nature microbiology

Article

https://doi.org/10.1038/s41564-024-01739-1

Multikingdom and functional gut microbiota markers for autism spectrum disorder

Received: 1 August 2023

Accepted: 20 May 2024

Published online: 08 July 2024

Check for updates

Qi Su[1,2,9], Oscar W. H. Wong[3,4,9], Wenqi Lu[1,2], Yating Wan[1,2,4], Lin Zhang[1,2], Wenye Xu[1,2], Moses K. T. Li[1], Chengyu Liu[1], Chun Pan Cheung[1,2], Jessica Y. L. Ching[1], Pui Kuan Cheong[1], Ting Fan Leung[4,5], Sandra Chan[3,4], Patrick Leung[4,6], Francis K. L. Chan[1,4,7] & Siew C. Ng[1,2,4,8]

Associations between the gut microbiome and autism spectrum disorder (ASD) have been investigated although most studies have focused on the bacterial component of the microbiome. Whether gut archaea, fungi and viruses, or function of the gut microbiome, is altered in ASD is unclear. Here we performed metagenomic sequencing on faecal samples from 1,627 children (aged 1–13 years, 24.4% female) with or without ASD, with extensive phenotype data. Integrated analyses revealed that 14 archaea, 51 bacteria, 7 fungi, 18 viruses, 27 microbial genes and 12 metabolic pathways were altered in children with ASD. Machine learning using single-kingdom panels showed area under the curve (AUC) of 0.68 to 0.87 in differentiating children with ASD from those that are neurotypical. A panel of 31 multikingdom and functional markers showed a superior diagnostic accuracy with an AUC of 0.91, with comparable performance for males and females. Accuracy of the model was predominantly driven by the biosynthesis pathways of ubiquinol-7 or thiamine diphosphate, which were less abundant in children with ASD. Collectively, our findings highlight the potential application of multikingdom and functional gut microbiota markers as non-invasive diagnostic tools in ASD.

Autism spectrum disorder (ASD) is a heterogeneous neurodevelopmental disorder characterized by social, cognitive and behavioural impairments[1–3]. Although the cause of ASD is unknown, it is believed to relate to a complex interplay between genetic and environmental factors[4–6]. In the past decade, the gut microbiome has been shown to play a central role in modulating the gut–brain axis by regulating neuroimmune networks and directly communicating with the brain, and may contribute to the development of ASD[7]. Preclinical studies have shown that children with ASD had an altered gut microbiota composition and delayed development of the gut microbiota[8–13]. Furthermore, the transfer of faecal microbiota from individuals with ASD into germ-free mice promoted autistic-like behaviour[14], whereas faecal microbiota transplantation from healthy individuals to children with ASD resulted in improvements in symptoms[15,16]. So far, most studies have focused on the bacteria component of the gut microbiota and revealed many, albeit inconsistent, variations in microbial diversity and composition in children with ASD[17–19]. Metagenomic sequencing technologies have enabled the study of other microbial communities including archaea,

[1]Microbiota I-Center (MagIC), Hong Kong SAR, China. [2]Department of Medicine and Therapeutics, The Chinese University of Hong Kong, Hong Kong SAR, China. [3]Department of Psychiatry, The Chinese University of Hong Kong, Hong Kong SAR, China. [4]The D.H. Chen Foundation Hub of Advanced Technology for Child Health (HATCH), The Chinese University of Hong Kong, Hong Kong SAR, China. [5]Department of Paediatrics, The Chinese University of Hong Kong, Hong Kong SAR, China. [6]Department of Psychology, The Chinese University of Hong Kong, Hong Kong SAR, China. [7]Centre for Gut Microbiota Research, The Chinese University of Hong Kong, Hong Kong SAR, China. [8]Li Ka Shing Institute of Health Sciences, State Key Laboratory of Digestive Disease, Institute of Digestive Disease, The Chinese University of Hong Kong, Hong Kong SAR, China. [9]These authors contributed equally: Qi Su, Oscar W. H. Wong. e-mail: siewchienng@cuhk.edu.hk

https://doi.org/10.1038/s41564-024-01739-1

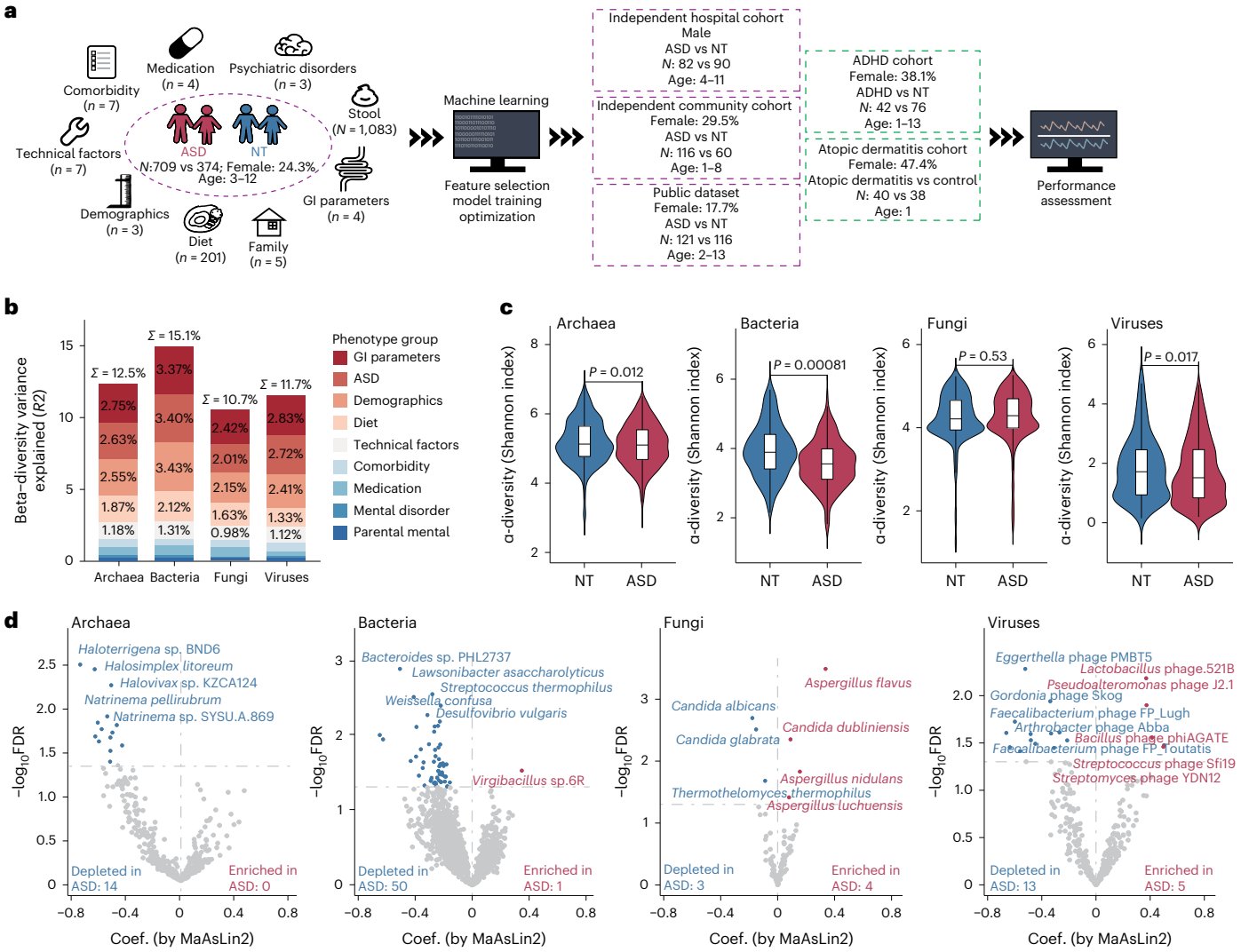


Fig. 1 | Associations between ASD and faecal microbiome composition. **a**, Graphical summary of the cohorts and overview of available metadata (n = number of variables collected, N = sample size). **b**, Variance in multikingdom (archaea, bacteria, fungi and viruses) microbiome composition explained by phenotype groups in multivariate PERMANOVA analysis. **c**, Alpha diversity in multikingdom (archaea, bacteria, fungi and viruses) microbiome measured by Shannon index of children with ASD (red, n = 709) and children considered neurotypical (NT, blue, n = 374). P value (two-sided test) was calculated using MMUPHin. Data are shown in boxplots as the median (centre line), 25th and 75th percentiles (box limits), and 5th and 95th percentiles (whiskers). **d**, Volcano plots show the associations between multikingdom (archaea, bacteria, fungi and viruses) species and ASD calculated using MaAsLin2 after adjusting for significant confounders. Associations with FDR < 0.05 were considered as significant and marked in red (enriched in ASD) or blue (depleted in ASD). Top-ranked species are labelled with name.

fungi and viruses that also colonize the human gut, and these dark matters may play a key role in the pathogenesis of ASD[20–22].

In this study, we explored multikingdom analyses of gut archaea, bacteria, fungi, viruses and their genes and functions, presented metagenomic analyses of 1,627 children considered neurotypical or with ASD, with extensive phenotype data, and validated our findings in public datasets of 237 faecal metagenomes.

Results

Study characteristics

A total of 1,627 children (aged 1–13 years, 24.4% female) from five independent cohorts were recruited in this study (Fig. 1a and Extended Data Table 1). Extensive phenotypic data (236 factors) were collected including age, sex, body mass index (BMI), diet, medication, comorbidity, concomitant psychiatric disorders, gastrointestinal (GI) symptoms (including stool consistency assessed by Bristol stool form score, BSFS), family characteristics and technical factors related to sample collection, storage and processing (Fig. 1a). All faecal samples were processed using the same standardized protocol to reduce the batch effects caused by technical factors. In total, we obtained over 10 terabytes of sequence data at an average depth of 6.34 gigabases for each metagenome (Extended Data Fig. 1). In the discovery cohort, we performed metagenomic sequencing on faecal samples from 709 children with ASD and 374 children considered neurotypical (age 3–12 years, 24.3% female). In an independent hospital cohort, we sequenced a total of 172 faecal samples from 82 children with ASD and 90 children considered neurotypical (aged 4–11 years, Fig. 1a). We also included a community cohort with younger children consisting of 116 children with ASD and 60 children considered neurotypical (aged 1–8 years, 29.5% female, Fig. 1a) to validate the findings in different age groups. A total of 237 faecal metagenomes (aged 2–13 years, 17.7% female) from published datasets were included in the analysis for external validation (Fig. 1a). Two additional non-ASD cohorts of children with attention deficit hyperactivity disorder (ADHD, n = 118)

https://doi.org/10.1038/s41564-024-01739-1

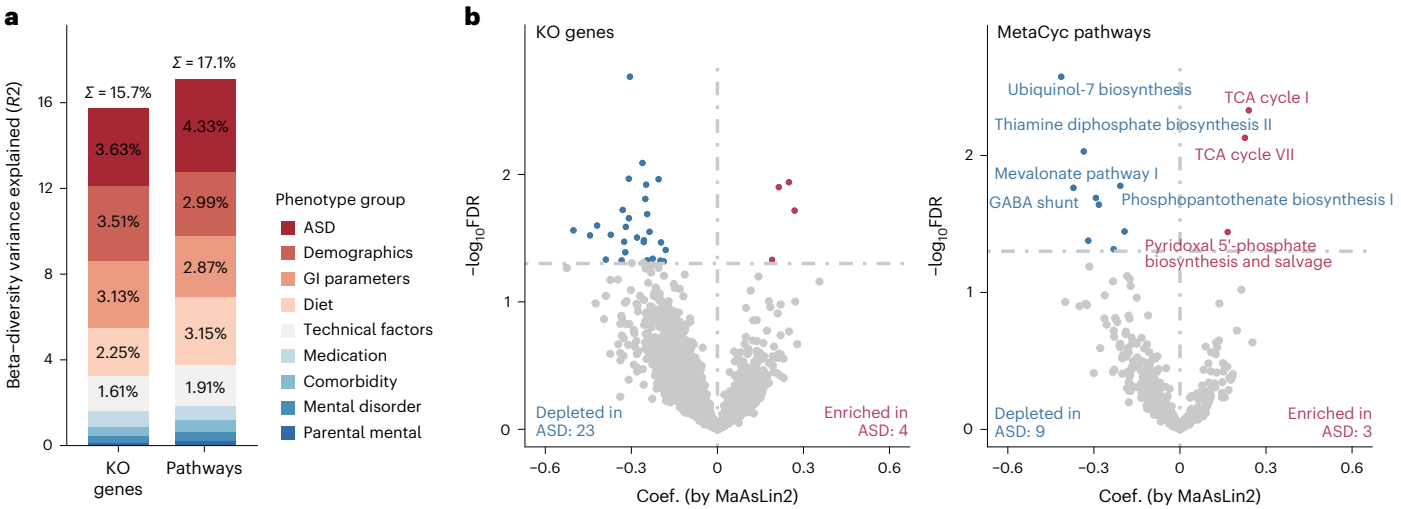


Fig. 2 | Associations between ASD and faecal microbiome functions. **a**, Variance in microbiome function (microbial pathways and genes) explained by phenotype groups in multivariate PERMANOVA analysis. **b**, Volcano plots show the associations between microbiome functions and ASD calculated using MaAsLin2 after adjusting for significant confounders. Associations with FDR < 0.05 were considered as significant and marked in red (enriched in ASD) or blue (depleted in ASD). Top-ranked pathways are labelled with name.

and atopic dermatitis (n = 78) were used to evaluate the specificity of our findings (Fig. 1a).

ASD-associated microbial species in four kingdoms

As gut microbiota composition is largely shaped by environmental and host factors[23,24], we analysed the impact of 236 host factors on the gut microbiome composition to determine potential confounders (Fig. 1a). In our discovery cohort (age 3–12 years, n = 1,083, 24.3% female), these host factors in combination explained 12.5%, 15.1%, 10.7% and 11.7% of the interindividual microbiome variation for archaea, bacteria, fungi and viruses, respectively (Fig. 1b). Among all host and dietary factors studied, a total of 21 factors showed a significant impact on gut microbiota composition, including ASD, age, sex, BMI, 3 GI parameters, 15 dietary factors and sequencing batch (Extended Data Fig. 2), therefore these factors were adjusted in all subsequent analyses. We next assessed changes in gut microbiota diversity between children considered neurotypical and children with ASD. Children with ASD showed a decrease in the diversity of archaea, bacteria and viruses compared with children considered neurotypical (Fig. 1c). A total of 14 archaeal, 51 bacterial, 7 fungal and 18 viral species showed differential abundances between children considered neurotypical and children with ASD (Fig. 1d). The relative abundance of 80 out of 90 identified microbial species was found to be significantly decreased in children with ASD compared with children considered neurotypical (Fig. 1d). This finding was most pronounced for the bacterial communities, where 50 bacterial species were depleted in children with ASD whereas only one bacteria species was enriched (Fig. 1d). Alterations in bacterial species in children with ASD were driven by the depletion of *Streptococcus thermophilus* and short-chain fatty acids-producing bacteria, such as *Bacteroides* sp. PHL2737 and *Lawsonibacter asaccharolyticus*.

Alterations of gut microbiota function in ASD

At functional level, host phenotype factors explained 17.1% and 15.7% of the variation in microbiome pathways and microbial genes, respectively (Fig. 2a). A diagnosis of ASD ranked as the top factor accounting for the variation in both microbiome pathways and microbial genes (Fig. 2a). A total of 19 host and diet factors showed a significant impact on the gut microbiome function, including ASD, age, sex, BMI, 2 GI parameters, 12 dietary factors and sequencing batch (Extended Data Fig. 3). After adjusting for these confounders, we identified 27 differential Kyoto Encyclopedia of Genes and Genomes (KEGG) orthology (KO) genes: 23 were decreased and 4 were increased in children with ASD compared with children considered neurotypical (Fig. 2c). At the pathway level, 12 differential pathways were noted, including 9 pathways showing negative associations and 3 showing positive associations with ASD (Fig. 2b). We found that biosynthesis pathways of ubiquinol-7 and thiamine diphosphate were reduced in children with ASD compared with children considered neurotypical (Fig. 2b). Ubiquinol exhibits antioxidant activity and has been identified to be capable of improving symptoms in children with ASD[25,26]. Impairment of thiamine diphosphate synthesis has been associated with ASD and other mental disorders in both animal and human studies[27–30]. We also observed a negative association between ASD and the shunt pathway of 4-aminobutyric acid (GABA). GABA is a major inhibitory neurotransmitter of the mammalian central nervous system and has been associated with ASD in previous studies[31–33].

Single microbial kingdom markers for ASD diagnosis

Bacterial markers for ASD diagnosis have been explored in several studies[8,9,34]. However, the performance of archaea, fungi, viruses, microbial genes (KO families) or function pathways has not been explored in ASD. To avoid discrimination bias driven by imbalanced sample size and residual confounders, we constructed a matched subcohort of children with ASD (n = 301, 95 girls and 206 boys) and children considered neurotypical (n = 301, 95 girls and 206 boys) using one-to-one pairing algorithm[24,35], based on the discovery cohort (n = 1,083, Fig. 3a). After matching, there was no difference in these confounders between children considered neurotypical and children with ASD within the matched subcohort (Extended Data Fig. 4). These results illustrated a comparable environmental and host background in children considered neurotypical or with ASD, thus our model is less likely to have been affected by these important factors[24] (Fig. 3a).

Within the matched cohort, we first tested the accuracy of models using single-kingdom markers to distinguish children with ASD from children considered neurotypical. Among all single-kingdom markers, the microbial pathway model displayed the strongest predictive ability to detect ASD with an average area under curve (AUC) of 0.87 (Fig. 3b), followed by microbial genes (AUC 0.86), bacteria (AUC 0.85), archaea (AUC 0.76), fungi (AUC 0.74) and viruses-based (AUC 0.68) models (Fig. 3b). Collectively, our findings demonstrate that faecal microbiome markers from different kingdoms provide promising predictive capabilities for ASD diagnosis.

https://doi.org/10.1038/s41564-024-01739-1

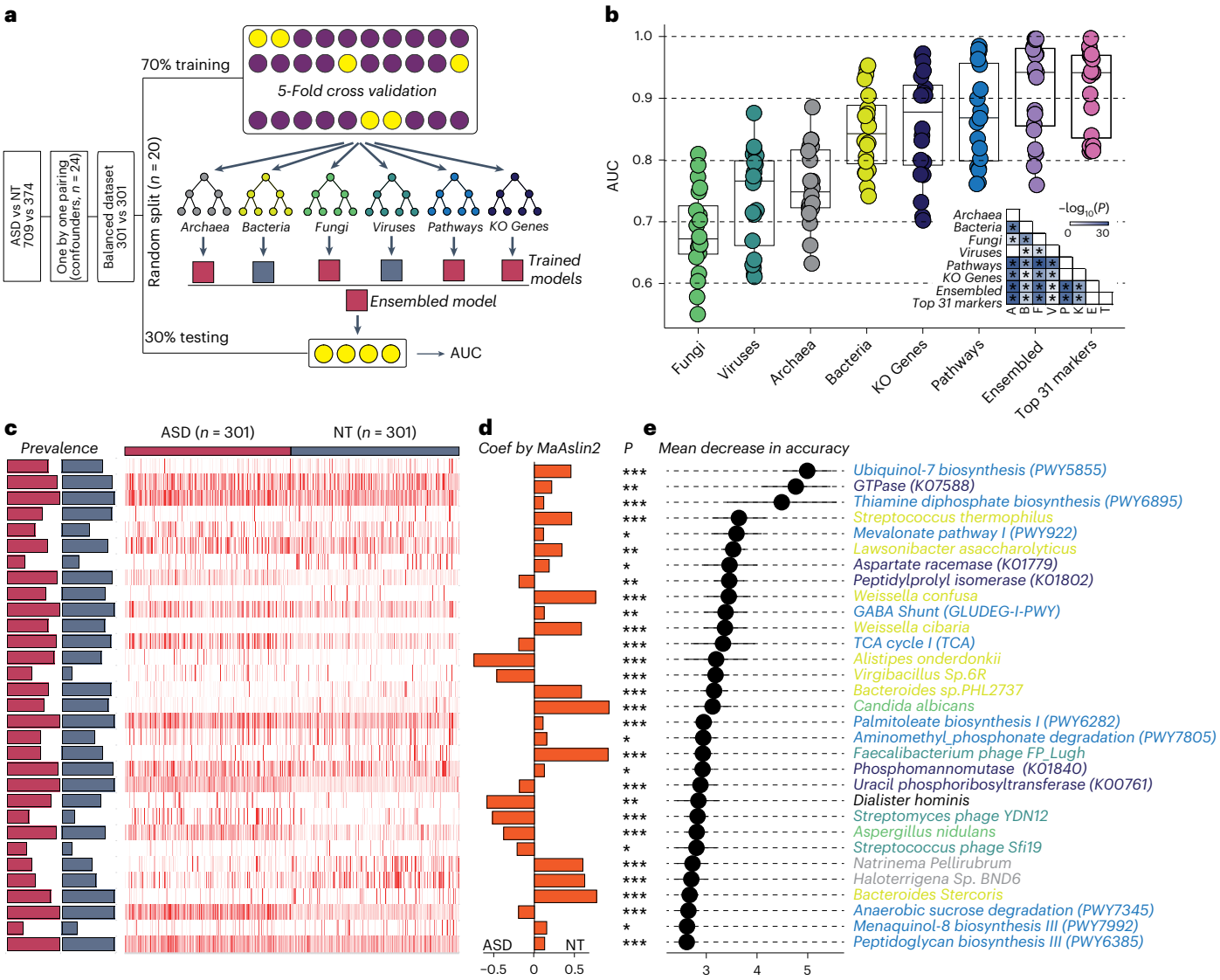


Fig. 3 | Random forest models for the diagnosis of ASD. a, Schematic diagram of random forest model development. The balanced dataset of children with ASD ($n = 301$) versus children considered neurotypical (NT, $n = 301$) was constructed from the original discovery cohort ($n = 1,083$) using one-by-one pairing algorithm, accounting for 24 significant confounders. A train–test sample split of 70% for training and 30% for testing was utilized, and the testing data were then used to estimate the accuracy of the random forest model. This process was randomly repeated 20 times to obtain a distribution of random forest prediction evaluations on the testing set. The input data are a vector with six components: archaea, bacteria, fungi, viruses, microbial pathways and KO genes. **b**, Box-and-whisker plot displaying the distribution of AUC scores yield from 20 random repeats of model training and test. Differences between groups were evaluated using two-sided Wilcoxon rank-sum test. Data are shown as median (centre line), 25th and 75th percentiles (box limits), and 5th and 95th percentiles (whiskers). **c**, Prevalence and relative abundance of the top 31 markers employed by the random forest model in the balanced dataset of children with ASD ($n = 301$) and children considered neurotypical (NT, $n = 301$). **d**, Associations between 31 markers with ASD assessed using MaAsLin2 (two-sided test, *P* value adjusted using the FDR method). *FDR < 0.05, **FDR < 0.01, ***FDR < 0.001. Exact FDR values are provided as Source data. **e**, Mean decrease in accuracy of the 31 markers.

Multikingdom microbial markers for ASD diagnosis

We next explored the performance of the model combining multikingdom features. We found that the ensembled model showed superior performance (average AUC 0.91) for the diagnosis of ASD compared with models based on single-kingdom features (Fig. 3b). These results confirmed that the multikingdom faecal microbiome biomarker panel had a higher diagnostic performance for detecting ASD than single-kingdom panels. To identify the minimal number of microbiome markers that achieve the highest accuracy, we included identified markers consecutively into the model according to their ranking, and finally, a total of 31 microbial features showed an AUC of 0.91 for the diagnosis of ASD (Fig. 3b). The prevalence and relative abundance of these 31 markers differed significantly between children considered neurotypical and children with ASD (Fig. 3c). Based on MasAsLin2, 21 markers were significantly depleted whereas 10 markers were significantly enriched in children with ASD (Fig. 3d). We re-analysed the importance of these 31 features and observed that the accuracy of the model was driven predominantly by the ubiquinol-7 biosynthesis pathway, the GTPase and the thiamine diphosphate biosynthesis pathways (Fig. 3e), which support their potential role in ASD pathogenesis. In addition, the depletion of several bacteria, including *Streptococcus thermophilus, Lawsonibacter asaccharolyticus, Weissella confuse*, *Weissella cibaria* and *Bacteroides* sp. PHL2737, were also the top-ranked microbial features contributing to the diagnostic accuracy (Fig. 3e). Collectively, our analysis shows that a 31-faecal microbiome marker panel represents a potentially promising non-invasive tool for the diagnosis of ASD.

https://doi.org/10.1038/s41564-024-01739-1

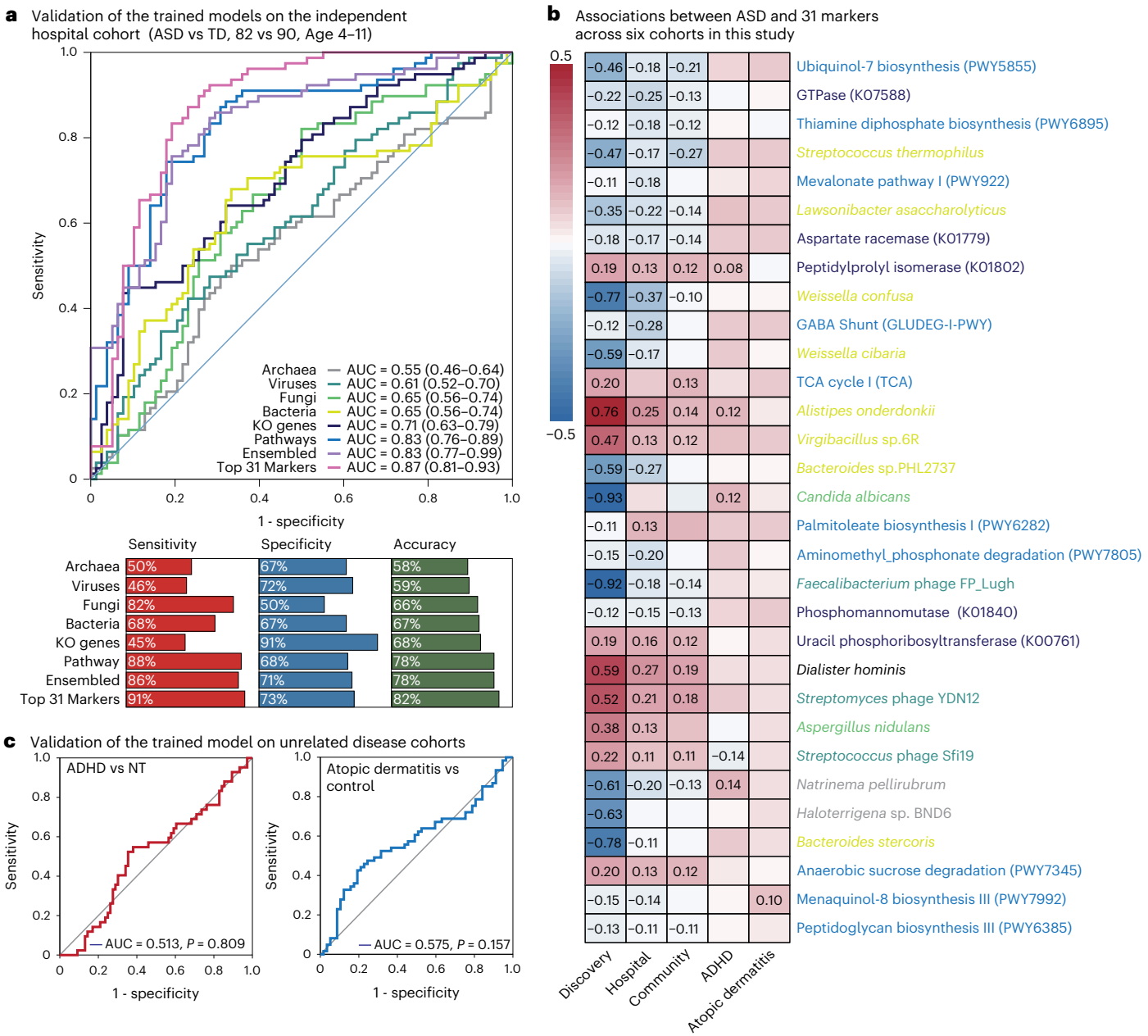


Fig. 4 | Validation of the random forest models. a, AUC (95% CI) of random forest models employing different features in the validation cohort. The sensitivity, specificity and accuracy were calculated on the basis of the Youden index. AUCs were calculated after adjustment for technical factors and available covariates including age, sex, BMI, BSBF, functional constipation and defecation disorders. **b**, Associations between ASD and 31 identified faecal microbiome markers across five cohorts calculated using MaAsLin2 (two-sided test). The Coef. value was only marked when $P < 0.05$. **c**, AUC of model using 31 markers tested in the independent ADHD and atopic dermatitis cohorts. *P* values were calculated using Wilcoxon rank-sum test (two-sided test).

Validations in independent cohorts and public datasets

To externally validate the diagnostic value and to avoid over-optimistic reporting of diagnostic accuracy, we tested the 31-marker panel in an independent hospital cohort. We found that our models maintained an AUC ranging from 0.55 to 0.87 (Fig. 4a). Among them, the ensembled model using 31 markers ranked first in AUC at 0.87, with a sensitivity of 91% and a specificity of 73% (Fig. 4a). The relative abundance of the 28 out of 31 markers remained significantly different between children considered neurotypical and children with ASD (Fig. 4b). Furthermore, our models showed an AUC ranging from 0.61 to 0.89 in a younger subset of children in this validation cohort ($n = 31$, 6 years old or younger), where the ensembled model using 31 markers again achieved the highest accuracy (Extended Data Fig. 5). We also tested whether this panel could be applied to predict the risk of ASD in another younger cohort (116 children with ASD and 60 children considered neurotypical, 1–8 years old, 29.5% female, Extended Data Table 2). The trained model achieved an AUC of 0.89, with a relatively balanced performance for males and females. When reducing the age range to 6 years or younger and 4 years or younger, the model showed AUCs of both 0.91. We tested the associations between these 31 markers and ASD in this younger cohort stratified by age and confirmed that most of these associations were still reproducible (Fig. 4b and Extended Data Fig. 6). Taken together, these results demonstrate the robustness of our trained model and the 31-marker panel across ages, sexes and cohorts.

To further test the reproducibility of our 31-marker multikingdom panel across different populations, we integrated 237 shotgun

faecal metagenome data from six public datasets from Asia, Europe and America[36–41] (Extended Data Fig. 7a). The panel showed an AUC of 0.78 in differentiating children with ASD from children considered neurotypical, with a sensitivity of 65.30% and a specificity of 72.4% (Extended Data Fig. 7b). More importantly, it showed comparable performance for males and females, confirming the applicability of our model to both sexes (Extended Data Fig. 7b). Overall, this 31-marker multikingdom panel may be relevant across different populations and geographical locations.

Specificity of the multikingdom marker panel

Considering shared microbiota alterations across different diseases[23,42], it is important to verify the disease specificity for our identified microbial biomarkers panel to ensure a low false positive rate for the diagnosis of ASD. For this purpose, we assessed our trained model in two non-ASD cohorts in children with attention deficit hyperactivity disorder (ADHD, $n = 118$) or atopic dermatitis ($n = 78$; Fig. 1a). ADHD and atopic dermatitis have been reported to be associated with alterations in the gut microbiota[43–46]. AUC values of our marker panel were lower in children with atopic dermatitis or ADHD (AUC 0.51) and atopic dermatitis (AUC 0.58; Fig. 4c). In addition, only 17 of the tested participants were predicted to have ASD by the trained model, reflecting a low false positive rate of 8.7%. Overall, these results support the specificity of the 31-biomarker multikingdom panel for ASD.

Decreased ubiquinol-7 and thiamine diphosphate in ASD

Previous studies suggested that ubiquinol improves symptoms in children with ASD[25,26]. Decreased concentrations of plasma thiamine (vitamin B1) and its related metabolites, such as thiamine diphosphate, have been implicated in ASD[47–49]. However, the reasons underlying these observations are still unclear. We found that the relative abundance of the ubiquinol-7 biosynthesis pathway and the thiamine diphosphate biosynthesis pathway predominantly drove the accuracy of our diagnosis model and exhibited a consistent reduction in children with ASD across three cohorts compared with children considered neurotypical (Fig. 4b). A total of 17 enzymes were involved in these two pathways, and most of them were depleted in children with ASD across different cohorts (Extended Data Fig. 8). Altogether, these findings highlight that decreased abundance of ubiquinol-7 and thiamine diphosphate biosynthesis genes in the gut microbiota appeared to be strongly associated with ASD.

Discussion

Most studies have primarily focused on gut bacterial alterations in ASD[50–52]. Recently, investigations have revealed the critical roles of non-bacterial microorganisms, such as archaea, fungi and viruses in the gut–brain axis[21,22]. However, they are rarely explored in ASD. In this study, we performed a comprehensive analysis of the multikingdom and functional microbiome using over 1,600 metagenomes across 5 independent cohorts in children. We showed that archaeal, fungal, viral species and functional microbiome pathways could also separate children with ASD from children considered neurotypical. We demonstrated that a model based on a panel of 31 multikingdom markers achieved high predictive values for ASD diagnosis. The reproducible performance of the models across ages, sexes and cohorts highlights their potential as promising diagnostic tools for ASD.

We discovered a series of bacterial and non-bacterial markers and profiled their associations with ASD. Among them, we observed several beneficial bacteria, such as *Streptococcus thermophilus*[53], *Weissella confusa*[54] and *Weissella cibaria*[55,56] that exhibited negative associations with ASD. Several bacterial markers for ASD have also been reported in previous studies, such as *Bacteroides stercoris*[57]. However, very few studies have looked at the association between ASD and archaea, fungi and viruses, therefore these markers we reported require further exploration in future studies. In addition, we found that specific microbial functions may contribute to ASD pathogenesis via the deregulation of ubiquinol and thiamine diphosphate biosynthesis[47–49]. Ubiquinol and thiamine-related metabolites play crucial roles in mental health and neural signal transduction[29,58]. Our findings provided further evidence that thiamine diphosphate biosynthesis of the gut microbiome may also serve as a therapeutic target in the future.

Although several studies have attempted to identify reproducible microbial biomarkers for ASD[8,9,34], adequate validations in different cohorts are scarce. The lack of agreement across studies raises the question of whether microbial results obtained so far reflect intrinsic biological differences across cohorts, experimental biases or inadequate statistical power to preclude meaningful comparisons. On the basis of our ASD metagenomic dataset, we systematically evaluated the impact of host variables and technical factors on gut microbiota, and we fully adjusted the identified confounders throughout the analysis. Also, a matching algorithm for developing the machine learning model was adopted from a previous study which showed that mismatched host variables induce significant overestimation of AUC for binary classification of human diseases, as the unmatched confounders will reduce the machine learning model's ability to focus on the disease itself and therefore decrease its robustness across cohorts[24]. Furthermore, although the method of undersampling the majority class and then adding the environmental/technical factors as additional covariates in a mixed model is traditional, it may be more challenging to apply clinically given that data other than faecal microbiome will need to be collected and are harder to interpret biologically in real-life settings; therefore, a solely faecal microbiome-based random forest model was employed to achieve the prediction of ASD in this study. Finally, we showed superior diagnostic accuracy for models constructed using multikingdom and function markers than single-kingdom markers, in particular the addition of microbial pathways which has also been shown to have high diagnostic and prognostic value in other human diseases[59,60]. Moreover, models based on the 31-marker panel achieved very high predictive values for ASD diagnosis with robust performance across cohorts, ages and sexes. Although there is no disease-modifying treatment available for ASD, early identification of the disorder for comprehensive evaluation and initiation of training has been shown to lead to better social and behavioural outcomes. Our model has promising potential for clinical application and warrants further exploration. Moreover, the 31-marker panel included several archaea, fungi and viruses, highlighting the pivotal roles of non-bacterial microorganisms as diagnostic ASD biomarkers. The correlations among multikingdom species may also be developed into complex microbiota assemblies that serve as ecological drivers in the pathogenesis of ASD. However, the functional capabilities of these multikingdom associations have not been investigated and deserve further analysis.

There has been a heated debate about whether ASD-associated gut dysbiosis is driven solely by dietary preferences[12,61]. Our results showed that diet has an impact on the gut microbiome in children with ASD. However, ASD-associated microbiome alterations including microbial diversity and composition remained present after correction for dietary factors in our analysis. At the same time, to avoid our diagnostic model being misled by environmental and host factors, we have constructed a fully matched training cohort on the basis of a one-by-one pairing algorithm[24,35], which therefore reduced the influence of confounding factors and might explain the robustness of our model across cohorts. We have also included an analysis of two common childhood diseases known to have an association with gut microbiome alterations, atopic dermatitis and ADHD, and demonstrated that our 31-marker panel remained specific to ASD diagnosis.

Analysis of children cohorts that are heterogeneous in lifestyle, ethnicity and location presents a distinct opportunity for studying ASD-associated microbiome. By combining multiple small hospital-based and community cohorts of potentially low generalizability, we were able to obtain a better representation of the spectrum

of ASD cases and controls. However, this study has some limitations. Our understanding of how the gut microbiome is linked to dietary preferences, host immunity and GI and ASD behavioural symptoms is limited in cross-sectional studies, hence current data restrict our ability to perform causal inference. Although we did not analyse the effect of genotype on the microbiome, previous studies have identified genes that are at high risk for ASD[6,62]. Prospective studies of genetic markers in combination with microbiome panels are needed to establish whether they can further elevate the diagnostic accuracy to predict ASD earlier for disease prevention. In addition, although a previous study has fully characterized the matching algorithm[24], it may also introduce unknown biases into the results, thus independent validation and testing for the trained models are necessary. We have initially demonstrated the robustness of our model across cohorts and ages, with a low sensitivity to unrelated diseases including ADHD and atopic dermatitis, but a prospective validation trial of our models across the entire spectrum of neurodevelopmental conditions, common physical diseases of children and geographical locations would be highly needed to further demonstrate the universality and clinical applicability of these markers. Lastly, although our model shows good performance across ages, sexes and cohorts as well as the public dataset, there might be a residual possibility of unmeasured confounders or batch effects which are hard to avoid and may potentially lead to inflated performance, hence independent third-party validation is warranted before clinical application.

In conclusion, this study presents a highly specific multikingdom microbial panel for non-invasive diagnosis of ASD. Development of reproducible microbiome biomarkers and accurate disease predictive models from combined analyses of heterogeneous ASD forms the basis for future clinical diagnostic tests and hypothesis-driven mechanistic studies.

Methods

Ethics statement

This research complies with ethics regulations, with protocol approved by the Joint Chinese University of Hong Kong-New Territories East Cluster Clinical Research Ethics Committee (the Joint CUHK-NTEC CREC). Written consent was obtained from the children's parents.

Study population

Discovery ASD cohort. Children considered neurotypical or with ASD aged less than 12 years of age were recruited from the Child and Adolescent Psychiatric Clinic of the New Territory East Cluster (NTEC) of the Hospital Authority from December 2021 to December 2023. The NTEC serves one of the largest populations in Hong Kong, of which the Child and Adolescent Psychiatric Clinic receives constant referrals from the Child Assessment Centre, school education psychologists and private medical doctors for the assessment and treatment of children with ASD. The diagnosis of ASD was made by child psychiatrists on the basis of the Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition (DSM-5) diagnostic criteria[63]. Children considered neurotypical that were matched on age and sex to children with ASD were invited from participants of an ongoing territory-wide psychiatric epidemiological study (CREC Ref. No. 2018.497) in the same Child and Adolescent Psychiatric Clinic where over 6,000 random community-based youths were assessed on the presence of psychiatric disorders using the Diagnostic Interview Schedule for Children (DISC-5). To exclude the possibility of ASD from these community participants, children considered neurotypical were first screened using the Chinese Autism Spectrum Quotient Child Version[64] (AQ-Child, for children up to 11 years old). Exclusion criteria for children considered neurotypical included a positive family history of ASD among first-degree relatives, screening positive in the AQ-Child and the presence of psychiatric disorder according to the DISC-5. Exclusion criteria for both groups included mental retardation, neurological disorders, psychosis, depressive disorders or other major medical illness and avoidant/restrictive food intake disorder assessed by a 3-day food record questionnaire. Participants using probiotics, antidepressants, anti-epileptics and those who had exposure to antibiotics within 1 month before entering the study were also excluded. Children's parents or legal guardians were informed of the nature of the study and written informed consent was obtained. The study protocol followed the Declaration of Helsinki.

After recruitment, parent/guardian-filled questionnaires and medical records were used to profile each participant in the following areas: (1) social demographics: family composition, education level of parents, household income, number of siblings; (2) physical parameters of body height and weight; (3) presence of common physical diseases in children such as atopic dermatitis and asthma ($n = 14$); (4) common co-occurring psychiatric disorders such as ADHD ($n = 7$); (5) the presence of functional gastrointestinal disorder, defined according to the Rome IV Criteria, including three broad categories of functional nausea and vomiting disorder (FNVD), functional abdominal pain disorders (FAPD), functional defecation disorders (FDD) and stool consistency assessed by the BSFS ($n = 4$); (6) medication history including the use of psychiatric medications ($n = 7$); (7) parental parameters ($n = 5$) and (8) dietary patterns (assessed by food consumption within 3 months, $n = 201$). Finally, 709 children with ASD and 374 children considered neurotypical (24.3% female) of Chinese ethnicity aged between 3 and 12 years old were recruited.

Independent hospital ASD cohort. To independently validate the performance of the selected markers and trained models, we performed metagenomics analysis on an independent historical hospital cohort recruited in 2018 from the Child and Adolescent Psychiatric Clinic and consisting of 172 boys (90 children considered neurotypical and 82 participants with ASD, 4–11 years old). The neurotypical boys screened negative on the AQ-Child with parent-reported absence of psychiatric diagnosis. Otherwise, the inclusion and exclusion criteria were the same as those of the discovery cohort. It was also ascertained by the research team that there was no overlapping of participants between the historical cohort and the discovery cohort as they were recruited from the same Psychiatric Clinic.

Independent community ASD cohort. To validate the accuracy of the selected markers and trained models for ASD detection in the community setting, an independent community cohort consisting of children considered neurotypical or with ASD of younger ages was recruited from March 2022 to December 2023. These children were recruited from the families of service users in a community-based educational support programme. Other than a self-reported diagnosis of ASD, the parents had to submit a formal medical certificate from a qualified psychiatrist, paediatrician or psychologist to confirm the ASD diagnosis before the child was recruited as the case group. Children considered neurotypical not related to the case group were recruited as the control group according to the same inclusion and exclusion criteria as above. Finally, 116 children with ASD and 60 children considered neurotypical (29.5% female) of Chinese ethnicity aged between 3 and 12 years old were recruited.

Independent ADHD cohort. To validate the specificity of the selected markers and trained models for ASD detection, an independent cohort consisting of children with ADHD and children considered neurotypical was recruited from the community under the same study and sampling frame as the independent ASD community cohort. Participants with a confirmed diagnosis of ADHD based on the ADHD rating scale (ADHD-RS)[65] and Test of Variables of Attention (T.O.V.A)[66] were recruited as the case group. Children considered neurotypical not related to the case group were recruited as the control group according to the same inclusion and exclusion criteria as above.

Finally, 42 children with ADHD and 76 children considered neurotypical (38.1% female) of Chinese ethnicity aged between 1 and 13 years old were recruited.

Independent atopic dermatitis cohort. To validate the specificity of the selected markers and trained models for ASD detection, an independent cohort consisting of children with or without atopic dermatitis was recruited. Children were clinically examined by a paediatrician for signs of atopic dermatitis on the basis of Hanifin and Rajka criteria[67], and those with a diagnosis of atopic dermatitis were recruited as the case group. Children without any sign of atopic dermatitis were recruited as the control group. Written consent was obtained from the children's parents. The study was approved by the Joint CUHK-NTEC CREC (CREC Ref. No. 2016.637). Finally, 40 children with atopic dermatitis and 38 controls (47.4% female, 1 year old) of Chinese ethnicity were recruited.

HK$200 worth of supermarket coupons were given to each participant. No statistical methods were used to predetermine sample sizes, but our sample sizes are similar to those reported in previous publications[59,68].

Phenotype data cleaning and preprocessing

Most host variables for comorbidities, unrelated mental disorders, GI parameters and medication history were defined as binary with a positive and negative class. Beyond binary variables, the stool consistency was assessed uisng BSFS, which is a categorical variable from 1 to 7 (ref. 69). Within parental parameters, educational level was defined as a categorical variable from 'primary', 'secondary' to 'tertiary or above', and household income was divided into 'below HK$10,000', 'HK$10,001–HK$30,000', 'HK$30,001–HK$50,000' and 'over HK$50,000'. Potential confounding factors with continuous values were transformed into discrete variables as quartiles.

Stool sample collection

Faecal samples were collected at home by all participants using tubes containing Norgen preservative media (63700, Norgen Biotek) prepared by investigators. The Norgen preservative can preserve and allow safe transportation of microbial DNA and RNA at ambient temperature and eliminated sample variability. Stool samples were delivered to the hospital within 24 h of collection and stored at −80 °C refrigerators until further processing. We have previously shown that data on the gut microbiota composition generated from faecal samples collected using this preservative medium was comparable to data obtained from fresh samples that were immediately stored at −80 °C[70]. Also, the storage times were comparable between cases and controls in all involved cohorts.

Stool DNA extraction and sequencing

To avoid potential batch effects from sample processing and sequencing, all faecal samples from each cohort were handled in a random sequence according to the same protocol. Briefly, after removing the preservative media, microbial DNA was isolated with the Qiagen DNeasy PowerSoil Pro kit, according to manufacturer instructions. After the quality control procedures using Qubit 2.0, agarose gel electrophoresis and Agilent 2100, extracted DNA was subjected to DNA libraries construction, completed through the processes of end repairing, adding A to tails, purification and PCR amplification, using Illumina DNA Prep (M) Tagmentation. Libraries were subsequently sequenced on our in-house Illumina NovaSeq ssytem (150 base pairs, paired-end). ZymoBIOMICS Microbial Community Standard (D6300, ZYMO Research) and ZymoBIOMICS Microbial Community DNA Standard (D6306) were used as positive controls during DNA extraction, library construction, sequencing and quality assessment, while resequencing was performed when abnormal signals were detected.

Sequencing data preprocessing

Raw sequence data were quality filtered using Trimmomatic (v.39) to remove the adaptor, low-quality sequences (quality score <20) and reads shorter than 50 base pairs[71]. The remaining reads were mapped to the mammalian genome (hg38, felCat8, canFam3, mm10, rn6, susScr3, galGal4 and bosTau8 were collected from UCSC Genome Browser at https://genome.ucsc.edu/) and bacterial plasmids (National Center for Biotechnology Information (NCBI) RefSeq database accessed in January 2023 at https://www.ncbi.nlm.nih.gov/refseq/), complete plastomes (NCBI RefSeq database accessed in January 2023 at https://www.ncbi.nlm.nih.gov/refseq/) and UniVec sequences (NCBI RefSeq database accessed in January 2023 at https://www.ncbi.nlm.nih.gov/refseq/) using bowtie2 (v.2.4.2)[72]; matching reads that were potentially host-associated and laboratory-associated sequences were removed as contaminant reads using KneadData v.0.6. GNU parallel (v.3.0) was used for parallel analysis jobs to accelerate data processing.

Microbial taxonomic and functional profiles

According to previous studies[59,73], the taxonomic classification of bacteria, archaea, fungi and viruses was assigned to metagenomic reads using Kraken 2 (v.2.1.2), an improved metagenomic taxonomy classifier that utilizes *k*-mer-based algorithms[74]. A custom database consisting of bacterial, archaeal and viral reference genomes from the NCBI RefSeq database (accessed in January 2023) and fungal reference genomes from the NCBI RefSeq database (accessed in January 2023), FungiDB (http://fungidb.org) and Ensemble (http://fungi.ensembl.org, accessed in January 2023) was built using Jellyfish (v.2.3.0) by counting distinct 31-mers in the reference libraries, with each *k*-mer in a read mapped to the lowest common ancestor of all reference genomes with exact *k*-mer matches[75]. Thereafter, each query was classified to a specific taxon with the highest total *k*-mer hits matched by pruning the general taxonomic trees affiliated with the mapped genomes. Bracken (v.2.5.0) was used to accurately estimate taxonomic abundance[76], especially at the species and genus level, based on Kraken 2. The read counts of species were converted into relative abundance for further analysis. Microbiome functional pathways and KO gene families were profiled using HUMAnN (v.3.0) and transformed into relative abundance before analysis. The microbiome data were transformed via the centred log-ratio (CLR) transformation using the geometric mean of relative abundances of these features as the CLR denominator to break the compositionality of the data and normalize skewed distributions of microbiome features using the R package 'compositions' (v.2.0-5)[77]. After transformation, the data distribution was assumed to be normal, but this was not formally tested.

Calculation of microbiome–phenotype associations

The proportion of variance in microbiome composition that can be explained by phenotypes was simultaneously calculated by permutational multivariate analysis of variance (PERMANOVA) using distance matrices (adonis) implemented in the adonis function of the R package vegan (v.2.6-4)[78] following a previous study[23]. Analysis was performed on the microbiome beta-diversity (Bray–Curtis distance matrix calculated using relative abundances of microbial species) and phenotypes using the adonis2 implementation of the PERMANOVA algorithm in R with 9,999 permutations. To calculate the proportion of microbiome functional potential explained by phenotypes, an equivalent analysis was performed on the Bray–Curtis distance matrix calculated using relative abundances of MetaCyc microbial biochemical pathways and KO genes. The total proportion of microbiome composition variance and function explained by groups of phenotypes was calculated by multivariate adonis analyses. The analyses were performed for each kingdom separately, and all factors that have a significant impact on any kingdom are presented and employed for the subsequent analysis.

The diversity of microbiome was measured using Shannon index and richness (defined by the observed number of species), and the difference in microbiome diversity between children with ASD and children considered neurotypical was calculated using MMUPHin (v.1.18.0)[79] after adjustment for the above-identified significant confounders. The associations between each microbiome feature (microbial taxa at species level, MetaCyc pathways and KO genes) and ASD were calculated using MaAsLin2 (v.1.4.0)[80], which relies on general linear models and offers a variety of data exploration, normalization and transformation methods. To correct for potential confounders, the association analysis included all phenotypes that showed significant association ($P < 0.05$) with microbiome composition or function in the above analyses.

Subcohort of children considered neurotypical or with ASD

To avoid prediction bias driven by the imbalanced sample size of children with ASD ($n = 709$) and children considered neurotypical ($n = 374$) in the original discovery cohort and to account for potential confounders, we constructed a fully paired subcohort with equal number of cases and controls for marker selection and model training. The method used for building such a fully paired cohort with a balanced sample size of children considered neurotypical and with ASD was adopted from previous studies[24,35]. Briefly, the pairing algorithm was constructed on the basis of all phenotypes that showed significant association ($P < 0.05$) with microbiome composition or function in the above univariate analyses. The pairwise Euclidean distances were computed between the children considered neurotypical ($n = 709$) and participants with ASD ($n = 374$) on the basis of the above set of matching variables that were normalized to zero-mean and unit variance (centred and scaled). Subsequently, a child with ASD and the closest neurotypical child were removed from the selection group and then added to the subcohort. The selection process was successive until no children considered neurotypical remained in the selection group. If multiple ASD samples shared the closest distance with children considered neurotypical, a random ASD sample was selected and moved into the subcohort. The one-by-one pairing was done independently by sex. Finally, a total of 301 children considered neurotypical (95 girls and 206 boys) and 301 children with ASD (95 girls and 206 boys) were employed in the subcohort with a balanced number of cases and controls.

Random forest binary classifier

Machine learning binary classifier used random forest through the package randomForest (v.4.7-1.1) in R (4.1.3), as this algorithm has been shown to outperform, on average, other learning tools for microbiota data in previous studies[42,59]. Normalized abundance tables were used to train the model. Machine learning models were first trained on the randomly selected training set (70%, 5-fold cross-validation) and then applied to the withheld test set (30%) to assess performance. Then, we tuned hyperparameters (for example, mtry, ntree, nodesize, maxnodes) using the caret package (v.6.0-88) on the basis of the model performance in the test set to avoid overfitting issues. Finally, with the best combination of hyperparameters, the random split of the cohort was repeated 20 times to obtain a distribution of random forest prediction evaluations on the test set, and the mean AUC value was calculated accordingly for visualization of results. The highest-ranked and frequently selected microbial features were considered predictive signatures for further interpretation. We retrieved prediction performance for each feature using the same training datasets. The trained models were then tested in the independent validation cohorts to assess their robustness, which was assessed by AUC calculated using predicted probabilities. Considering the potentially unbalanced covariates between children with ASD and children considered neurotypical in the independent hospital cohort and community cohort, corresponding adjusted AUCs were calculated after adjustment for technical factors and available covariates including age, sex, BMI, BSBF, functional constipation and defecation disorders using the R package 'pROC' (v.1.18.5).

Public data download and processing

For the construction of the external dataset, a total of 353 raw shotgun faecal metagenomes were obtained from six independently published studies across four countries[36–41]. After removing studies with sequencing depth less than 4 GB ($n = 32$) and those without metadata ($n = 84$), a total of 237 samples were included in the final analyses. Quality filtration and species-/function-level profiling were performed according to the above standardized process. The trained model using 31 markers was then tested in the public dataset to assess its robustness as assessed by AUC calculated using predicted probabilities. Adjusted AUCs were calculated after adjustment for available covariates, including age, sex and batch effects (defined by different studies), from public datasets using the R package 'pROC' (v.1.18.5).

Statistical analysis

No statistical method was used to predetermine sample size. No data were excluded from the analyses. Statistical analyses were done using R v.4.1.3. Considering that microbial data are sparse with a non-normal distribution, relevant statistics using relative abundance were performed using the ggpubr (v.0.6.0) package (https://github.com/kassambara/ggpubr) with non-parametric tests, such as the Wilcoxon rank-sum test. Multiple hypothesis testing corrections were done using the false discovery rate (FDR) method.

Reporting summary

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

Data availability

The metagenomic sequencing data generated in this study have been deposited in the NCBI Sequence Read Archive database under accession code PRJNA943687. The publicly available raw sequencing data were downloaded through the retrieved accession numbers from cited papers, including GSE113540, PRJNA516054, PRJNA782533, CRA004105, PRJEB23052 and CRA001746. Mammalian genomes including hg38, felCat8, canFam3, mm10, rn6, susScr3, galGal4 and bosTau8 are available in UCSC Genome Browser at https://genome.ucsc.edu. Bacterial plasmids, complete plastomes, UniVec sequences and reference genomes database consisting of bacterial, archaeal and viral reference genomes are available in NCBI RefSeq database at https://www.ncbi.nlm.nih.gov/refseq. Fungal reference genomes are available in NCBI RefSeq database (https://www.ncbi.nlm.nih.gov/refseq), FungiDB (http://fungidb.org) and Ensemble (http://fungi.ensembl.org). Source data are provided with this paper.

Code availability

All software used are from publicly available sources. Codes used for the microbiome analyses or figures are available via GitHub at https://github.com/qsu123/ASD_multi-kingdom_diagnosis ref. 81.

References

1. Lord, C., Elsabbagh, M., Baird, G. & Veenstra-Vanderweele, J. Autism spectrum disorder. *Lancet* **392**, 508–520 (2018).
2. Masi, A., DeMayo, M. M., Glozier, N. & Guastella, A. J. An overview of autism spectrum disorder, heterogeneity and treatment options. *Neurosci. Bull.* **33**, 183–193 (2017).
3. Matin, B. K. et al. Contributing factors to healthcare costs in individuals with autism spectrum disorder: a systematic review. *BMC Health Serv. Res.* **22**, 604 (2022).
4. Tordjman, S. et al. Gene × environment interactions in autism spectrum disorders: role of epigenetic mechanisms. *Front. Psychiatry* **5**, 53 (2014).

5. Geschwind, D. H. Autism: many genes, common pathways? *Cell* **135**, 391–395 (2008).
6. Satterstrom, F. K. et al. Large-scale exome sequencing study implicates both developmental and functional changes in the neurobiology of autism. *Cell* **180**, 568–584.e23 (2020).
7. Taniya, M. A. et al. Role of gut microbiome in autism spectrum disorder and its therapeutic regulation. *Front. Cell. Infect. Microbiol.* **12**, 915701 (2022).
8. Wang, W. & Fu, P. Gut microbiota analysis and in silico biomarker detection of children with autism spectrum disorder across cohorts. *Microorganisms* **11**, 291 (2023).
9. Wan, Y. et al. Underdevelopment of the gut microbiota and bacteria species as non-invasive markers of prediction in children with autism spectrum disorder. *Gut* **71**, 910–918 (2022).
10. Lou, M. et al. Deviated and early unsustainable stunted development of gut microbiota in children with autism spectrum disorder. *Gut* **71**, 1588–1599 (2022).
11. Liu, Z. et al. Gene variations in autism spectrum disorder are associated with alteration of gut microbiota, metabolites and cytokines. *Gut Microbes* **13**, 1–16 (2021).
12. Yap, C. X. et al. Autism-related dietary preferences mediate autism–gut microbiome associations. *Cell* **184**, 5916–5931.e17 (2021).
13. Jacobson, A., Yang, D., Vella, M. & Chiu, I. M. The intestinal neuro-immune axis: crosstalk between neurons, immune cells, and microbes. *Mucosal Immunol.* **14**, 555–565 (2021).
14. Sharon, G. et al. Human gut microbiota from autism spectrum disorder promote behavioral symptoms in mice. *Cell* **177**, 1600–1618.e17 (2019).
15. Kang, D. W. et al. Microbiota transfer therapy alters gut ecosystem and improves gastrointestinal and autism symptoms: an open-label study. *Microbiome* **5**, 10 (2017).
16. Kang, D. W. et al. Long-term benefit of microbiota transfer therapy on autism symptoms and gut microbiota. *Sci. Rep.* **9**, 5821 (2019).
17. Andreo-Martínez, P., Rubio-Aparicio, M., Sánchez-Meca, J., Veas, A. & Martínez-González, A. E. A meta-analysis of gut microbiota in children with autism. *J. Autism Dev. Disord.* **52**, 1374–1387 (2022).
18. Ho, L. K. H. et al. Gut microbiota changes in children with autism spectrum disorder: a systematic review. *Gut Pathog.* **12**, 6 (2020).
19. Bezawada, N., Phang, T. H., Hold, G. L. & Hansen, R. Autism spectrum disorder and the gut microbiota in children: a systematic review. *Ann. Nutr. Metab.* **76**, 16–29 (2020).
20. Qu, A. et al. Children with autism show differences in the gut DNA virome compared to non-autistic children: a case control study. *BMC Pediatr.* **23**, 174 (2023).
21. Weiner, A., Turjeman, S. & Koren, O. Gut microbes and host behavior: the forgotten members of the gut-microbiome. *Neuropharmacology* **227**, 109453 (2023).
22. Wan, Y. et al. Alterations in fecal virome and bacteriome virome interplay in children with autism spectrum disorder. *Cell Rep. Med.* **5**, 101409 (2024).
23. Gacesa, R. et al. Environmental factors shaping the gut microbiome in a Dutch population. *Nature* **604**, 732–739 (2022).
24. Vujkovic-Cvijin, I. et al. Host variables confound gut microbiota studies of human disease. *Nature* **587**, 448–454 (2020).
25. Gvozdjáková, A. et al. Ubiquinol improves symptoms in children with autism. *Oxid. Med. Cell. Longev.* **2014**, 798957 (2014).
26. Crane, F. L., Löw, H., Sun, I., Navas, P. & Gvozdjáková, A. Plasma membrane coenzyme Q: evidence for a role in autism. *Biologics* **8**, 199–205 (2014).
27. Zhang, G. et al. Thiamine nutritional status and depressive symptoms are inversely associated among older Chinese adults. *J. Nutr.* **143**, 53–58 (2013).
28. Vignisse, J. et al. Thiamine and benfotiamine prevent stress-induced suppression of hippocampal neurogenesis in mice exposed to predation without affecting brain thiamine diphosphate levels. *Mol. Cell. Neurosci.* **82**, 126–136 (2017).
29. Wiley, K. D. & Gupta, M. Vitamin B1 (Thiamine) Deficiency. in *StatPearls* (ed. Gupta, M.) (StatPearls Publishing, 2023).
30. Hasler, M., Fideli, Ü. S., Susi, A. & Hisle-Gorman, E. Examining the relationship between autism spectrum disorder and neural tube defects. *Congenit. Anom.* **63**, 100–108 (2023).
31. Tokatly Latzer, I. et al. Autism spectrum disorder and GABA levels in children with succinic semialdehyde dehydrogenase deficiency. *Dev. Med. Child Neurol.* **65**, 1596–1606 (2023).
32. Adak, P., Banerjee, N., Sinha, S. & Bandyopadhyay, A. K. Gamma-aminobutyric acid type A receptor variants are associated with autism spectrum disorders. *J. Mol. Neurosci.* **73**, 237–249 (2023).
33. Hollestein, V. et al. Excitatory/inhibitory imbalance in autism: the role of glutamate and GABA gene-sets in symptoms and cortical brain structure. *Transl. Psychiatry* **13**, 18 (2023).
34. Xu, Y. et al. Leveraging existing 16S rRNA microbial data to define a composite biomarker for autism spectrum disorder. *Microbiol. Spectr.* **10**, e0033122 (2022).
35. Su, Q. et al. Gut microbiome signatures reflect different subtypes of irritable bowel syndrome. *Gut Microbes* **15**, 2157697 (2023).
36. Dan, Z. et al. Altered gut microbial profile is associated with abnormal metabolism activity of Autism Spectrum Disorder. *Gut Microbes* **11**, 1246–1267 (2020).
37. Kovtun, A. S., Averina, O. V., Alekseeva, M. G. & Danilenko, V. N. Antibiotic resistance genes in the gut microbiota of children with autistic spectrum disorder as possible predictors of the disease. *Microb. Drug Resist.* **26**, 1307–1320 (2020).
38. Nirmalkar, K. et al. Shotgun metagenomics study suggests alteration in sulfur metabolism and oxidative stress in children with autism and improvement after microbiota transfer therapy. *Int. J. Mol. Sci.* **23**, 13481 (2022).
39. Tong, Z. et al. Implications of oral streptococcal bacteriophages in autism spectrum disorder. *npj Biofilms Microbiomes* **8**, 91 (2022).
40. Wang, M. et al. Alterations in gut glutamate metabolism associated with changes in gut microbiota composition in children with autism spectrum disorder. *mSystems* **4**, e00321 (2019).
41. Zhang, M. et al. A quasi-paired cohort strategy reveals the impaired detoxifying function of microbes in the gut of autistic children. *Sci. Adv.* **6**, eaba3760 (2020).
42. Su, Q. et al. Faecal microbiome-based machine learning for multi-class disease diagnosis. *Nat. Commun.* **13**, 6818 (2022).
43. Chan, C. W. H., Leung, T. F., Choi, K. C., Tsui, S. K. W. & Chan, J. Y. W. Effects of gut microbiome and environment on the development of eczema in Chinese infants. *Medicine* **99**, e20327 (2020).
44. Kang, Y., Cai, Y. & Pan, W. Change in gut microbiota for eczema: implications for novel therapeutic strategies. *Allergol. Immunopathol.* **46**, 281–290 (2018).
45. Taş, E. & Ülgen, K. O. Understanding the ADHD–gut axis by metabolic network analysis. *Metabolites* **13**, 592 (2023).
46. Wang, L., Xie, Z., Li, G., Li, G. & Liang, J. Two-sample Mendelian randomization analysis investigates causal associations between gut microbiota and attention deficit hyperactivity disorder. *Front. Microbiol.* **14**, 1144851 (2023).
47. Anwar, A. et al. Quantitation of plasma thiamine, related metabolites and plasma protein oxidative damage markers in children with autism spectrum disorder and healthy controls. *Free Radic. Res.* **50**, S85–s90 (2016).
48. Rashaid, A. H. B. et al. Profiling plasma levels of thiamine and histamine in Jordanian children with autism spectrum disorder (ASD): potential biomarkers for evaluation of ASD therapies and diet. *Nutr. Neurosci.* **26**, 842–849 (2023).

49. Obrenovich, M. E., Shola, D., Schroedel, K., Agrahari, A. & Lonsdale, D. The role of trace elements, thiamin (e) and transketolase in autism and autistic spectrum disorder. *Front. Biosci.* **7**, 229–241 (2015).
50. Liu, J. et al. Alteration of gut microbiota: new strategy for treating autism spectrum disorder. *Front. Cell Dev. Biol.* **10**, 792490 (2022).
51. Alharthi, A., Alhazmi, S., Alburae, N. & Bahieldin, A. The human gut microbiome as a potential factor in autism spectrum disorder. *Int. J. Mol. Sci.* **23**, 1363 (2022).
52. Settanni, C. R. et al. Gastrointestinal involvement of autism spectrum disorder: focus on gut microbiota. *Expert Rev. Gastroenterol. Hepatol.* **15**, 599–622 (2021).
53. Guidetti, C. et al. Randomized double-blind crossover study for evaluating a probiotic mixture on gastrointestinal and behavioral symptoms of autistic children. *J. Clin. Med.* **11**, 5263 (2022).
54. El-Mekkawy, R. M., Hamour, N. E., Hassanein, W. A. & Allam, A. A. Evaluation of the antibacterial activity of *Weissella confusa* K3 cell-free supernatant against extended-spectrum βeta lactamase (ESBL) producing uropathogenic *Escherichia coli* U60. *Saudi J. Biol. Sci.* **30**, 103595 (2023).
55. Ahmed, S. et al. The *Weissella* genus: clinically treatable bacteria with antimicrobial/probiotic effects on inflammation and cancer. *Microorganisms* **10**, 2427 (2022).
56. Kwak, S. H., Cho, Y. M., Noh, G. M. & Om, A. S. Cancer preventive potential of kimchi lactic acid bacteria (*Weissella cibaria*, *Lactobacillus plantarum*). *J. Cancer Prev.* **19**, 253–258 (2014).
57. West, K. A. et al. Multi-angle meta-analysis of the gut microbiome in autism spectrum disorder: a step toward understanding patient subgroups. *Sci. Rep.* **12**, 17034 (2022).
58. Cooper, J. R. & Pincus, J. H. The role of thiamine in nervous tissue. *Neurochem. Res.* **4**, 223–239 (1979).
59. Liu, N. N. et al. Multi-kingdom microbiota analyses identify bacterial–fungal interactions and biomarkers of colorectal cancer across cohorts. *Nat. Microbiol.* **7**, 238–250 (2022).
60. Ananthakrishnan, A. N. et al. Gut microbiome function predicts response to anti-integrin biologic therapy in inflammatory bowel diseases. *Cell Host Microbe* **21**, 603–610.e3 (2017).
61. Morton, J. T., Donovan, S. M. & Taroncher-Oldenburg, G. Decoupling diet from microbiome dynamics results in model mis-specification that implicitly annuls potential associations between the microbiome and disease phenotypes—ruling out any role of the microbiome in autism (Yap et al. 2021) likely a premature conclusion. Preprint at *bioRxiv* https://doi.org/10.1101/2022.02.25.482051 (2022).
62. Rolland, T. et al. Phenotypic effects of genetic variants associated with autism. *Nat. Med.* **29**, 1671–1680 (2023).
63. First, M. B. Diagnostic and statistical manual of mental disorders, 5th edition, and clinical utility. *J. Nerv. Ment. Dis.* **201**, 727–729 (2013).
64. Wong, P. P., Wai, V. C., Chan, R. W., Leung, C. N. & Leung, P. W. Autism-spectrum quotient-child and autism-spectrum quotient-adolescent in Chinese population: screening autism spectrum disorder against attention-deficit/hyperactivity disorder and typically developing peers. *Autism* **25**, 1913–1923 (2021).
65. Pappas, D. ADHD Rating Scale-IV: checklists, norms, and clinical interpretation. *J. Psychoeduc. Assess.* **24**, 172–178 (2006).
66. Memória, C. M. et al. Applicability of the Test of Variables of Attention - T.O.V.A in Brazilian adults. *Dement. Neuropsychol.* **12**, 394–401 (2018).
67. Hanifin, J. M. & Rajka, G. Diagnostic features of atopic dermatitis. *Acta Derm. Venereol.* **60**, 44–47 (1980).
68. Tito, R. Y. et al. Microbiome confounders and quantitative profiling challenge predicted microbial targets in colorectal cancer development. *Nat. Med.* **30**, 1339–1348 (2024).
69. Chumpitazi, B. P. et al. Bristol Stool Form Scale reliability and agreement decreases when determining Rome III stool form designations. *Neurogastroenterol. Motil.* **28**, 443–448 (2016).
70. Chen, Z. et al. Impact of preservation method and 16S rRNA hypervariable region on gut microbiota profiling. *mSystems* **4**, e00271-18 (2019).
71. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
72. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
73. Lu, J. et al. Metagenome analysis using the Kraken software suite. *Nat. Protoc.* **17**, 2815–2839 (2022).
74. Wood, D. E., Lu, J. & Langmead, B. Improved metagenomic analysis with Kraken 2. *Genome Biol.* **20**, 257 (2019).
75. Marçais, G. & Kingsford, C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* **27**, 764–770 (2011).
76. Lu, J., Breitwieser, F. P., Thielen, P. & Salzberg, S. L. Bracken: estimating species abundance in metagenomics data. *PeerJ Comput. Sci.* **3**, e104 (2017).
77. van den Boogaart, K. G. & Tolosana-Delgado, R. "compositions": a unified R package to analyze compositional data. *Comput. Geosci.* **34**, 320–338 (2008).
78. Dixon, P. VEGAN, a package of R functions for community ecology. *J. Veg. Sci.* **14**, 927–930 (2003).
79. Ma, S. et al. Population structure discovery in meta-analyzed microbial communities and inflammatory bowel disease using MMUPHin. *Genome Biol.* **23**, 208 (2022).
80. Mallick, H. et al. Multivariable association discovery in population-scale meta-omics studies. *PLoS Comput. Biol.* **17**, e1009442 (2021).
81. Su, Q. ASD_multi-kingdom_diagnosis. *GitHub* https://github.com/qsu123/ASD_multi-kingdom_diagnosis (2024).

Acknowledgements

We thank F. Mo, C. Shea and C. Ho for facilitating the clinical participants' enrolment. This study was supported by InnoHK (F.K.L.C, S.C.N.), the Government of Hong Kong, Special Administrative Region of the People's Republic of China, The D. H. Chen Foundation (F.K.L.C., S.C.N.), and the New Cornerstone Science Foundation through the New Cornerstone Investigator Program (S.C.N.).

Author contributions

Q.S. and O.W.H.W. conceived the study, ran analyses and drafted the paper. W.L., Y.W., L.Z., W.X., M.K.T.L., C.L. and T.F.L. contributed to part of metagenomic sequencing. C.P.C., J.Y.L.C. and P.K.C. contributed to participant recruitment, sample collection and biobank management. S.C. and P.L. contributed to participant recruitment and clinical assessment. F.K.L.C. contributed to the study design and data interpretation. S.C.N. oversaw the entire study and contributed to the study design, data analysis and interpretation, and paper writing. All authors gave final approval for the version to be published.

Competing interests

F.K.L.C. is a Board Member of CUHK Medical Centre. He is a co-founder, non-executive Board Chairman, honorary Chief Medical Officer and shareholder of GenieBiome Ltd. He receives patent royalties through his affiliated institutions. He has received fees as an advisor and honoraria as a speaker for Eisai Co. Ltd., AstraZeneca, Pfizer Inc., Takeda Pharmaceutical Co., and Takeda (China) Holdings Co. Ltd. S.C.N. has served as an advisory board member for Pfizer, Ferring, Janssen, and Abbvie and received honoraria as a speaker for Ferring, Tillotts, Menarini, Janssen, Abbvie, and Takeda. She has received research grants through her affiliated institutions from

Olympus, Ferring, and Abbvie, and is a scientific co-founder and shareholder of GenieBiome Ltd. She also receives patent royalties through her affiliated institutions. Q.S., L.Z., Y.W., F.K.L.C. and S.C.N. are named inventors of patent applications (no. 63/533,871, US, 2023) held by the CUHK and MagIC that cover the therapeutic and diagnostic use of microbiome. The remaining authors declare no competing interests.

Additional information

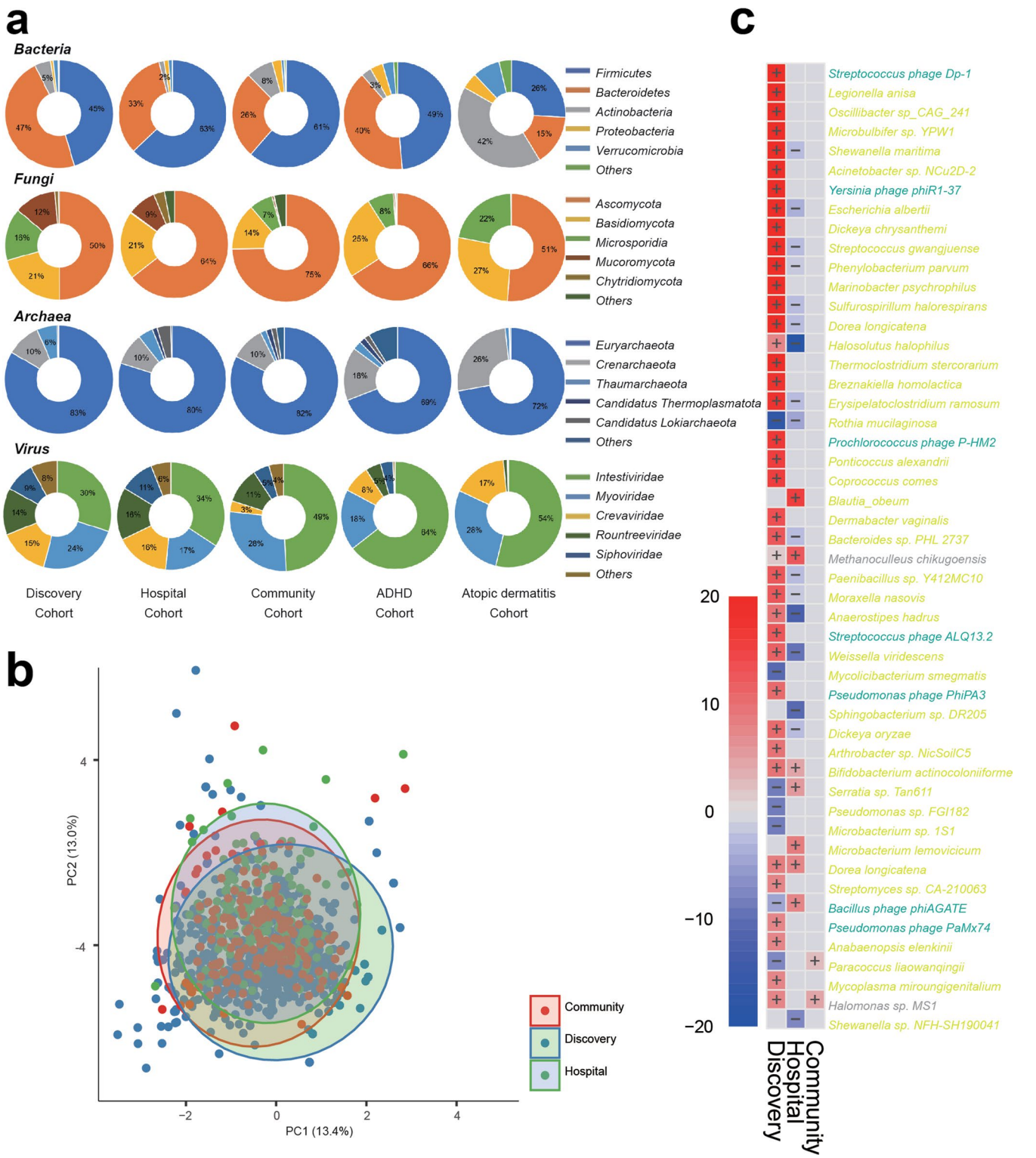
Extended data is available for this paper at https://doi.org/10.1038/s41564-024-01739-1.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41564-024-01739-1.

Correspondence and requests for materials should be addressed to Siew C. Ng.

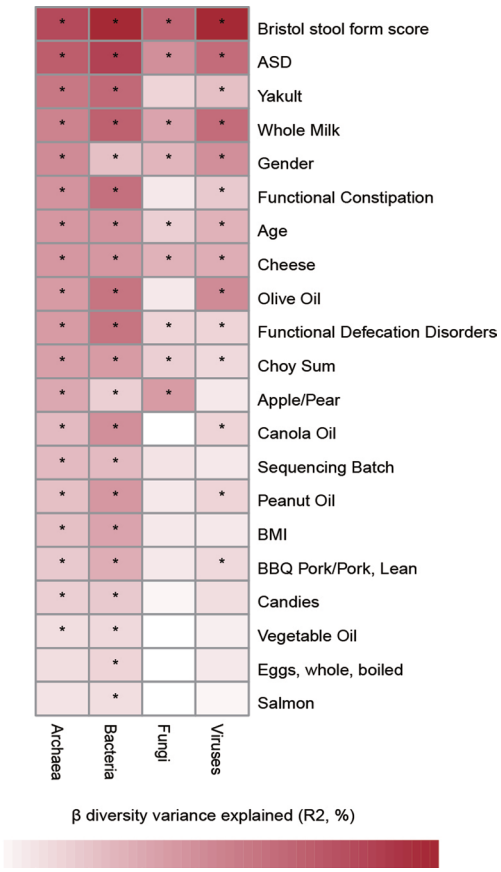
Peer review information *Nature Microbiology* thanks Ivan Vujkovic-Cvijin and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

https://doi.org/10.1038/s41564-024-01739-1



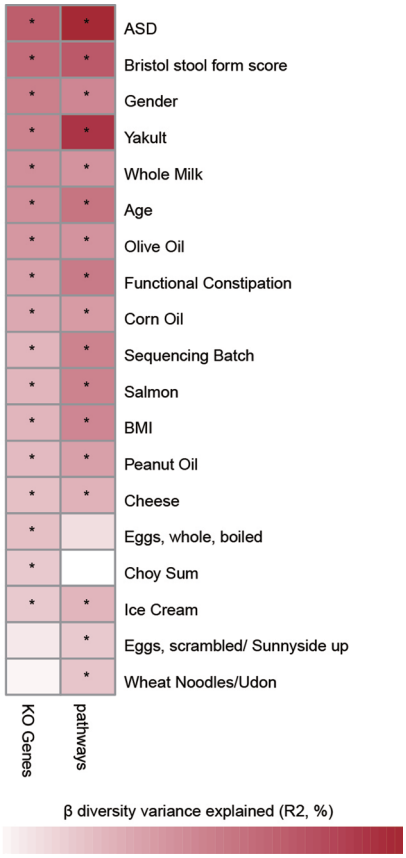
Extended Data Fig. 1 | Overview of microbial composition in four kingdoms across cohorts. a, Composition of archaea, bacteria and fungi was shown at phylum level and composition of virus was shown at family level. Top 5 abundant phyla or families are shown in the pie chart and others are summed into others. **b**, Principal components analysis of gut microbiome composition across different cohorts. **c**, Differential taxa across different cohorts identified by MaAsLin2.

https://doi.org/10.1038/s41564-024-01739-1



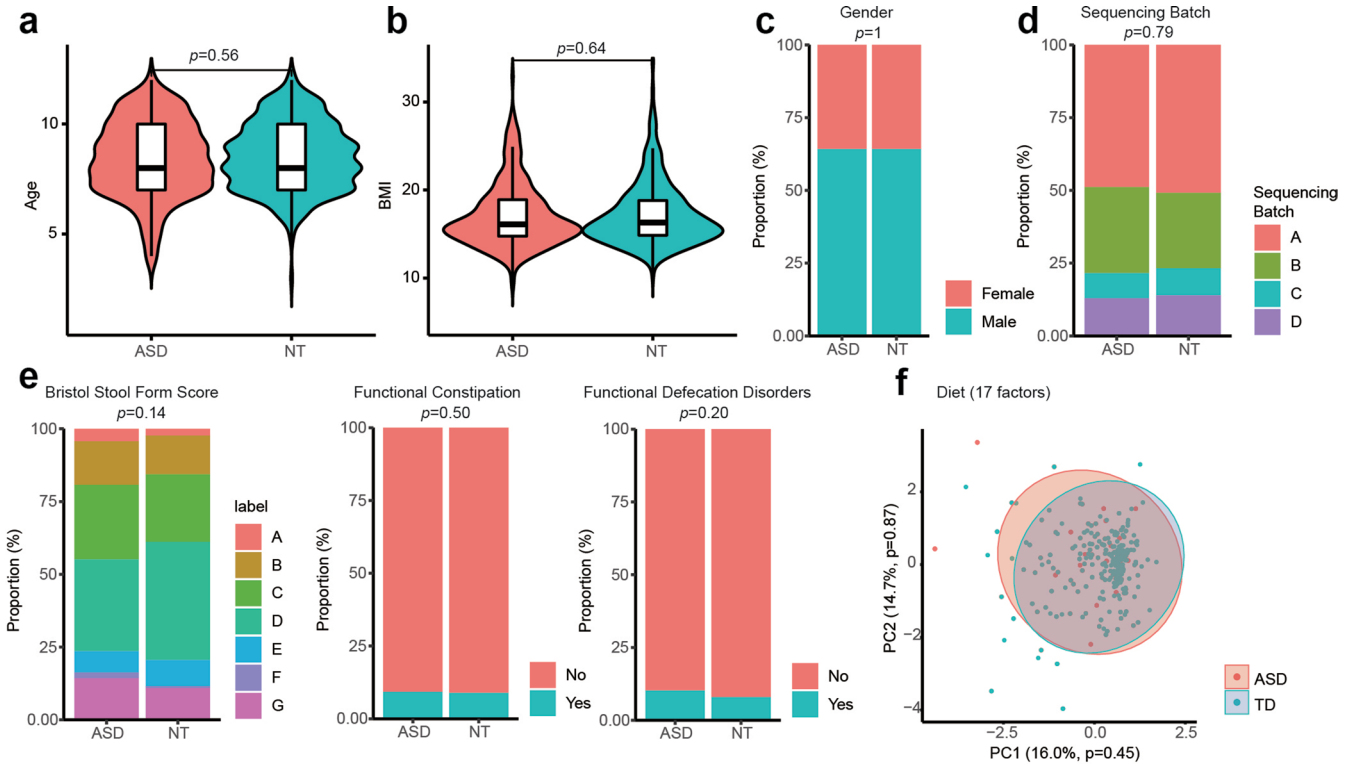
Extended Data Fig. 2 | The influence of individual phenotypic factor on the multi-kingdom microbiome (archaea, bacteria, fungi and viruses) composition assessed by multivariate PERMANOVA analysis (two-sided test) in the discovery cohort (709 ASD and 374 NT). Phenotypic factors with a significant ($p < 0.05$) influence on each kingdom were marked with *. ASD, autism spectrum disorder; NT, neurotypical.

https://doi.org/10.1038/s41564-024-01739-1



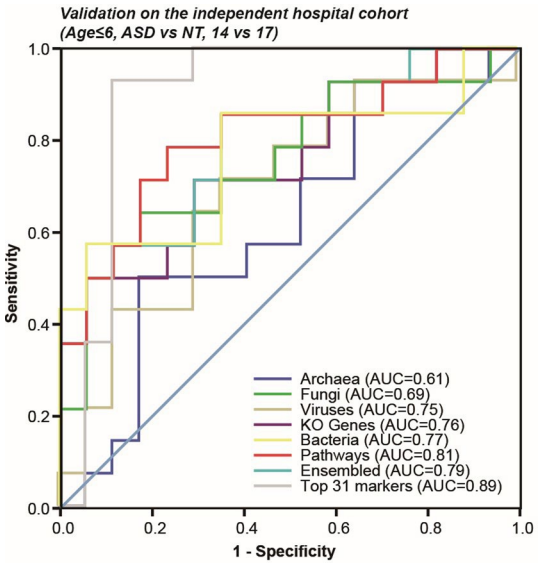
Extended Data Fig. 3 | The influence of individual phenotypic factor on the microbiome function (KO genes and MetaCyc pathways) assessed by multivariate PERMANOVA analysis (two-sided test) in the discovery cohort (709 ASD and 374 NT). Phenotypic factors with a significant ($p < 0.05$) influence on each kingdom were marked with *. KO, Kyoto Encyclopedia of Genes and Genomes orthology. ASD, autism spectrum disorder; NT, neurotypical.

https://doi.org/10.1038/s41564-024-01739-1



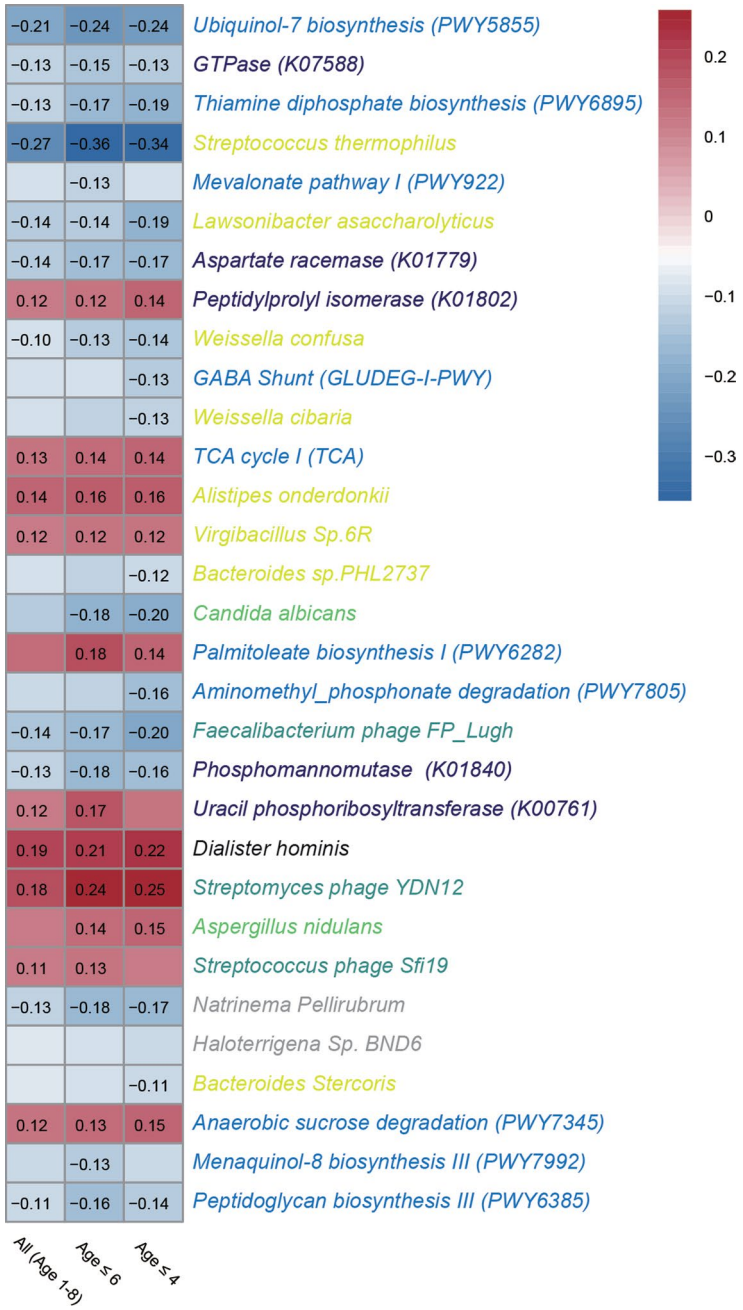
Extended Data Fig. 4 | Matched confounders between ASD (n = 301) and NT (n = 301) in the balanced sub-cohort constructed by one-by-one pairing algorithm, including age (a), BMI (b), sex (c), sequencing batch (d), GI parameters (e) and dietary factors (f). Continuous data were compared using two-sided Mann-Whitney U test (two-sided test) and shown via the median (centre line), 25th and 75th percentiles (box limits) and the 5th and 95th percentiles (whiskers). Categorical variables were presented as proportions and were compared using two-sided Chi-squire test or Fisher's exact test (expected count<5). Dietary patterns were assessed by principal components analysis. BMI, body mass index; GI, gastrointestinal. ASD, autism spectrum disorder; NT, neurotypical.

https://doi.org/10.1038/s41564-024-01739-1



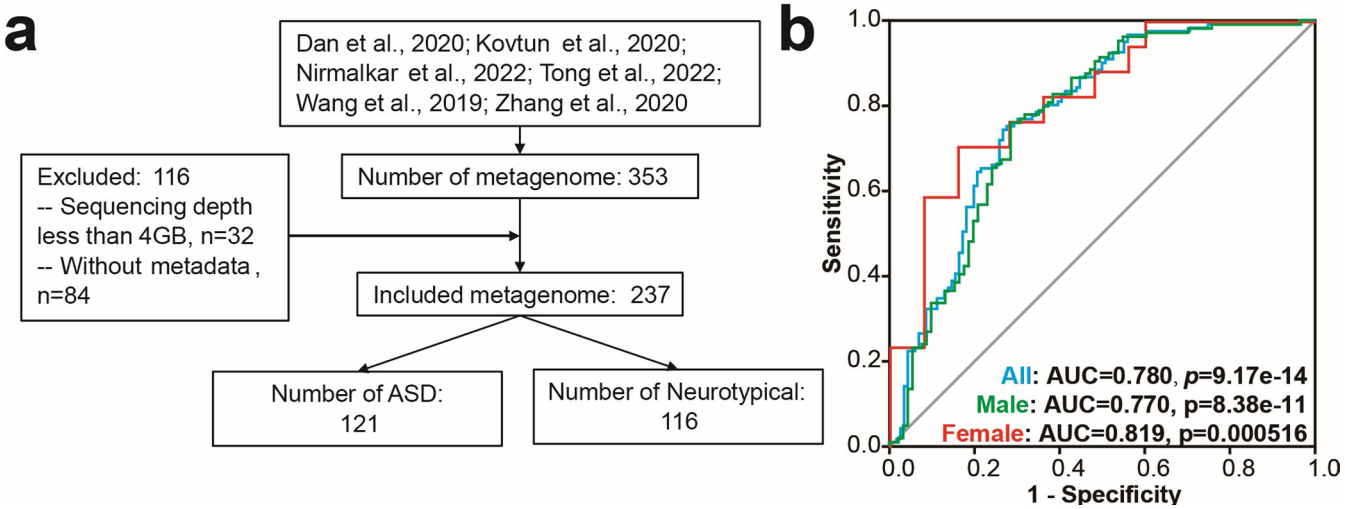
Extended Data Fig. 5 | Validation of trained models in children under 6 years of age (14 ASD and 17 NT) in the independent hospital cohort. AUC were calculated after adjustment for technical factors and available covariates including age, sex, body mass index, Bristol stool form scale, functional constipation and defecation disorders. ASD, autism spectrum disorder; NT, neurotypical.

https://doi.org/10.1038/s41564-024-01739-1



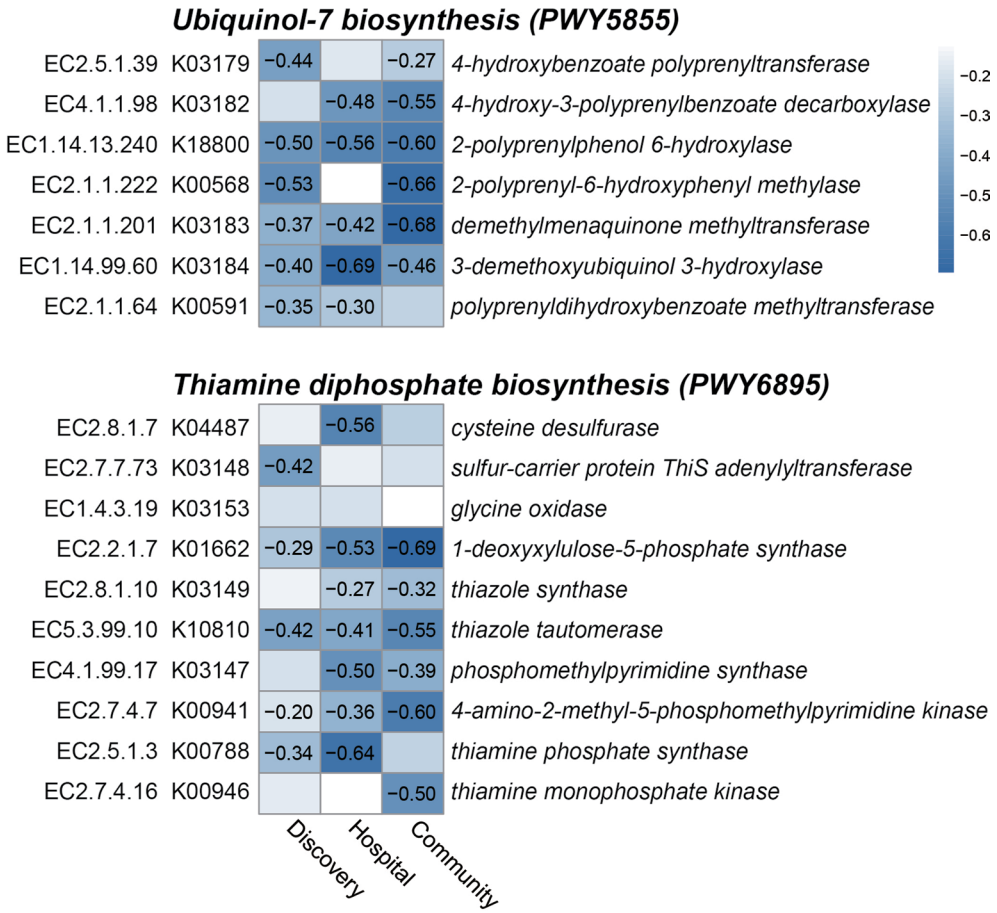
Extended Data Fig. 6 | Associations between ASD and 31 markers in the independent community cohort (stratified by age) assessed by MaAsLin 2 (two-sided test). The Coef value of each association was only marked when p value less than 0.05. ASD, autism spectrum disorder; NT, neurotypical.

https://doi.org/10.1038/s41564-024-01739-1



Extended Data Fig. 7 | Validation of model using 31 markers on the public dataset. a, Construction of an external validation cohort of ASD from six published studies. **b**, Area under the Curve of model using 31 markers tested in the public dataset. AUC were calculated after adjustment for available covariates from public datasets including age, sex, and batch effects (defined by different studies). P values were calculated by Wilcoxon rank-sum test (two-sided test). ASD, autism spectrum disorder; NT, neurotypical.

https://doi.org/10.1038/s41564-024-01739-1



Extended Data Fig. 8 | Decreased abundance of ubiquinol-7 and thiamine diphosphate biosynthesis genes in ASD children across different cohorts. Associations between ubiquinol-7 and thiamine diphosphate biosynthesis genes and ASD were assessed by MaAsLin 2 (two-sided test). The Coef value of each association was only marked when p value less than 0.05. ASD, autism spectrum disorder; NT, neurotypical.

https://doi.org/10.1038/s41564-024-01739-1

Extended Data Table 1 | Demographics of subjects recruited in this study

Cohorts	Group	n	Sex	Age, ave (IQR)
Discovery	ASD	709	Female, 13.4%	8 (6-9)
	NT	374	Female, 44.9%	9 (8-10)
Hospital	ASD	82	Male	8 (5-11)
	NT	90	Male	8 (5-11)
Community	ASD	116	Female, 27.6%	3 (2-6)
	NT	60	Female: 33.3%	3 (2-6)
ADHD	ADHD	42	Female, 26.2%	8 (7-10)
	NT	76	Female, 44.8%	5 (3-7)
Atopic dermatitis	Case	40	Female, 47.5%	1 (1-1)
	Control	38	Female, 47.4%	1 (1-1)

Notes: ASD, autism spectrum disorder; NT, neurotypical; ADHD, attention deficit hyperactivity disorder; ave, average; IQR, interquartile range.

https://doi.org/10.1038/s41564-024-01739-1

Extended Data Table 2 | Validation of model using 31 markers on the independent community cohort

	All (Age 1-8)	Age≤6	Age≤4
ASD, n	116	88	46
NT, n	60	42	27
AUC	0.89	0.91	0.91
Sensitivity	75.9%	90.9%	86.9%
Specificity	86.7%	83.7%	85.2%
Accuracy	79.5%	88.6%	86.3%
False Positive Rate	13.3%	16.2%	14.8%
False Negative Rate	24.1%	9.1%	13.0%
Positive Prediction Value	91.7%	91.9%	90.9%
Negative Prediction Value	65.0%	81.8%	79.3%

Notes: ASD, autism spectrum disorder; NT, neurotypical; AUC, area under curve. AUC were calculated after adjustment for technical factors and available covariates including age, gender, BMI, BSBF, functional constipation and defecation disorders.

nature portfolio

Corresponding author(s): Siew C Ng

Last updated by author(s): May 5, 2024

Reporting Summary

Nature Portfolio 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 Portfolio 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	For full details see Methods; Bcl2fastq (v2.20), Trimmomatic (v39), Bowtie2 (v2.4.2), KneadData (v0.6), Kraken 2 (v2.1.2), Bracken (v2.5.0), HUMAnN (v3.0), GNU parallel (v3.0) were used to process the microbiome sequencing data. R packages including compositions (v2.0-5), vegan (v2.6-4), randomForest (v4.7-1.1), pROC (v1.18.5), ggpubr (v0.6.0), MaAsLin2 (v1.4.0), MMUPHin (v1.18.0), caret (v6.0-88) were used to analyze the microbiome profiling data.
Data analysis	All software used were from publicly available sources. Open-source codes and scripts used for the microbiome analyses or figures are available at the GitHub repository (https://github.com/qsu123/ASD_multi-kingdom_diagnosis).

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 Portfolio 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 description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

The metagenomic sequencing data generated in this study have been deposited in the NCBI Sequence Read Archive database under accession code PRJNA943687. The public available raw sequencing data were downloaded through the retrieved accession numbers from cited papers, including GSE113540, PRJNA516054, PRJNA782533, CRA004105, PRJEB23052 and CRA001746. Mammalian genome including hg38, felCat8, canFam3, mm10, rn6, susScr3, galGal4 and bosTau8 were available in UCSC Genome Browser at https://genome.ucsc.edu. Bacterial plasmids, complete plastomes, UniVec sequences and reference genomes database consisting of bacterial, archaeal and viral reference genomes were available in NCBI RefSeq database at https://www.ncbi.nlm.nih.gov/refseq. Fungal reference genomes were available in NCBI RefSeq database (https://www.ncbi.nlm.nih.gov/refseq), FungiDB (http://fungidb.org) and Ensemble (http://fungi.ensembl.org). Source data are provided with this paper.

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	A total of 1,627 children (aged 1-13 years, 24.4% female) from five independent cohorts were recruited in this study. The information was collected during clinical interviews and verified from the clinical management system (CMS) of hospital authority in Hong Kong
Reporting on race, ethnicity, or other socially relevant groupings	No socially relevant groupings were involved in this study.
Population characteristics	We reported this information in Extended Data Table 1, Figure 1a, Extended Figure 2 and 3. Extensive phenotype data (236 factors) were collected and analyzed . 24 factors including age, gender, BMI, sequencing batch, Bristol Stool Form Score, a diagnosis of functional constipation or functional defecation disorders and 17 dietary factors, including beverages (Yakult), dairy products (cheese and whole milk), eggs (whole boiled eggs and scrambled/sunnyside up eggs), fruits (apple/pear), grains (wheat noodles/udon), meats (barbeque pork/lean pork, salmon), cooking oil (peanut oil, olive oil, corn oil, vegetable oil and canola oil), snacks (candies and ice cream) and vegetables (choy sum) were treated as covariates in analysis.
Recruitment	Discovery ASD cohort ASD and neurotypical children aged less than 12 years of age were recruited from the Child and Adolescent Psychiatric Clinic of the New Territory East Cluster (NTEC) of the Hospital Authority from Dec-2021 to Dec-2023. The NTEC serves one of the largest populations in Hong Kong, of which the Child and Adolescent Psychiatric Clinic receives constant referrals from the Child Assessment Centre, school education psychologists and private medical doctors for the assessment and treatment of children with ASD. The diagnosis of ASD was made by child psychiatrists based on the Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition (DSM-5) diagnostic criteria64. Neurotypical children that were matched on age and sex to children with ASD were invited from participants of an ongoing territory-wide psychiatric epidemiological study (CREC Ref. No.: 2018.497) in the same Child and Adolescent Psychiatric Clinic where over 6000 random community-based youths were assessed on the presence of psychiatric disorders using the Diagnostic Interview Schedule for Children (DISC-5). To exclude the possibility of ASD from these community subjects, neurotypical children were first screened using the Chinese Autism Spectrum Quotient Child Version65 (AQ-Child, for children up to 11 years old). Exclusion criteria for neurotypical children included a positive family history of ASD amongst their first-degree relatives, screened positive in the AQ-Child, and the presence of psychiatric disorder according to the DISC-5. Exclusion criteria for both groups included mental retardation, neurological disorders, psychosis, depressive disorders or other major medical illness, and avoidant/restrictive food intake disorder assessed by a 3-day food record questionnaire. Subjects using probiotics, antidepressants, anti-epileptics and those who had exposure to antibiotics within one month before entering the study were also excluded. Children's parents or legal guardians were informed of the nature of the study and written informed consent was obtained. The study protocol followed the Declaration of Helsinki. After recruitment, parents/guardian-filled questionnaires and medical records were used to profile each participant in the following areas: 1) Social demographics: family composition, education level of parents, household income, number of siblings; 2) Physical parameters of body height and weight; 3) Presence of common physical diseases in children such as atopic dermatitis and asthma (n=14); 4) Common co-occurring psychiatric disorders such as ADHD (n=7); 5) The presence of functional gastrointestinal disorder was defined according to the Rome IV Criteria: including three broad categories of functional nausea and vomiting disorder (FNVD), functional abdominal pain disorders (FAPD) and functional defecation disorders (FDD) and stool consistency assessed by the BSFS (n=4); 6) medication history including the use of psychiatric medications (n=7); 7) parental parameters (n=5) and 8) dietary patterns (assessed by food consumption within three months, n=201). The study was approved by the Joint Chinese University of Hong Kong of New Territories East Cluster Clinical Research Ethics Committee (CUHK-NTEC CREC, Ref. No.: 2021.550). Finally, 709 ASD and 374 neurotypical children (24.3% female) of Chinese ethnicity aged between 3 and 12 years old were recruited. Independent hospital ASD cohort To independently validate the performance of the selected markers and trained models, we performed metagenomics

analysis on an independent historical hospital cohort which was recruited in 2018 from the Child and Adolescent Psychiatric Clinic and consisted of 172 boys (90 neurotypical children and 82 subjects with ASD, 4-11 years old). The neurotypical boys were screened negative on the AQ-Child with parent-reported absence of psychiatric diagnosis. Otherwise, the inclusion and exclusion criteria were the same as the discovery cohort. It was also ascertained by the research team that there was no overlapping of subjects between the historical cohort and the discovery cohort as they were recruited from the same Psychiatric Clinic. Written consent was obtained from the children's parents. The study was approved by the Joint CUHK-NTEC CREC (Ref. No.: 2018.497).

Independent community ASD cohort
To validate the accuracy of the selected markers and trained models for ASD detection in the community setting, an independent community cohort consisting of ASD and Neurotypical children of younger ages was recruited from Mar-2022 to Dec-2023. These children were recruited from the families of service users in a community-based educational support programme. Other than a self-reported diagnosis of ASD, the parents had to submit a formal medical certificate from a qualified psychiatrist, pediatrician or psychologist to confirm the ASD diagnosis before the child was recruited as the case group. Neurotypical children not related to the case group were recruited as the control group according to the same inclusion and exclusion criteria as above. Written consent was obtained from the children's parents. The study was approved by the Joint CUHK-NTEC CREC (Ref. No.: 2021.146). Finally, 116 ASD and 60 neurotypical children (29.5% female) of Chinese ethnicity aged between 3 and 12 years old were recruited.

Independent ADHD cohort
To validate the specificity of the selected markers and trained models for ASD detection, an independent cohort consisting of children with ADHD and neurotypical children was recruited from the community under the same study and sampling frame as the independent ASD community cohort. Subjects with a confirmed diagnosis of ADHD based on the ADHD rating scale (ADHD-RS)66 and Test of Variables of Attention (T.O.V.A)67 were recruited as the case group. Neurotypical children not related to the case group were recruited as the control group according to the same inclusion and exclusion criteria as above. Written consent was obtained from the children's parents. The study was approved by the Joint CUHK-NTEC CREC (CREC Ref. No.: 2021.146). Finally, 42 ADHD and 76 neurotypical children (38.1% female) of Chinese ethnicity aged between 1 and 13 years old were recruited.

Independent atopic dermatitis cohort
To validate the specificity of the selected markers and trained models for ASD detection, an independent cohort consisting of children with or without atopic dermatitis was recruited. Children were clinically examined by a pediatrician for signs of atopic dermatitis by pediatricians based on Hanifin and Rajka criteria68, and those with a diagnosis of atopic dermatitis were recruited as the case group. Children without any sign of atopic dermatitis were recruited as the control group. Written consent was obtained from the children's parents. The study was approved by the Joint CUHK-NTEC CREC (CREC Ref. No.: 2016.637). Finally, 40 children with atopic dermatitis and 38 controls (47.4% female, 1 year old) of Chinese ethnicity were recruited.

HKD200 worth supermarket coupons were given to each participant.

Ethics oversight

The study was approved by The Joint Chinese University of Hong Kong – New Territories East Cluster Clinical Research Ethics Committee (The Joint CUHK-NTEC CREC). All subjects provided written informed consent.

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

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

No statistical methods were used to pre-determine sample sizes, but our sample sizes are similar to those reported in previous publications.

Liu, N.N., et al. Multi-kingdom microbiota analyses identify bacterial-fungal interactions and biomarkers of colorectal cancer across cohorts. Nat Microbiol 7, 238-250 (2022).
Tito, R.Y., et al. Microbiome confounders and quantitative profiling challenge predicted microbial targets in colorectal cancer development. Nature Medicine (2024).

Data exclusions

1,627 samples were successfully sequenced and passed the quality assessment (read depth > 10 million), thus none were excluded from the analyses.

Replication

Machine learning models were first trained on the randomly selected training set (70%, 5-fold cross-validation) and then applied to the withheld test set (30%) to access the performance. The random split of the cohort was repeated 20 times to obtain a distribution of random forest prediction evaluations on the test set, and the mean AUC value was calculated accordingly for visualization of results. The trained models were then tested in the independent validation cohorts to assess their robustness.

For independent validation using publicly available datasets, we integrated 237 shotgun stool metagenome data from 6 published studies

	from 4 countries, and our trained model showed satisfactory performance. All attempts at replication were successful.
Randomization	The random split of the cohort was repeated 20 times to obtain a distribution of random forest prediction evaluations on the test set, and the mean AUC value was calculated accordingly for visualization of results. The trained models were then tested in the independent validation cohorts to assess their robustness.
Blinding	The conventional blinding (as used in clinical trials or intervention studies) was not relevant for this study because this study did not include any interventions.

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
☒	☐ Clinical data
☒	☐ Dual use research of concern
☒	☐ Plants

Methods

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