Accelerated Article Preview

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

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

This is a PDF file of a peer-reviewed paper that has been accepted for publication.
Although unedited, the content has been subjected to preliminary formatting. Nature
is providing this early version of the typeset paper as a service to our authors and
readers. The text and figures will undergo copyediting and a proof review before the
paper is published in its final form. Please note that during the production process
errors may be discovered which could affect the content, and all legal disclaimers
apply.

Article

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

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 to what extend and how host genetics contributes to this variation is paramount yet has proven difficult as few associations have been replicated, particularly in humans[2]. We herein 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 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 N-acetyl-galactosaminyl-transferase gene underpinning the ABO blood group in humans. We show that this deletion is a ≥3.5 million years old trans-species polymorphism under balancing selection. We demonstrate that it decreases the concentrations of N-acetyl-galactosamine in the gut thereby reducing the abundance of *Erysipelotrichaceae* that can import and catabolize N-acetyl-galactosamine. Our results provide very strong evidence for an effect of host genotype on the abundance of specific bacteria in the intestine combined with insights in the molecular mechanisms that underpin this association. They pave the way towards identifying the same effect in human rural populations.

It is increasingly recognized that comprehensive understanding of the physiology and pathology of organisms requires integrated analysis of the host and its multiple microbiota[1]. In humans, intestinal microbiota composition is associated with physiological and pathological parameters including HDL cholesterol, fasting glucose levels and body mass index (BMI)[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, which 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 be partially heritable. While studies in rodents support this[7], evidence has been shallower in humans. Initial reports didn't reveal a higher microbiota resemblance between monozygotic than dizygotic twins suggesting limited impact of host genotype[8]. Better-powered studies provided evidence for a significant impact 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 difficult to replicate[2,10–14]. The analysis of larger human cohorts is needed to gain 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 observe a strong impact of host genotype on microbiota composition and identify a locus with large effect on the abundance of specific taxa by controlling the concentration of N-acetyl-galactosamine in the gut thereby affecting some of the species that use this metabolite as a carbon source.

Results

A mosaic pig population to study complex traits

We generated a large (> 7,500) mosaic population by intercrossing offspring of 61 F0 founders from four Chinese and four western breeds for >10 generations (STable 1, Extended Data Fig. 1). Animals were reared in uniform housing and feeding conditions. We measured >200 phenotypes (pertaining to body composition, physiology, disease resistance

and behavior), transcriptome, epigenome and chromatin interaction data on multiple tissues, plasma metabolome, and microbiome data in up to 954 F6 and 892 F7 animals. F0 were whole-genome sequenced at average 28.4-fold depth, and F6 and F7 at average 8.0-fold depth. We called genotypes at 23.8 million SNPs and 6.4 million indels with MAF ≥ 0.03 (>1/100-bp). The nucleotide diversity (π) (i.e. 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 *H. sapiens* and *H. neanderthalensis* (~3x10^{-3})[15], while that between a Chinese and a European breed approached halve that between human and chimpanzee (~ 4.3x10^{-3})[16]. The proportion of the eight founder genomes in F6 and F7 ranged from 11.2% to 14.7% at genome level, and from 4.9% to 22.1% at 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 feces at days 25 (suckling period), 120 (growing period) and 240 (slaughter day), as well as cecal and ileal content (F6 & F7) and cecal and ileal mucosal scrapings (F7 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) (STable 2.1). Tags were rarefied and clustered in 32,032 OTUs. 12,054 OTUs amounting to 98.7% of reads were kept and annotated to 41 phyla, 87 classes, 149 orders, 207 families, 360 genera and 150 species. Average microbiota composition of the 12 data series indicated remarkable consistency across F6 and F7, yet marked differences between sample types (Fig. 1a&b, STable 2.2). Even at family-level, some taxa were found to be nearly sample type-specific (Extended Data Fig. 2). The proteobacteria *Enterobacteriaceae, Pseudomonadaceae, Pasteurellaceae*, the firmicutes *Clostridiaceae, Peptostreptococcaceae, Bacillaceae, Leuconostocaceae*, and the actinobacteria *Microbacteriaceae* were at least ten times more abundant in ileal than in any other sample type. Amongst those, *Leuconostocaceae* were nearly digesta-specific, while *Pseudomonadaceae* were nearly mucosa-specific. The Bacteroidetes *Odoribacteraceae* and *Rikenellaceae* were at least ten times more abundant in day 25 feces than in any other sample type. The firmicutes *Christensenellaceae* were nearly ten times more abundant in feces than in any other sample type. This confirms that limiting the analysis of the intestinal microbiota to adult feces can only provide a partial view of its complexity and the factors that determine it[17]. α-diversity of feces was lower at day 25 than at days 120-240, reminiscent of the enrichment of the intestinal flora between child- and adult-hood in humans[8,18]. It was lower for ileal content than for cecal 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 feces. Variation of pair-wise Bray-Curtis dissimilarities was highest for ileal content which had the lowest α-diversity (Fig. 1d, STable 2.3). Microbiota composition of pig D240 feces was more similar to that of human than of mice feces (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: (i) we restricted the analyses to full-sib litter-mates raised in the same environment, and (ii) we confronted genetic similarity and microbiota dissimilarity across generations (F6 and F7). 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 F6 and F7 jointly (STable 3.1, Fig. 2c). 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 fecal (average D240: 10.2%) than for content samples (average IC: 5.7%), and lower for OTUs (average: 6.6%) than for higher taxonomic levels (average: 7.5%) (Fig. 2d). The correlation between h^2 estimates obtained separately in F6 and F7 was positive and significant for D120 (p = 4.2x10^{-5}), D240 (p = 2.2x10^{-16}) and CC (p= 2.3x10^{-3}), supporting genuine genetic effects (Extended Data Fig. 3). We established a list of 55 most likely heritable taxa/OTUs based on congruent h^2 estimates in F6 (≥0.15), F7 (≥0.15) and across F6 and F7 (≥0.10)(STable 3.1). It included the order *Campylobacterales* in D240 feces, the species *Bacteroides coprophilus* in D25 feces, and 53 OTUs of which two (D240, CC) were assigned to *Christensenellaceae*, five to *Ruminococcaceae* (4x D240, CC), one (D25) to *Ruminococcus*, one (D25) to *Phascolarctobacterium*, and one (D240) to *RF32*, i.e. taxons characterized by h^2 ≥0.10 in human studies[19]. With the exception of RF32, all 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, while highly heritable traits may be very polygenic. To gain insight in the genetic architecture of gut microbiota composition in this population we performed 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) (STable 4.1). This yielded 1,527 genome-wide significant signals (≤5x10^{-8}). For these we performed meta-analyses across sample types, separately in F6 and F7 (Extended Data Fig. 4). We identified six signals exceeding the experiment-wide discovery threshold (1.5x10^{-12}) in one cohort and the experiment-wide replication threshold (0.007) in the other (Fig. 3a, STable 4.2 and 4.3). All lead SNPs mapped within 3,037 base pairs from each other (chromosome 1) and were in high LD (Fig. 3b). They affected two OTUs (OTU476 and OTU327) as well as genus p-75-a5 to which OTU476 is assigned. All three are *Erysipelotrichaceae*.

To determine whether this miQTL affects other taxa, we plotted the F6 and F7 association log(1/p-values) for lead SNP 1_272907239 and the 8,490 studied OTUs. OTU476 and OTU327 were standing out, highly significant in F6 and F7 (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), with sign 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. Of note, 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) feces, and h^2 estimates for OTU327 in D120 feces (0.13) (STable 4.4).

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

A 3.5 MY-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], 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 intra-chromosomal recombination between SINEs. We showed by RNA-Seq and eQTL analysis in 300 F7 cecum samples that the deleted allele produces reduced amounts (~1/3rd of normal) of a 7.4-Kb transcript (with alternative 8th and 9th exons) encoding a protein that lost 62% of its amino-acids including 7/8 active sites[25] (Fig. 4a). Our results indicate that the 2.3 Kb deletion creates a null allele (henceforth referred to as "*O*" allele) and that it is

a very strong candidate for the causative miQTL mutation (see also Extended Data Fig. 5 & 6).

We examined the effect of *AO* genotype on the abundance of OTU476, OTU327 and p-75-a5. It showed that (i) the effect of the *A* allele is dominant over the *O* allele (STable 5.1), and (ii) that the effect manifests in D120 and D240 feces, cecal content as well as mucosa, but not in D25 feces and D240 ileal content and mucosa (Fig. 4b and STable 5.1). In these samples (D120, D240, CC, CM), *AO* genotype explained 7.9%, 3.2% and 6.6% of the variation in abundance for OTU476, OTU327 and genus p-75-a5 (STable 5.2). Of note, the abundance of OTU476 and OTU327 was highest in cecal content (~0.92% and ~0.02% of reads in *AA/AO* animals, and ~0.47% and ~0.003% of reads in *OO* animals; Extended Data Fig. 5).

In primates, the *ABO* locus is under strong balancing selection that has perpetuated identical-by-descent (IBD) alleles segregating in humans, gibbons and Old-World monkeys for tens of millions of years[26]. We analyzed the sequences of the 61 F0 (*Sus scrofa domestica*), 18 wild boars (9 Asian, 9 European)(*Sus scrofa*), one Indonesian wild boar from Sumatra (*Sus scrofa vittatus*), one Visayan warty pig from the Philippines (*Sus cebifrons*), one 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 have diverged from a common *Sus scrofa* ancestor ~1 million years ago (MYA), *Sus scrofa* and *Sus scrofa vittatus* ~1.5 MYA, *Sus scrofa* and *Sus cebifrons/verrucosus* ~3.5 MYA, and *Sus scrofa* and *Phacochoerus africanus* ~10 MYA[27]. The 2.3Kb deletion segregated in all eight F0 breeds, in all Asian and European/American wild-boar populations, and in *Sus cebifrons* (STable 5.3). The sequence of the deletion breakpoint was identical in all samples confirming the IBD 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 *Sus cebifrons* lay outside the cluster of *Sus scrofa O* alleles (Fig. 4c, Extended Data Fig. 7). Further supporting balancing selection, the *AO* gene showed a marked drop 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 exceedingly long periods and broad geographic 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 particularly 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, because of the abundance of intestinal mucus, the "*O*" allele might reduce the concentration of GalNAc in the intestine, thereby reducing the growth of bacterial species dependent on this sugar. We measured the concentration of free HexNAc (= GalNAc, GlcNAc and ManNAc isomers) in caecum content of 278 D240 animals (124 F7 and 154 Duroc x Landrace x Large White) by LC-MS/MS. Intestinal HexNAc in pigs comprises approximately equal proportions of the GlcNAc and GalNAc isomers[30]. Dosage of the *A* allele increased ($p_{add\_model} = 5.7 \times 10^{-9}$) caecal HexNAc concentrations (Fig. 5a). Assuming that the increase is only due to GalNAc (given the enzymatic activity of the *A* allele), *AA* animals have at least twice the amount of free GalNAc compared to *OO* (Fig. 5a, Extended Data Fig. 9). Non-zero values in OO animals are primarily due to GlcNAc/ManNAc, as well as likely 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). It is noteworthy that the effect of the A allele appears dominant on bacterial abundance, yet additive on GalNAc concentrations (p-value of LR-test against additive model = 0.12). This indicates that the additional

increase in GalNAc availability in *AA* animals (vs *AO* animals) does not further favor growth of OTU476, OTU327 and p-75-a5.

Showing that the *O* deletion (i) reduces caecal GalNAc concentrations on the one hand, and (ii) reduces the abundance of OTU476, OTU327 and p-75-a5 on the other, 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 according to which it is the change in GalNAc concentrations that causes the change in bacterial abundance makes a testable prediction. We may expect in this case to observe a "residual" correlation between GalNAc concentrations and bacterial abundance "within" *AO* genotype, i.e. by blocking the effect of 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, Extended Data Fig. 9).

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

miQTL-responsive bacteria utilize GalNAc

To be used as carbon source by intestinal bacteria, GalNAc needs (i) to be released from the glycan structures by secreted glycoside hydrolases (GHs), (ii) to be imported across the bacterial membranes by dedicated transport systems (TR), and (iii) to be converted into intermediates of central metabolism by a specific catabolic pathway (CP). While some bacteria have both GHs and TR/CP for specific monosaccharides, other only have 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 (i) isolated two bacterial strains (4-8-110 and 4-15-1) with V3-V4 sequence similarity of 100% and 99.8% with OTU476 and sequenced their genome on an Oxford Nanopore Promethion, and (ii) 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 utilization (TR/CP)[28,31–36]. These encode (i) 11 components of three GalNAc transporter systems (AgaPTS: *agaE, agaF, agaV, agaW;* TonB dependent transporter: *omp, agaP, agaK;* GnbPTS: *gnbA, gnbB, gnbC, gnbD*), (ii) two GalNAc-6P deacetylases (*agaA, nagA*), (iii) two galactosamine-6P (GalN-6P) isomerase and/or deaminases (*agaI, agaS*), (iv) three tagatose-6P kinases (*pfka, lacC, fruK*), (v) four tagatose-1,6-PP aldolases or aldolase subunits (*gatY-kbaY, gatZ-kbaZ, lacD, fba*), and (vi) two regulon repressors (*agaR, gntR*), for a total of six essential pathway constituents (Fig. 5d, STable 6). Genes involved in the utilization 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 capacity of bacterial species to utilize GalNAc. The first (pathway score) counted the number of essential steps in GalNAc utilization (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 utilization steps that could be fulfilled by orthologues that were clustered in the genome, i.e. forming a potential operon. We used *agaS* as anchor gene to establish the regulon score[28]. The first striking observation was that at least one orthologue of *agaS* was found in the two (=100%) 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* than for other MAGs ($p_{pathway}$=2.0e-16 and $p_{regulon}$=2.0e-16), and for the two OTU476-like strains than for

Erysipelotrichaceae and non-*Erysipelotrichaceae* MAGs combined ($p_{pathway}$=2.2e-3 and $p_{regulon}$=1.2e-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% of 3,113, demonstrating the uncommon status of OTU476-like strains with regards to GalNAc utilization. Neither the 4-15-1 nor 4-8-110 genome encode GHs 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 by GC-MS. After 13 hours of culture, ~29% of glyceraldehyde-3-P (GAP) and ~25% of dihydoxyacetone phosphate (DHAP) (the products of conversion of tagatose-1, 6-PP by the gatZ aldolase) were labelled with ^{13}C, demonstrating the utilization of GalNAc by strain 4-8-110. Approximately 23% of 3-phosphoglycerate and ~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 GalNAc addition on the relative growth of 4-8-110 and *E. coli*, *in vivo*. We gavaged mice either with GalNAc (200mg/kg 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 sacrificed the mice at day 12 and measured the relative abundance of *E. coli* and 4-8-110 by 16S rRNA sequencing in caecum and feces. 4-8-110 was in essence not detectable in PBS-gavaged mice, while it accounted for an average of 0.9% of reads in caecum content (p = 0.0079) and 0.7% of reads in feces (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. Why is the chromosome 1 miQTL not affecting these species? The organization of their GalNAc operons 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 ~10Kb) and on the same strand, compatible with poly-cistronic mRNAs enabling coregulated expression. In contrast, the ORFs of the GalNAc clusters of the two OTU476-like strains and at least one studied *Erysipelotrichaceae* are spanning respectively ~50 and ~30Kb, and 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, contrary 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 utilization only when needed[33,35,36], but may rather 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 analyzed their transcriptome by RNA Seq. The GalNAc regulon of *E. coli* was tightly regulated, with near complete absence of transcription in the absence of GalNAc, and ~300-fold upregulation in its presence. In the OTU476-like 4-15-1 strain, on the contrary, GalNAc TR/CP genes were expressed at nearly equal levels with and without GalNAc, as predicted. 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 (~5-10-fold) 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. While GalNAc addition 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 ≤ 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 (STable 7).

Discussion

In humans, *ABO* genotype affects susceptibility to various viral (including SARS-CoV-1&2), bacterial and protozoan pathogens[20,42,43], and possibly the composition of the intestinal microbiota[27]. Invoked mechanisms are usually immune, including pathogen adhesion, toxin binding, soluble decoys and natural antibodies[26,44]. We provide strong evidence that the miQTL reported in this work acts "simply" by affecting intestinal GalNAc concentrations, and thereby the growth of GalNAc-utilizing bacteria. A puzzling finding is that not all species capable of utilizing 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 behavior of the GalNAc operon 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 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) because wasting energy transcribing and translating useless genes. By tightly regulating expression of their operon in response to ambient GalNAc availability, species like *E. coli* may fair equally well in the gut of *AA/AO* as in that of *OO* pigs, hence not be affected by the miQTL. We note that, in addition to the distinct organization of their GalNAc operon, the transcriptome of the two studied OTU476-like strain responds more dramatically to GalNAc addition than that of *E. coli*. The precise significance of this observation remains to be established but may point towards alternative or additional factors that underpin differential sensitivity to the miQTL.

Null *FUT2* alleles segregate in humans, precluding synthesis of the type I H antigen in homozygotes, and hence epistatically of the type I A and/or B antigens in non-O individuals[20,21] (Extended Data Fig. 8). Although an *H117R FUT2* 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 *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* vs *A* and *O* individuals[45], a subsequent study in a larger cohort (n=1,503) could not detect experiment-wide significant effects of *ABO* genotype on gut microbiota composition[46]. More recently, an effect of *ABO* genotype was reported on the abundance of OTUs assigned respectively to *Faecalobacterium* and *Bacteroides* in a cohort of ~9,000 Germans[21], but this was not confirmed in a subsequent study (~18,000 individuals)[47]. We examined the effect of ABO blood group on the abundance of ~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 *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. **Note:** 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*, Online ahead of print https://doi.org/10.1038/s41564-020-0743-8 (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., Davenport, E. R., Beaumont, M., Jackson, M. A., Knight, R., Ober, C., Spector, T. D., Bell, J. T., Clark, A. G. & Ley, R. E. 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., Cuesta-Seijo, J. A., Striebeck, A., Lafont, B., Palcic, M. M. & Vidal, S. 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 Animal 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 Genomics* **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., Monedero, 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 prokaryotes 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. Rahfeld, 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://www.medrxiv.org/content/10.1101/2020.05.31.20114991v1 (2020).
44. Blancher, A. Evolution of the ABO supergene family. *ISBT Science Series* **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., Goodrich, J. K., Cell, J. T., Spector, T., Ley, R. E. & Clark, A. G. ABO antigen and secretor statuses are not associated with gut microbiota composition in 1,500 twins. *BMC Genomics* **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 Biology* **20**, 16 (2019).
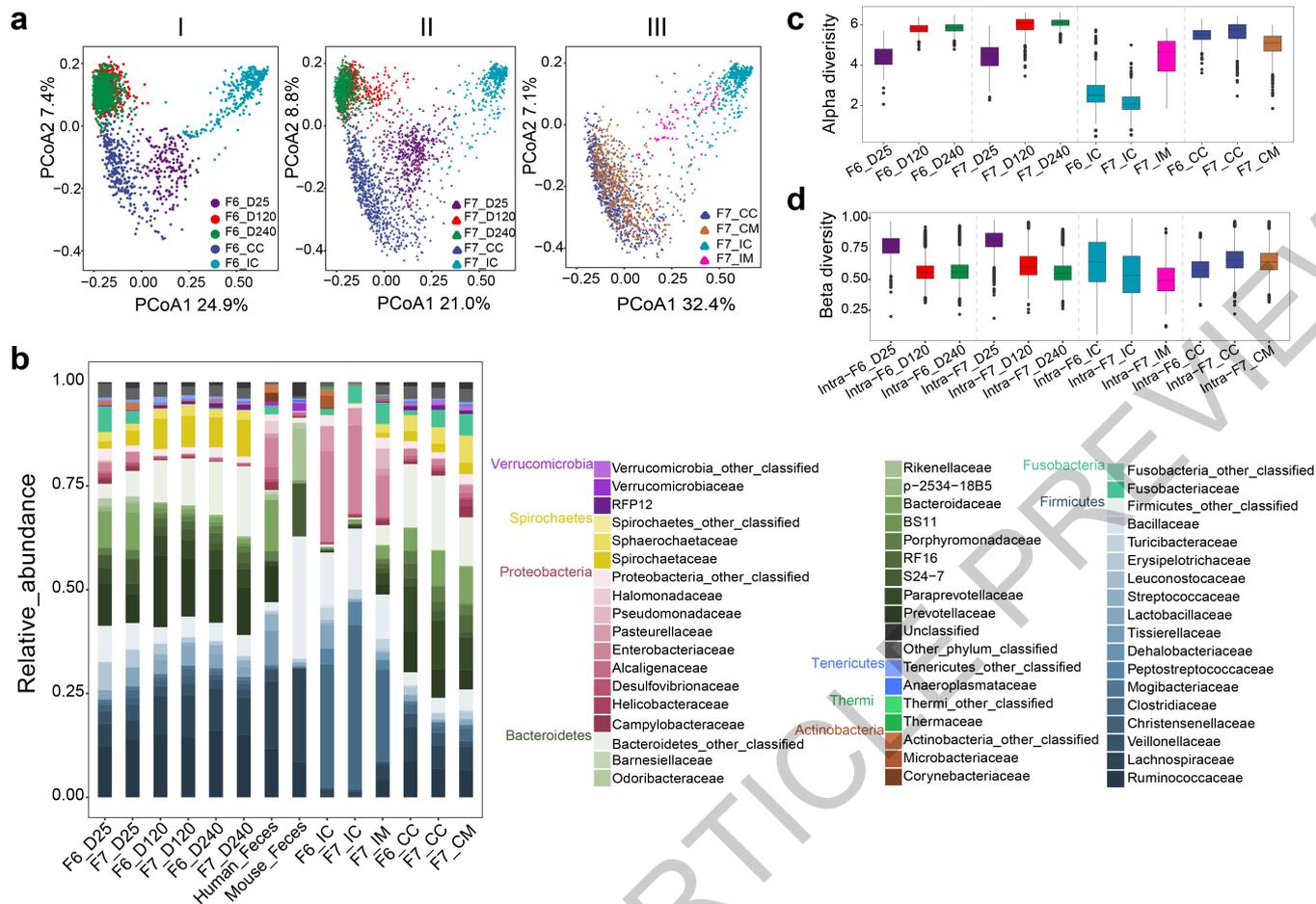
Fig. 1 | Intestinal microbiota of the healthy pig. (a) Joint Principal Coordinate Analysis (PCoA) of 5,110 16S rRNA profiles. (I) F6: day 25 feces (D25, mauve), day 120 feces (D120, red), day 240 feces (D240, green), ileal content (IC, light blue), cecum content (CC, dark blue). (II) F7: as in I. (III) F7: ileal content (IC, light blue), cecum content (CC, dark blue), ileal mucosa (IM, pink), cecal mucosa (CM, brown). **(b)** Average microbiota composition of the 12 data series. Taxa colored by phylum and family within phylum, highlighting 43 families amongst the top 15 in at least one data series. Names of the corresponding phyla and families are in the legend. Average composition of 106 human feces and 6 mouse feces (C57BL/6). **(c)** α-diversities (Shannon's index) for the 12 data series color-labelled as in A. Sample sizes are as in STable 2.1. Boxplots show median (line in box), interquartile range (IQR, box), 1.5x IQR span (whiskers), and outliers (dots). **(d)** β-diversities (Bray-Curtis distance) for the 12 data series colored as in A. Distances computed for all sample. Sample numbers and boxplots as in C.
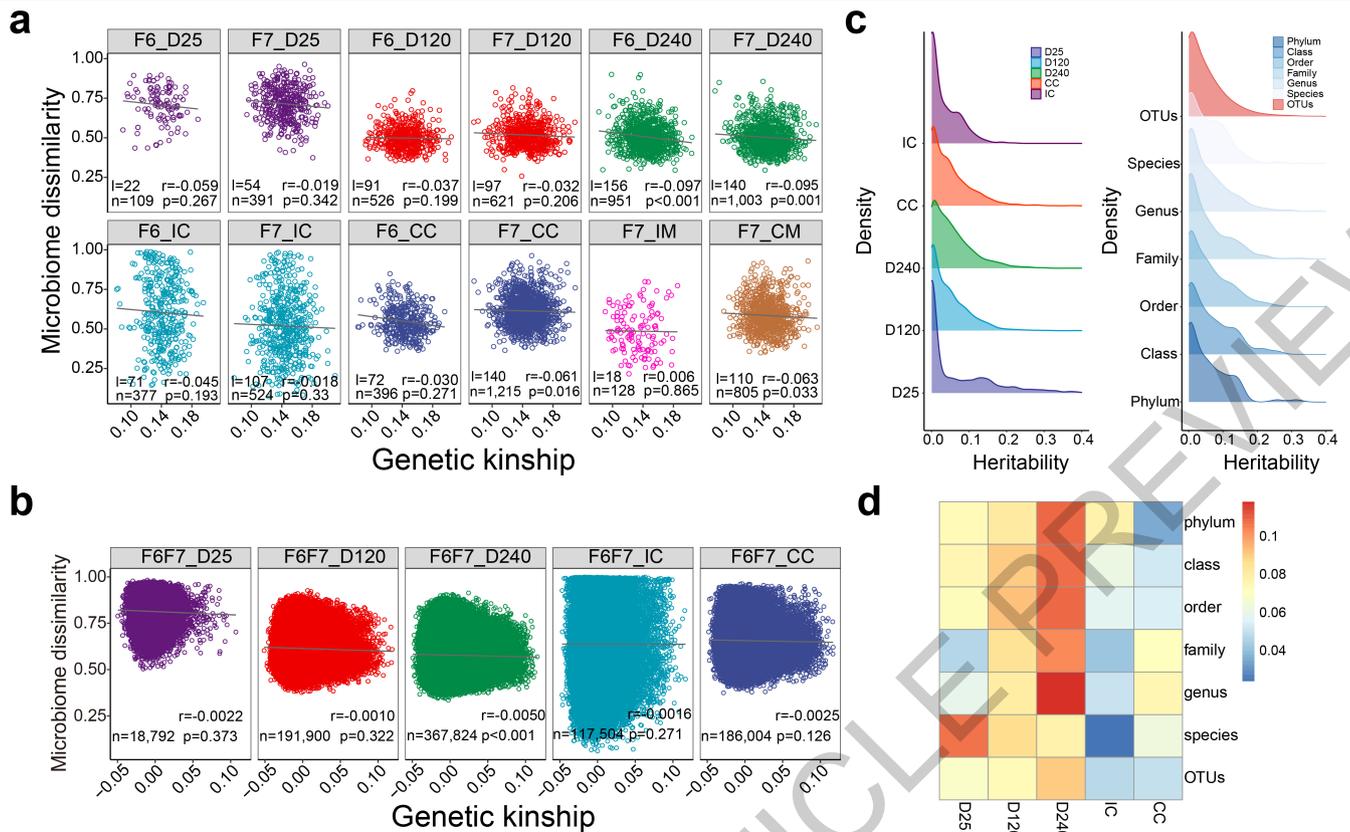
Fig. 2 | Heritability of microbiota composition in mosaic pigs.
(a) Correlation between genome-wide kinship (Θ) and microbiome dissimilarity (Bray-Curtis distance) within litter. Correlations (Spearman r) were measured separately for the 12 data series. P-values (one-sided) were computed by permutation. Adjusted r-values were below the 50-ties percentile of the permutation values for 11 of 12 (p = 0.0029). Empirical p-value (one-sided) of r was ≤ 0.05/12=0.004 (i.e. Bonferroni corrected) for two. p-values were combined across the 12 data series yielding an overall p-value of $3×10^{-4}$. Number of litters (*l*) and animal pairs (*n*) used are given for each data series. **(b) Correlation between genome-wide kinship (Θ) and microbiome**
dissimilarity (Bray-Curtis distance) across generations. We considered all possible pairs of F6 and F7 animals (not including sow-offspring pairs). Analyses conducted for the five traits measured in both F6 and F7. *r, p,* and *n* are as in (A). r-values were below the 50-ties percentile of the permutation values for the five analyzed sample types (p = 0.03). Empirical p-value (one-sided) of r was ≤ 0.05/5=0.01 (i.e. Bonferroni corrected) for one. p-values for the five sample types were combined yielding an overall p-value of 0.013. **(c) Frequency distribution of heritabilities of individual taxa** sorted by sample type (left) or taxonomic level (right). Values obtained by joint analysis of F6 and F7. **(d)** Total heritabilities[2] computed by sample type and taxonomic level.
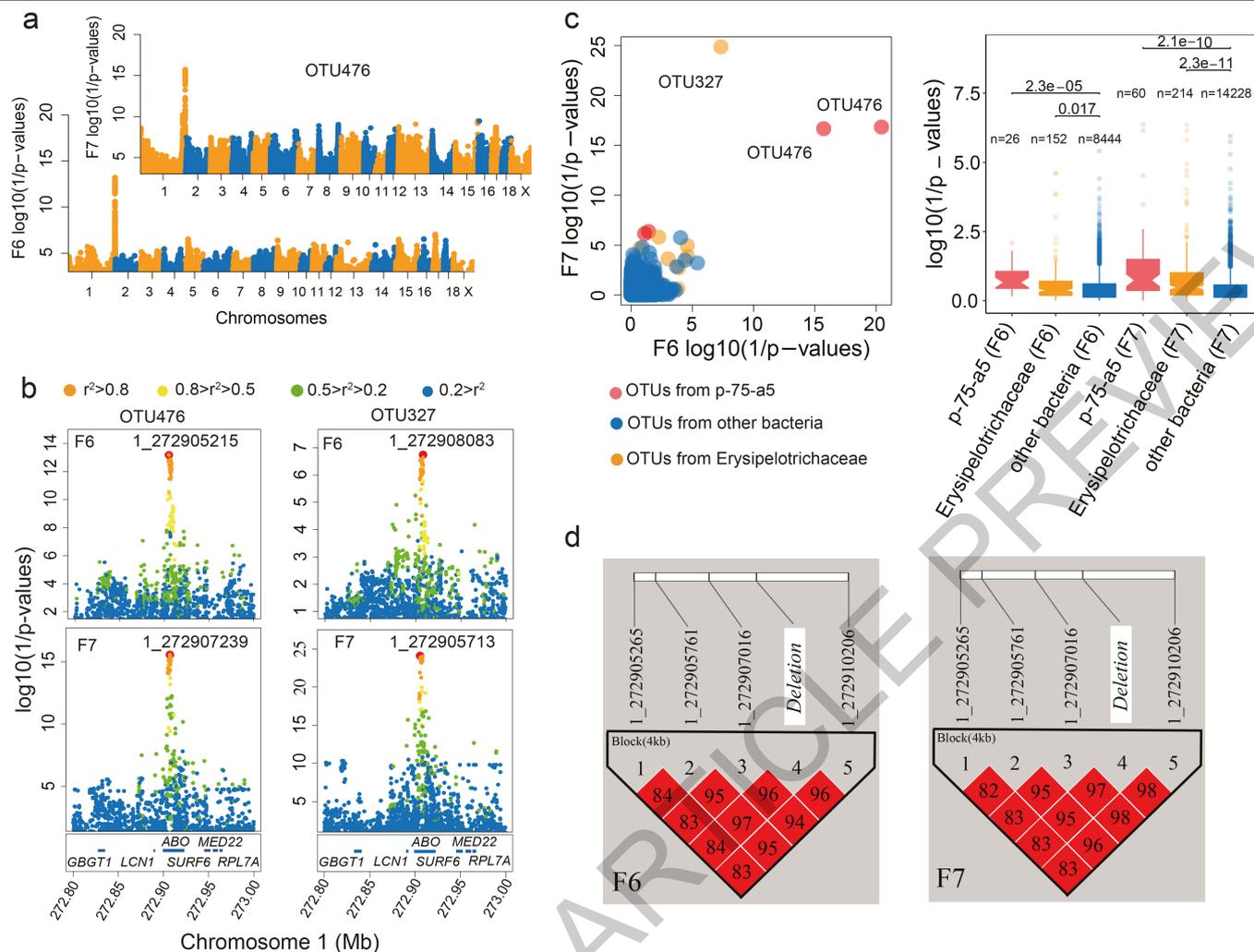
Fig. 3 | A miQTL affecting *Erysipelotrichaceae* species. (a) Result of genome-wide meta-analysis (across sample types) in F6 and F7 for OTU476. Reported log transformed p-values are nominal (i.e. not corrected for multiple testing). **(b)** Local zooms (chromosome 1: 272.8-273Mb) for OTU476 and OTU327 in F6 and F7. Log transformed nominal p-values as in A. **(c)** (Left) Log(1/p) values in F6 (x-axis) and F7 (y-axis) for association between SNP 1_272907239 and abundance of 8,490 OTUs for all sample types and two analyses methods (abundance and presence/absence, explaining the two OTU476 values). OTUs that belong to p-75-a5 (respectively *Erysipelotrichaceae*) are shown in red (respectively yellow). (Right) Comparing the distribution of association (1_272907239) p-values for p-75-a5 and *Erysipelotrichaceae* OTUs with other OTUs in F6 and F7. Boxplots as in Fig. 1c. Notches in boxes correspond to 95% confidence intervals of the medians. Distributions were compared using Wilcoxon's sum-of-rank test. P-values (nominal) of the comparisons are given above the horizontal lines. **(d)** Linkage disequilibrium (r^2) between the four top SNPs and the 2.3Kb ABO deletion in F6 and F7.
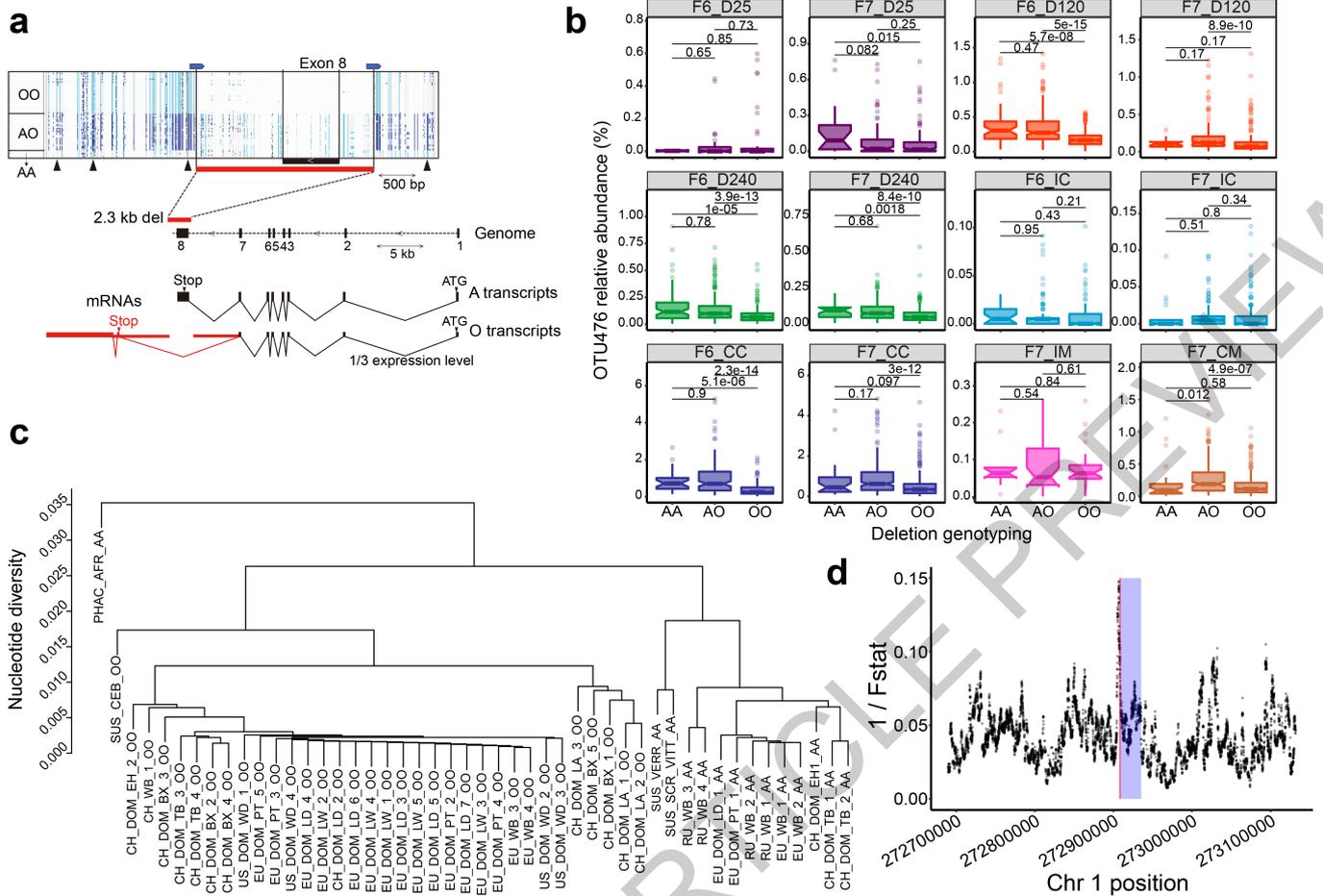
Fig. 4 | A 3.5 MY-old deletion in the pig ABO orthologue causes the miQTL. (a) Structure of the porcine AO blood group gene. IGV view of the genotypes of the 61 F0 for 145 variants in a ~5 Kb interval spanning the 2.3 Kb deletion. Two upper red rectangles: 2.3 Kb deletion. Light blue: homozygous alternate; dark blue: heterozygous; gray: homozygous reference. Horizontal blue arrows mark SINEs that have mediated intra-chromosomal recombination. Vertical black arrows mark top variants from Fig. 3b. Effect of the 2.3 Kb deletion on acetyl-galactosaminyl transferase transcripts: (i) creation of alternate exon 8 and 9, and (ii) reduction of transcript levels to ~1/3th of normal. **(b) Effect of *AO* genotype (*AA, AO* or *OO*) on abundance of OTU476.** (i) The effect of the *A* allele is dominant over that of the *O* allele, and (ii) the miQTL effect is detected in cecum (content and mucosa) and in day 120 and 240 feces. Sample sizes as in

STable 4.1. Boxplots as in Fig. 3d. **(c)** UPGMA dendrogram based on sequence similarity between 14 AA and 34 *OO* animals in a 5-Kb window centered on 2.3Kb deletion. PHAC_AFR: common warthog, SUS_VERR: Javan warty pig, SUS_CEB: Visayan warty pig, SUS_SCR_VII: Sumatran wild boar, CH/RU/EU_WB: Chinese/Russian/European wild boars, CH/EU/AM_DOM: Chinese/European/American domestic pigs. Breed acronyms as in ED 1. **(d) Peak of reduced population differentiation coinciding with the 2.3 Kb deletion (red) in the porcine *AO* gene (blue)**. X-axis: position on chromosome 1. Y-axis: 1/(mean F statistic) for all variants in a 2Kb sliding window. F statistic computed as the ratio of the "between-breed mean squares" and the "within-breed mean squares" for the dosage of *O* allele.
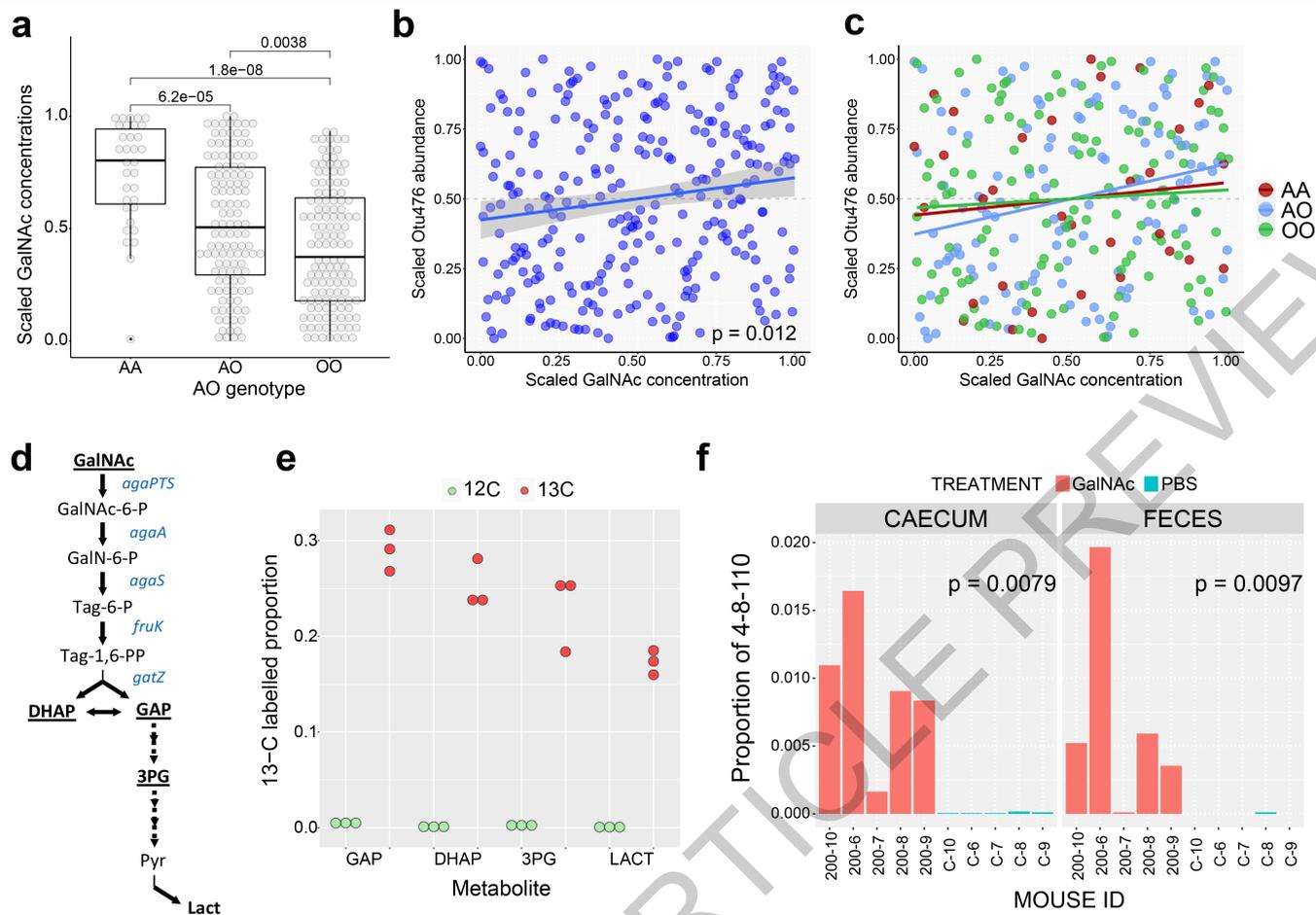
Fig. 5 | The miQTL acts by increasing GalNAc concentrations and affects GalNAc utilizing bacteria. (a) Effect of *AO* genotype on GalNAc concentrations in caecal content (n_{AA}=33, n_{AO}=118, n_{OO}=127). Concentrations corrected for batch effect and scaled between 0 and 1 to equalize residual variance. P-values (two-sided and nominal) for genotype contrasts computed using Wilcoxon's test. Boxplots as in Fig. 1c. **(b)** Correlation between GalNAc concentration and OTU476 abundance "within *AO* genotype". Areas Under the Curve (AUC) for GalNAc corrected for batch effect and *AO* genotype and scaled as above. P-value (nominal, two-sided) of Spearman's correlation is given (p = 0.012). Shaded area corresponds to the 95% confidence region for the regression fit. **(c)** Same as in (B), with animals colored by *AO* genotype. **(d)** GalNAc transport and catabolic pathway in OTU476-like strains. GalNAc: N-acetylgalactosamine, 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 in blue. Metabolites considered in the metabolic flux analysis are in bold. **(e)** Proportion of [13]*C*-labelled metabolites determined by GC-MS in 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. Proportion of 16S rRNA reads mapping to the 4-8-110 reference rRNA sequence versus *E. coli* rRNA sequence (i.e. 1 minus the proportions shown in the figure correspond to reads mapping to the E.coli 16SrRNA) in caecum content and feces of 10 germ free mice (Kunming line) inoculated by gavage with pure culture of 4-8-110 and *E. coli* and force-fed with GalNAc (red bars) versus phosphate-buffered saline (PBS) (green bars). P-values (nominal, two-sided, uncorrected) of the difference in abundance determined using Wilcoxon test.

Fig. 6 | GalNAc operon organization and transcriptome response of miQTL responsive bacteria. (a) GalNAc operon organization and local transcriptome response to GalNAc addition in OTU476-like strains and *E. coli*. Upper: GalNAc operon organization: Identified Open Reading Frames (ORFs) represented as colored boxes. Genes implicated in GalNAc import and catabolism 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 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. **Lower:** Local gene expression levels (FPKM) with (dark color) and without (light color) addition of GalNAc in the growth medium. The colors for the ORFs with known function (GalNAc gene: red, other gene: blue) are as in the upper panels. The error bars are SE from three replicates (individual values shown as dots). **(b) Global transcriptome response to GalNAc addition in OTU476-like strains and *E. coli*:** Log2-fold expression change 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. Red: GalNAc genes. Blue: other genes. Insets: corresponding QQ plots, showing the near absence of effects on genes other than GalNAc regulon in *E. coli* versus widespread response in OTU476-like strains. P-values used to generate the QQ plots are nominal, obtained with DISEQ2.

Article

Methods

Animal rearing and sample collection

This study focused on the sixth (F6) and seventh (F7) 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 F0 generation. For each Chinese breed, the boars were mated with the ewes of one European breed, and the sows with the boars of another European breed to produce the F1 generation. Thus, every Chinese and every European breed is parent breed of two distinct F1 hybrid combinations each, for a total of eight F1 combinations (BX-LW, BX-PT, LA-PT, LA-LD, TB-LD, TB-WD, EH-WD, EH-LW). The F2 generation was obtained by mating each F1 hybrid combination with two others that did not share parental breeds for a total of eight F2 combinations (BX-LW x LA-PT, BX-PT x LA-LD, LA-PT x TB-LD, LA-LD x TB-WD, TB-LD x EH-WD, TB-WD x EH-LW, EH-WD x BX-LW, EH-LW x BX-PT). Every F2 combination was obtained by reciprocally crossing an average of 4 boars from one F1 combination with an average of 7.25 sows from the other. The F3 generation was obtained by mating each of the eight F2 hybrid combinations with the only complementary F2 combination that did not share any parental breeds for a total of four F3 combinations (BX-LW-LA-PT x TB-LD-EH-WD, BX-PT-LA-LD x TB-WD-EH-LW, LA-PT-TB-LD x EH-WD-BX-LW, LA-LD-TW-WD x EH-LW-BX-PT) expected to each have ~12.5% of their genome from each of the founder breeds. Every F3 combination was obtained by reciprocally crossing an average of 7 boars from one F2 combination with an average of 10.8 sows from the complementary one. The F4, F5, F6 and F7 generations were obtained by intercrossing 57 boars x 75 sows (F3->F4), 62 boars x 97 sows (F4->F5), 85 boars x 170 sows (F5->F6), and 82 boars x 111 sows (F6->F7)(STable 1).

All F6 and F7 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 ~46 days of age. Litters were transferred to 12-pig fattening pens with automatic feeders (Osborne Industries, US), 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. Fecal samples were manually collected from the rectum of experimental pigs at the ages of 25, 120 and 240 days, dispensed in 2ml tubes, flash frozen in liquid nitrogen, and stored at -80 °C. Animals were slaughtered at day 240. Ileum and cecum were sealed at both ends with a sterile rope and extracted from the carcass. Within 30 min after slaughter, ileal and cecal luminal content were collected (F6 and F7 animals), ileum and cecum rinsed with sterile saline solution, and samples of ileal and cecal mucosa scraped with a sterile microscopic slide (F7 animals only). Approximately one gram 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 STable 2.1. All the animals included in the analyses were healthy and did not receive any antibiotic treatment within one month of sample collection. All procedures involving animals were approved by the Ethics Committee of the Jiangxi Agricultural University (No. JXAU2011-006).

Genotyping by sequencing of the F0, F6 and F7 generations

Genomic DNA was extracted from ear punches using a standard phenol-chloroform-based DNA extraction protocol. DNA concentrations were measured using a Nanodrop-1000 instrument (Thermo Scientific, USA), and DNA quality of all samples assessed by agarose (0.8%) gel electrophoresis. Genomic DNA was sheared to 300-400 bp fragment size. 3'-ends were adenylated and indexed primers ligated. Libraries were amplified by PCR using Phusion High-Fidelity DNA polymerase (NEB, USA) following the recommendations of the manufacturer (Illumina, US). The libraries were loaded on Illumina X-10 instruments (Illumina Inc., San Diego, CA) for 2×150 bp paired-end sequencing by Novogene (Beijing, China). We removed reads with quality score ≤ 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 *Sus scrofa* reference genome assembly 11.1[50] using BWA (v0.7.17)[51]. Bam files of mapped reads were sorted by chromosome position using SAMTools (v1.6)[52]. Indel realignment and marking of duplicates were done with Picard (v2.21.4, http://broadinstitute.github.io/picard). Individual genotypes were called from BAM files using Platypus (v0.8.1)[53]. Individual genotypes were merged into a single VCF file using PLINK (v1.9)[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 minor allele frequencies (MAF) < 0.03 were removed.

Computing nucleotide diversities

Nucleotide diversities between pairs of breeds were computed from variant frequencies as follows:

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

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 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 F6 and F7 generation at genome and chromosome level

We estimated the proportion of the genome of the eight founder breeds in the F6 and F7 generation following Coppieters et al[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, etc … as well as a frequency in the F6 (or F7) 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 F6 (or F7) as P_1, of breed 2 in generation F6 (or F7) as P_2, etc … We estimated the values of P_1, P_2, etc. … using a set of linear equations:

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

$$\vdots$$

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

$$\vdots$$

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

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 feces, luminal content and mucosal scrapings using the QIAamp Fast DNA stool Mini Kit following the manufacturer's recommendations (Qiagen, Germany). DNA concentrations were measured using a Nanodrop-1000 instrument (Thermo Scientific, USA), and DNA quality 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 an Illumina MiSeq platform (Illumina, USA) at Major bio (Shanghai, China). 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 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 ~40,888 (Suppl. Table 2.1). Chimeric reads were removed using USEARCH (v.7.0.1090)[59]. Sequence data were rarefied to 19,632 tags, i.e. the lowest number of tags per sample. Tags were clustered in operational taxonomic units (OTUs) with VSEARCH (v.2.8.1)[60] using 97% as 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 (v13.5) database and the RDP classifier (v2.2)[61]. Principal coordinate analysis (PCoA) was performed with the "ape" and "vegan" R packages (v3.5.3) 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 (v3.5.3). The mouse fecal microbiome data were from Cheema et al[63]. The human fecal microbiome data were 16S rRNA data from 106 healthy individuals (Shagam et al., in preparation).

Measuring the heritability of microbiota composition (all taxa combined)

We first estimated the impact 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 (see hereafter) using the SNP genotypes at the above-mentioned 30.2 million DNA variants using either GEMMA (v0.97)[64] or GCTA (v1.26)[65]. Both programs yielded estimates of Θ with same distribution after standardization, albeit different raw values. We herein report results obtained with GEMMA (v0.97). 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 (v3.5.3). We first performed this analysis for each sample type and generation separately within litter, i.e. only considering pairs of full-sibs born within the same litter, hence in essence following Visscher et al[67]. We then performed the analysis across the F6 and F7 generations. The pairs of individuals considered were all F6-F7 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 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 (v3.5.3)[68]. P-values for the heritability estimates were computed using the associated *lrt_h2* function[68]. We first estimated heritabilities using information from the F6 and F7 generation jointly ("F6 plus F7"). The model included "individual animal" (i.e. polygenic effect), "dam", "litter", "pen" and "batch" as random effects and "generation" as fixed effect. The additive genetic relationship matrix determining covariances between "individual animal" effects was computed with GCTA (v1.26)[65]. This "F6 plus F7" analysis extracts information from the correlation between kinship and phenotypic resemblance (i) between F6 animals ("within F6"), (ii) between F7 animals ("within F7"), and (iii) between F6 and F7 animals ("between F6 and F7"). To evaluate whether these three sources of information were consistent, we also estimated – for each taxon/OTU - the three heritabilities separately. The "within F6" 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 "F6 plus F7" analysis. The "between F6 and F7" heritabilities were estimated by regressing squared difference in abundance on additive relationship following Haseman-Elston[69]:narrow sense heritability was estimated as $-\hat{\beta}/(2\hat{\sigma}_P^2)$ were $\hat{\beta}$ is the least square regression coefficient and $\hat{\sigma}_P^2$ an estimate of the phenotypic variance (see Extended Data Fig. 3). The "total heritability" of the intestinal microbiome was further computed from the heritabilities and abundance of individual taxa/OTU following Rothschild et al[2].

Mapping microbiota QTL (miQTL)

miQTL were mapped using the GenABEL R package (v3.5.3)[70] applying two models following Turpin et al[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 (i.e. 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 F6, 23 for F7) and the three first genomic principal components as fixed covariates. GWAS were conducted separately for each taxon × data series combination and p-values 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 (v3.0)[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 (i.e. across all tested SNPs) average ($\bar{Z}$) and standard deviation (σ_Z) of the Z score. The p-value of Z scores was (conservatively) computed by assuming that $(Z - \bar{Z})/\sigma_Z$ 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, Beijing, China). A 40 kb SMRTbell DNA library (Pacific Biosciences of California, CA, USA) was prepared using BluePippin for DNA size selection (Sage Science, MA, USA) and then

sequenced on a PacBio Sequel platform (Pacific Biosciences of California, CA, USA) with P6/C4 chemistry at Novogene Biotech, Beijing, China. We obtained a total of 18,148,470 subreads with N50 length of 17,273 bp. Additionally, a paired-end library with insert size of 350-bp was constructed and sequenced on an Illumina Novaseq 6000 PE150 platform (2x150bp reads) at Novogene Biotech, Beijing, China. PacBio reads were self-corrected using Canu (v1.7.1) before assembly with Flye (v2.4.2)[72]. Errors in the primary assembly were first corrected using PacBio subreads using racon (v1.4.10)[73], and Illumina paired-end reads were then mapped to the contigs using bwa-mem[74] to polish the contigs using Pilon (v1.23, Broad Institute, MA, USA)[75]. Lastz (v1.02.00)[76] and Minimap2 (v2.17-r941)[77] were used to compare the Bamaxiang contig and the 40k sequence spanning the ABO gene of the Sus scrofa Build 11.1 reference genome.

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

We designed two pairs of primers to genotype the deletion in the F6 and F7 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′- TTGCCTAAGTCTACCCCTGTGC-3′). The second pair of primers was located in exon 8 (FP2: 5′-CGCCAGTCCTTCACCTACGAAC-3′, RP2: 5′-CGGTTCCGAATCTCTGCGTG-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, Japan) under thermocycle conditions of 94 °C for 4 minutes, 35 × (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 minutes on a PE 9700 thermal cycler (Applied Biosystem, USA).

RNA seq and eQTL analysis

A total of 300 cecum tissue samples from F7 pigs which also had microbiota and genotype data were used to extract total RNA with TRIzol™ (Invitrogen, USA) following the manufacturer's manual. Total RNA was electrophoresed on 1% agarose gels. RNA purity and integrity were assessed using an eNanoPhotometer® spectrophotometer (IMPLEN, USA) and a Bioanalyzer 2100 system (Agilent Technologies, USA). A Qubit3.0 Fluorometer (Life Technologies, CA, USA) was used to measure RNA concentrations. 2-µg total RNA of each sample were used to constructed RNA sequencing libraries, using the NEBNext® UltraTMR NA Library Prep Kit for Illumina (NEB, USA) following the manufacturer's protocol. Briefly, Oligo (dT) magnetic beads (Invitrogen, USA) 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, USA). After purification, cDNA was end-repaired, and index codes and sequencing adaptors ligated. After PCR amplification, purification and quantitation, the libraries were sequenced on a 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 (v020201)[78]. Samtools (v1.6)[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 (v1.6.4)[79]. The expression abundance of the ABO gene was normalized to fragments per kilobase of exon model per million mapped reads (FPKM). Gender and batch were treated as covariates to correct for gene expression levels, and the corrected residuals used for subsequent analyses. GEMMA (v0.97)[64] was used to analyze the association of ABO expression level with genome-wide variants using a linear mixed model.

Whole-genome sequencing and bioinformatic analysis for wild boars, Sus verrucosus and Sus cebifrons

The genomes of six Russian wild boars, one Sumatran wild boar, and one African warty hog were sequenced on an Illumina HiSeq X Ten platform at Novogene Biotech, Beijing, China. Additionally, six Chinese wild boars were sequenced in a previous study[80], and we downloaded the genome sequence 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 (v4.2)[81]. We replaced the ABO gene of the Sus 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 Sus scrofa reference genome (build 11.1) using BWA (v0.7.17)[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, hence circumventing the need to phase the corresponding genotypes. We defined windows of varying size (0.5 to 50Kb) centered around the 2.3 Kb deletion. For all pairs of individuals, we computed a running sum over all 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 pair-wise 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:

$$BMS = \left(\sum_{i=1}^{B} \sum_{j=1}^{n_i} \left(\overline{y}_i - \overline{y}_T \right)^2 \right) / (B - 1)$$

$$WMS = \left(\sum_{i=1}^{B} \sum_{j=1}^{n_i} \left(y_{ji} - \overline{y}_i \right)^2 \right) / (N_T - B)$$

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 $\overline{y}_T$ is the average allelic dosage in the entire data set. We computed the average of the corresponding F-statistic for all variants within a sliding window of fixed physical size (f.i. 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 across variants (and taking the inverse) following Nei[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, cecum, 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, CA, USA). RNA concentration was measured using Qubit® RNA Assay Kit in Qubit® 2.0 Flurometer (Life Technologies, USA). 1-µg total RNA of each sample were used to constructed RNA sequencing libraries. Sequencing libraries were generated using TruSeq RNA Library Preparation Kit (Illumina, USA) following manufacturer's recommendations and index codes were added to attribute sequences to each sample. Briefly,

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. Remaining overhangs were converted into blunt ends via exonuclease/polymerase activities. After adenylation of 3' ends of DNA fragments, Illumina adaptors were ligated. In order to select cDNA fragments of preferentially 350-400 bp in length, the library fragments were purified with AMPure XP system (Beckman Coulter, USA). PCR was performed with Phusion High-Fidelity DNA polymerase, Universal PCR primers and Index (X) Primer. PCR products were purified (AMPure XP system) and library quality was assessed on an Agilent Bioanalyzer 2100 system. Clustering of the index-coded samples was performed on a cBot Cluster Generation System using TruSeq PE Cluster Kit v3-cBot-HS (Illumia, USA) according to the manufacturer's instructions. Sequencing was performed on an Illumina Novaseq platform and 150 bp paired-end reads were generated. Reads were filtered 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 (v2.2.1)[85]. Samtools (v1.6)[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 (v1.6.4)[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 via function dist(), hclust(), as.dendrogram() and set() implemented in R package stats and dendextend.

Testing the effect of AO genotype on other phenotypes
The associations between the 2.3Kb ABO deletion and 150 traits were calculated in the F6 and F7 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 F6 and F7 generations as follows: $Z = Z1W1 + Z2W2/(W1 + W2)$, where $Z1 = b1/SE1$ and $Z2 = b2/SE2$, $W1 = 1/(SE1)^2$ and $W2 = 1/(SE2)^2$, where the subscripts 1 and 2 denote F6 and F7 generations, respectively; b1, b2, SE1 and SE2 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 F6 and F7 generation, while 150 traits that were shared in F6 and F7 generations were used for meta-analysis.

Determination of the concentration of N-acetyl-galactosamine in cecal 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 cecal lumen samples. Samples of cecal content were collected at D240 for 124 F7 and 154 animals from a Duroc x Landrace x Large White commercial population. The samples were lyophilized, grounded into powder by a Mixer Mill MM 400 (30 Hz, 1 min) (Retsch, Germany), and stored at -80 °C until use. Approximately 15 mg of powder was weighted and extracted with 500 μL of 70% methanol/water. The mixture was vortexed for 10 min, and centrifuged at -16,000 g and 4 °C for 10 min. 300 μL of supernatant 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 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, USA) in an ExionLC™ AD System (AB Sciex, USA). Linear ion trap (LIT) and triple quadrupole (QQQ) scans were performed on an Applied Biosystems 6500 Triple Quadrupole (QTRAP® 6500) equipped with an ESI Turbo Ion-Spray interface. The operation was carried out under positive ion

mode and controlled by Analyst software (v1.6.3) (Sciex, USA). The ESI source parameters were set as follows: ion source: ESI+, source temperature: 550 °C, ion spray voltage (IS): 5500 V (Positive), curtain gas (CUR): 35 psi. The GalNAc was analyzed using multiple reaction monitoring (MRM). 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 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
Fecal 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 (pH7.0) in an anaerobic glovebox (Electrotek, UK), which contained 10% hydrogen, 10% carbon dioxide and 80% nitrogen. The fecal suspension was diluted 10^{-6}, 10^{-7}, 10^{-8} and 10^{-9} -fold, and plated on GAM medium (Nissui Pharmaceutical, Japan). 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 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 harvested at the period of logarithmic growth. Genomic DNA was extracted using the Blood & Cell Culture DNA Midi Kit (Qiagen, Germany) following the manufacturer's protocol. Libraries for whole-genome sequencing of the strains were constructed and sequenced on an ONT PromethION (Oxford Nanopore Technology, UK) at NextOmics (Wuhan, China). To correct sequencing errors, a library for second-generation sequencing was also constructed for each of the two strains and sequenced (2x100 bp) on a BGISEQ platform (BGI, China). Bioinformatic analyses of sequencing data were performed following Kolmogorov et al[72]. and Hunt et al[86]. In brief, after quality control, the sequence data were assembled with flye (v2.6)[72] with parameter: --nano-raw, and the assembled genomes were corrected by combining the Oxford Nanopore data with the second-generation sequencing data using pilon under default parameter. The encoded genes were predicted using prodigal (v2.6.3) (parameter: -p none-g 11)[87].

MAGs assembly
A total of 92 fecal samples from eight pig populations, four intestinal locations and different ages were used for metagenomic sequencing and construction of metagenome-assembled genomes (MAGs). Microbial DNA was extracted as described above. The libraries for metagenomic sequencing were constructed following the manufacturer's instructions (Illumina, USA), with an insert size of 350 base pairs (bp) for each sample, and 2x150 bp paired-ends sequenced on a Novaseq 6000 platform. Raw sequencing data were filtered to remove adapter sequences and low-quality reads using fastp (v0.19.41)[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

(v1.1.3) 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 with two different binning algorithms '--metabat2 --maxbin2' using the metaWRAP package (v1.1.1)[90]. The bins (metagenomic assembly genomes, 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 (v1.1.1). Metagenomic sequences were further assembled to optimize MAGs using the reassemble_bins module of metaSPAdes (v3.15.3) in the metaWRAP pipeline. CheckM (v1.0.12) 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 (v2.3.2) 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 (v1.1.1). The FASTA file of amino acid sequences translated from coding genes was used to perform KEGG annotation using Ghost KOALA tool (v2.2)[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 and 5c were generated using custom made perl (v5.10.1) and R scripts (v3.5.3). Pathway and regulon scores were computed using a custom-made perl script. Both scores included (i) 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), (ii) one point for GalNAc deacetylase activity (having an orthologue of agaA and/or nagA), (iii) one point for GalN deaminase/isomerase (having on orthologue of agaS), (iv) one point for tagatose-6-P kinase (having an orthologue of pfkA and/or lacC and/or fruK), and (v) 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 colocalisation of the genes on one and the same sequence contig. The effect of MAG-type (OTU476-like, Erysipelotrichaceae and others), -completion, -contig, -genome size on pathway and regulon scores were estimated using the R lm function, and were 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 the 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.

Determination of metabolic flux ratios from ^{13}C - experiments by gas chromatography-mass spectrometry

^{13}C-labelled GalNAc (N-acetyl-D-[UL-13C6]galactosamine) (99% isotopic purity of ^{13}C) was bought from ZZBIO CO., LTD (Shanghai, China). The 4-8-110 strain was cultured in 6 ml of GAM medium (Nissui Pharmaceuticals, Japan) adding 1g/L of ^{13}C-labelled GalNAc or regular GalNAc (as control) in an anaerobic glovebox (Electrotek, UK) in triplicate as described above. After 13 h of culture, the cultures were centrifuged for 5 min at 1500 × g. Bacterial cells were re-suspended 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 mins, and then thawed on ice. We added 0.4 ml of chloroform, vortexed for 30 seconds, centrifuged for 10 min at 17,530 g (4 °C), transferred the supernatant to new 1.5 ml tubes, and dried at -105 °C in a FreeZone Freeze Dryers (Labconco, USA). Metabolites

were derivatized for GC/MS analysis as follows. First, 70μl 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) was added and the samples were re-incubated 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 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 and then to 280 °C at 20 °C /min with a final hold at this temperature for 2 min. The GC flow rate with helium carrier gas was 50 cm/s. The GC column used was a 20 m x 0.25 mm x 0.25 mm Rxi-5ms. 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 1 kV detector. GC/MS data were analyzed to determine isotope labeling. To determine ^{13}C labeling, 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 7x7 matrix-based probabilistic methods as described[95–97] and implemented in MATLAB (Release R2021a). Labeling results are expressed as average fraction the particular compound that contains isotopic label from the particular precursor.

Gavage experiment

4-8-110 and an E. coli strain isolated from cecal content (with 4-15-1 and 4-8-110) were cultured in GAM medium (Nissui Pharmaceuticals, Japan) for 13 and 4 hours, respectively, in an anaerobic glovebox (Electrotek, UK) as described above. OD's 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 kindly provided by Huazhong Agricultural University (Wuhan, China). All experimental procedures involving mice were approved by the Ethics Committee in Huazhong Agricultural University (HZAUMO-2021-0077). Two groups (A and B) of five females each were housed in two separate cages (temperature, 25 ± 2°C; humidity, 45-60%; lighting cycle, 12 h/day; light hours, 06:30-18:30). Mice were gavaged 100μl of PBS (group A) or 100μl of PBS + 80g/L GalNAc (corresponding to 8mg/100 μl or ~200mg/kg of live weight) (group B) directly in the mouth with a sterile syringe twice per day (9:00 am and 16:00 pm) for 10 consecutive days (days 2 to 11). All mice were further gavaged with 150μl of the 4-8-110 glycerol stocks (see above) once per day (16:00 pm) for five consecutive days (days 3 to7), and with 150μl of the E. coli glycerol stocks (see above) once per day (16:00 pm) for three consecutive days (days 5 to7). Mice were sacrificed 5 days after the last bacterial inoculation (on day 12, based on[98]), and caecal content and feces collected for each mouse. Bacterial DNA was extracted using the QIAamp fast DNA stool mini kit (Qiagen, Germany). 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, UK) containing 10% hydrogen, 10% carbon dioxide and 80% nitrogen at 37 °C. 4-ml samples were pipetted after gentle shaking to measure optical density (OD) at 600nm using a UV spectrophotometer (Yoke instrument, China). 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 harvested at the end of the exponential growth phase (respectively 11 and 4 hours). Total RNA was extracted from bacterial cells using TRIzol® Reagent following the manufacturer's manuals (Thermofisher scientific, USA), and genomic DNA was removed using DNase I (TaKara, China). The quantity and quality of total RNA were evaluated using a 2100 Bioanalyser (Agilent, USA) and NanoDrop-2000 (Thermofisher scientific, USA). The RNA samples with OD260/280=1.8-2.0, OD260/230≥2.0, RIN≥6.5, 28S:18S≥1.0, ≥100ng/μl, and ≥2μg were used to construct libraries for RNA sequencing.

Construction of libraries and Sequencing. The libraries for RNA sequencing were constructed following TruSeq™ RNA sample preparation Kit (Illumina, USA) using 2μg of total RNA. In brief, ribosomal RNA (rRNA) was removed from total RNA by Ribo-Zero Magnetic kit (epicenter, USA). All mRNAs were broken into short fragments (200nt) by adding fragmentation buffer. Double-stranded cDNA was synthesized using a SuperScript double-stranded cDNA synthesis kit (Invitrogen, CA, USA) with random hexamer primers (Illumina, CA, USA). cDNA was subjected to end-repair, phosphorylation and 'A' base addition according to library construction protocol (Illumina, USA). The constructed libraries were sequenced on a Novaseq 6000 platform with 2×150bp pair-end strategy. Base-calling and quality value calculations were performed using the Illumina GA Pipeline (v1.6) (Illumina, USA).

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 raw data with a Perl script. The clean sequences were mapped to the reference genomic sequences of two OTU476 like strains (generated by ourselves as described above) and *E. coli* (accession number: 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 (v1.6.4). FPKM (Fragments per kilobase of exon model per million mapped reads) method was used to calculate the gene expression level. We used DESeq2 package[100] in R (v3.5.3) to test for differential expression.

Analyzing the effect of ABO genotype on abundance of Erysipelotrichaceae in humans

Samples. The data used correspond to the previously described CEDAR cohort[101]. It included 300 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. 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 prior to donation in agreement with the recommendations of the declaration of Helsinki for experiments involving human subjects.

Sequencing. For microbiota analysis, DNA was extracted from biopsies using the QIAamp DNA Stool Mini Kit (QIAgen, Germany). Three 16S rRNA amplicons corresponding respectively to the V1-V2, V3-V4 and V5-V6 variable regions were generated in separate PCR reactions and subjected to paired-end (2x300bp) NGS sequencing on a MiSeq instrument (Illumina, USA) following the standard protocol at the GIGA genomics core facility.

Data processing. Reads were QV20 trimmed from the 3' end, demultiplexed, primer sequences removed using the bbduk tool (BBMap – Bushnell B. – sourceforge.net/projects/bbmap/). Reads mapping to the human genome were eliminated using the BBTools suite (v38.82). The corresponding pipeline was constructed using Snakemake (v7.0.1)[102]. Further analyses were performed using QIIME 2 (2018.11)[103]. The paired-end reads were denoised and joined using the DADA2 plugin (v1.16)[104] using batch-specific trimming length parameters yielding 9.1±2.0K amplicon sequence variants (ASVs) per run for V1-V2, 4.5±1.6K for V3-V4 and 6.8±0.67K for V5-V6 amplicon. ASVs mapping to known contaminant taxa as well as ASVs with abundance negatively correlated with coverage depth were removed. Samples that 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)[105]. 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 (Shagam, 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[106].

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 (i) ABO blood group (A, B, AB or O), (ii) secretor status, (iii) sex, (iv) smoking status, (v) age and (vi) BMI. 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 sequencing data were submitted to the GSA database under accession numbers: CRA006230, CRA006239, CRA006240 and CRA006216, respectively. The genotype data was deposited at the GVM under the GSA database under accession number: GVM000310. The GWAS summary statistics is available through Figshare with https://doi.org/10.6084/m9.figshare.19313960. The whole-genome sequencing data of experimental pigs have been deposited in the GSA database under accession number: GRA006383. The source data are available in the repository: 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* (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 Bioinformatics* **19**, 68 (2018).

69. Haseman, J. K. & Elston, R. C. The investigation of linkage between a quantitative trait and a marker locus. *Behavior Genetics* **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 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. *Doctor thesis, The Pennsylvania State University* (2007).

77. Li, H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics, 18 (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 Bioinformatics* **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 Bioinformatics* **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 rartios from ¹³C-experiments and gas chromatography-mass spectrometry data: protocol and principles. *Methods in Molecular Biology* **358**, 177–197 (2007).

97. van Winden, W. A. et al. Correcting mass isotopomer 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. Köster, J. & Rahmann, S. Snakemake: a scalable bioinformatics workflow engine. *Bioinformatics* **28**, 2520–2522 (2012).

103. 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).

104. Callahan, B. et al. DADA2: High-resolution sample inference from Illumina amplicon data. *Nat Methods* **13**, 581–583 (2016).

105. Ghodsi, M., Liu, B. & Pop, M. DNACLUST: accurate and efficient clustering of phylogenetic marker genes. *BMC Bioinformatics* **12**, 271 (2011).

106. 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).

107. Srivastava, A. et al. Genomes of the mouse collaborative cross. *Genetics* **206**, 537–556 (2017).

108. Yu, N., Zhao, Z., Fu, Y.-X., Sambuughin, N., Ramsay, M., Jenkins, T., Leskinen, E., Patthy, L., Jorde, L. B., Kuromori, T. & Li, W.-H. Global patterns of human DNA sequence variation in a 10-kb region on chromosome 1. *Mol. Biol. Evol.* **18**, 214–222 (2001).

109. The 1000 Genomes project Consortium. A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061–1073 (2010).

110. 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).

111. Charlier, C., Li, W., Harland, C., Littlejohn, M., Coppieters, W., Creagh, F., Davis, S., Druet, T., Faux, P., Guillaume, F., Karim, L., Keehan, M., Kadri, N. K., Tamma, N., Spelman, R. & Georges, M. NGS-based reverse genetic screen for common embryonic lethal mutations compromising fertility in livestock. *Genome Res* **26**, 133–1341 (2016).

112. Georges, M., Charlier, C. & Hayes, B. Harnessing genomic information for livestock improvement. *Nat Rev Genet* **20**, 135–156 (2019).

113. 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).

114. Huang, H. et al. Fine-mapping inflammatory bowel disease loci to single-variant resolution. *Nature* **547**, 173–178 (2017).

115. 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).

116. 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).

117. 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).

118. Robinson, M. G., Tolchin, D. & Halpern, C. Enteric bacterial agents and the ABO blood groups. *Am J Hum Genet* **23**, 135–145 (1971).

119. 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).

120. 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* **71**, 182 (1977).

121. 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).

122. Chaudhuri, A. & De, S. Cholera and blood groups. *Lancet* **ii**, 404 (1977).

123. Boren, T. et al. Attachment of Helicobacter pylori to human gastric epithelium mediated by blood group antigens. *Science* **262**, 1892–1895 (1993).

124. Lindesmith, L. et al. Human susceptibility and resistance to Norwalk virus infection. *Nat Med* **9**, 548–553 (2003).

125. Galili, U. Evolution of α1,3Galactosyltransferase and the α-Gal epitope. In: Galili U, Avila JL (eds) α-Gal and anti-Gal. *Subcellular Biochemistry* 32, 1-23 Springer, Boston, MA (1999).

126. Prather, R. S., Shen, M. & Dai, Y. Genetically modified pigs for medicine and agriculture. *Biotechnology and Genetic Engineering Reviews* **25**, 245–266 (2008).

127. Kelly, R. J., Rouquier, S., Giorgi, D., Lennon, G. G. & Lowe, S. B. Sequence and expression of a candidate for the human Secretor blood group α(1,2) fucosyltransferase gene (*FUT2*). *J Biol Chem* **270**, 4640–4649 (1995).

128. Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., Peplies, J. & Glöckner, F. O. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res* **41**, D590–D596 (2013).

129. DeSantis, T. Z., Hugenholtz, P., Larsen, N., Rojas, M., Brodie, E. L., Keller, K., Huber, T., Dalevi, D., Hu, P. & Andersen, G. L. Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Appl Environ Microbiol* **72**, 5069–5072 (2006).

Author contributions HY analyzed the 16S rRNA sequence data, performed GWAS, meta-analyses and local association analyses, computed heritabilities of individual taxa, contributed to ABO genotyping and analyzed the effect of the 2.3 Kb deletion on taxa abundance. JW analyzed 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 cecal lumen, analyzed the GalNAc import and utilization pathway in the MAGs, and contributed to ABO genotyping. XH participated in 16S rRNA sequencing (F6) and GWAS (F6). YZ performed metagenome sequencing analysis, analyzed the GalNAc import and utilization pathway in MAGs, analyzed the RNA seq data from cecum samples, and contributed to ABO genotyping. YZ 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 F6 and F7 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. ML assisted with the isolation of the OTU476-like strains, the GalNAc feeding experiments, and genotyping of the ABO gene. QL assisted with measuring the concentrations of GalNAc in cecal lumen. SK, MH, HF, SF, XX, HJ, SC and JG assisted with the experiments. ZZ, ZW, HG and YH assisted with the preparation of genotype data from whole-genome sequence data and conducted the analysis of the Nanopore data of the ABO region. JM assisted with the construction of the mosaic population. HA assisted with the bioinformatic analysis of the ABO region, de novo assembly of the A allele, and evolutionary analysis of the ABO alleles. LS analyzed the effect of ABO genotype on intestinal microbiota composition in humans. WC assisted in the analysis of the sequence data for the trans-species polymorphisms. CaCh supervised the characterization of the ABO gene and the 2.3 Kb deletion and the corresponding haplotype structure in the F0, F6 and F7 population and for the trans-species polymorphism. BY 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 F6 and F7 generation and the definition of expected mapping resolution. MG supervised the bioinformatic and statistical analyses, performed bioinformatic and statistical analyses, and wrote the paper. CC codesigned the study, supervised experiments, supervised bioinformatic and statistical analyses of gut microbiome and wrote the paper. LH 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.
Reprints and permissions information is available at http://www.nature.com/reprints.
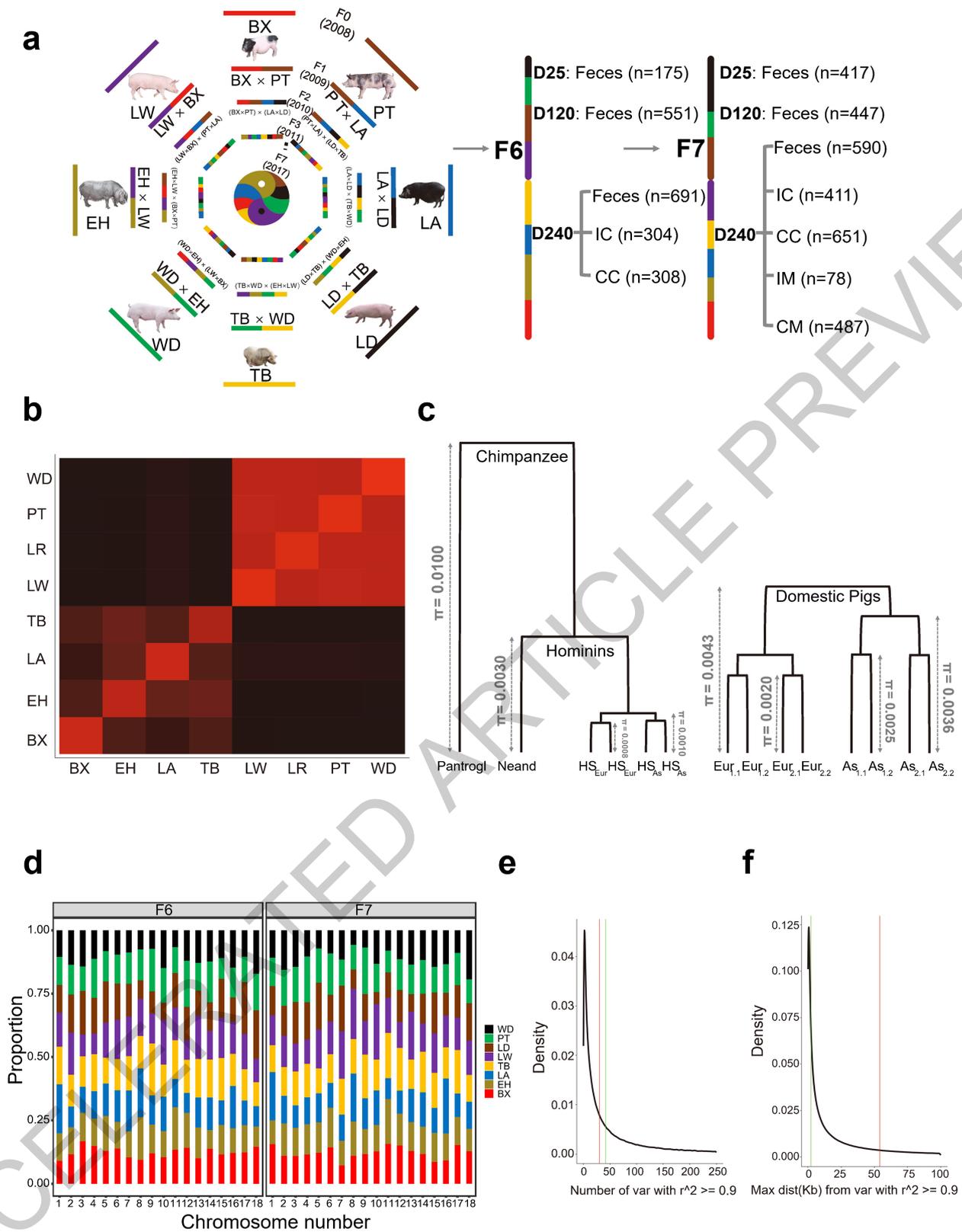
Extended Data Fig. 1 | See next page for caption.

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 feces (D25, D120, D240), luminal content of the ileum (IC) and cecum (CC), and mucosal scrapings in the ileum (IM) and cecum (CM). BX: Bamaxiang, EH: Erhualian, LA: Laiwu, TB: Tibetan, LW: Large White, LD: Landrace, PT: Piétrain, WD: White Duroc. (b) Average similarity ($1 - \pi$) between allelic sequences sampled within and between the eight founder breeds. The color 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[107]. (c) Comparison of the average nucleotide diversity (π, 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[108,109]. 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[110–112]. 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.[108]). 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*[113], 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[113]. (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[114]. 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).

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 feces and cecum 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 color-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[115], and human data from Fig. 6 in Vuik et al[116]. 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 *Streptocaccaceae*.
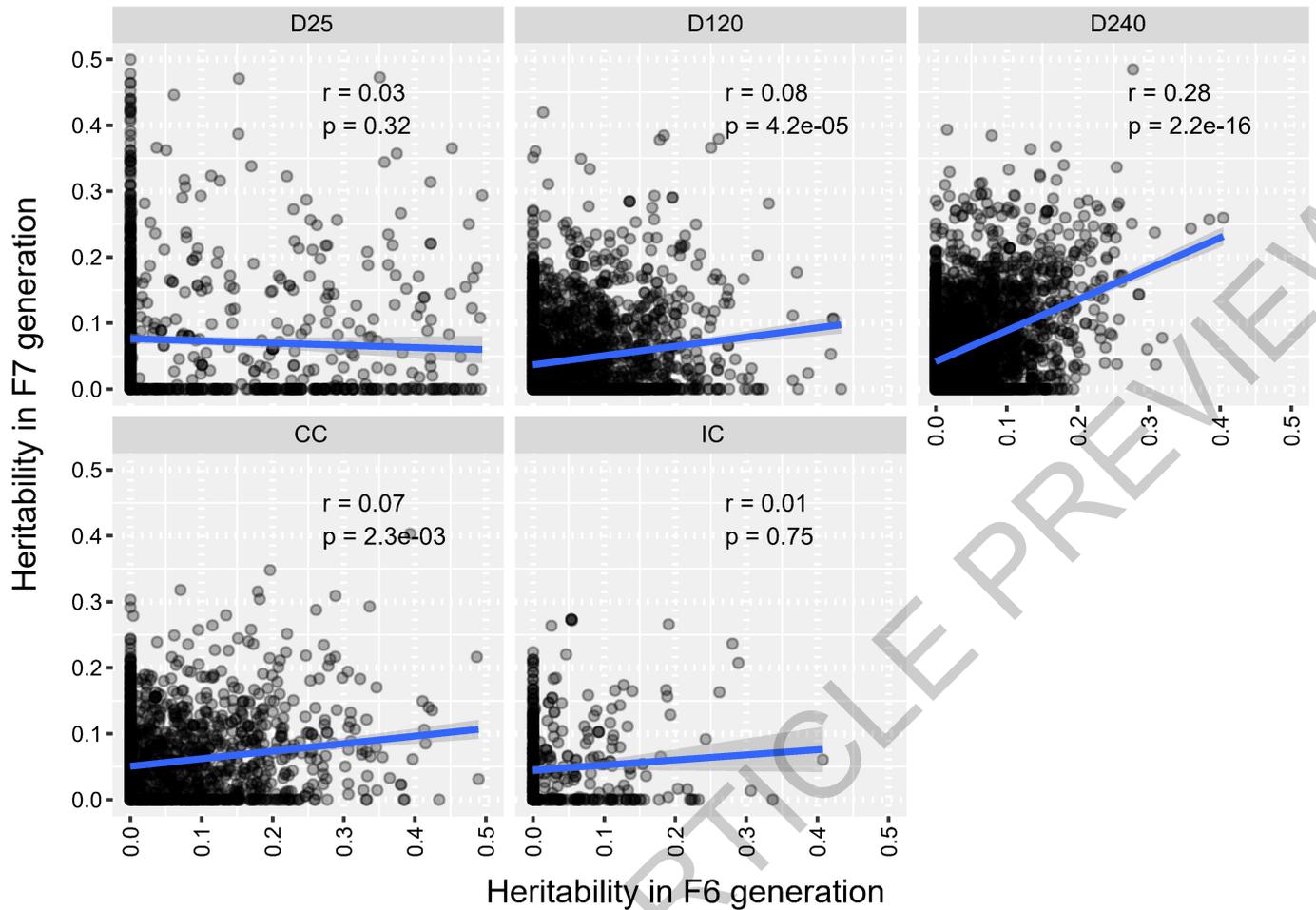
Extended Data Fig. 3 | Evaluating the heritability of intestinal microbiota composition in the mosaic pig population. Correlation between heritability estimates of taxa/OTUs in F6 and F7 generation by sample type (D25, D120, D240, CC and IC). Correlation coefficients (r) and associated p-values (p) were 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).
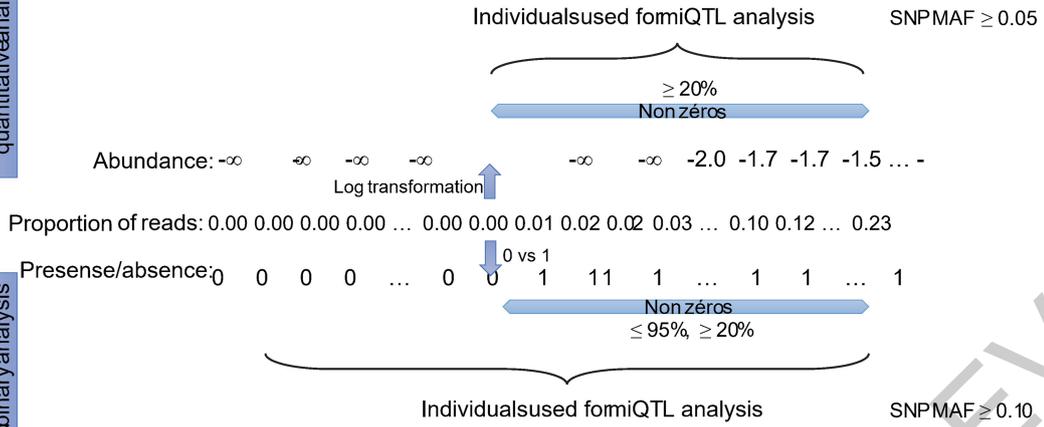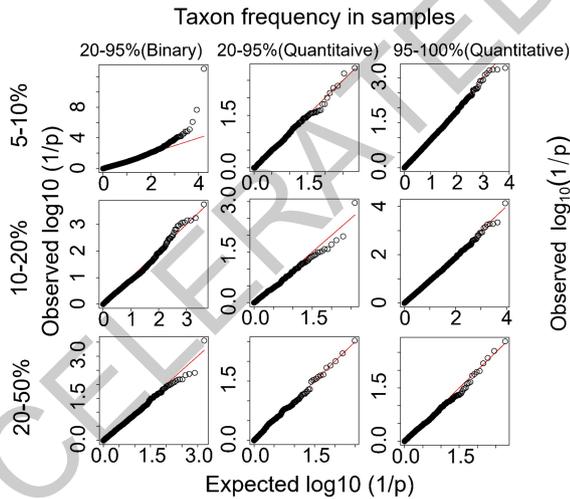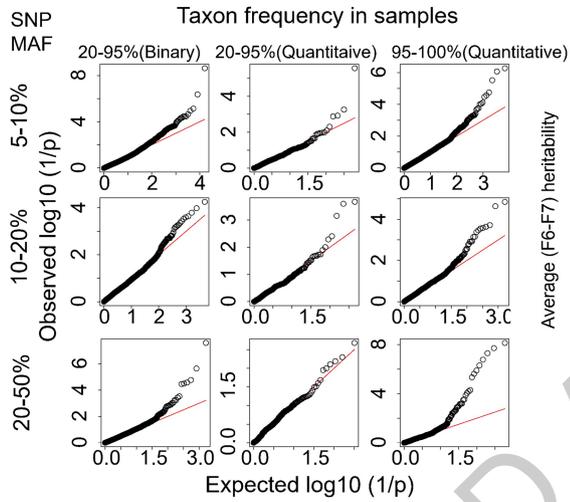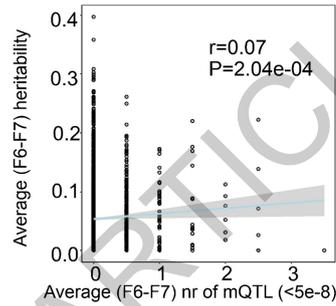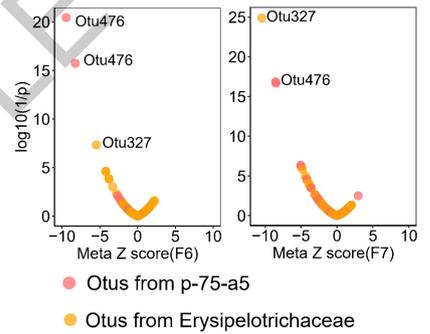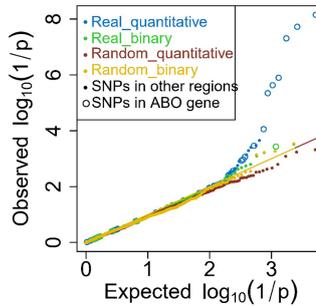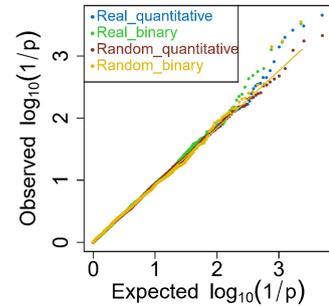
Extended Data Fig. 4 | See next page for caption.

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 x 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 fecal 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 2.3Kb deletion ($p = 1.9 \times 10^{-43}$) 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.
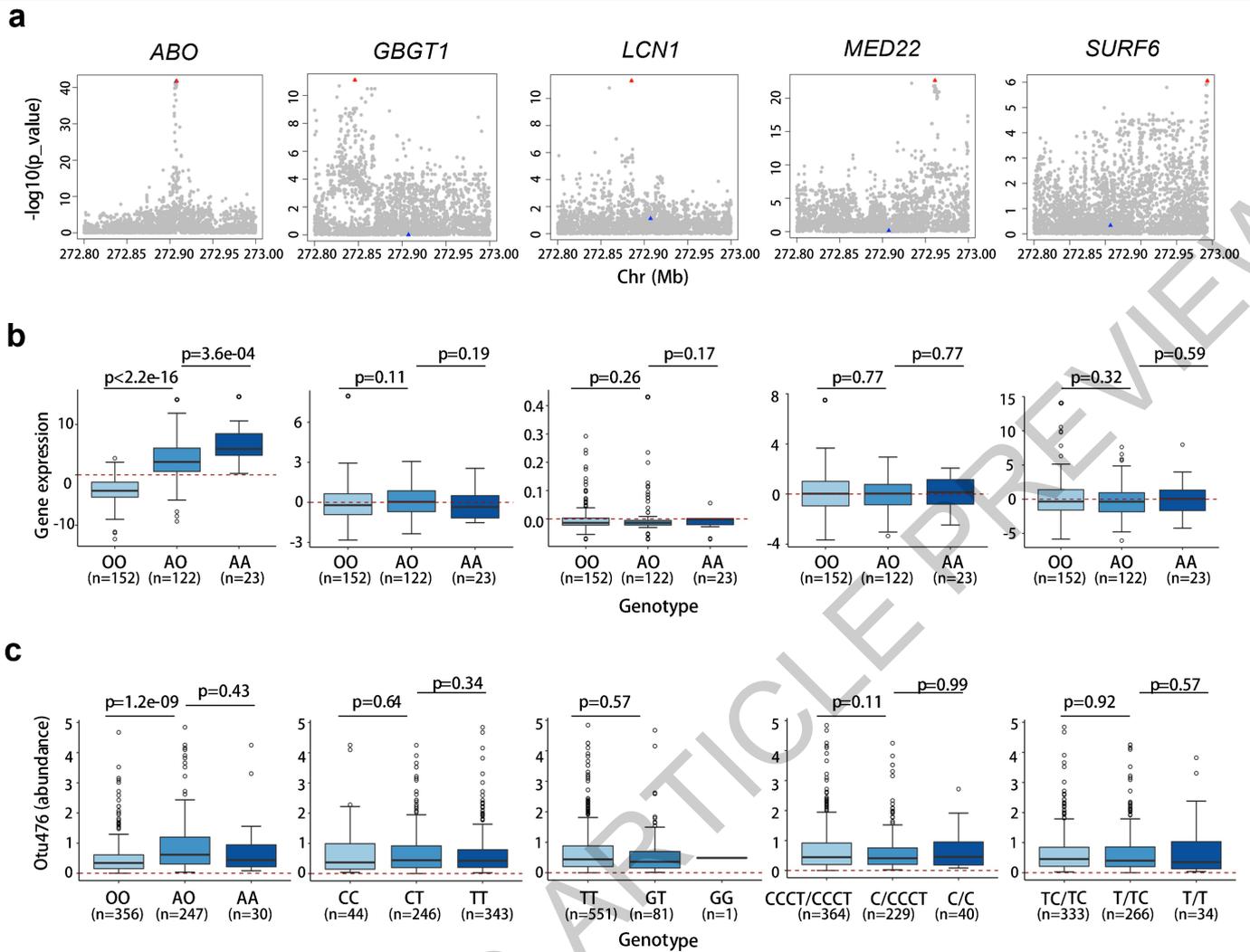
Extended Data Fig. 6 | cis-eQTL analyses in the vicinity of the chromosome 1 miQTLK supports the causality of the 2.3 Kb deletion. (a) Cis-eQTL analysis for the porcine N-acetyl-galactosaminyl transferase ("ABO"), *GBTG1*, *LCN1* (=*OBP2B*), *MED22* and *SURF6* genes in caecum. The blue triangle corresponds 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.
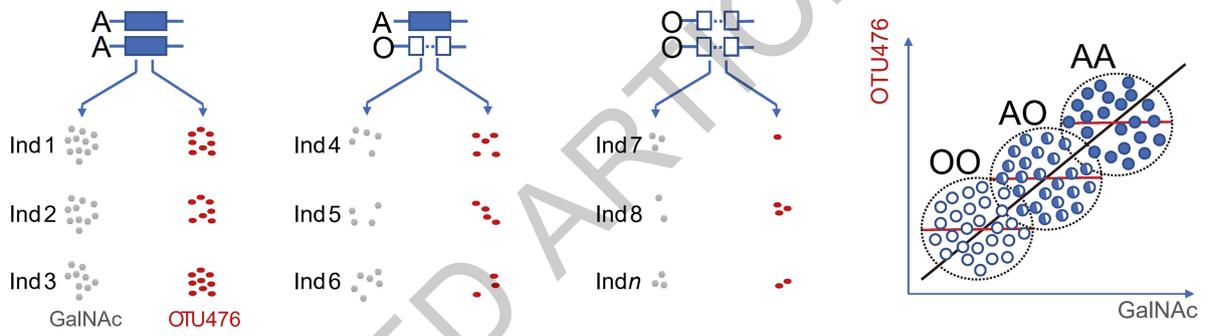
Extended Data Fig. 7 | See next page for caption.

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) centered 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 centered 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 (centered on the position of the 2.3Kb 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 ($r^2 ≥ 0.9$) 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[117], *E. Coli* and *Salmonella* enteric infection[118], SARS-CoV-1[42], SARS-CoV-2[43] and schistosomiasis[119–121], while being a possible risk factor for cholera[122], *H. pylori*[123] and norovirus infection[124]. 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 behavior. 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).
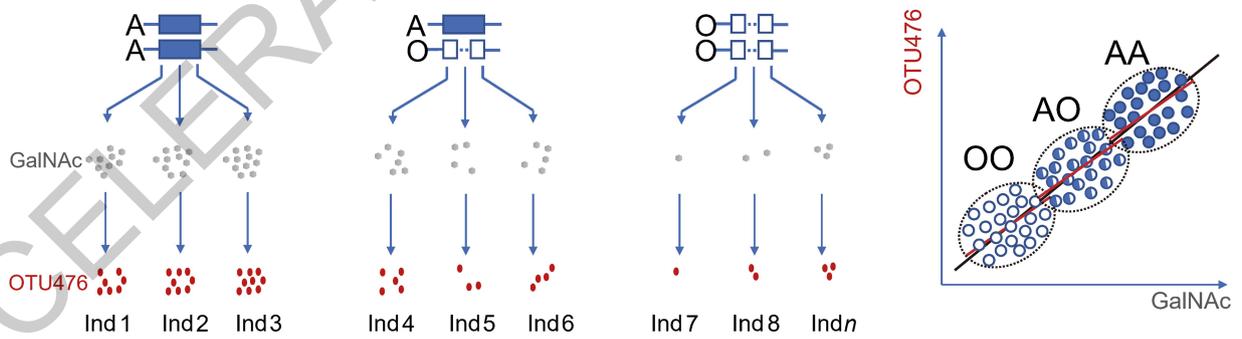
a
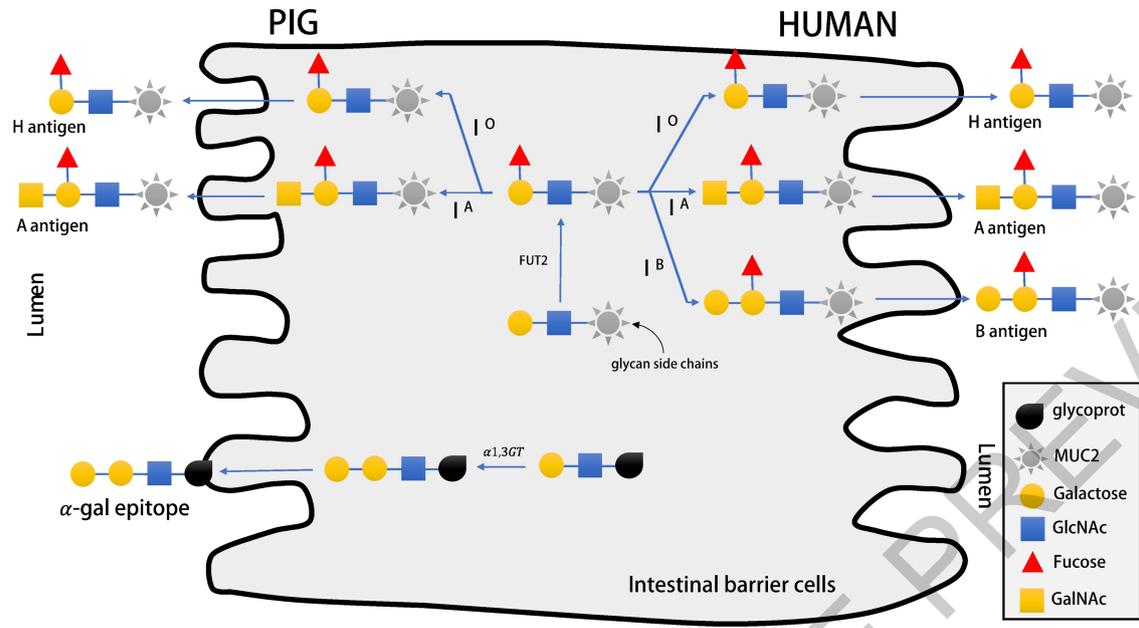
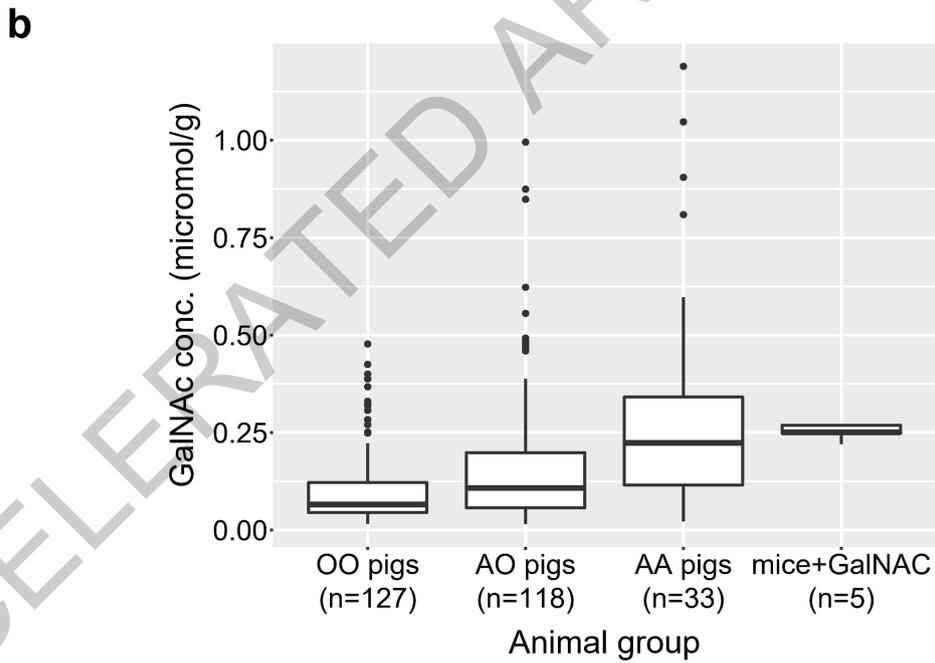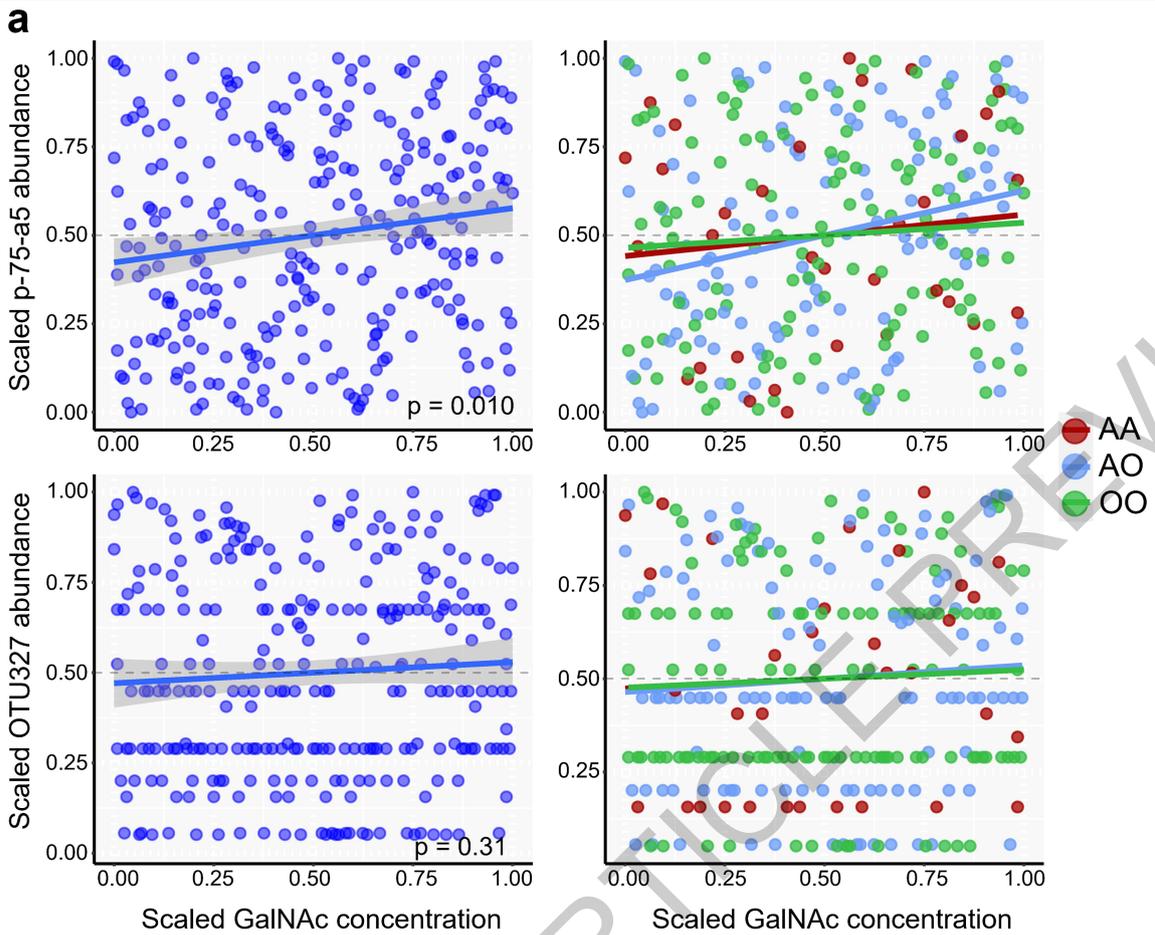

b1

b2

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 (Gal*α*1-3(Fuc*α*1-2)Gal*β*1-4GlcNAc-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" (Gal*α*1-3Gal*β*1-4GlcNAc-R), which results from the addition of a galactose to the Gal*β*1-4GlcNAc-R precursor by a *α*1,3galactosyltransferase 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[125,126–129].
(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.
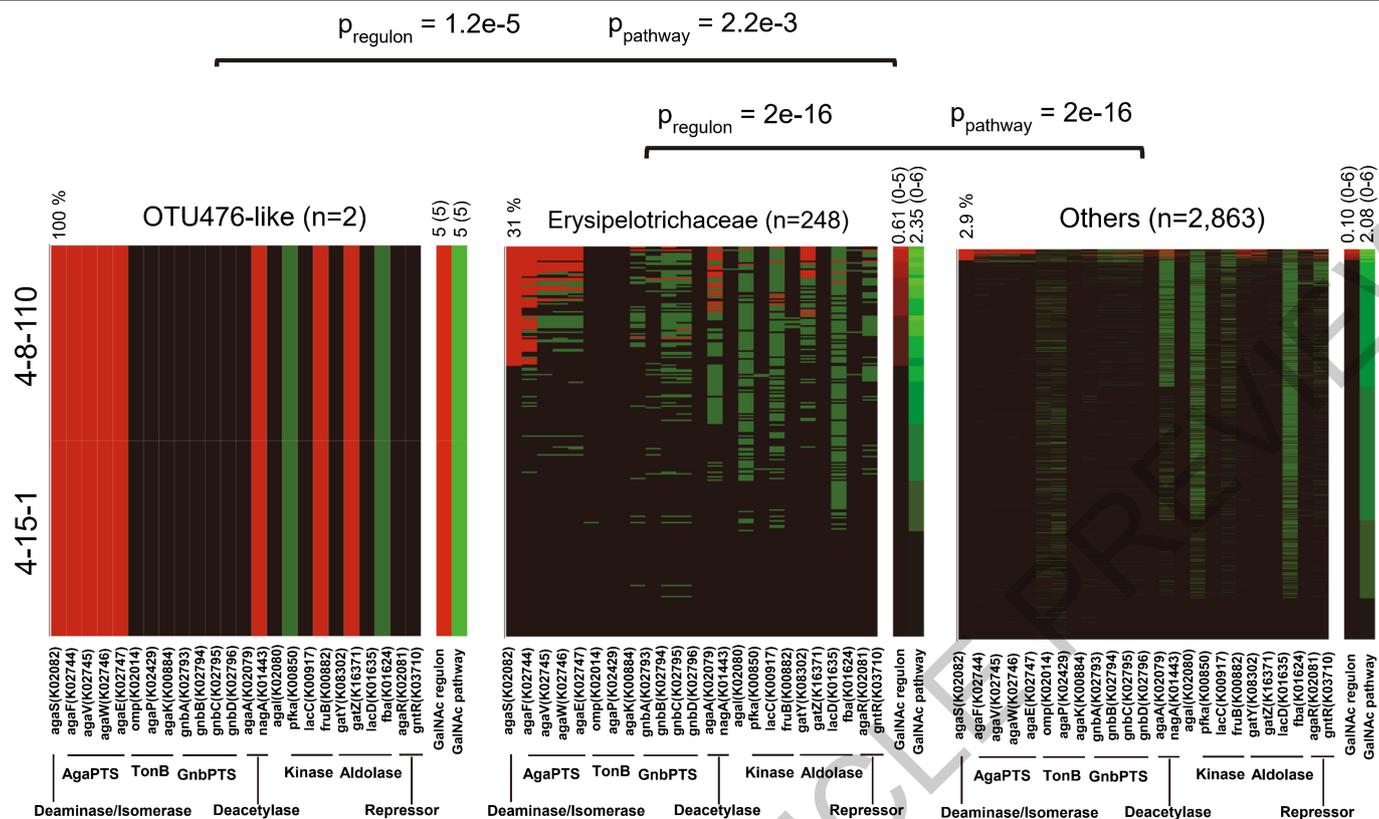
Extended Data Fig. 10 | The chromosome 1 miQTL affects bacteria with a functional GalNAc import and catabolic pathway. Presence anywhere in the genome (green), presence in close proximity to *agaS* (red), or absence (black) of the orthologues of 24 genes implicated in the GalNAc TR/CP pathway in the 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.
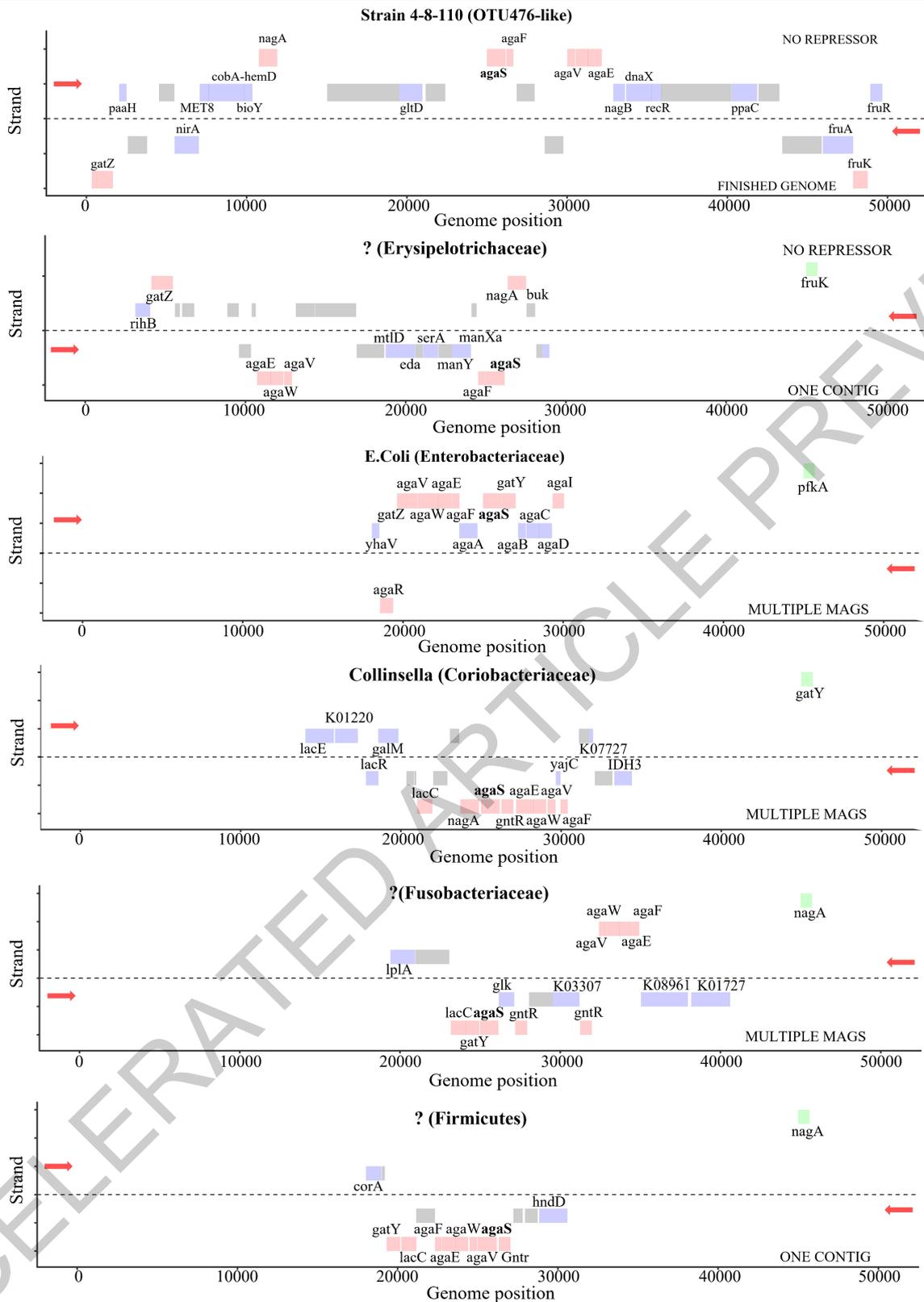
Extended Data Fig. 11 | Different GalNAc operon structure and transcriptome response in miQTL-sensitive versus -insensitive GalNAc utilizing bacteria. Maps of GalNAc "operons" in one of the two OTU476-like strains (NB: The organization of the GalNAc gene cluster was identical in both 4-15-1 and 4-8-110 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 colored 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).
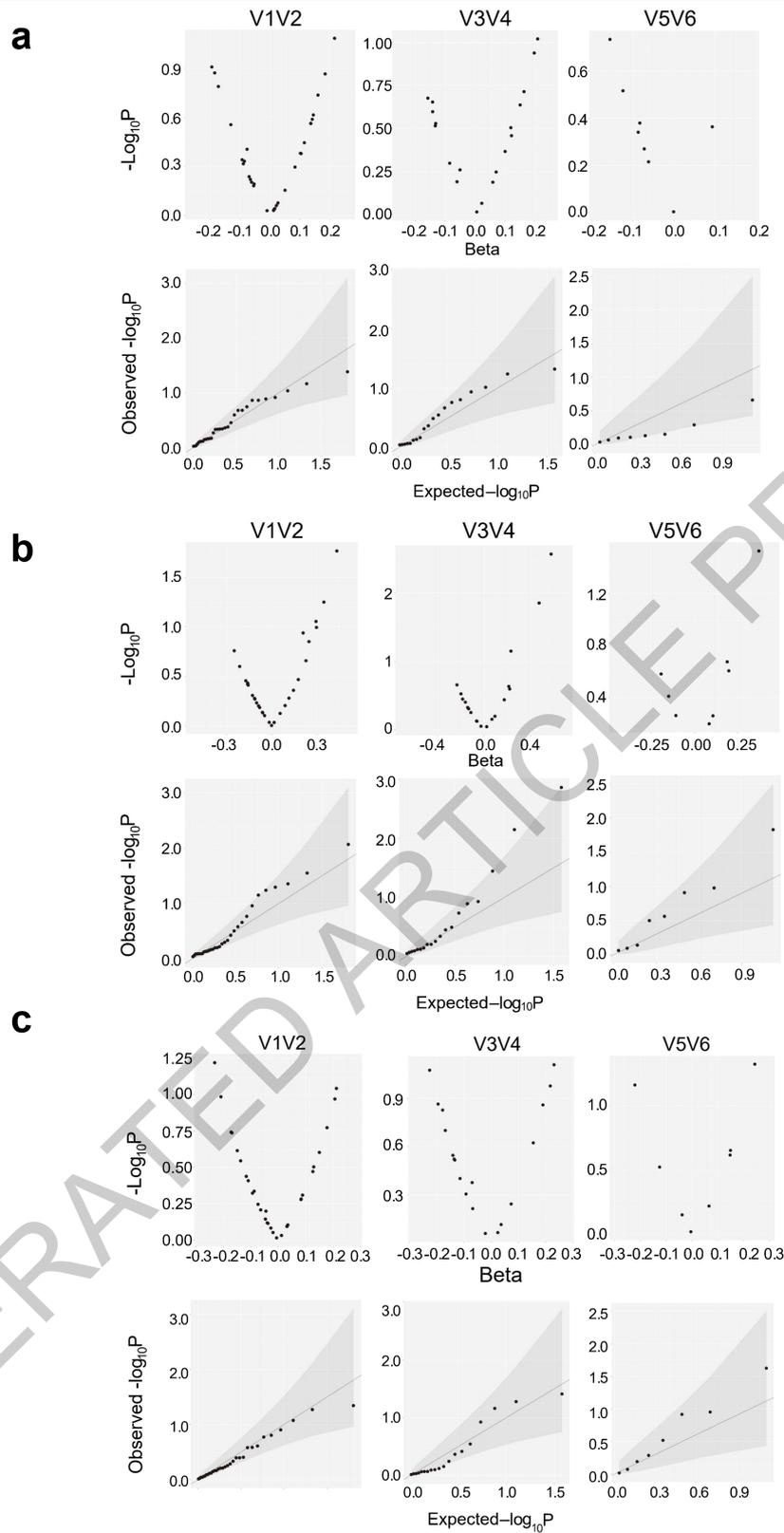
Extended Data Fig. 12 | No effect of ABO genotype on intestinal Erysipelotrichaceae abundance in human. Volcano and QQ plots for 43 (V1-V2), 20 (V3-V4) and 9 (V5-V6) OTUs classified as *Erysipelotrichaceae* for the contrasts (a) [AA, AO and AB] versus [BB, BO and OO], (b) [BB, BO and AB] versus [AA, 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.

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 (v3.5.3); 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.