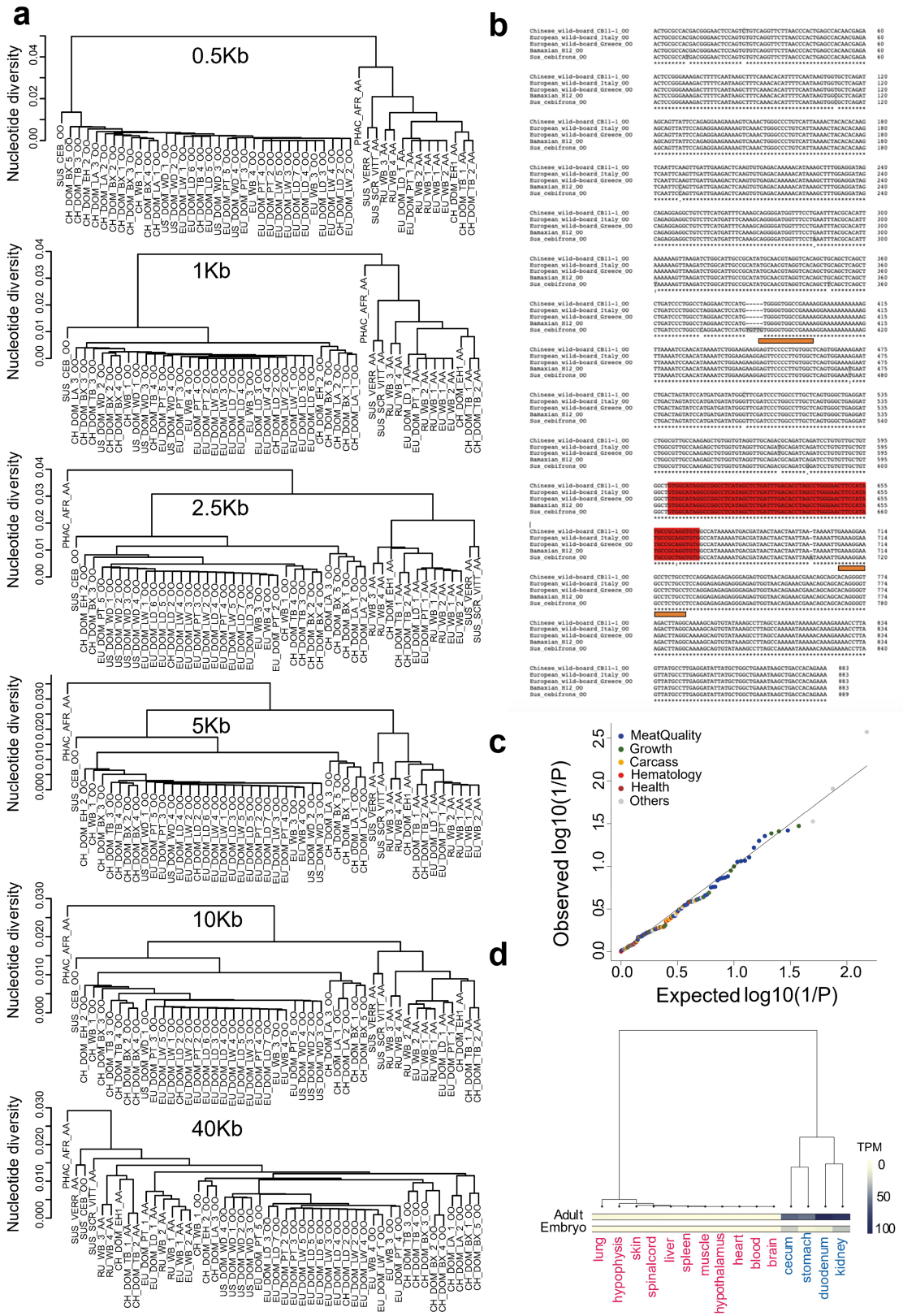


Extended Data Fig. 5 | See next page for caption.

Extended Data Fig. 5 | The chromosome 1 miQTL is caused by a 2.3 kb deletion in the orthologue of the human ABO gene. (a) Breakpoints of the 2.3 kb deletion showing the role of a duplicated SINE sequence in mediating an intra-chromosomal recombination. (b) Illustrative example of allelic balance for the *cG146C* SNP in an AA homozygote and of allelic imbalance for the same SNP in an AO heterozygote. (c) (Upper) eQTL analysis for the porcine AO gene maximizing at the exact position of the 2.3 kb deletion ($p = 1.9x10^{-43}$) and showing the additive effect of the A allele increasing transcript levels ~3-fold (inset; FPKM: Fragments Per Kilobase of transcript per Million mapped reads). The "n's" correspond to the number of animals of each genotype available for analysis. Boxplots are as in Fig. 1c. (Lower) Genome wide eQTL scan for the porcine ABO gene showing the strong cis-eQTL signal on chromosome 1. eQTL analysis was conducted with GEMMA (v0.97)[64]. Reported log-transformed p-values are nominal and two-sided. (d) Effect of N-acetyl-galactosaminyl transferase genotype (AA, AO or OO) on abundance of OTU327 and p-75-a5 in the twelve data series. Absence of an effect of N-acetyl-galactosaminyl transferase genotype (AA, AO or OO) on abundance of *E. coli* in the twelve data series. Sample sizes are as in STable 4.1. Boxplots are as in Fig. 3d. (e) Abundance of OTU476, OTU327 and p-75-a5 in the twelve data series. Violin plots with indication of the median. Numbers (n's) are as in STable 4.1. See also Supplemental discussion 2.

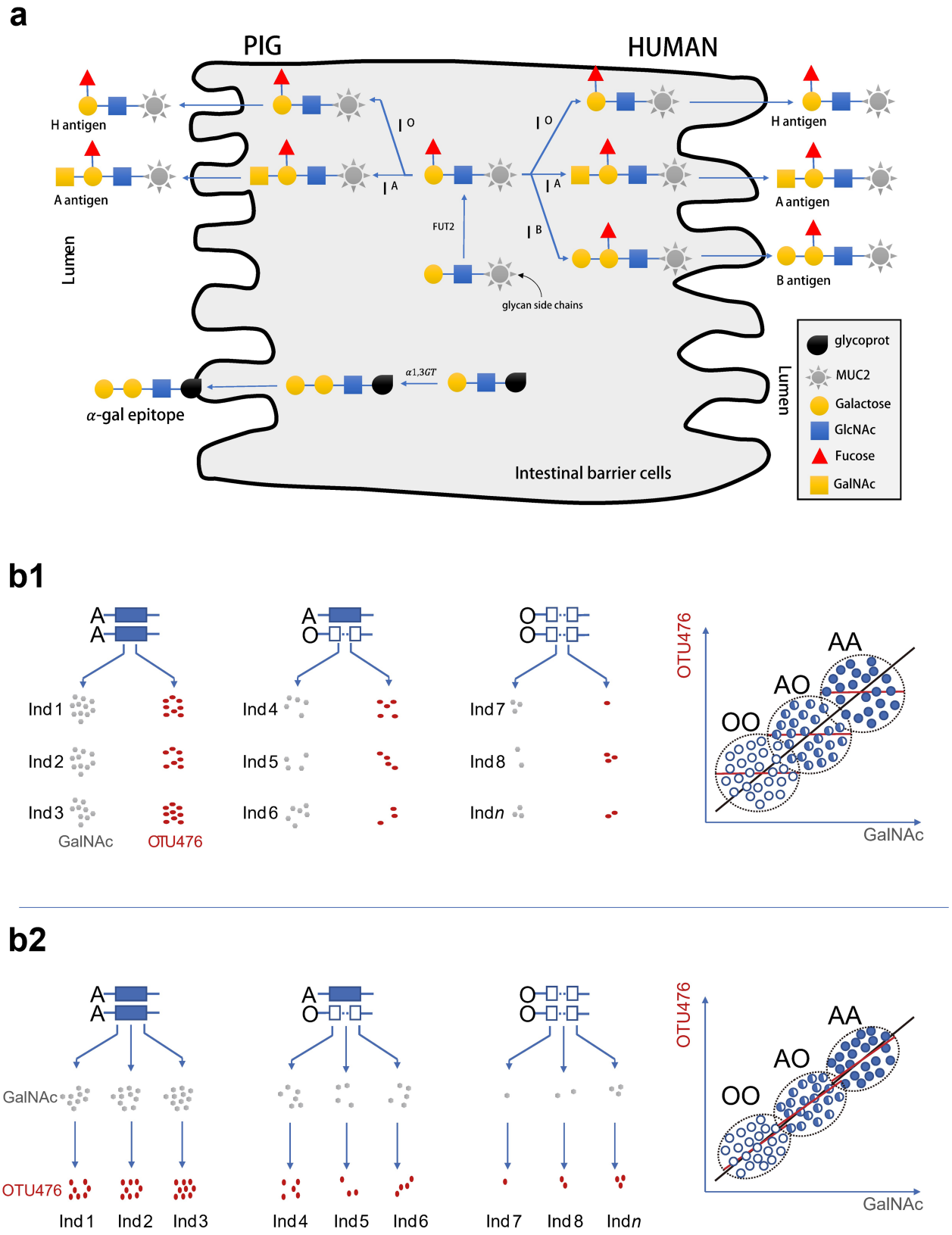


Extended Data Fig. 6 | cis-eQTL analyses in the vicinity of the chromosome 1 miQTLK supports the causality of the 2.3 kb deletion. (a) Cis-eQTL analysis for the porcine N-acetyl-galactosaminyl transferase ("ABO"), *GBTG1*, *LCN1* (=*OBP2B*), *MED22* and *SURF6* genes in caecum. The blue triangle corresponds to the top SNP for the miQTL. The red triangles correspond to the top SNPs for the respective cis-eQTL. Only for N-acetyl-galactosaminyl transferase are blue and red variants the same. eQTL analyses were conducted with GEMMA (v0.97)[64]. Reported log-transformed p-values are nominal and two-sided. (b) Effect of *AO* genotype on the expression levels of the corresponding genes in caecum. There was no evidence for an effect of AO genotype on the expression of any of these genes other than ABO. The number of AA, AO and OO samples available for cis-eQTL analysis for each gene are given (n). Boxplots are as in Fig. 1c. We tested the difference in gene expression level between pairs of genotype classes using a two-sided t-test. (c) Effect of the top cis-eQTL SNPs (blue triangles in A) on OTU476 abundance. Only the top cis-eQTL SNPs for ABO has an effect on OTU476 abundance. The number of AA, AO and OO samples available for miQTL analysis for each gene are given (n). Boxplots are as in Fig. 1c. We tested the difference in bacterial abundance between pairs of genotype classes using a two-sided t-test.



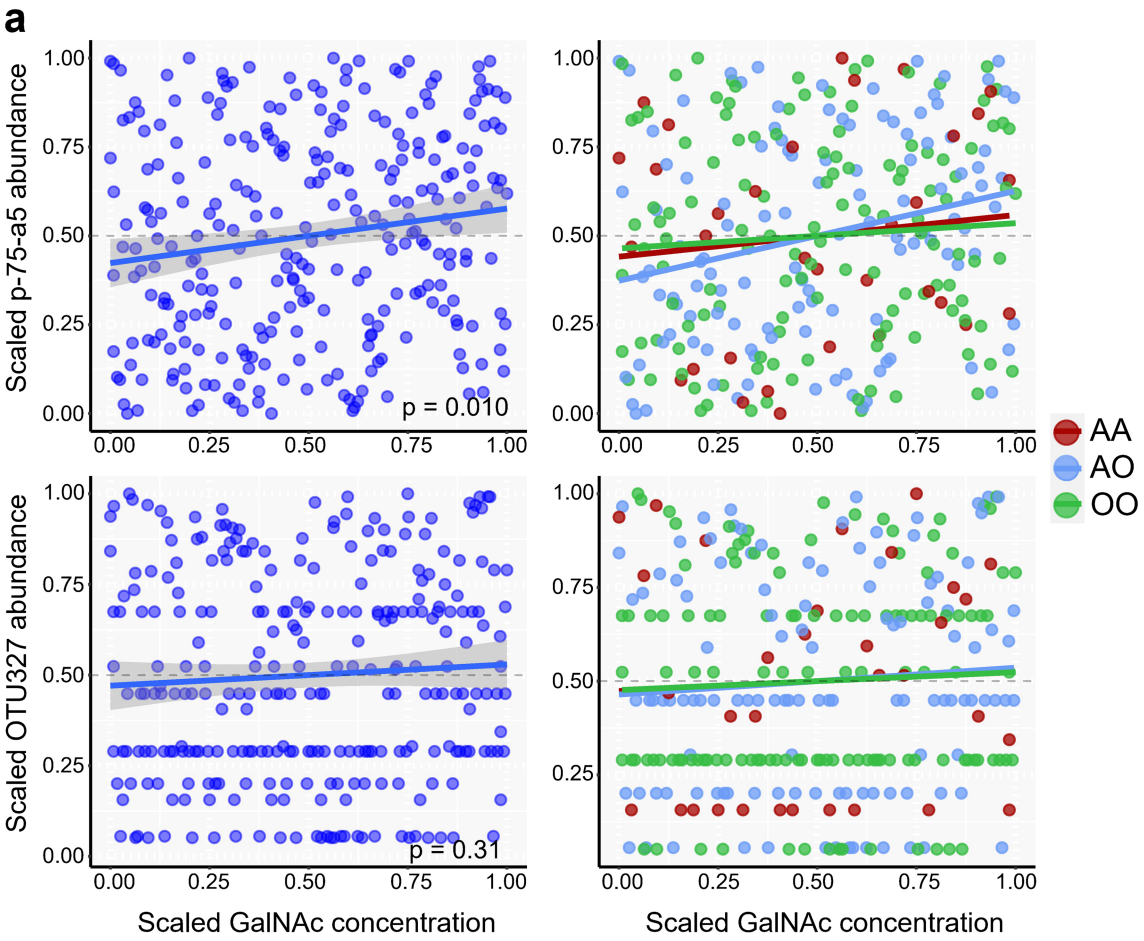
Extended Data Fig. 7 | See next page for caption.

Extended Data Fig. 7 | The 2.3 kb deletion in the orthologue of the human ABO gene is 3.5 million years old and under balancing selection. (a) UPGMA tree based on nucleotide diversities between 14 AA and 34 OO animals in windows of increasing size (0.5 to 40 kb) centred on the 2.3 kb deletion in the porcine N-acetyl-galactosaminyl transferase gene (porcine O allele). PA: *Phacochaerus Africanus*, SC: *Sus cebifrons*, SV: *Sus verrucosus*, SU: *Sus scrofa vittatus*, CB: Chinese wild boar, RB: Russian wild boar, EB: European wild boar, ERH: Erhualian, BX: Bamaxiang, T: Tibetan, LA: Laiwu, LR: Landrace, LW: Large White, PI: Piétrain, WD: White Duroc. Context: To gain additional insights in the age of the porcine *O* allele, we generated phylogenetic trees of the A and O alleles of 14 AA and 34 OO animals including domestic pigs, wild boars, Visayan and Javanese warty pigs, and common African warthog. Examination of their local SNP genotypes (50K window encompassing the ABO gene) reveals traces of ancestral recombinations between O and A haplotypes as close as 300 and 800 base pairs from the proximal and distal deletion breakpoints, respectively, as well as multiple instances of homoplasy that may either be due to recombination, gene conversion or recurrent de novo mutations. On their own, these signatures support the old age of the O allele. We constructed UPGMA trees based on nucleotide diversity for windows ranging from 500 bp to 40 kb centred on the 2.3 kb deletion. Smaller windows have a higher likelihood to compare the genuine ancestral O versus A states, yet yield less robust trees because they are based on smaller number of variants. Larger windows will increasingly be contaminated with recombinant A-O haplotypes blurring the sought signal. Indeed, for windows ≥ 20 kb or more, the gene tree corresponds to the species tree, while for windows ≤ 15 kb the tree sorts animals by AA vs OO genotype. For all windows ≤ 15 kb the *Sus cebifrons* O allele maps outside of the *Sus scrofa* O allele supporting a deep divergence (rather than hybridization) and hence the old age of the O allele. Of note, for windows ≤1.2 kb, the warthog A allele is more closely related to the *Sus* A alleles than to the *Sus* O alleles (ED7A). This suggests that the O allele may be older than the divergence of the *Phacochoerus* and *Sus* A alleles, i.e. >10 MYA. It will be interesting to study larger numbers of warthog to see whether the same 2.3 kb deletion exists in this and other related species as well. (b) Alignment of ~900 base pairs of the O alleles of domestic pigs (Bamaxian), European and Asian wild boars, and *Sus cebufrons* demonstrating that these are identical-by-descent. The SINE element that is presumed to have mediated the recombinational event that caused to 2.3 kb deletion is highlighted in red. Context: To further support their identity-by-descent we aligned ~900 base pairs (centred on the position of the 2.3 kb deletion) of the O alleles of domestic pig, European and Asian wild boars and *Sus cebifrons*. The sequences were nearly identical further supporting our hypothesis. It is noteworthy that the old age of the "O" allele must have contributed to the remarkable mapping resolution (≤3 kb) that was achieved in this study. In total, 42 variants were in near perfect LD ($r^2 \geq 0.9$) with the 2.3 kb deletion in the F0 generation, spanning 2,298 bp (1,522 on the proximal side, and 762 on the distal side of the 2.3 kb deletion). This 2.3 kb span is lower than genome-wide expectations (17th percentile), presumably due to the numerous cross-overs that have accrued since the birth of the 2.3 kb deletion that occurred in the distant past. Yet the number of informative variants within this small segment is higher than genome-wide average of (57% percentile) also probably due at least in part to the accumulation of numerous mutations since the remote time of coalescence of the A and O alleles (see Fig. 1d in main text). (c) *QQ* plots for the effect of AO genotype on 150 phenotypes pertaining to meat quality, growth, carcass composition, hematology, health, and other phenotypes in the F6 and F7 generation. P-values were obtained using a mixed model followed by meta-analysis (weighted Z score) across the F6 and F7 generations as described in Methods. log-transformed p-values used for the *QQ* plot are nominal and two-sided. Context: Our findings in suidae are reminiscent of the trans-species polymorphism of the ABO gene in primates attributed to balancing selection[26]. The phenotype driving balancing selection remain largely unknown yet a tug of war with pathogens is usually invoked: synthesized glycans may affect pathogen adhesion, toxin binding or act as soluble decoys, while naturally occurring antibodies may be protective[20,44]. In humans, the O allele may protect against malaria[118], *E. Coli* and *Salmonella* enteric infection[119], SARS-CoV-1[42], SARS-CoV-2[43] and schistosomiasis[120–122], while being a possible risk factor for cholera[123], *H. pylori*[124] and norovirus infection[125]. Whatever the underlying selective force, it appears to have operated independently in at least two mammalian branches (primates and suidae), over exceedingly long periods of time, and over broad geographic ranges, hence pointing towards its pervasive nature. To gain insights in what selective forces might underpin the observed balanced polymorphism, we tested the effect of porcine AO genotype on >150 traits measured in the F6 and F7 generations pertaining to carcass composition, growth, meat quality, hematological parameters, disease resistance and behaviour. No significant effects were observed when accounting for multiple testing, including those pertaining to immunity and disease resistance. (d) Expression profile of the AO gene in a panel of adult and embryonic porcine tissues (own RNA-Seq data).



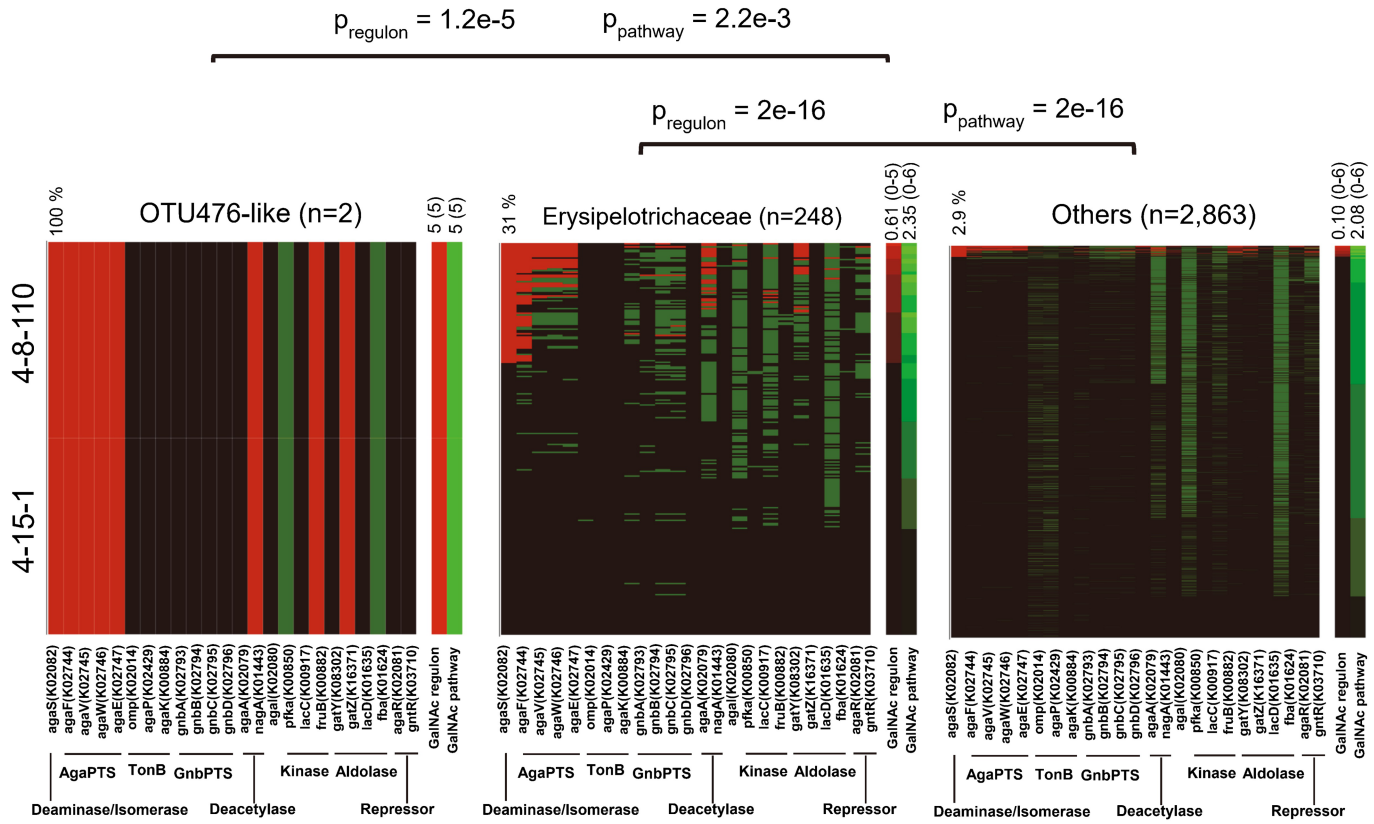
Extended Data Fig. 8 | See next page for caption.

Extended Data Fig. 8 | The chromosome 1 miQTL affects caecal N-acetyl-D-galactosamine (GalNAc) concentrations which are correlated with the abundance of *Erysipelotyrichaceae* species within *AO* genotype: theory. (a) ABO and **α**-gal epitopes in pigs and human. The glycosyltransferase gene located on 9q34.2 and underpinning the human ABO blood group is characterized in most human populations by three major alleles: (i) I^A encoding a α-3-N-acetyl-D-galactosaminyltransferase that is adding GalNAc to H and Lewis antigens (yielding the A antigen) on various glycoproteins including mucins secreted in the intestinal lumen, (ii) I^B encoding a α-3-D-galactosyltransferase that is adding galactose to the same antigens (yielding the B antigen), and (iii) the inactive I^O null allele that precludes expression of either the A and/or the B antigen. Mutations in the fucosyltransferase 2 gene (*FUT2*) preclude formation of the H antigen on secreted proteins and hence the detection of A and B antigens in secretions[20]. The pig orthologue of the human ABO glycosyltransferase gene is located on the telomeric end of porcine chromosome 1q, and is characterized by two major alleles: (i) the *A* allele, encoding a α-3-N-acetyl-D-galactosaminyltransferase that is adding GalNAc to H and Lewis antigens, similar to the human I^A allele, and (ii) the O allele corresponding to a null allele as a result of a 2.3 kb deletion similar to the human I^O allele[24]. Thus, the B antigen (Galα1-3(Fucα1-2)Galβ1-4GlcNAc-R) is not observed in pig populations. However, what is found abundantly on the surface of cells in many tissues is the so-called "α-gal epitope" (Galα1-3Galβ1-4GlcNAc-R), which results from the addition of a galactose to the Galβ1-4GlcNAc-R precursor by a α1,3galactosyltransferase encoded by the *GGTA1* gene. The orthologue of the GGTA1 gene is non-functional in human and Old World non-human primates, which, however, have high titers of circulating anti-α-gal antibodies contributing to acute rejection of xenografts[126,127]. (b) Identifying whether changes in GalNAc concentration are the cause of the observed changes in abundance of Erysipelotrichaceae species by searching for a correlation between the two phenotypes "within AO genotype". (b1) If *AO* genotype is associated with the abundance of Erysipelotrichaceae species and GalNAc concentrations by virtue of different molecular mechanisms (for instance because they involved distinct causative mutations albeit in linkage disequilibrium, or because the gene has an as of yet unknown other activity that is causing the change in bacterial abundance, independently of its glycosyltransferase activity), there is no reason to expect a correlation between bacterial abundance and GalNAc concentration within *AO* genotype (red horizontal lines in the dotted circles). There is of course a correlation across genotypes that is due to the fact that AO genotype has a (direct or indirect) effect on both phenotypes. (b2) If, on the other hand, *AO* genotype causes the change in GalNAc concentration (which is very likely given its known enzymatic activity) which then causes the change in the abundance of Erysipelotrichaceae species, one can expect that bacterial abundance and GalNAc concentration will be correlated, also within *AO* genotype, as indicated by the sloped red lines within the dotted ellipses. This is what is observed with the real data.

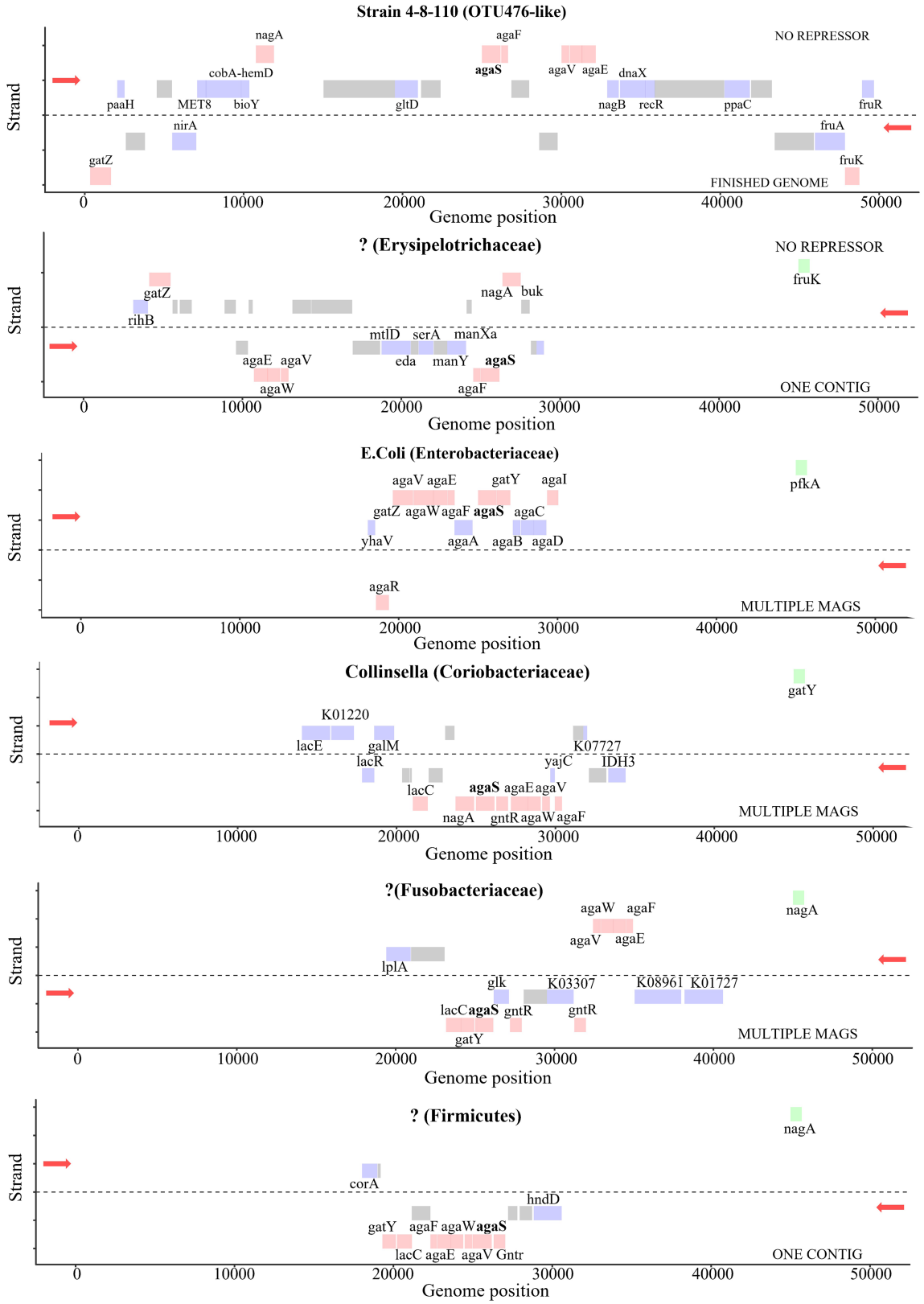


Extended Data Fig. 9 | See next page for caption.

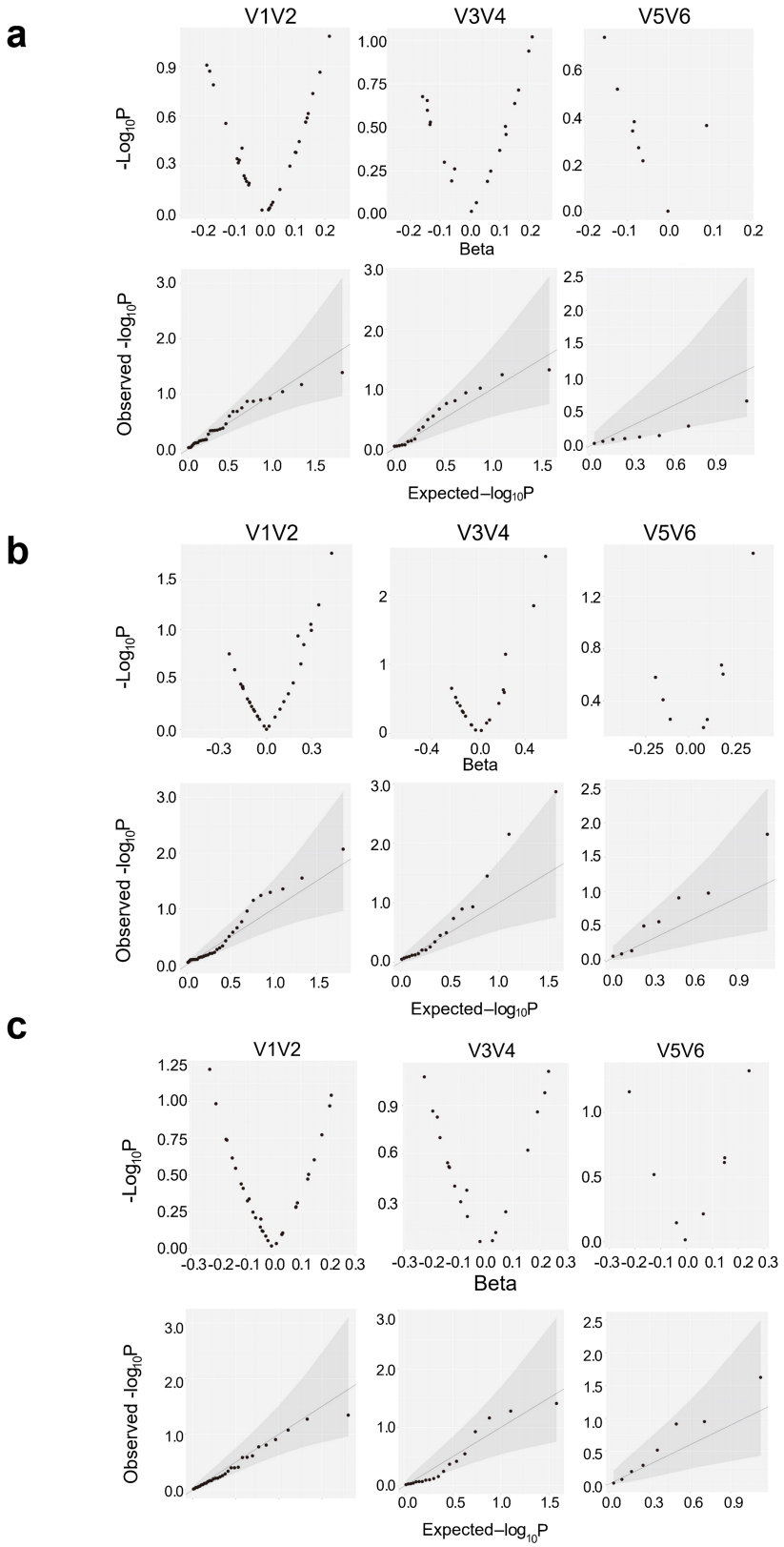
Extended Data Fig. 9 | The chromosome 1 miQTL affects caecal N-acetyl-D-galactosamine (GalNAc) concentrations which are correlated with the abundance of *Erysipelotyrichaceae* species within *AO* genotype: results. (a) Positive correlation between caecal GalNAc concentrations and bacterial abundance (upper panels: p-75-a5; lower panels: OTU327) "within AO genotype". GalNAc concentrations and bacterial abundances were corrected for batch effects and *AO* genotype and scaled between 0 and 1 to equalize residual variance. Correlations were computed using all samples jointly and Spearman's rank-based test; corresponding p-values (nominal; two-sided) are given (left panels). Regression lines are shown for the different AO genotypes separately (right panels); all of them are positive. Note that the scatter plots for p-75-a5 are not identical but very similar to those for OTU476 (Fig. 5b, c). This is because OTU476 accounts for most of the p-75-a5 genus in caecum content (see also Extended Data Fig. 5). These data can therefore not be considered to be independent. The shaded areas correspond to the 95% confidence regions for the regression fit. (b) Comparison of the free GalNAc concentrations in caecal content of OO, AO and AA pigs as well as in caecal content of germ-free mice gavaged with 200mg/kg GalNAc. Concentrations were determined in freeze-dried caecal content powder using LC-MS/MS. Number of analyzed samples are given (n). Boxplots are as in Fig. 1c.



Extended Data Fig. 10 | The chromosome 1 miQTL affects bacteria with a functional GalNAc import and catabolic pathway. Presence anywhere in the genome (green), presence in close proximity to *agaS* (red), or absence (black) of the orthologues of 24 genes implicated in the GalNAc TR/CP pathway in the genome of (i) two OTU476 like strains (4-15-1 and 4-8-110), (ii) 248 MAGs assigned to the Erysipelotrichaceae family, and (iii) 2,863 MAGs assigned to other bacterial families. The two lanes on the right of the three panels correspond to the Regulon (red) and Pathway (green) score respectively. Both scores range from 0 (black) to 6 (bright red or green). Means (range) for the corresponding dataset are given on top. P-values (nominal, two-sided, uncorrected) of the pathway and regulon scores were computed using a linear model described in Methods.



Extended Data Fig. 11 | Different GalNAc operon structure and transcriptome response in miQTL-sensitive versus -insensitive GalNAc utilizing bacteria. Maps of GalNAc "operons" in one of the two OTU476-like strains (NB: The organization of the GalNAc gene cluster was identical in both 4-15-1 and 4-8-110 strains), and six MAGs assigned respectively to an Erysipelotrichaceae, *E. coli* (an Enterobacteriaceae), a *Collinsella* (a *Coriobacteriaceae*), a *Fusobacteriaceae*, a *Firmicutes* and a *Clostridium*. Identified Open Reading Frames (ORFs) are represented as coloured boxes. Genes implicated in GalNAc import and catabolism are in red if they are part of the cluster and in green if located elsewhere in the genome. Genes with a known function unrelated to GalNAc are in blue. ORFs with uncharacterized gene product in gray. Gene acronyms are given next to the corresponding boxes. ORFs transcribed from the top (respectively bottom) strand are above (below) the dotted line. The respective transcriptional directions are marked by the arrows. The source of information used to confirm the map order is given (finished genome, multiple MAGs, single contig).



Extended Data Fig. 12 | No effect of ABO genotype on intestinal Erysipelotrichaceae abundance in human. Volcano and *QQ* plots for 43 (V1-V2), 20 (V3-V4) and 9 (V5-V6) OTUs classified as Erysipelotrichaceae for the contrasts (a) [AA, AO and AB] versus [BB, BO and OO], (b) [BB, BO and AB] versus [AA, AO and OO], and (c) [OO] versus [all others]. The shaded areas correspond to the 95% confidence intervals of the spread of the *QQ* plot under the null hypothesis of no QTL. The actual points are always within these intervals precluding us to reject the null hypothesis. P-values (nominal, two-sided) were computed using the linear model described in Methods and hereafter. See also Supplemental discussion 3.

nature research

Corresponding author(s): Lusheng Huang, Congying Chen, Michel Georges

Last updated by author(s): Mar 7, 2022

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Genotype data of the mosaic population: BWA(v0.7.17); Samtools (v1.6); Picard (v2.21.4); Platypus (v0.8.1); Plink (v1.90); Beagle (v.40). 16S rRNA data of the mosaic population: Trimmomatic (v.0.39); FLASH (v.1.2.11); USEARCH (v.7.0.1090); VSEARCH (v.2.8.1); Greengenes (v13.5); RDP classifier (v2.2). Metagenome data: fastp (v0.19.41); BWA (v.0.7.17); MEGAHIT (v1.1.3); metaWRAP (v1.1.1); dRep (v2.3.2); metaSPAdes (v3.15.3); CheckM(v1.0.12). RNA sequencing data of cecum tissues: STAR (v020201); Samtools (v1.6); FeatureCounts (v1.6.4). Nanopore sequencing data of Erysipelotrichaceae strains: Flye (v2.6); Prodigal (v2.6.3). PacBio sequencing data of Bamaxiang pig: Canu (v1.7.1); Flye (v2.4.2); racon (v1.4.10); Pilon (v1.23); bwa-mem (0.7.17-r1188); Lastz (v1.02.00); Minimap2 (v2.17-r941). Whole-genome sequencing data for wild boars, Sus verrucosus, Sus cebifrons, six Russian wild boars, one Sumatran wild boar, and one African warty hog : GATK(v4.2), BWA (v0.7.17). RNA sequencing data of E. coli and two Erysipelotrichaceae strains: Bowtie2 (V 2.4.2), FeatureCounts (v1.6.4); R (v 3.5.1) "DESeq2". Determination of the concentration of N-acetyl-galactosamine in cecal lumen: ExionLCTM AD System, Applied Biosystems 6500 Triple Quadrupole (QTRAP® 6500), Analyst software (1.6.3). metabolic flux measurement: A Shimadzu QP-2010 Ultra GC-MS.
Data analysis	Estimating the contribution of the eight founder breeds: "lm" in R (v3.5.3) ; PCoA: "vegan" and "ape" in R (v3.5.3); alpha-diversity: mothur (v1.43.0); Heritability: lme4QTL in R (v3.5.3) ; genome-wide kinship: GEMMA (v0.97); GCTA (v1.26); Spearman's rank correlations: "corrtest" function in R (v3.5.3); Microbiome dissimilarity: "vegan" in R (v3.5.3); mGWAS: "GenABEL"in R (v3.5.3); GWAS meta-analysis: METAL (v3.0), Perl (v5.10.1); eQTL analysis: GEMMA (v0.97); Phylogenetic analysis of the O alleles in the Sus genus: GATK (v4.2), "hclust" in R (3.5.3); Population differentiation: ANOVA in R (v3.5.3); Profiling ABO gene expression: HISAT2 (v2.2.1), Samtools (v1.6), FeatureCounts (v1.6.4), R(v3.5.3); MAG bioinformatic analyses: Ghost KOALA (v2.2); PhyloPhlAn (v.0.99); R (v3.5.3); Perl (v5.10.1); Association analysis of ABO blood group: R (v3.5.3) "lm". RNA sequencing data analysis of E. coli and OTU-476 like strain: Perl (v5.10.1); Bowtie2 (v2.4.2); FeatureCounts (v1.6.4); R (v 3.5.1) "DESeq2". metabolic flux data analysis: MATLAB (Release R2021a). 16S rRNA data of human: bbduk tool (BBMap –

Bushnell B. –sourceforge.net/projects/bbmap/); BBTools (38.82); Snakemake (7.0.1); QIIME 2 (2018.11); DADA2 (v1.16); DNACLUST (v. r3). KEGG pathway analysis: Ghost KOALA tool (v2.2). The custom codes developed in the study are available in the repository: https://github.com/yanghuijxau/Manuscript-microbiota-ABO.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

All the 16S rRNA sequencing data, the metagenomics sequence data and the RNA sequencing data were submitted to the GSA database with accession numbers: CRA006230, CRA006239, CRA006240 and CRA006216. The genotype data was deposited at the GVM under the GSA database with accession numbers: GVM000310, and the link: http://bigd.big.ac.cn/gvm/getProjectFile?t=307e8d7e.
The whole genome sequences of experimental pigs are available at: http://jxlab.jxau.edu.cn/. The source data are available in the repository: https://github.com/yanghuijxau/Manuscript-microbiota-ABO. The GWAS summary statistics is available through Figshare with doi: 10.6084/m9.figshare.19313960.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	This is an observational study. We used all the samples that were collected during the study. The detailed information about the sample size was described as follows. This study focused on the sixth (F6) and seventh (F7) generations of a mosaic population generated from 61 F0 founders. A total of 954 and 892 individuals were generated for F6 and F7, respectively. Among them, 836 F6 and 668 F7 pigs had genotypes data. A total of 5,110 feces and intestinal content samples from F6 and F7 were used for 16S rRNA sequencing, including feces at days 25, 120 and 240, as well as cecal and ileal content (F6 & F7) and mucosal scrapings (F7 only) at day 240 (7 traits, 12 data series). Three hundred cecum tissue samples from F7 pigs were collected and used for RNA sequencing. Ninety-two samples from eight pig populations, four intestinal locations and different ages were used for metagenomic sequencing. Cecum content samples from 278 pigs at the age of 240 days were used to determine the concentration of N-acetyl-galactosamine by targeted LC-MS/MS. Ten germ-free female mice were used to gavage experiments. Six samples were used for metabolic flux analysis of GalNAc. We confirmed that our sample size was well powered to answer the questions in this research. The details were described in the paper and supplementary information.
Data exclusions	Samples with both whole-genome resequencing data and 16S rRNA sequencing data were retained for GWAS analysis. Samples with whole-genome sequencing data and RNA sequencing data were used for eQTL analysis.
Replication	We used the F6 and F7 generations as two experimental pig populations. As two independent populations, Experimental replication was performed by repeating microbiome composition analysis heritability estimate and GWAS analysis. Furthermore, a GWAS meta-analysis was performed in the F6 and F7 generations.
Randomization	Randomization was not relevant to this study and was not employed. We controlled for potential sources of confounding by (a) animals were reared in standardized housing and feeding conditions. (b) collection method of samples was unified. (c) including covariates encoding gender, sample collection batches, and the top three principal component of the host genotypes
Blinding	Blinding is not relevant to this study and was not employed because this is an observational study.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☐	☒ Animals and other organisms
☐	☒ Human research participants
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Animals and other organisms

Policy information about studies involving animals; ARRIVE guidelines recommended for reporting animal research

Laboratory animals	A total of 954 (449♂, 505♀) and 892 (494♂, 398♀) individuals were generated for the sixth (F6) and seventh (F7) generation of a mosaic population constructed with four indigenous Chinese pig breeds including Erhualian (2♂, 2♀; at the age of 1-3 years), Bamaxiang (2♂, 2♀; at the age of 1-3 years), Tibetan (2♂, 2♀; at the age of 1-3 years) and Laiwu (2♂, 2♀; at the age of 1-3 years), and four commercial European/American pig breeds including Landrace (2♂, 2♀; at the age of 1-3 years), Large White (2♂, 2♀; at the age of 1-3 years), Duroc (2♂, 2♀; at the age of 1-3 years) and Piétrain (2♂, 2♀; at the age of 1-3 years). These F6 and F7 pigs were slaughtered at the age of 240 days and phenotyped more than 200 traits throughout the whole growth stage. Ten germ-free female mice (Kunming line, 6 weeks of age) were used for gavage experiment. Mice were housed in two separate cages (temperature, 25 ± 2 centigrade; humidity, 45-60%; lighting cycle, 12 h/day; light hours, 06:30-18:30) with free access to water and food. Cecum contents from 278 pigs were used to determine GalNAc concentration.
Wild animals	The ear tissues of six Russian wild boars, one Sumatran wild boar, and one African warty hog were collected and performed the whole-genome resequencing. These eight wild boars were captured using tranquilizer guns and collected ear tissue immediately. In addition, feces and intestinal content samples were collected from six Chinese wild boars. These six wild boars were also captured using tranquilizer guns. After anesthetized and transported to the laboratory, all six wild boars were slaughtered by bleeding after electrical stunning. We didn't know the exact sex and age of these wild boars. All experiments involving wild boars were permitted by Wildlife conservation organization.
Field-collected samples	All F6 and F7 pigs were born and reared at the experimental farm of the National Key Laboratory for swine Genetic Improvement and Production Technology, Jiangxi Agricultural University (Nanchang, Jiangxi). Piglets remained with their mother during the suckling period and were weaned at ~46 days of age. Litters were transferred to 12-pig fattening pens (~20 m2/pen) with automatic feeders (Osborne Industries, US), minimizing splitting and merging of litters. The farm houses was under natural temperature and photoperiod. All pigs were fed twice per day with formula diets. Water was available ad libitum from nipple drinkers. Fecal samples were manually collected from the rectum of experimental pigs at the ages of 25, 120 and 240 days in farm houses. Cecum content samples were collected from standard commercial slaughter house within 30 min after slaughter.
Ethics oversight	All procedures involving animals were performed according to the guidelines for the care and use of experimental animals established by the Ministry of Agricultural and Rural Affairs and the Ethics Committee in Jiangxi Agricultural University. The experiment Protocol involving mice was approved by the Ethics Committee in Huazhong Agricultural University (HZAUMO-2021-0077). The experimental protocol involving humans was approved by the ethics committee of the University of Liège Academic Hospital. Informed consent was obtained prior to donation in agreement with the recommendations of the declaration of Helsinki for experiments involving human subjects.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Human research participants

Policy information about studies involving human research participants

Population characteristics	The data used correspond to the previously described CEDAR cohort (Momozawa, Y. et al. IBD risk loci are enriched in multigenic regulatory modules encompassing putative causative genes. Nat Commun 9, 2427, 2018). It comprised 323 healthy individuals of European descent including 182 women and 141 men averaging 56 years of age (range: 19-86).
Recruitment	The CEDAR1 cohort corresponds to healthy individuals of European descent that were visiting the University Hospital (CHU) from the University of Liège as part of a national screening campaign for colon cancer. Yet individuals with polyps or other diseases were excluded from the cohort. We therefore are not aware of any biases that may have affected the results.
Ethics oversight	The experimental protocol was approved by the ethics committee of the University of Liège Academic Hospital. Informed consent was obtained prior to donation in agreement with the recommendations of the declaration of Helsinki for experiments involving human subjects.

Note that full information on the approval of the study protocol must also be provided in the manuscript.