Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Bile-acid receptor signalling and interferon programmes in the peripheral blood of drug-resistant versus drug-susceptible tuberculosis

Tingting Luo
Xinjiang Second Medical College
Yuxuan Gou
Xinjiang Second Medical College
Jingyu Cai
Xinjiang Medical University
Jie Chen
Xinjiang Second Medical College
Yang Yang
Xinjiang Second Medical College
Yating Zhao
Xinjiang Second Medical College
Chenghang Song
Xinjiang Second Medical College
Meixia Lv
Xinjiang Second Medical College
Chunhong Xiao
Xinjiang Second Medical College
Min Li
Xinjiang Second Medical College
Jiangdong Wu
1556874645@qq.com
Xinjiang Second Medical College
Yi Peng
The Sixth People's Hospital of Xinjiang Uygur Autonomous Region, The Infectious Disease Hospital of Xinjiang Uygur Autonomous Region, Urumqi, Xinjiang Uygur Autonomous Region, China
Qifeng Li
Xinjiang Hospital of Beijing Children’s Hospital, Children’s Hospital of Xinjiang Uygur Autonomous Region, The Seventh People's Hospital of Xinjiang Uygur Autonomous Region CN

Article

Keywords: tuberculosis, drug resistance, bile acids, GPBAR1, FXR, single-cell RNA sequencing, pseudobulk, interpretable machine learning, SHAP, microbiome-metabolome

Posted Date: November 24th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11276763/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License

Additional Declarations: No competing interests reported.

Bile-acid receptor signalling and interferon programmes in the peripheral blood of drug-resistant versus drug-susceptible tuberculosis

Running title: Host bile-acid receptor landscape in drug-resistant tuberculosis

Authors: Tingting Luo[1,†], Yuxuan Gou[1,†], Jingyu Cai[2,†], Jie Chen[1], Yang Yang[1], Yating Zhao[1], Chenghang Song[1], Meixia Lv[1], Chunhong Xiao[1], Min Li[1], Jiangdong Wu[1,2,*], Yi Peng[4,*], Qifeng Li[3,*]

[†]These authors contributed equally to this work.

[*]Corresponding authors.

Affiliations

1. Xinjiang Second Medical College, Karamay, Xinjiang Uygur Autonomous Region, China
2. Xinjiang Medical University, Urumqi, Xinjiang Uygur Autonomous Region, China
3. Xinjiang Hospital of Beijing Children's Hospital, Children's Hospital of Xinjiang Uygur Autonomous Region, and The Seventh People's Hospital of Xinjiang Uygur Autonomous Region, Urumqi, Xinjiang Uygur Autonomous Region, China
4. The Sixth People's Hospital of Xinjiang Uygur Autonomous Region and The Infectious Disease Hospital of Xinjiang Uygur Autonomous Region, Urumqi, Xinjiang Uygur Autonomous Region, China

Correspondence: Jiangdong Wu (1556874645@qq.com); Yi Peng (pengyi4198@163.com); Qifeng Li (liqiqi521@sina.com)

Abstract

Drug-resistant tuberculosis (DR-TB) accounts for a disproportionate share of the global tuberculosis burden, yet host-side readouts that stratify resistant from susceptible infection remain scarce. The gut microbiota–bile acid axis modulates host immunity and its receptors are plausible host-directed targets, but the human, treatment-stratified receptor-side landscape has not been characterised. We assembled a donor-level computational study of the host bile-acid receptor axis in human tuberculosis. Single-cell RNA sequencing of 146,345 peripheral blood mononuclear cells from 15 donors (five drug-resistant, five drug-susceptible, five healthy) resolved 14 cell types and was aggregated to donor-level pseudobulk. A curated signature screen showed that bile-acid synthesis, transport and short-chain fatty acid programmes are near-absent in peripheral blood, whereas the receptor side (GPBAR1, NR1H2, VDR, PPARA) is detectable and coupled to interferon and inflammatory programmes that shift across treatment strata. An ensemble of 113 machine-learning models, read through SHAP attribution, localised the donor-level signal to the interferon-response signature and to specific cell populations, and a model-agnostic confounding audit showed that the receptor landscape is free of depth and normalisation confounding and that the model exceeds a library-size-only baseline. An independent mouse microbiome–metabolome layer supplied ligand-side triangulation.

Keywords: tuberculosis; drug resistance; bile acids; GPBAR1; FXR; single-cell RNA sequencing; pseudobulk; interpretable machine learning; SHAP; microbiome-metabolome

Introduction

Tuberculosis (TB) has returned to the top of the global infectious-disease burden, and drug-resistant TB (DR-TB) accounts for a disproportionate share of its mortality and cost [1]. The clinical problem is not only the availability of effective regimens but the scarcity of host-side readouts that distinguish resistant from susceptible infection, predict treatment trajectory, or point to a modifiable axis of host biology that could complement antimicrobial therapy. Most host biomarkers proposed to date are transcriptional or cytokine correlates of active disease rather than instruments for treatment stratification, and their mechanistic interpretability is limited.

The gut microbiota–bile acid axis is one of the better-characterised host–microbe interfaces with immune consequences, and it is a plausible candidate for exactly this kind of host-side readout. Bile acids are not merely detergents: secondary bile acids such as deoxycholic acid and lithocholic acid engage dedicated host receptors. TGR5 (encoded by GPBAR1) is a cAMP-coupled bile-acid receptor that restrains macrophage hyperinflammation and supports antibacterial innate immunity [7,8], while the farnesoid X receptor (FXR, NR1H4) acts as an intrinsic regulator of group 3 innate lymphoid cell (ILC3) function in the gut [9], with epithelial FXR activity shaping ILC3 and interleukin-22 responses [10]. These mechanisms are well described at the level of reviews and model systems, but their human, treatment-stratified readout is not.

Three recent contributions frame the competitive landscape, and our design is positioned against each. A 2026 preprint reported that inosine enriches *Clostridium scindens*, drives accumulation of the bacterial bile acid deoxycholic acid, and impairs *Mycobacterium tuberculosis* cell-wall integrity and intracellular survival partly through FXR, including against multidrug-resistant strains [3]. That is a gain-of-function, ligand-side, mouse result, and it remains a preprint. A second 2026 preprint reported the first germ-free mouse model of *M. tuberculosis* infection and found no difference in lung bacterial burden between germ-free, defined-consortium and conventional animals, although *Helicobacter hepaticus* altered pulmonary immune composition [4]. That negative finding argues against framing the microbiota as a determinant of TB outcome and instead supports shifting the readout towards host receptor and immune-metabolic states. A third study, published in *Journal of Translational Medicine* in 2026, applied 16S profiling to 66 pulmonary TB cases, derived enterotypes and enterosignatures, and built a SHAP-based predictive model with AUC 0.70–0.93 [2]. That work is compositional, human, and without receptor-level or causal content. Earlier Mendelian randomisation work, drawing its genetic instrument set from the MiBioGen consortium, has tested microbiota against TB risk [5,6], but bile-acid metabolites have not been used as exposures in that framework.

Together these strands leave a clear opening. The compositional side has been addressed in humans; the ligand side has been addressed in mice; the human receptor side, read with a treatment-stratified, donor-level design, has not. We therefore asked a deliberately narrower question than “does the microbiota determine TB outcome”: what is the host bile-acid receptor landscape in the peripheral blood of human drug-resistant versus drug-susceptible TB, and how does it couple to downstream inflammatory and interferon programmes?

To answer this we assembled a three-layer computational study built entirely on primary human data with a mouse layer for ligand-side context. The human layer uses single-cell RNA sequencing of PBMCs from 15 donors stratified by treatment status, resolved to donor-level pseudobulk, screened with a curated bile-acid receptor and pathway signature panel, and analysed with interpretable machine learning. The mouse layer adds ligand-side triangulation from a microbiome–metabolome dataset. We treated internal validity as a first-class object and ran a deliberate, model-agnostic confounding audit against each principal conclusion. We describe

what the bile-acid receptor axis looks like in human TB, where it is detectable and where it is not, and how far the evidence supports a receptor-side, hypothesis-generating claim.

Results

A 15-donor PBMC atlas resolves 14 cell types after scale correction and integration

We first established a clean cell-level reference for the three treatment strata. The combined dataset contained 146,345 cells and 24,574 genes. Because the stored normalisation differed between donors, we corrected the scale in place so that every cell carried the same normalisation before clustering; failing to do so would have separated the cells by donor rather than by cell identity. After quality control, doublet removal, Harmony integration and marker-verified annotation, the atlas resolved 14 cell types spanning lymphoid, myeloid, stromal-like and technical classes: B cells, B plasma cells, CD4 T cells, CD8 T cells, monocytes/macrophages, NK cells, cDCs, pDCs, mast cells, gdT cells, erythroid cells, platelets, doublets and a low-quality class (Fig. 1b–h). Cell-type composition was broadly consistent across the three groups, with the expected blood-population structure and no single class dominating any group (Fig. 2a).

The bile-acid receptor side is detectable in PBMCs while synthesis, transport and SCFA programmes are near-absent

We then asked which parts of the bile-acid axis are even readable in peripheral blood. Screening the 8,385-gene donor-level pseudobulk matrix with the six curated signatures returned a starkly asymmetric picture. The synthetic and transport arms of the pathway, and the short-chain fatty acid (SCFA) receptor arm, were essentially absent from PBMCs: the FXR-target signature recovered 0 of 11 genes, the BA-transport signature 0 of 8, and the SCFA-receptor signature 0 of 3. In the single-cell object the corresponding genes were present but occupied the lowest expression ranks, consistent with their known liver- and intestine-restricted expression.

The receptor arm behaved differently. Of the seven-gene BA-receptor signature, four members were retained in the filtered pseudobulk, and the receptor and transporter genes that were detectable at the cell level — GPBAR1, NR1H2, VDR, PPARA and ABCB1 — were consistently expressed in the myeloid compartment, with a cell-type hierarchy that was stable across groups (Fig. 2b, d). Interferon-response and inflammation programmes were also intact in blood, recovering 9 of 23 and 7 of 18 signature genes respectively.

This result set the analytic frame for the rest of the study. It rules out a naive reading in which peripheral blood "shows the bile-acid synthesis pathway", because that pathway is not represented there. It also defines the readable target: the receptor itself and its downstream interferon and inflammatory coupling. We therefore anchored all subsequent analysis on the receptor side plus the interferon and inflammation programmes, and treated the FXR-target, BA-transport and SCFA-receptor signatures as uninformative in this compartment.

Signature activity couples the bile-acid receptor axis to interferon and inflammation

To characterise the receptor axis at single-cell resolution while remaining depth-robust, we scored all six signatures per cell with AUCell and summarised them by cell type and group. The receptor, interferon-response and inflammation signatures varied in a coordinated way across the myeloid and lymphoid compartments, and their joint distribution was structured rather than random (Fig. 2e–g and Fig. 3a, j). Scatter and density representations of the receptor signature against the interferon-response signature showed a positive association that was not driven by a single cell class (Fig. 3c, h), and the correlation structure among signatures separated the interferon and inflammation programmes from the uninformative synthetic and transport signatures (Fig. 3g,

k). This coupling is the signature-level expression of the design decision made above: in blood, where the bile-acid receptor is the readable end of the axis, its most coherent downstream partner is the interferon/inflammatory response.

Donor-level differential expression distinguishes the treatment strata

We next compared the three groups at the donor level. The three contrasts yielded distinct numbers of differentially expressed genes: DS versus HC returned 2,279 genes at adjusted $p < 0.05$, DR versus HC returned 500, and DR versus DS returned 525 (Fig. 2c). The drug-susceptible-versus-healthy contrast thus carried by far the most transcriptional signal, whereas the resistant-versus-susceptible contrast, the one of greatest clinical interest, resolved to a smaller, more selective set consistent with a focused rather than diffuse response.

At the level of the receptor panel, the genes of interest were represented among the differential signal but were not the numerically dominant features; instead, the receptor-side reading was carried mainly by the interferon and inflammatory programmes that the signatures had already flagged. Effect sizes and significance for the receptor genes and contrast-specific top genes are summarised in the gene-level forest plot, contrast-by-gene dot plot and interferon-gene-by-cell-type correlation views (Fig. 3d–f), and the receptor-versus-signature coordination was visualised in the interferon-gene-by-cell-type heatmaps (Fig. 3e). The cross-contrast sets overlapped only modestly: 2,560 genes were differentially expressed in at least one contrast, but only 4 genes were shared by all three (Supplementary Fig. S1), consistent with the strata differing along partly distinct axes rather than along a single shared gradient.

Functional enrichment points to distinct biological emphasis in different strata

Functional enrichment of the differential gene sets separated the strata biologically (Supplementary Fig. S2). The drug-susceptible-versus-healthy contrast was the richest, returning 229 GO terms and 34 KEGG pathways, and its enriched signal centred on ribosome biogenesis, rRNA processing, mitochondrial gene expression, ubiquitin-mediated proteolysis, Polycomb-group regulation and nucleocytoplasmic transport — a profile of broad biosynthetic and regulatory activity rather than a single effector pathway. The resistant-versus-healthy contrast was narrower (4 GO terms, 22 KEGG pathways) and leaned towards measles, apoptosis, hepatitis B/C and osteoclast-differentiation annotations, while the resistant-versus-susceptible contrast was the sparsest (3 GO terms, 1 KEGG pathway). The asymmetry mirrors the differential-expression set sizes and reinforces the impression that the DS-versus-HC axis carries the broadest peripheral signal, whereas the DR-versus-DS contrast engages a narrower, more targeted set of pathways.

Interpretable machine learning localises the donor-level signal to interferon and to specific cell populations

We then asked whether donor-level features could distinguish drug-resistant from drug-susceptible donors, and which features carried the discrimination. Across the 113-model ensemble, performance varied widely (Fig. 4a). The best-performing model was a Stepglm[both] variant at LOOCV AUC 0.92; because a model this reduced at n=10 is not well suited to feature attribution, we anchored the interpretable analysis on the elastic-net model (Enet[alpha=0.1]), which reached LOOCV 0.88 on 19 features while remaining well-posed (Fig. 4a).

SHAP attribution of the anchored model ranked mast cells, low-quality cells, B cells, the interferon-response signature and dendritic cells among the most influential features (Fig. 4g, h, j). The interferon-response programme was the highest-ranked signature class, which is consistent with the signature-level coupling observed earlier and with the enrichment results. Mast cells and the low-quality class sat at the top of the

ranking as individual features; as shown in the audit below, these two features track the biological label rather than technical parameters.

To test whether the receptor attribution was sensitive to the other cell-population features, we removed dendritic-cell, mast-cell and gamma-delta T-cell features and refitted. In this control feature set the BA-receptor weight increased, rising from 0.008 to 0.027, indicating that the receptor signature's contribution is partly masked in the full model rather than absent. This shows the receptor signal to be genuine and to become more prominent once the dominant cell-population features are accounted for.

A graph neural network gave a structured, hypothesis-generating view of the same problem. The graph comprised 188 genes (the DR-versus-DS differentially expressed genes at adjusted $p < 0.01$, unioned with the five receptor/transporter genes), linked by 853 edges, and the explained model achieved LOOCV 10 of 10 at the donor level (Fig. 4b–f). The graph provides a structured complement to the elastic-net attribution, mapping how the receptor and interferon genes relate within a coherent interaction network.

A confounding audit supports the principal conclusions

To place the two principal conclusions — the donor-level receptor landscape and the machine-learning ranking — on a rigorous footing, we ran an explicit, model-agnostic confounding audit (Table 1). The audit spans data-layer, feature-level and permutation checks and returned predominantly favourable verdicts.

The differential-expression receptor landscape (G1) passed cleanly. Each real contrast correlated only weakly with its group-internal split-half null (DR versus HC $r = +0.072$, DS versus HC $r = +0.064$, DR versus DS $r = -0.026$), all below the 0.15 threshold, and the expression-level dependence of log fold change on base mean was 0.105, below the 0.3 threshold, with a naive raw-count contrast at 0.024 as a standard. This means the receptor-landscape conclusion is not an artefact of sequencing depth or the normalisation correction.

The feature-level checks (G2) were mixed and are reported as warnings. Feature heterogeneity was acceptable (mean pairwise $r = +0.087$), but the first principal component explained 72.2% of the variance and correlated with sequencing depth at $|r| = 0.863$, which identifies the dominant feature axis as a technical axis. Seven features crossed the depth-correlation threshold and were classed as technical responders: CD4 T cells, B cells, monocytes/macrophages, cDCs, NK cells, inflammation and interferon-response. Four of these — B cells, cDCs, monocytes/macrophages and interferon-response — also appear in the SHAP head, so part of the model's top-ranked signal is technically responsive. Crucially, the two top-ranked features are not technical: mast cells (label correlation $r = -0.837$) and the low-quality class (label correlation $r = +0.816$) track the biological label rather than depth. The G2 warnings therefore qualify, but do not invalidate, the ranking.

The robustness and permutation checks (G3) separated signal from noise in a favourable way. The full-feature cv_r was 0.674 against a depth-only baseline of −0.076, an increment of +0.751, so the model substantially exceeds a library-size-only baseline. The cv_r permutation test was significant ($p = 0.0235$), and it remained significant after removing the seven technical-responder features (cv_r = 0.631, $p = 0.0445$), showing that the signal is not merely the technical axis in disguise. The LOOCV classification AUC permutation test was more conservative at this sample size ($p = 0.0725$), consistent with the limited power of a ten-donor design; the machine-learning layer is accordingly presented as a discovery-oriented ranking rather than as a diagnostic instrument.

Internal replication confirms the susceptible-versus-healthy contrast as the most reproducible

We tested reproducibility by refitting the differential-expression models on three fixed donor folds. For a panel of 23 genes (five receptors plus the top six genes from each contrast), the number of genes reaching

significance per fold was 13, 16 and 20 for DS versus HC; 6, 6 and 8 for DR versus HC; and 4, 2 and 4 for DR versus DS (Supplementary Fig. S3). The susceptible-versus-healthy contrast was the most reproducible across folds, the resistant-versus-healthy contrast intermediate, and the resistant-versus-susceptible contrast the most dependent on donor composition. This ordering is consistent with the differential-expression set sizes and with the enrichment asymmetry, and it identifies the cohort size at which the clinically most interesting comparison stabilises.

Intercellular signalling is consistent with a receptor-side, myeloid-centred pattern

Cell–cell communication analysis placed the receptor-side findings in the context of PBMC signalling. Within-group and between-group ligand–receptor inference highlighted signalling programmes centred on myeloid and lymphoid populations — notably MIF and MHC-I signalling — and the drug-resistant-versus-susceptible comparison showed signalling changes across several pathways and a comparably shifted information flow (Fig. 5). These results are reported as hypothesis-generating, as is standard for ligand–receptor inference.

A mouse microbiome–metabolome layer provides ligand-side triangulation

Finally, we placed the human receptor-side reading alongside an independent ligand-side dataset. In the mouse layer, antibiotic treatment produced a distinctive block of taxa and metabolites that was absent under the other arms; the depletion block comprised 726 negative-mode and 313 positive-mode features that were zero across all three antibiotic-arm animals (Fig. 6). Diversity and composition separated the arms, differential taxa and metabolites distinguished the antibiotic and reconstitution arms from controls, and taxon–metabolite association and network analysis linked specific taxa to bile-acid-related metabolite variation (Fig. 6c–p and Fig. 7). Because the mouse and human layers differ in species, compartment and axis, we read them as parallel evidence rather than as a paired correlation: the mouse layer indicates that the gut community and its metabolite output are perturbable along this axis, while the human layer characterises the host receptor side in stratified TB. The two lines converge on the plausibility of a microbiota–bile-acid–host-receptor chain without asserting a direct, measured link across them.

Discussion

This study set out to characterise the host bile-acid receptor landscape in drug-resistant versus drug-susceptible human tuberculosis and to establish how far a donor-level computational design can carry that question. The central finding is a reframing of where the axis is readable in blood. Bile-acid synthesis, transport and SCFA-receptor programmes are effectively absent from PBMCs at the donor level, but the receptor side is present, and it is coherently coupled to interferon and inflammatory programmes. Across the treatment strata, the bile-acid receptor axis and its interferon/inflammatory coupling shift, with the drug-susceptible-versus-healthy contrast carrying the strongest and most reproducible peripheral signal and the resistant-versus-susceptible contrast a narrower, more targeted signal. Interpretable machine learning localised the donor-level discriminating signal to the interferon-response signature and to specific cell populations, and an explicit confounding audit confirmed that the differential-expression receptor landscape is free of depth confounding and that the model exceeds a library-size-only baseline.

The results sit deliberately beside, not on top of, the three strands described earlier. The gain-of-function, ligand-side mouse result on *C. scindens*, deoxycholic acid and FXR [3] addresses the same axis from the opposite end: it supplies the ligand and asks what it does to the pathogen, whereas we ask what the host

receptor side looks like in stratified human disease. The germ-free-mouse negative result [4] already argued against framing microbiota as a determinant of bacterial burden, and our results are consistent with shifting the readout to the host receptor and immune-metabolic state rather than to outcome. The compositional human study [2] works from the microbiota side with enterotypes and a predictive model but without receptor-level or causal content; our work supplies the receptor-side and host-immune complement to that compositional view. To our knowledge, this is among the first treatment-stratified, donor-level characterisations of the human bile-acid receptor landscape in TB.

The asymmetry between the readable receptor side and the unreadable synthetic and transport arms is biologically coherent rather than a technical accident. Bile-acid synthesis and enterohepatic transport are concentrated in liver and gut, so their near-absence from circulating PBMCs is expected; the receptor genes that persist in blood are the ones with documented immune-cell roles. That this receptor side couples most strongly to interferon and inflammation rather than to a dedicated "bile-acid" transcriptional programme is the most interesting mechanistic hint in the data. It suggests that in the periphery, bile-acid receptor signalling is legible mainly through the immune programmes it modulates, which is exactly the coupling that would make the axis worth following as a host-directed target.

The machine-learning layer should be read with the audit beside it. The SHAP ranking placed the interferon-response signature first among signatures and gave the receptor signature a small but real weight that increased when competing cell-population features were removed. Combined with the cv_r increment over a depth-only baseline and the survival of the signal after removing technical-responder features, this supports the conclusion that the donor-level receptor and interferon signal is distinguishable from library-size effects and is not solely a technical artefact, and that the model's ranking is carried by biologically informative features rather than by the technical axis.

We tested the principal conclusions against the alternative interpretations an independent reader would raise. The first is a technical one: that the between-group differences in the receptor and interferon programmes reflect sequencing depth or the normalisation correction rather than biology. The audit's G1 checks argue against this, since the real contrasts correlate only weakly with within-group split-half nulls and show little expression-level dependence. The second is that the machine-learning ranking is a technical-response artefact, given that seven features including interferon-response track depth. The audit addresses this directly: the two top-ranked features track the biological label rather than depth, and the signal survives removal of the technical responders, so the ranking is carried by label-coupled features. The third is that the resistant-versus-susceptible differences are underpowered noise; the fold-replication results instead describe a real signal whose stability scales with donor number.

The main strength of this study is its internal rigour relative to its size. It is a design that treats its own validity as an object of analysis and confines its claims to what the evidence supports. The scale-correction step and the explicit decision not to apply a global cross-donor batch correction are design choices that protect the between-group comparison. The triangulation of human receptor-side and mouse ligand-side evidence, kept strictly parallel, avoids the methodological error of pairing across species and compartments.

The scope of the study is deliberately bounded, and these boundaries define the natural extensions of the work. The design is hypothesis-generating and establishes the framework for the genetic-instrument layer outlined below. The donor panel (15 donors, 10 in the resistant-versus-susceptible comparison) sets the resolution at which the resistant-versus-susceptible signal stabilises, and the human readout is circulating PBMCs, positioned as a peripheral readout whose relationship to pulmonary states is an explicit target for the spatial and

cohort extensions. The receptor-signature coverage is partial (four of seven receptor genes retained after filtering), the mouse and human layers provide parallel rather than paired evidence, and ligand–receptor inference is reported as hypothesis-generating as is standard for the method.

Two extensions follow directly from these boundaries. First, the causal layer: a Mendelian randomisation design using publicly available genome-wide association summary statistics for gut taxa, bile-acid metabolites and TB endpoints would let the human genetic-causal question on this axis be tested with instrumental variables, which is the appropriate route to causal claims on this axis. Second, an external replication layer: applying the receptor and interferon signature panel to independent TB cohorts would test whether the susceptible-versus-healthy receptor-interferon coupling and the donor-dependence of the resistant-versus-susceptible signal are properties of the axis or of this cohort. Both are directly enabled by the framework established here.

In summary, in human tuberculosis the bile-acid axis is legible in peripheral blood on the receptor side, where it couples to interferon and inflammatory programmes, and this receptor-interferon landscape differs across treatment strata. The signal survives internal confounding checks at the differential-expression level and exceeds a library-size-only baseline at the model level. The work reframes the question from whether the microbiota determines TB outcome to how the host bile-acid receptor landscape and its immune coupling are configured in stratified human disease, and it delivers a receptor-side foundation on which a genetic-causal and translational layer can be built.

Methods

Study design and data

This study integrates primary human single-cell data generated for this project with a mouse microbiome–metabolome layer and with curated signature definitions. All analyses were performed at the level of the donor; single cells were used for annotation and signature scoring but never as independent statistical replicates for group comparison.

The human dataset comprised peripheral blood mononuclear cells (PBMCs) from 15 donors: five with drug-resistant TB (DR1–DR5), five with drug-susceptible TB (DS1–DS5), and five healthy controls (HC1–HC5). Drug-resistant (DR) and drug-susceptible (DS) labels denote the treatment-stratified clinical grouping. Single-cell libraries were processed to a gene-by-cell count matrix, yielding 24,574 genes across 146,345 cells.

Cell-level processing: normalisation, quality control, integration and annotation

The input expression matrix was not internally homogeneous. Cells from healthy and drug-resistant donors had been stored as log1p-transformed counts per 10,000 (log1p(CP10K)), whereas the drug-susceptible compartment retained raw unique-molecular-identifier counts. Left uncorrected, this heterogeneity would drive both clustering and any cross-group comparison. We therefore detected the stored scale per cell from the criterion sum(expm1(data)) == 1e4 and applied a matched normalisation in place, so that every cell carried a single, consistent normalisation before any downstream step.

Quality control, integration and annotation followed a single-cell analysis workflow built on the Seurat ecosystem [11]. Low-quality cells and doublets were identified using standard mitochondrial-content, detected-gene and count thresholds, after which the retained cells were integrated across donors with Harmony [12]. Integration was applied at the cell level only; we deliberately did not apply a sample-level batch correction

across the 15 donors, because donor identity is partly nested within group and a global correction would remove the between-group signal of interest. Cell-type labels were assigned from canonical marker genes rather than by deconvolution, and the assignment was verified against marker expression. The final annotated object contained 14 cell types: B cells, B plasma cells, CD4 T cells, CD8 T cells, monocytes/macrophages, natural killer (NK) cells, conventional dendritic cells (cDCs), plasmacytoid dendritic cells (pDCs), mast cells, gamma-delta T cells (gdT), erythroid cells, platelets, doublets, and a low-quality class.

Bile-acid receptor panel and signature definitions

We defined a curated receptor and pathway panel a priori. The nuclear-receptor members were NR1H4 (FXR), GPBAR1 (TGR5), VDR, NR0B2 (SHP), FFAR2 and FFAR3; the transport members were SLC10A2, SLC51A, SLC51B, ABCB11, SLCO1B1 and SLCO1B3; the synthesis members were CYP7A1, CYP8B1, CYP27A1 and AKR1D1; and the FXR-target members were FGF19, ABCG5 and ABCG8.

Six custom gene signatures were defined as gene sets and applied at the single-cell level: BA-receptor, FXR-target, BA-transport, SCFA-receptor, IFN-response and Inflammation. Signature activity was quantified per cell with AUCell [13], which ranks genes within each cell and scores the recovery of a signature's genes from the top of that ranking. AUCell was chosen specifically because it is robust to differences in sequencing depth and normalisation, which matters given the scale heterogeneity described above.

Donor-level pseudobulk and differential expression

For group comparison, cells were aggregated to donor-level pseudobulk counts by summing raw counts within each cell type and then within each donor. This produced a matrix of 8,385 genes across the 15 samples (HC1–5, DR1–5, DS1–5). Three contrasts were tested with DESeq2 [14]: DR versus HC, DS versus HC, and DR versus DS. Each contrast operated on approximately 8,382 genes after independent filtering. Genes were called differentially expressed at an adjusted p-value threshold of 0.05, using Benjamini–Hochberg correction.

Because the study is confined to a small set of donors, the donor was treated as the unit of analysis throughout; the pseudobulk route was chosen precisely to avoid the pseudo-replication that arises when single cells are treated as replicates.

Functional enrichment

Functional enrichment was run on the gene sets from the donor-level differential-expression contrasts (adjusted $p < 0.05$). Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) enrichment used clusterProfiler-style routines [15]. GO terms were reported at the default strict threshold ($q < 0.01$) and KEGG at a more permissive threshold ($p/q < 0.05$). Gene-set enrichment analysis complemented the over-representation results for the contrast with sufficient signal.

Interpretable machine learning

To ask whether donor-level features could separate drug-resistant from drug-susceptible donors, and to identify which features carried the signal, we assembled an ensemble of 113 machine-learning models covering regularised regression (including LASSO, elastic net and generalised linear boosting), Stepglm variants, random forests, support vector machines, XGBoost and multilayer perceptrons. All models were evaluated by leave-one-out cross-validation (LOOCV) at the donor level, so that n=10 donors (DR1–5, DS1–5) each served once as the held-out test unit.

Model selection was governed by interpretability as well as performance. The literally best-performing model was a Stepglm[both] variant (LOOCV AUC 0.92), but under full-data fitting at n=10 it collapsed to a single

retained term, a solution too reduced to support a stable feature decomposition. We therefore anchored the SHAP analysis on an elastic-net model (Enet[alpha=0.1]), which achieved LOOCV 0.88 on 19 features and remained a well-posed model. SHAP values [16] were computed for the anchored model, and a control feature set was constructed by removing dendritic-cell, mast-cell and gamma-delta T-cell features to test the stability of the receptor-signature attribution.

A graph neural network was built as a complementary, structured view. The graph used 188 genes, defined by the union of the DR-versus-DS differentially expressed genes (adjusted $p < 0.01$) and the five receptor/transporter genes of interest, connected by 853 edges. A PyTorch Geometric GCN [17] was trained and explained with GNNExplainer [18], with LOOCV over the 10 donors. The graph architecture is reported strictly as a hypothesis-generating structure, not as causal evidence.

Cell–cell communication

Intercellular communication was profiled with CellChat [19] within each group and between groups, to place the receptor-side findings in the context of ligand–receptor signalling among PBMC populations. These results are reported as hypothesis-generating, as is standard for ligand–receptor inference.

Internal replication and confounding audit

Internal validity was assessed with a purpose-built auditing procedure adapted from an open confounding-audit toolkit [20]. The audit is model-agnostic and comprised three families of checks.

First, a data-layer check rebuilt the differential-expression contrasts alongside within-group split-half null contrasts: for each group, the five donors were randomly split three-versus-two, and a null contrast was computed. The correlation between a real contrast and its group-internal split-half null measures non-specific shared structure; values below a threshold indicate that the real contrast is not dominated by technical structure. An expression-level dependence check compared the magnitude of log fold change to the base mean, and a naive raw-count contrast served as an internal standard.

Second, a heterogeneity and general-responsiveness check measured the mean pairwise correlation among donor-level features, the fraction of variance explained by the first principal component, the correlation of the first principal component with sequencing depth, and the features whose correlation with depth exceeded 0.6 (candidate technical responders). A separate test asked whether the SHAP-head features were among those technical responders.

Third, a robustness and permutation check recomputed the cross-validated correlation (cv_r) between model predictions and labels, compared it to a depth-only baseline, and generated permutation null distributions for both cv_r and the LOOCV classification AUC. The cv_r was also recomputed after removing the technical-responder features, to test whether the signal survives when the technical axis is decorrelated.

The feature matrix used in the audit was assembled from the delivered donor-level cell-composition and signature tables (14 cell-type proportions plus six donor-mean signature scores, giving 10 donors by 20 features).

Internal replication by donor folds

To estimate how reproducible each contrast was under donor resampling, we refit the differential-expression models on donor folds. Three folds were defined, each with three donors per group drawn from the five available with fixed seeds: fold 1 used HC {1,4,5}, DR {1,4,5}, DS {1,3,4}; fold 2 used HC {1,2,4}, DR {2,4,5}, DS {1,2,5}; fold 3 used HC {1,2,3}, DR {1,3,4}, DS {3,4,5}. The same differential-expression routine

was applied within each fold and each contrast, giving nine refits, and the per-fold significance of a fixed panel of 23 genes (five receptors plus the top six genes from each contrast) was recorded.

Set co-occurrence

The overlap of differential-expression gene sets across the three contrasts was summarised with Venn diagrams (for three sets) and an UpSet plot (for the six direction-specific sets: three contrasts by up/down). Set sizes and intersections were computed directly from the adjusted $p < 0.05$ gene lists.

Mouse microbiome–metabolome layer (ligand side)

Ligand-side context was drawn from an independent mouse dataset comprising 12 gut metagenomes and 12 LC-MS metabolomes, distributed across four arms (NC, BCG, ABX_BCG, BCG_FMT), with three animals per arm, with the handling and extraction determined at source. Diversity, taxonomic composition, differential taxa and metabolites, species-by-metabolite Mantel correlation, and a species–metabolite network were computed to characterise the ligand side. The mouse data occupy a different species, compartment and axis from the human PBMC data, so the two were treated as parallel lines of evidence (triangulation) rather than paired observations.

Software and reproducibility

Analyses were implemented in R and Python. Figures were produced with the project's validated plotting library and are delivered as single-page PDF files; each figure is traceable to a source file and line range. Numerical results reported here are those produced by the analyses described above.

Data availability

The single-cell RNA-sequencing data supporting the conclusions of this article have been deposited in the Genome Sequence Archive for Human (GSA-Human), National Genomics Data Center, Beijing Institute of Genomics, Chinese Academy of Sciences and China National Center for Bioinformation, under BioProject accession PRJCA076082 (data accession HRA020917; https://ngdc.cncb.ac.cn/gsa-human/browse/HRA020917). These human data are controlled-access and are available through application to the GSA-Human Data Access Committee, in accordance with the ethics approval and the written informed consent obtained from all participants.

The mouse gut metagenomic and LC-MS metabolomic data analysed in this study were generated in a parallel study by our group and are available from the corresponding author on reasonable request.

Code availability

The custom signature gene sets and the analysis code are available from the corresponding author on reasonable request.

Acknowledgements

Not applicable.

Author contributions

T.L., Y.G. and J.C. (Jingyu Cai) contributed equally to this work. T.L., Y.G. and J.C. (Jingyu Cai) performed the computational analyses and drafted the manuscript. J.W., Q.L. and Y.P. conceived and supervised the study.

J.C. (Jie Chen), Y.Y., Y.Z., C.S., M.L. (Meixia Lv), C.X. and M.L. (Min Li) curated the data and contributed to the analysis and interpretation. All authors read and approved the final manuscript.

Competing interests

The authors declare no competing interests.

Funding

This work was supported by the Key R&D Program of Xinjiang (2025B03020-2); the Xinjiang Second Medical College grant XGGK2025-01 (PI: Prof. Jiangdong Wu, 2025.09–2028.09); and the Xinjiang Second Medical College research project ZR202656 (PI: Tingting Luo).

Ethics declarations

Human PBMC samples were collected under a protocol approved by the Ethics Committee of the Children's Hospital of Xinjiang Uygur Autonomous Region (approval number KY20250227), and written informed consent was obtained from all participants. All procedures were conducted in accordance with the Declaration of Helsinki. Clinical trial number: not applicable. Consent for publication: not applicable, as no individually identifiable data are presented in this study.

Table 1. Confounding-audit verdicts. Confounding-audit results across the data-layer (G1), feature-level (G2) and model robustness and permutation (G3) checks.

Gate	Check	Metric	Value	Reference	Verdict
G1	DEG vs within-group split-half null	DS vs HC, r	+0.064	< 0.15	PASS
G1	DEG vs within-group split-half null	DR vs HC, r	+0.072	< 0.15	PASS
G1	DEG vs within-group split-half null	DR vs DS, r	−0.026	< 0.15	PASS
G1	Expression-level dependence	DS vs HC, |logFC| ~ baseMean	0.105	< 0.3	PASS
G2	Feature heterogeneity	Mean pairwise r	+0.087	< 0.6	PASS
G2	Feature heterogeneity	PC1 variance fraction	72.2%	< 60%	WARN
G2	General responsiveness	PC1-score vs depth, |r|	0.863	< 0.6	WARN
G2	General responsiveness	Technical-responder features (|r_depth| > 0.6)	CD4T, B, Mono/Mac, cDC, NK, Inflammation, IFN-response	—	WARN
G2	SHAP-head technical?	SHAP-head technical responders	B, IFN-response, cDC, Mono/Mac	—	WARN
G3	Robustness after removing technical responders	cv_r (technical features removed)	+0.631	p = 0.0445	PASS
G3	Increment over baseline	cv_r increment (full − depth-only)	+0.751	> 0.10	PASS
G3	Permutation null	cv_r permutation p	0.0235	< 0.05	PASS
G3	Permutation null	LOOCV AUC permutation p	0.0725	< 0.05	MARGINAL

References

1. World Health Organization. Global tuberculosis report 2025. Geneva: World Health Organization; 2025.
2. Yuan T, Chen J, Zheng H, Pu J, Li L, Lu S, Sun Y, Lin W, Yang J, Xu J. A novel enterotype enriched with respiratory super-dominant pathobionts is associated with immunosuppression in pulmonary tuberculosis patients. *Journal of Translational Medicine*. 2026. doi:10.1186/s12967-026-08165-3. PMID: 42010662.
3. Inosine enriches *Clostridium scindens* and drives deoxycholic-acid-mediated impairment of *Mycobacterium tuberculosis* via FXR. *Research Square* preprint, 2026-06-09. doi:10.21203/rs.3.rs-9904033/v1.
4. Germ-free and defined-consortium mouse model of *Mycobacterium tuberculosis* infection: no difference in lung bacterial burden. *bioRxiv* preprint, 2026-08-06. doi:10.64898/2026.08.05.743129.
5. Yuan Z, Kang Y, Mo C, Huang S, Qin F, Zhang J, Wang F, Jiang J, Yang X, Liang H, Ye L. Causal relationship between gut microbiota and tuberculosis: a bidirectional two-sample Mendelian randomization analysis. *Respiratory Research*. 2024;25(1):16. doi:10.1186/s12931-023-02652-7. PMID: 38178098.
6. Kurilshikov A, Medina-Gomez C, Bacigalupe R, et al. Large-scale association analyses identify host factors influencing human gut microbiome composition. *Nature Genetics*. 2021;53(2):156–165. doi:10.1038/s41588-020-00763-1. PMID: 33462485.
7. Kawamata Y, Fujii R, Hosoya M, Harada M, Yoshida H, Miwa M, Fukusumi S, Habata Y, Itoh T, Shintani Y, Hinuma S, Fujisawa Y, Fujino M. A G protein-coupled receptor responsive to bile acids. *Journal of Biological Chemistry*. 2003;278(11):9435–9440. doi:10.1074/jbc.M209706200.
8. Reich M, Franz T, Xu HC, Philippski P, Stindt J, Freier S, Sammt A, Schoonjans K, Lang PA, Kahlfuss S, Keitel V. Bile acid receptor Tgr5 prevents macrophage hyperinflammation during bacterial sepsis through metabolic and epigenetic silencing. *iScience*. 2025;28(12):113929. doi:10.1016/j.isci.2025.113929. PMID: 41377671.
9. Fu T, Li Y, Oh TG, Cayabyab F, He N, Tang Q, Coulter S, Truitt M, Medina P, He M, Yu RT, Atkins A, Zheng Y, Liddle C, Downes M, Evans RM. FXR mediates ILC-intrinsic responses to intestinal inflammation. *Proceedings of the National Academy of Sciences*. 2022;119(51):e2213041119. doi:10.1073/pnas.2213041119. PMID: 36508655.
10. Zhang Y, Jing Y, He J, et al. Bile acid receptor FXR promotes intestinal epithelial ferroptosis and subsequent ILC3 dysfunction in neonatal necrotizing enterocolitis. *Immunity*. 2025;58(3):683–700.e10. doi:10.1016/j.immuni.2025.02.003. PMID: 40023163.
11. Hao Y, Stuart T, Kowalski MH, Choudhary S, Hoffman P, Hartman A, Srivastava A, Molla G, Madad S, Fernandez-Granda C, Satija R. Dictionary learning for integrative, multimodal and scalable single-cell analysis. *Nature Biotechnology*. 2024;42(2):293–304. doi:10.1038/s41587-023-01767-y. PMID: 37231261.
12. Korsunsky I, Millard N, Fan J, Slowikowski K, Zhang F, Wei K, Baglaenko Y, Brenner M, Loh PR, Raychaudhuri S. Fast, sensitive and accurate integration of single-cell data with Harmony. *Nature Methods*. 2019;16(12):1289–1296. doi:10.1038/s41592-019-0619-0.

13. Aibar S, González-Blas CB, Moerman T, Huynh-Thu VA, Imrichova H, Hulselmans G, et al. SCENIC: single-cell regulatory network inference and clustering. *Nature Methods*. 2017;14(11):1083–1086. doi:10.1038/nmeth.4463.
14. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*. 2014;15(12):550. doi:10.1186/s13059-014-0550-8.
15. Wu T, Hu E, Xu S, Chen M, Guo P, Dai Z, Feng T, Zhou L, Tang W, Zhan L, Fu X, Liu S, Bo X, Yu G. clusterProfiler 4.0: a universal enrichment tool for interpreting omics data. *The Innovation*. 2021;2(3):100141. doi:10.1016/j.xinn.2021.100141. PMID: 34557778.
16. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. In: *Advances in Neural Information Processing Systems 30 (NIPS 2017)*. 2017:4765–4774. arXiv:1705.07874.
17. Fey M, Lenssen JE. Fast graph representation learning with PyTorch Geometric. *ICLR 2019 Workshop on Representation Learning on Graphs and Manifolds*. 2019. arXiv:1903.02428.
18. Ying R, Bourgeois D, You J, Zitnik M, Leskovec J. GNNExplainer: generating explanations for graph neural networks. In: *Advances in Neural Information Processing Systems 32 (NeurIPS 2019)*. 2019:9240–9251. arXiv:1903.01973.
19. Jin S, Guerrero-Juarez CF, Zhang L, Chang I, Ramos R, Kuan CH, Myung P, Plikus MV, Nie Q. Inference and analysis of cell-cell communication using CellChat. *Nature Communications*. 2021;12(1):1088. doi:10.1038/s41467-021-21246-9. PMID: 33597522.
20. willow0077. isp-confound-toolkit: confound-aware diagnostics for foundation-model in-silico perturbation predictions (version v0.1.0). Zenodo; 2026. doi:10.5281/zenodo.20743524.

Figure legends

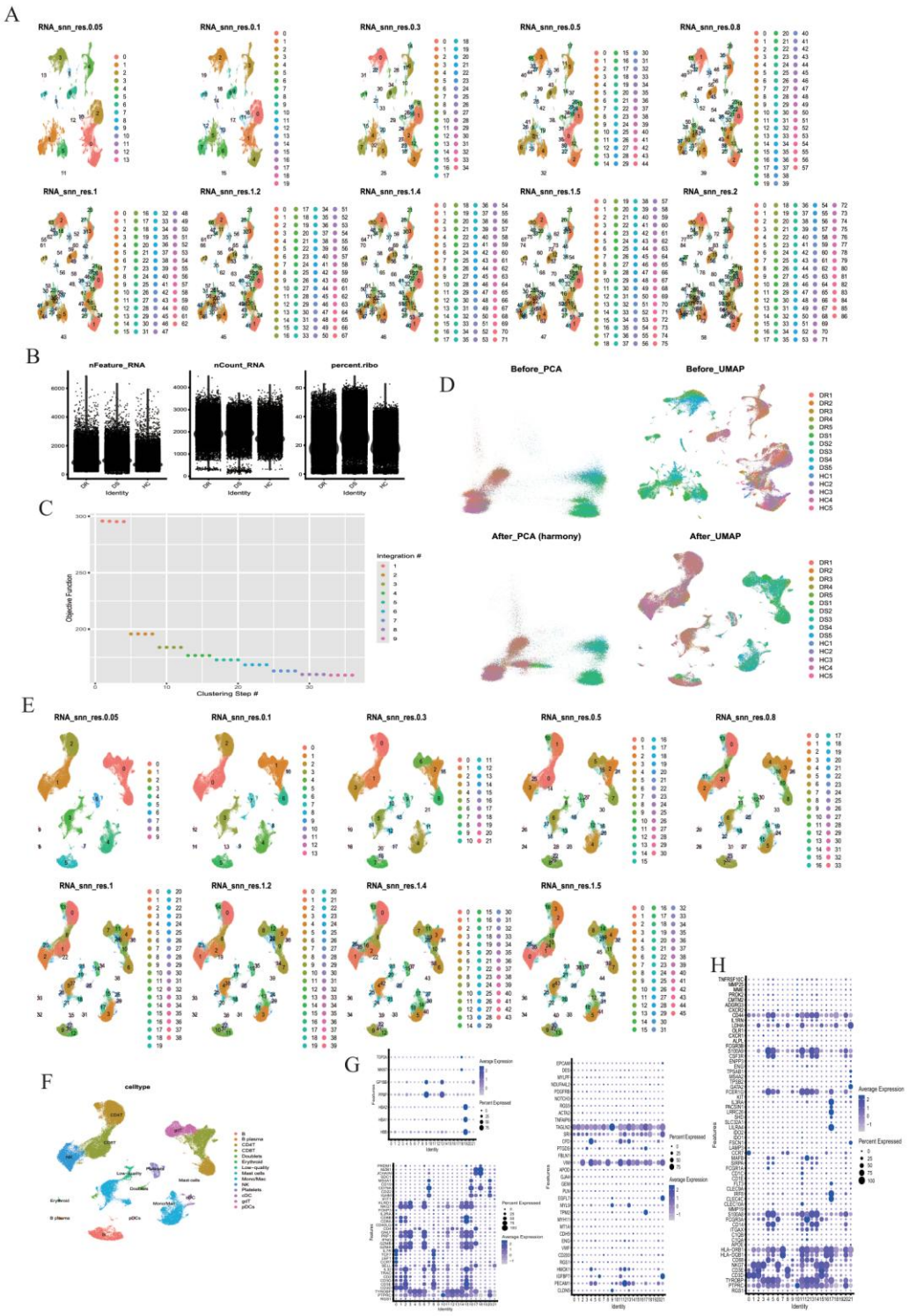


Fig. 1. Single-cell atlas of 15-donor PBMCs across treatment strata. (a) Cluster assignments across the lower end of the resolution sweep (RNA_snn_res 0.05–0.8). (b) Quality-control distributions of detected genes (nFeature_RNA), unique-molecular-identifier counts (nCount_RNA) and ribosomal content (percent.ribo), by group. (c) Harmony integration objective function across clustering steps for the integration runs. (d) PCA and UMAP embeddings before and after Harmony integration, coloured by donor (DR1–5, DS1–5, HC1–5). (e) Cluster assignments across the higher end of the resolution sweep (RNA_snn_res 1.0–2.0). (f) Annotated UMAP of the 14 cell types. (g) Dot plot of cycling and erythroid markers. (h) Cell-type marker dot plot (average expression and percent expressed).

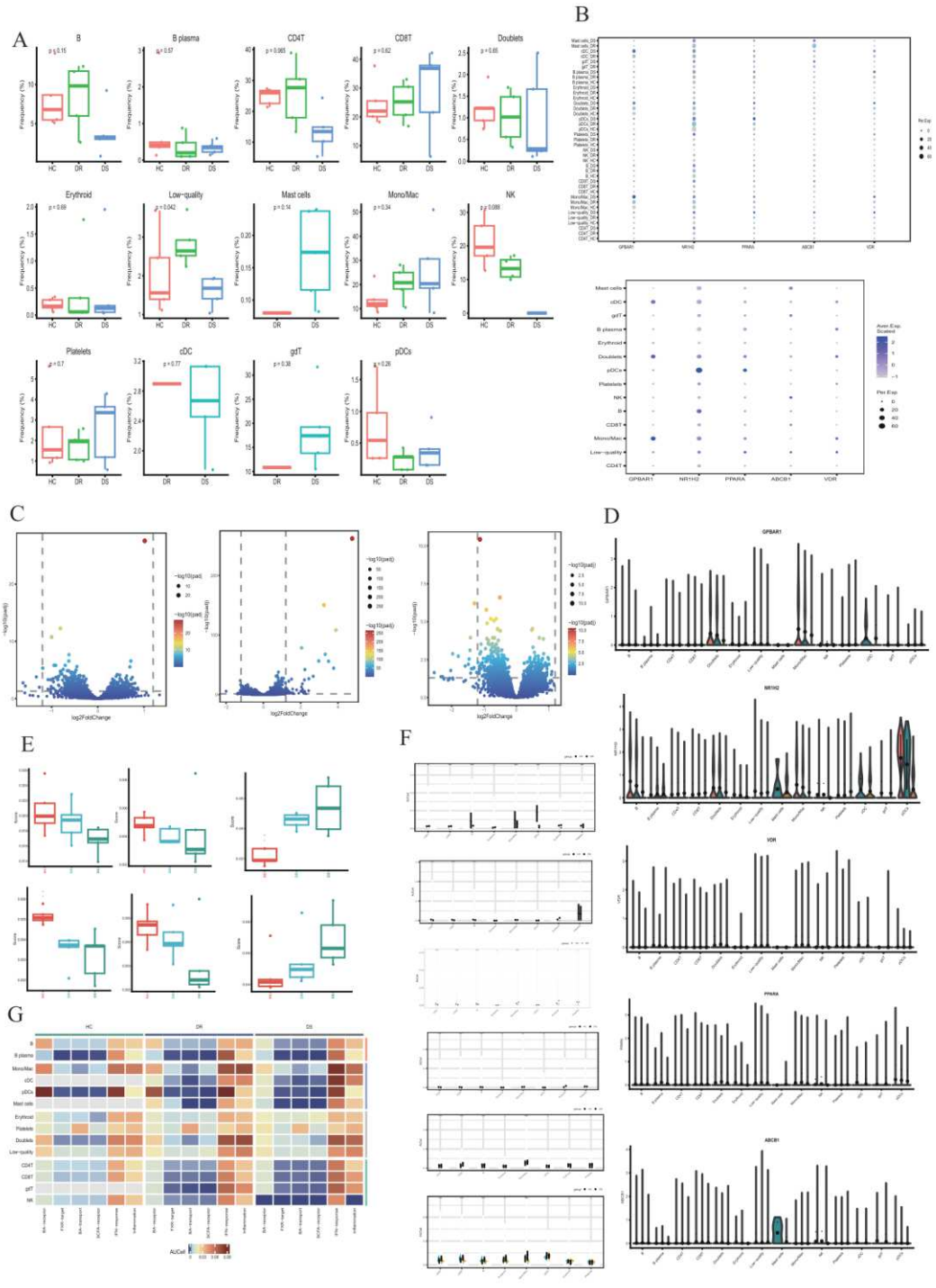


Fig. 2. The bile-acid receptor landscape and pathway-signature activity in PBMCs. (a) Per-cell-type frequency across groups, with Kruskal–Wallis p-values. (b) Receptor/transporter dot plot (GPBAR1, NR1H2, VDR, PPARA, ABCB1) across cell types, split by group. (c) Differential-expression volcano plots for the three contrasts. (d) Per-gene receptor/transporter expression by cell type (GPBAR1, NR1H2, VDR, PPARA, ABCB1). (e) Module-score boxplots of the six curated signatures by group. (f) Per-sample AUCell scores by signature with pairwise comparisons. (g) Cell-type-by-signature AUCell heatmap split by group.

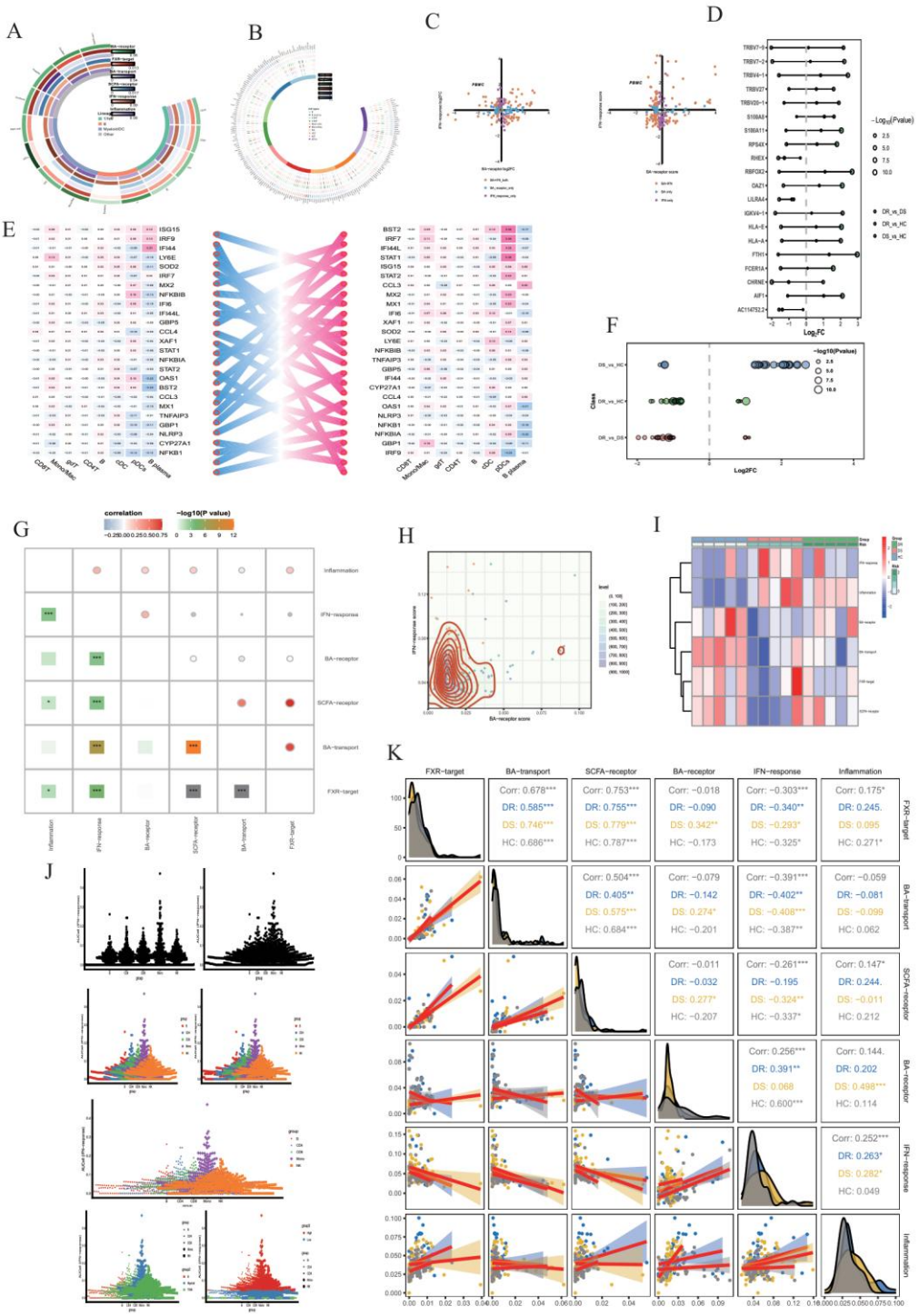


Fig. 3. Signature coupling and donor-level differential expression. (a) Circos-style summary of signature-by-cell-type association strength. (b) Circular dendrogram of the annotated cell subsets. (c) Receptor-versus-interferon log-fold-change scatter. (d) Gene-level forest plot of log2 fold change for DR versus HC and DR versus DS. (e) Interferon-gene-by-cell-type correlation heatmaps with gene-to-cell-type links. (f) Contrast-by-gene log2 fold-change dot plot. (g) Signature-by-signature correlation dot matrix. (h) Receptor-versus-interferon score density scatter. (i) Clustered signature heatmap across donors with group annotation. (j) Signature AUCell scores by cell type and group. (k) Pairwise scatter-matrix of the six signatures with per-group correlations.

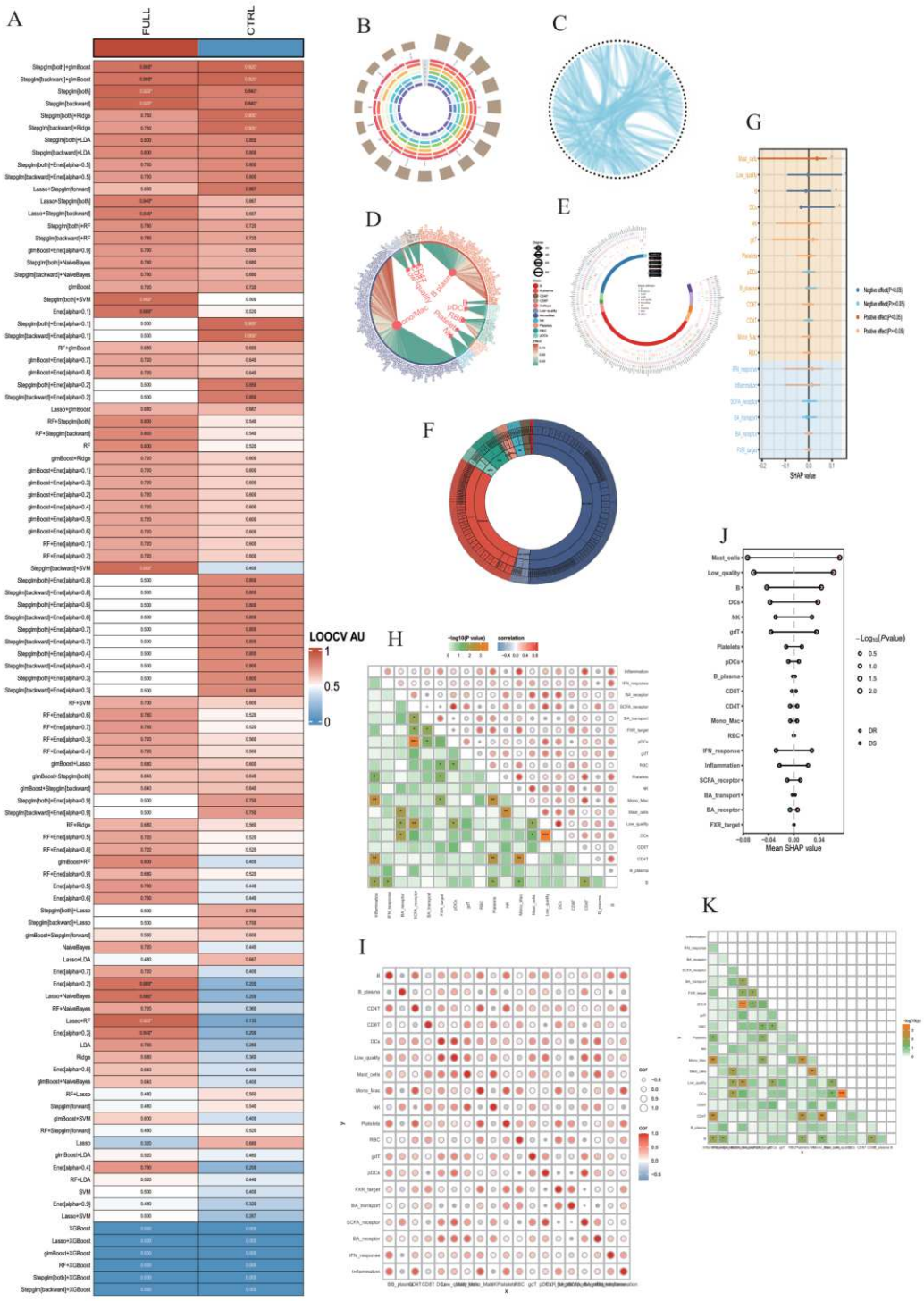


Fig. 4. Interpretable machine learning of drug-resistant versus drug-susceptible donors. (a) Leave-one-out cross-validation AUC across the 113-model ensemble, for the full and control feature sets. (b–f) Circular association and tree views of the donor-level feature space. (g) SHAP-effect forest across cell-population and signature features. (h) Cell-type-by-signature correlation summary with LOOCV-AUC annotation. (i) Feature-by-feature correlation circle-matrix. (j) Mean SHAP values for the drug-resistant and drug-susceptible groups. (k) Lower-triangular significance matrix of the feature correlations.

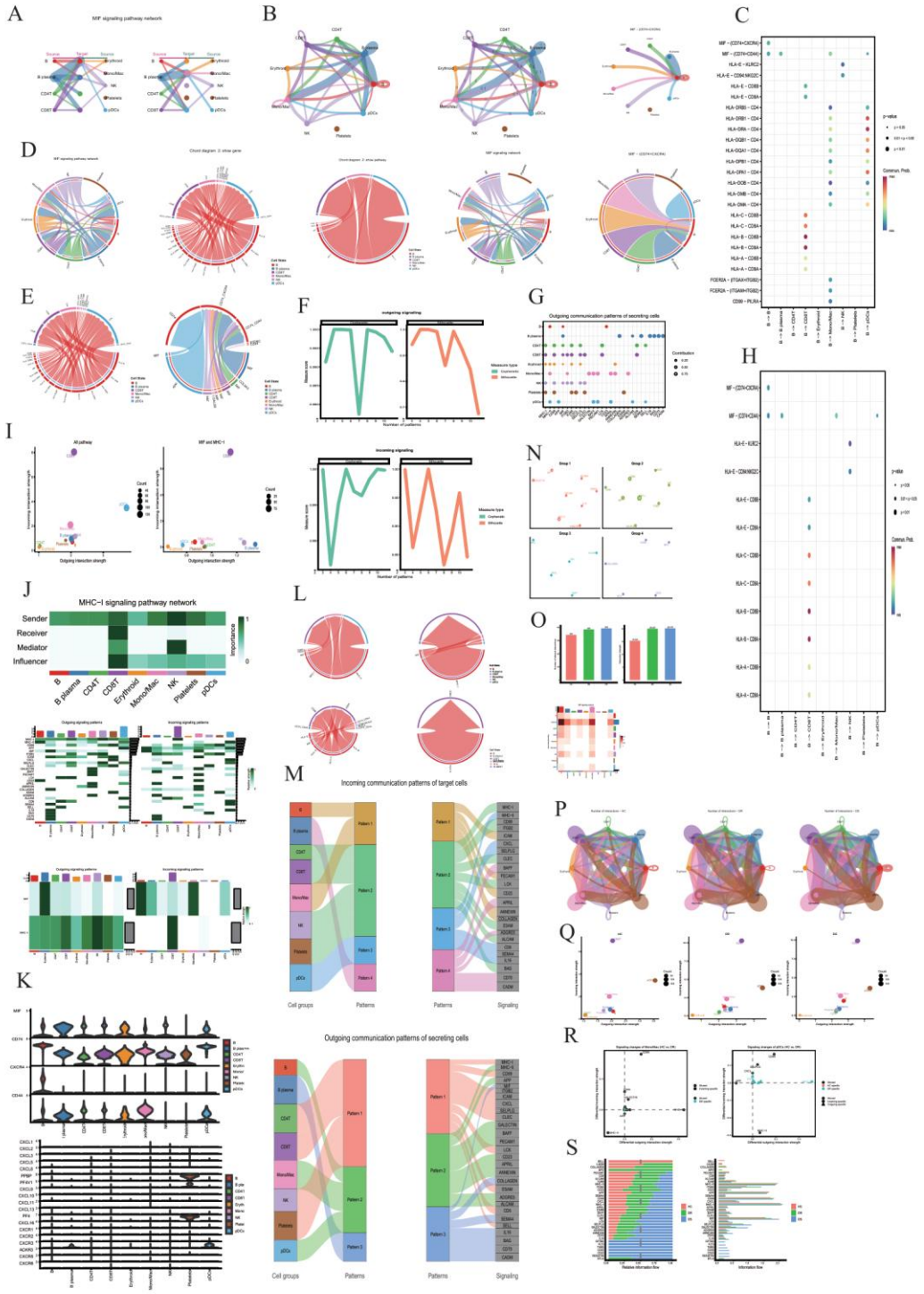


Fig. 5. Cell–cell communication. (a–e) MIF and MHC-I signalling networks and chord diagrams. (f–i) Signalling patterns and interaction-strength summaries. (j–n) Pathway importance, signalling-gene expression and pattern contribution. (o–s) Between-group differential interaction strength and information flow. Reported as hypothesis-generating.

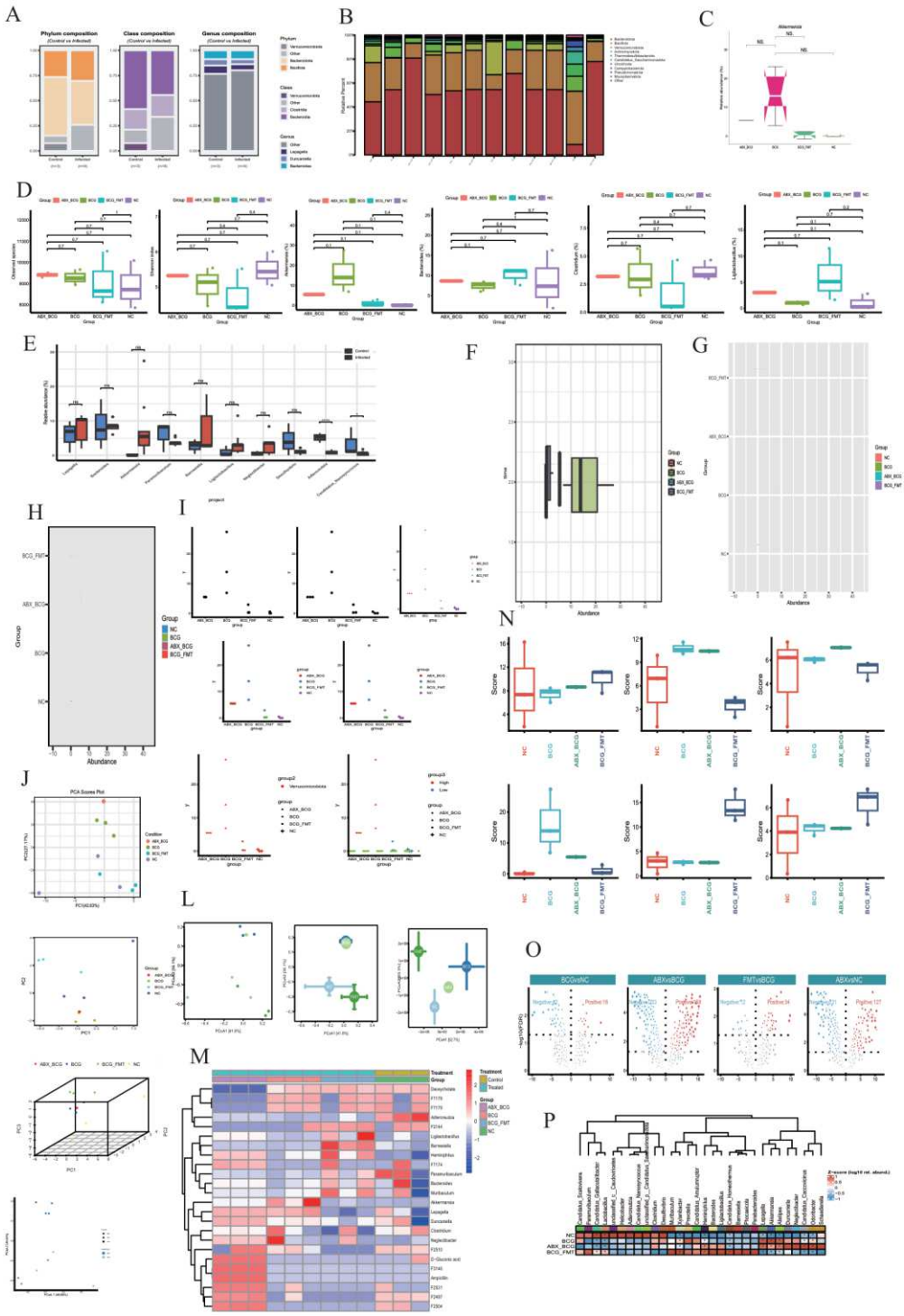


Fig. 6. Mouse gut microbiome and metabolome layer. (a–b) Phylum/class/genus composition and per-sample relative abundance. (c–i) Alpha diversity, per-genus abundance and ordination across the NC, BCG, ABX_BCG and BCG_FMT arms. (j–l) PCA/PCoA ordinations. (m–n) Taxon- and metabolite-level heatmaps with treatment and group annotation. (o) Differential taxa/metabolites for each contrast. (p) Sample dendrogram.

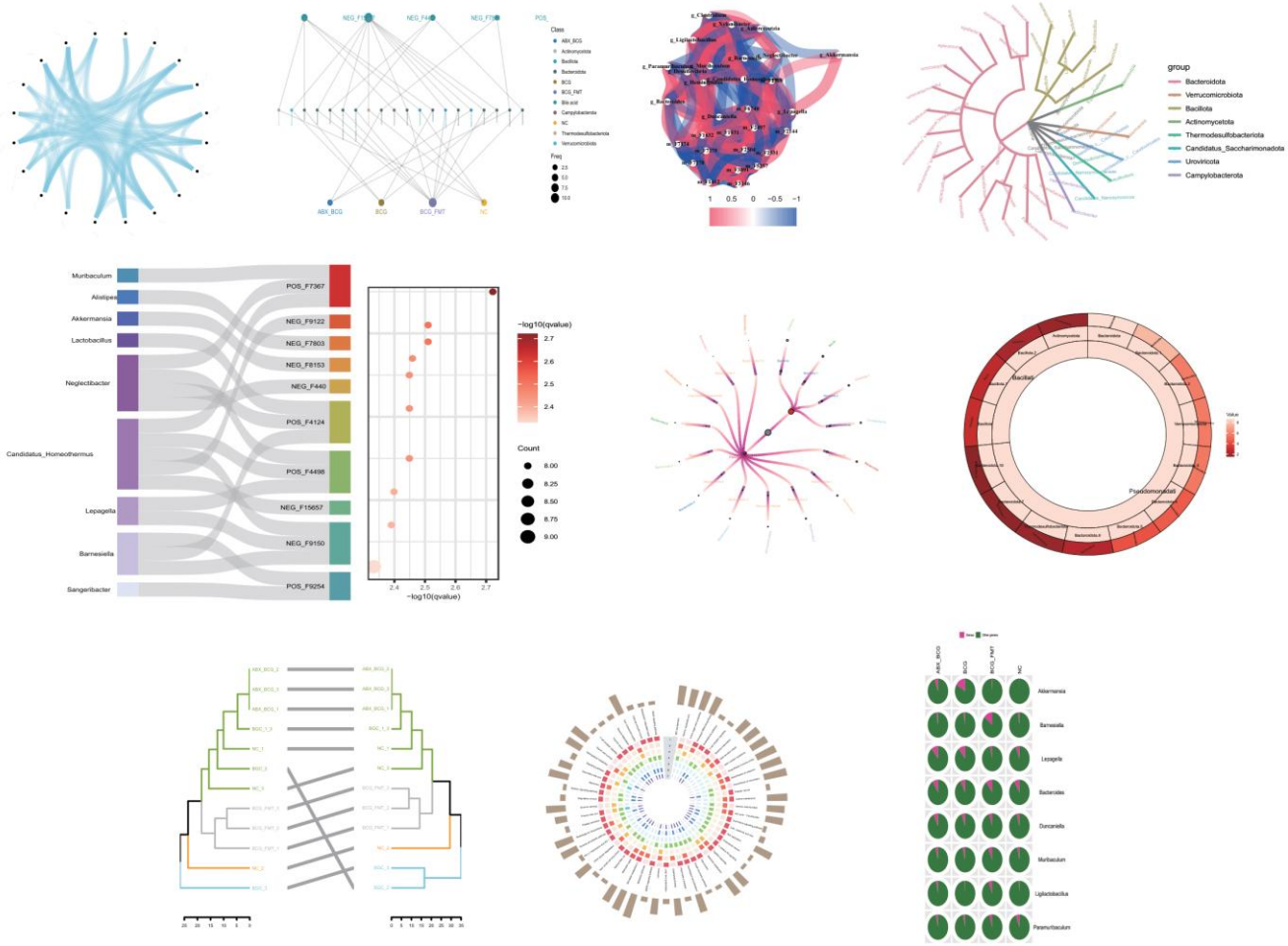


Fig. 7. Taxa–metabolite association structure in the mouse layer. Taxon association network, feature-by-group bipartite network, taxon correlation network, taxonomic trees, taxon-to-metabolite alluvial map, q-value dot plot and taxon-by-group pie grid.

Supplementary Fig. S1. Set co-occurrence of the differential-expression contrasts. Venn diagram of the three contrasts and UpSet plot of the six direction-specific sets (three contrasts by up/down).

Supplementary Fig. S2. Functional enrichment of the differential gene sets. GO and KEGG enrichment for DS versus HC (including advanced bubble and circle views), DR versus HC and DR versus DS.

Supplementary Fig. S3. Internal replication by donor folds. Per-fold significance of the 23-gene panel across the three contrasts (dumbbell views) and the nine-facet forest plot.