AutismCRC

Development and validation of systems genomics-based predictors for autism

FINAL REPORT

Dr Jacob Gratten
Chloe Yap
Anjali Henders
Leanne Wallace
Tiana McLaren
Professor Naomi Wray

June 2022

The University Of Queensland AUSTRALIA

mater research

AutismCRC Biobank

Australian Government
Department of Industry, Science, Energy and Resources

AusIndustry Cooperative Research Centres Program

Development and validation of systems genomics-based predictors for autism

FINAL REPORT

Dr Jacob Gratten
Mater Research | Autism CRC

Chloe Yap
University of Queensland | Autism CRC

Anjali Henders
University of Queensland | Autism CRC

Leanne Wallace
University of Queensland | Autism CRC

Tiana McLaren
University of Queensland | Autism CRC

Professor Naomi Wray
University of Queensland | Autism CRC

Project number: 1.042RC

ISBN: 978-1-922365-43-9

Citation: Gratten, J., Yap, C.X. Henders, A.K, Wallace, L., McLaren, T., Wray, N.R. (2022) Development and validation of systems genomics-based predictors for autism (Stage 1): Final Report: Brisbane, Autism CRC.

Copies of this report can be downloaded from the Autism CRC website **autismcrc.com.au**.

Acknowledgements

The authors acknowledge the financial support of the Cooperative Research Centre for Living with Autism (Autism CRC), established and supported under the Australian Government's Cooperative Research Centre Program, in addition to Mater Foundation and the National Health and Medical Research Council. Staff and non-staff in kind were provided by Autism CRC essential participant The University of Queensland (UQ). We sincerely thank all the families who generously contributed their time, data and biological samples to the Australian Autism Biobank, without whom this study would not have been possible. We thank all the members of the UQ Human Studies Unit, including Jolene Berry, Vandhana Bharti, Gemyka English, Shivangi Wani, Hailey Miranda, Jasmine Connell and Lorelle Nunn, along with all those who have contributed to establishing the Australian Autism Biobank, including Andrew Whitehouse, Valsamma Eapen, Cheryl Dissanayake, Helen Heussler, Paul A. Dawson, Dr Felicity Rose, Peter M. Visscher, Gail Alvares, Dominique Cleary, Claire Hafekost, Alexis Harun, Jodie Leslie, Rachel Grove, Helen Holdsworth, Mira Levis Frenk, Michaela Nothard, Rachel Jellett, Lauren P. Lawson, Melanie Muniandy, Feroza Khan, Anne Masi, Nisha E. Mathew, Dr Natalie Silove and the Child Development Unit team at Sydney Children's Hospital, Westmead, the staff at KU Marcia Burgess Autism Specific Early Learning and Care centre and KU Children's Services, Dr Nancy Sadka from the Olga Tennison Autism Research Centre at La Trobe University, and Dr Anja Ravine and the pathology services at Pathwest for their contributions in collecting and processing blood samples in Perth. Thanks also to the contributing institutions in the Australian Autism Biobank: University of Western Australia, Telethon Kids Institute, PathWest, University of New South Wales, Sydney Children's Hospital Network, LaTrobe University, Mater Medical Research Institute and Wesley Medical Research.

The Cooperative Research Centre for Living with Autism (Autism CRC)

The Cooperative Research Centre for Living with Autism (Autism CRC) is the world's first national, cooperative research effort focused on autism. Taking a whole-of-life approach to autism focusing on diagnosis, education and adult life, Autism CRC researchers are working with end-users to provide evidence-based outcomes which can be translated into practical solutions for governments, service providers, education and health professionals, families and people on the autism spectrum.

A note on terminology

We recognise that when referring to individuals on the autism spectrum, there is no one term that suits all people. In our published material and other work, when speaking of adults we use the terms 'autistic person', 'person on the autism spectrum' or 'person on the spectrum'. The term 'autistic person' uses identity first language, which reflects the belief that being autistic is a core part of a person's identity.

Autism Spectrum Disorder (ASD) is diagnostic terminology used by the healthcare sector, and is used in the context of a person being 'diagnosed with Autism Spectrum Disorder'.

Table of contents

1. Introduction

1.1 Background and Rationale

Autism is a highly heterogeneous condition, both clinically and genetically, which has complicated efforts to understand underlying causes and to identify diagnostic biomarkers. An important unresolved question in autism research is whether genomic information can assist in early and accurate diagnosis of autism. The objectives of this strategic project, commissioned in late 2015, were to receipt and process biospecimens for the Australian Autism Biobank ("Biobank"), derive and evaluate systems genomics-based predictors for autism using Biobank samples, and to integrate these with a clinical diagnostic instrument based on behavioural surveillance.

The strategy we proposed was to combine a genetic predictor derived from genome-wide association study (GWAS) summary data with one or more non-genetic predictors derived from genome-wide data on DNA methylation, gut metagenomics and metabolomics. The rationale for this approach is that although heritability (i.e., defined as the proportion of variance between individuals in the population for propensity for a diagnosis explained by genetic factors) estimates for autism are high (i.e., ~60-80%) – indicating that inherited genetic factors make a large and important contribution to the likelihood of autism diagnosis – because autism is relatively rare in the population (~1%), even a perfect genetic predictor (which we do not yet have) can never be a reliable diagnostic predictor. For this reason, predictors that potentially capture information on autism-relevant environmental exposures, such as those derived from DNA methylation (a type of epigenetic mark on DNA that alters gene expression), the gut microbiome or metabolites, may enable more accurate prediction when combined with a genetic predictor.

Genetic factors make an important contribution to variation in the likelihood that a person receives a diagnosis of autism spectrum disorder (ASD). This fact has driven interest in genetic studies of autism, because improved understanding of the genes involved may help to improve early and accurate diagnosis. Genetic factors associated with autism include rare and *de novo* (i.e., newly arising in the parental germ cells) variations that impact gene function, as well as genetic variations that are common in the broader population and that predominantly influence when and the extent to which autism-related genes are turned off and on. Rare autism-associated genetic variants can have a large impact on the probability of having autism in those that carry them, whereas common variations have very small effects individually. However, recent studies suggest that cumulatively, these common genetic factors are more important in autism than rare and *de novo* gene-altering mutations, which have formerly received most attention in autism research.

One of the most successful experimental designs for identifying common genetic variations associated with complex conditions (that is, those with genetic and environmental contributions) such as autism is the genome-wide association study (GWAS). This study design uses cost-effective and reliable microarray technology to characterise the genome, in a large number of people with and without a condition, at hundreds of thousands of different genetic markers (referred to as single nucleotide polymorphisms or “SNPs”) that are known to vary between people in the population. The frequency of each variation in groups on the spectrum and not on the spectrum is then compared using a statistical test, and those variations with differences that are highly unlikely to have been observed by chance are reported to be associated. When we perform a GWAS of autism, genetic variations that are identified are markers telling us that nearby genes are involved in the development of autism symptoms. Very large GWAS for autism have been performed by the international Psychiatric Genomics Consortium (PGC) – e.g., the latest study included 18,381 autistic individuals and 27,969 individuals without autism, with still larger studies planned for the future. Using the results from PGC GWAS for autism it is possible to generate polygenic scores (PGS) that summarise the estimated effect of common autism-associated genetic markers. Given sufficiently large PGC GWAS for autism, PGS’s may help to stratify individuals in the Australian Autism Biobank (AAB) into those with low likelihood of a diagnosis versus a higher likelihood. Analysis of data from SNP arrays can also identify large copy number variations (CNVs), some of which have been previously associated with autism.

The emphasis on genetic research in autism means that autism-related environmental factors are less well understood. Environmental exposures are typically investigated using epidemiological approaches, with varying levels of evidence from these studies for a contribution of advanced parental age, premature birth, very low birth weight, birth complications and maternal hypertension to autism diagnosis. An alternative approach to investigating environmental factors in autism is to characterise patterns of DNA methylation (a type of epigenetic modification) in the genome, which can be influenced by both genetics and environmental factors. DNA methylation in gene promoter and repressor regions influences the extent to which genes are turned on and off. For this reason, studying variation in DNA methylation is one potential approach to assessing the contribution of environmental factors to an autism diagnosis. Methods using array-based technology now enable cost-effective screening of hundreds of thousands of DNA methylation sites in the genome. The application of these methods to perform methylome-wide association studies (MWAS) in autism is in its infancy, but larger studies are expected to be published in the coming years. These studies have the potential to identify individual genes with differential methylation in autism, potentially due to diagnosis-related environmental exposures. It may also be possible, using an approach similar

to polygenic scoring (see above), to calculate polymethylation scores (PMS) in the AAB, which may capture environmental exposures relevant to autism.

A specific environmental factor engendering substantial interest in the autism community is the gut microbiome, which refers to the highly diverse community of bacteria and other microorganisms that inhabit the gastrointestinal tract. There is substantial evidence supporting an important role of this community in many aspects of human health, including development and maintenance of the immune system and potentially also brain-related phenotypes. Interest in the gut microbiome in autism stems primarily from three sources: first, epidemiological studies showing that rates of gastrointestinal conditions are high in autism; second, direct reports of autism-associated bacterial taxa from sequence-based analyses of stool-derived DNA; and third, studies in animal models, including faecal microbiota transplantation experiments in germ-free mice, suggesting that the microbiota are causally related to autism traits. These observations have led to a proliferation of microbiome-based interventions for autism, including the progression of clinical trials (albeit open label and non-blinded) for faecal microbiota transplantation in autism. However, despite the undeniable excitement, it is important to note that the evidence supporting an association of the microbiome with autism, and indeed for a causal effect of the microbiome on autism traits, remains weak. This is because published studies are small and many do not adequately account for known confounders of microbiome variation, including age, sex, diet and medications. Further studies in deeply phenotyped cohorts, such as the AAB, will be needed to more-definitively establish the relationship between the gut microbiome and autism.

Another area of active research in autism is metabolomics, which refers to the study of small molecules that are the by-products of our metabolism (“metabolites”), using liquid chromatography-mass spectrometry (LS-MS) and related methods. Metabolites such as proteins and lipids are intermediate markers that bridge the gap between genetic factors and the observable clinical traits of autism. These are likely to be important to study as they reflect dynamic processes, provide high-level resolution into biological mechanisms such as mitochondrial dysfunction, oxidative stress and inflammatory processes, and may help to identify biological sub-types of autism, for whom prognosis and treatments could be better tailored. Information from metabolomics and lipidomics screens in the AAB is expected to enhance our understanding of the underlying causes of Autism Spectrum Disorder and will hopefully contribute to Autism biomarker discovery and the development of genomic predictors.

A strength of our proposal is that the generation of SNP, DNA methylation, gut microbiome and lipidomics data will enable Autism CRC to utilise international developments in autism genetics and genomics in the coming 3-5 years, including for example, results from large-scale GWAS, and the

potential identification of DNA methylation, microbiome and lipidomics signatures for autism and/or autism-relevant environmental exposures. Another strength is the collection of diverse biological samples, including faeces and urine, which together with detailed clinical and lifestyle data, will contribute to a rich and multi-layered biological resource that will be the foundation for future research and discovery. A further strength is the collection of biospecimens and data in both children on the spectrum and siblings and unrelated children without a diagnosis, since this enables the derivation and validation of predictors that have clinical sensitivity and specificity, and so guards against generation of a diagnostic test that produces a high proportion of false positives.

An important qualification of this proposal is that the data and resources available to us in the Australian Autism Biobank study are insufficient to generate a diagnostic predictor for autism. This project represents only the first step towards that ultimate goal.

1.2 Specific Aims

Aim 1: Receipt and process biological samples (blood, faeces, urine, hair) from up to 1,200 autistic children, their parents and non-autistic siblings and age-matched controls as part of the Australian Autism Biobank ("Biobank").

Aim 2: Generate and analyse genome-wide single nucleotide polymorphism (SNP) data in all Biobank participants.

Aim 3: Generate and analyse additional genomics datasets, including DNA methylation, metabolomics and gut microbiome, in sub-sets of autistic children and age-matched non-autistic siblings and controls in the Biobank.

Aim 4: Develop and validate systems genomics-based predictors for autism and integrate these with clinical diagnostic instruments to improve early and accurate diagnosis.

2. Study design and Methods

2.1 Ethics

All families provided informed consented to be included in the study:

- NSW (N=264 child participants, 235 parents): Sydney Children's Hospital Network HREC, approval number HREC/14/SCHN/269.
- QLD (N=303 child participants, 239 parents): Mater Research HREC, approval number HREC/14/MHS/212; the University of Queensland, approval number 2014001079
- VIC (N=365 child participants, 329 parents): La Trobe University, approval number HEC16/104
- WA (N= 648 child participants, 662 parents): Princess Margaret Hospital for Children approval number 2014029EP; the University of Western Australia approval number RA/4/1/8184

2.2 Biospecimen receipt and processing

Procedures for receipt and processing of biospecimens, including blood (EDTA and SST tubes), saliva, stool, urine and hair have been fully described in Alvares *et al.* (2018)[1].

Briefly, blood collection included an EDTA tube for DNA (highest priority), SST tube for serum, and in some cases a PaxGene tube for RNA (Table 2.2.1). Where a blood collection was unsuccessful (for example due to child distress), a saliva collection (2 ml through spit or swab) was attempted. Samples (collected by paediatric phlebotomists or through hospital/pathology phlebotomy services) were transported at room temperature to the University of Queensland's Human Studies Unit (HSU) and immediately processed (time from collection to processing between 12 and 72 h): EDTA and SST tubes were centrifuged at 3000 rpm for 15 min to separate the individual components of plasma, red blood cells, buffy coat (EDTA) and serum (SST), with fractions stored in 2 mL screw cap tubes for long-term storage at −80 degrees Celsius (Table 2.2.2).

Stool samples (at-home collection of duplicate teaspoon-sized samples per child participant) suspended in 4mLs RNAlater™ were transported to the HUS and immediately processed by vigorous homogenisation prior to aliquoting into a total of 6 × 1mL samples for long-term storage at

−80ºC. Urine samples (~20mL per child participant) were cold shipped in batches from the clinical sites to the HSU for labelling and storage at –80ºC. Subsequently, each tube was thawed (once only), thoroughly mixed and aliquoted into 8x 1mL aliquots. Hair samples (approximately 10 strands per child participant) in aluminium foil were transported at room temperature to the HSU for labelling, and then transferred to long-term storage.

Table 2.2.1 Biospecimens collected per participant group.

Sample type	ASD probands	Parents	Siblings/Controls
Blood			
EDTA	6mL	9mL	6mL
SST	5mL	5mL	5mL
PaxGene	2.5mL	2.5mL	2.5mL
Stool	2x teaspoon	–	2x teaspoon
Urine	~20mL	–	~20mL
Hair	~10 strands	–	~10 strands

Table 2.2.2: Sample fractions generated per participant.

Sample fraction	Volume (mL)	Aliquots
Whole Blood	0.1	1
Plasma	1	2
Red Blood Cells	1	1
Serum	0.5	2
Stool	1	6
Urine	1.1	8
Hair	1 (unit)	1
DNA Stock	0.10 - 0.40	1
DNA Variable	~0.200	1

2.3 Genetics

2.3.1 SNP genotyping quality control (QC) and imputation

SNP genotyping was initially performed on a total of 2,491 Biobank participants (756 mothers, 505 fathers, 887 autistic children ("ASD"), 219 non-autistic siblings ("SIB"), 117 unrelated non-autistic controls ("UNR")) who provided a blood sample, using the Illumina Global Screening Array v1 and v2. We applied standard QC steps to filter low quality samples and less-reliable SNPs. After imputation to the Haplotype Reference Consortium[2] dataset (using the Sanger Imputation Service, with pre-phasing performed using EAGLE2 software [3]) and further QC, there were 2,478 participants and 7,068,672 SNPs (6,991,521 autosomal markers and 77,151 on chromosome X; minor allele frequency <0.01) for prediction analyses. One further child participant in the UNR group was excluded on the basis of called CNVs (see below), after which data from 2,477 participants remained in the QC-ed dataset. For full details of SNP QC and imputation see Yap et al. (2021)[4].

2.3.2 Ancestry assignment and genetic relationships

We inferred genetic ancestry for each individual (European, South Asian, East Asian, African, and "other") by cross-referencing the Biobank genotyping data to the 1000 Genomes reference dataset (i.e. using the first two principal components (PCs) of the genetic reference data, which capture major genetic differences between ancestries). For the European subset (n=1,973) of the Biobank, we calculated 20 PCs (representing genetic sub-populations among Europeans) using GCTAv1.92 [5,6] (based on n=255,861 common genotyped SNPs), which were subsequently included as covariates in the polygenic scoring (PGS) analyses (see below).

Familial relationships were inferred using pairwise identity-by-state estimation (with the PLINK1.9 – genome command, using a set of 92,546 independent genetic markers), and also by constructed a genetic relatedness matrix (GRM) from the SNP genotypes using GCTAv1.92 [5,6]. These relationships were cross-referenced against the reported relationships in the Biobank, and to match family members not otherwise linked by the ID system (e.g., where individuals within one family were recruited on different dates or via different assessment centres).

2.3.3 Association testing

We did not test for association between individual SNPs and autism (or any other measured trait) in the Biobank, because the sample is underpowered. To place the Biobank (total *n*=2477 after QC) in the context of contemporary genetic studies of autism, the most recent genome-wide

association study (GWAS) meta-analysis published by the Psychiatric Genomics Consortium (PGC) comprised 18,381 individuals with a diagnosis of autism and 27,969 controls[7]. That study identified 5 genome-wide significant SNPs and provided evidence – consistent with that for all other common psychiatric conditions[8] – that a substantial proportion of the heritability is explained by common SNPs. For this reason, larger GWAS meta-analyses are expected to yield many more genetic discoveries for autism.

2.3.4 Polygenic scoring (PGS)

We calculated Polygenic scores (PGS) for Autism Spectrum Disorder, IQ, chronotype and height (as a benchmarking trait) in all Biobank participants. PGS require two components: 1) estimates of the strength of association (i.e. weights) between SNPs used to generate the score and the trait of interest (e.g. autism diagnosis), obtained from analysis of an cohort independent of the target sample, and 2) individual-level genotype data (generated and QC-ed as discussed above) in the target sample.

To generate SNP weights, we used the state-of-the-art SBayesR [9,10] software (using default settings), taking as input a) GWAS summary statistics for Autism [11], IQ [12], chronotype [13] and height [9], and b) a genetic reference matrix that is used to adjust for correlations between SNPs that may affect the PGS score. For height only, there was an additional step to filter GWAS SNPs with the software package DENTIST [14], to improve convergence of the SBayesR algorithm.

To generate PGS for each trait, we multiplied the number of alleles carried by each Biobank individual by the SBayesR weights, using the PLINK --score function. We restricted analyses of the target dataset to the subset of participants of inferred European ancestry, to match the ancestry profile of the input GWAS summary statistics. The PGS scores were standardised by subtracting the mean and dividing by the standard deviation of the UKB controls (see below).

Selection of UK Biobank (UKB) controls: Given that the UNR group had minimal exclusion criteria and was small (n=117), we also included UK Biobank (UKB) controls with European genetic ancestry as an additional control group. These individuals were selected by taking the n=5 “most genetically similar” individuals from the UKB, for each of the autistic Biobank participants of European ancestry. We note that international efforts are underway to address the current under-representation of non-European ancestries in human genetic studies, including in autism, and thus we anticipate that in future, analyses such as those described here could encompass all major ancestries. To confirm that the UKB control sample was well matched genetically to the Biobank, we analysed PGS for height as a control trait. Height is considered a model genetic trait because it has high heritability and summary statistics from very large GWAS are available[9]. As expected,

there was no evidence for over-transmission of common genetic variation for height in the Biobank ASD group compared to SIB's, or for differences in height PGS between any Biobank group and the UKB controls. Thus, the UKB control group was well matched to the Biobank.

Between-group PGS differences: We tested for a mean difference in PGS for each trait between ASD, SIB and UNR Biobank groups using Z-tests.

Over-transmission of common genetic variation for autism: We tested for over-transmission of common genetic variation for autism from parents to their children with and without autism, using pTDT software [15], including n=330 individuals diagnosed with autism and n=145 undiagnosed siblings with parental genotyping data. Briefly, this tests for a difference in the child's PGS from the average of their parents' PGS (which represents the null). We also looked for evidence of assortative mating for autism only (that is, parents choosing their partner based on similar traits). For this, we tested for correlation of PGS between the parental pairs within the family data.

Relationships between PGS and phenotypes: We calculated correlations between PGS of multiple traits and various phenotypes of interest recorded within individuals of European ancestry in the AAB. We removed (via linear regression) the effect of age, sex and ancestry from the phenotypes, before calculating how much of the phenotypic variance is explained by PGS.

Multiple testing correction: We used the Bonferroni method to avoid false positives and correct for multiple testing across all PGS analyses (50 tests, $p \leq 0.05/50$ or $p \leq 1e\text{-}3$). Phenotype-PGS associations that were tested for both variance and correlation were counted as one test, as these statistics are mathematically related.

2.3.5 Copy number variant (CNV) calling

We followed the Psychiatric Genomics Consortium CNV analysis pipeline [16], with a few modifications. CNVs were identified in each Biobank individual using consensus calling from the PennCNV [17] and iPattern [18] software. Specifically, we merged CNVs output by each software that were likely relating to the same event, then obtained consensus CNV calls for each individual by intersecting CNV calls from both methods and retaining those with ≥50% overlap, ensuring that copy number (gain or loss) was matching in each method. A total of 10,752 consensus CNVs were identified using this approach.

Sample QC was performed using summary statistics from the PennCNV output. We removed n=137 samples that did not pass standard QC filters. We also applied a filter for samples where CNVs made up greater than 20% of any chromosome to exclude aneuploidy (ie., differences in

chromosome count). This step identified n=2 individuals – one participant from the autistic group with known Down syndrome (trisomy 21), and one participant with diagnosed Smith-Magenis syndrome whom had been included in the UNR group. We also set filters to retain only high-confidence, large (>20Kb), and rare (<1%) CNVs. After these QC steps, we excluded one additional individual for whom 21 CNVs had been called (the next-highest number of CNVs for an individual was four), leaving 885 CNVs from 723 individuals remaining for subsequent analysis.

CNV annotation: We annotated CNVs for cytobands (if ≥50% of the called CNV length lay within that cytoband) and genes (if the CNV overlapped with any coding exons within the gene) using the hg19 biomaRt download [19]. We checked for overlap between CNVs in the Biobank dataset and a total of 51 CNV regions selected from clinical databases (ClinGen [20] and DECIPHER [21,22]) on the basis of evidence for association with autism and ID. We divided these known ASD/ID CNVs into those with and without a critical gene. CNVs with a critical gene required overlap with any exon, whereas CNVs without a critical gene required ≥80% of the pathogenic CNV region. We also looked for overlap of the Biobank CNVs with the 102 genes identified by the largest whole exome sequencing study of autism to date [23], and 93 genes associated with developmental disorders from the Deciphering Development Disorders (DDD) study [24] – a total of 158 unique genes.

2.4 Gut microbiome

2.4.1 Metagenomics data

Microba Life Sciences was contracted to generate gut metagenomics sequence data from 249 stool samples, including n=199 from the Australian Autism Biobank 99 autistic (ASD), 51 non-autistic siblings (SIB), 49 unrelated non-autistic controls (UNR)) and n=49 from the Queensland Twin Adolescent Brain (QTAB, all UNR) study (Figure 2.4.1.1). Microba also performed annotation of bacterial species and metabolic pathway potential and undertook initial QC on the dataset[25]. Stool-derived DNA samples were sequenced to a target depth of 3Gb using 2x150bp Illumina chemistry.

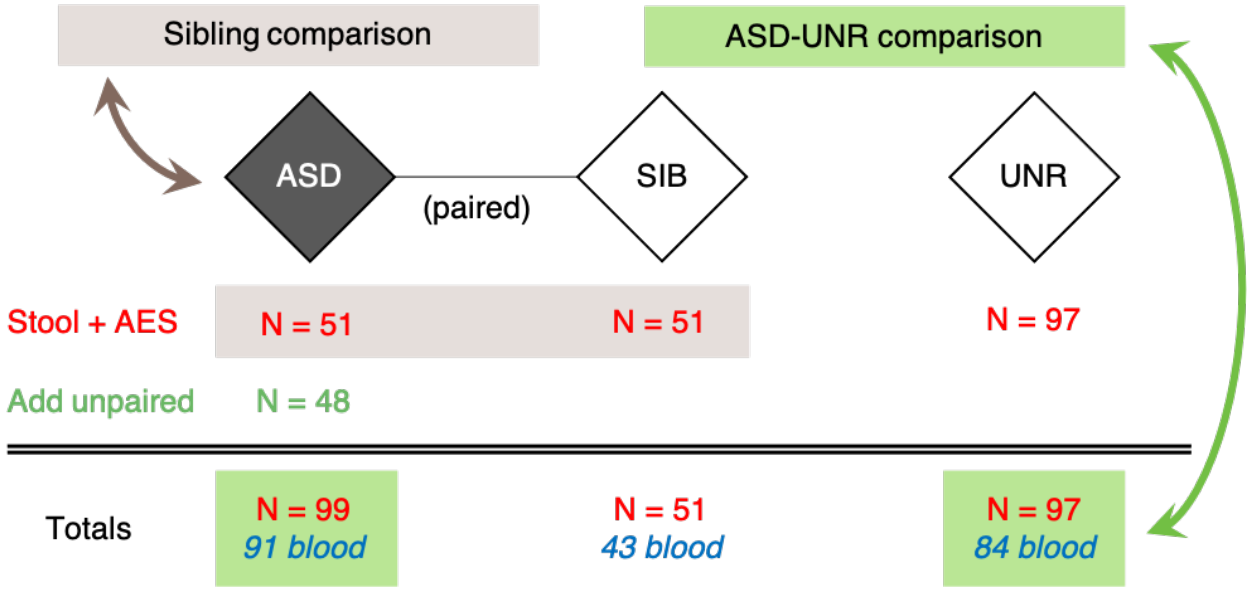


Figure 2.4.1.1 Schematic of the Australian Autism Biobank gut metagenomics pilot study.

2.4.2 Dietary data

Dietary data was collected in both the Biobank and QTAB cohorts – predominantly on the basis of parent-report – using the Australian Eating Survey (AES; toddler and children's versions) [26,27], which has been validated in the Australian population. Food-level intake data was available for n=245 of the 247 participants, and percent energy (pe) data was available for 246. The AES records frequencies of intake for 123 different foods, from which derived variables are generated, including percentage energy (pe) from each of 13 core (vegetables, fruit, meat, alternative proteins, grains, dairy) and non-core (sweet drinks, packed snacks, confectionery, baked products, takeaway, condiments, fatty meats) food groups; macronutrients (various carbohydrates, fats and proteins); micronutrients (various vitamins and minerals), and the Australian Recommended Food Score.

We used the dietary data in two ways. First, we used the food-level input to measure dietary diversity (n=245) using Shannon index – the same measure of alpha-diversity used in the microbiome analyses (see below). Second, we calculated principal components from the percent energy data (n=246; hereafter referred to as pe _PCs), to capture salient dietary features that may affect the microbiome, given that a strong relationship has been identified by others [28,29]. On the basis of their loadings onto the dietary items, the first 3 pe_PCs were interpreted as follows:

- PC1: diet high in plant-based diet (vegetables, fruit) / low in non-meat non-core foods (sweet drinks, packed snacks, confectionery, baked products, takeaway, fatty meats)
- PC2: high dairy diet / low in grains and takeaway
- PC3: high in meat (including fatty meats) / low in grains and dairy

2.4.3 Variance components analysis

Omics-relationship matrix (ORM) construction and OREML analysis

We performed variance component analysis using the software package OSCA [30]. This is useful in estimating the upper bound of how well the microbiome (or diet) dataset is able to predict a trait. Covariate choice depended on the focal phenotype, but universally included sex and age (except when age was the dependent phenotype), and in some cases, participant group or dietary pe_PCs. We treated common bacteria versus rare bacteria separately in this analysis, and we also generated ORMs on the basis of common and rare bacterial genes. This was motivated by the observation that there are some “core” taxa as well as “accessory” taxa which may have differing properties and roles.

2.4.4 Diversity analysis

We calculated two estimates of microbiome diversity in the Biobank dataset: alpha diversity and beta diversity. Alpha-diversity measures how diverse the microbiome community is within each sample. Richness is one alpha-diversity measure and quantifies how many different species were identified per individual. An extension to this is the Shannon Index, which accounts for both richness and evenness, which we calculated among bacterial species within our dataset. In contrast, beta-diversity refers to diversity between samples; that is, how different each individual is from other individuals. To calculate beta-diversity, we generated a weighted Unifrac index matrix, calculated using 1,054 bacterial species (i.e., not including archaea). We used this beta-diversity matrix to quantify differences in microbiome profiles between ASD, SIB and UNR groups.

2.4.5 Differential abundance analysis

Species-level taxonomic data: We tested whether any of n=607 common bacterial species were significantly more or less prevalent in the ASD group versus other groups. For this, we used ANCOMv2.1 [31] (implemented in R: https://github.com/FrederickHuangLin/ANCOM) as it is robust to statistical assumptions[32].

Gene-level functional data: We then took differentially abundant bacteria and tested whether any of the genes that they encoded were differentially abundant. In this analysis, we tested 4,950 genes from a single differentially abundant species.

2.4.6 Linear models

We generated linear models to test for associations between diversity measures and potentially-contributing phenotypic (Bristol Stool Score, ADOS-2/G repetitive and restricted behaviour (RRB) score, Social Responsiveness Scale t-score (SRS), Short Sensory Profile raw sensor score (SSP)) and biological measures (polygenic scores (PGS) for Autism Spectrum Disorder [11], ADHD-ASD-TS cross-trait [33] and neuroticism [34], and CD4+ T-cell proportions). To account for multiple testing, we performed Benjamini-Hochberg false discovery rate (FDR) correction.

2.5 DNA methylation

2.5.1 DNA methylation data and quality control

This study used whole blood samples from n=468 children and adolescents recruited into the Australian Autism Biobank[35] and Queensland Twin Adolescent Brain project. Participants were assigned to three groups: ASD, SIB and UNR (Figure 2.5.1.1). We used the Illumina EPIC Human Methylation array, which assays over 850,000 CpG methylation probes across the genome. We used the meffil[36] software package for normalisation. We also performed standard QC filters to identify reliable probes and target those that were more likely to have an effect and used this reduced subset for downstream analyses.

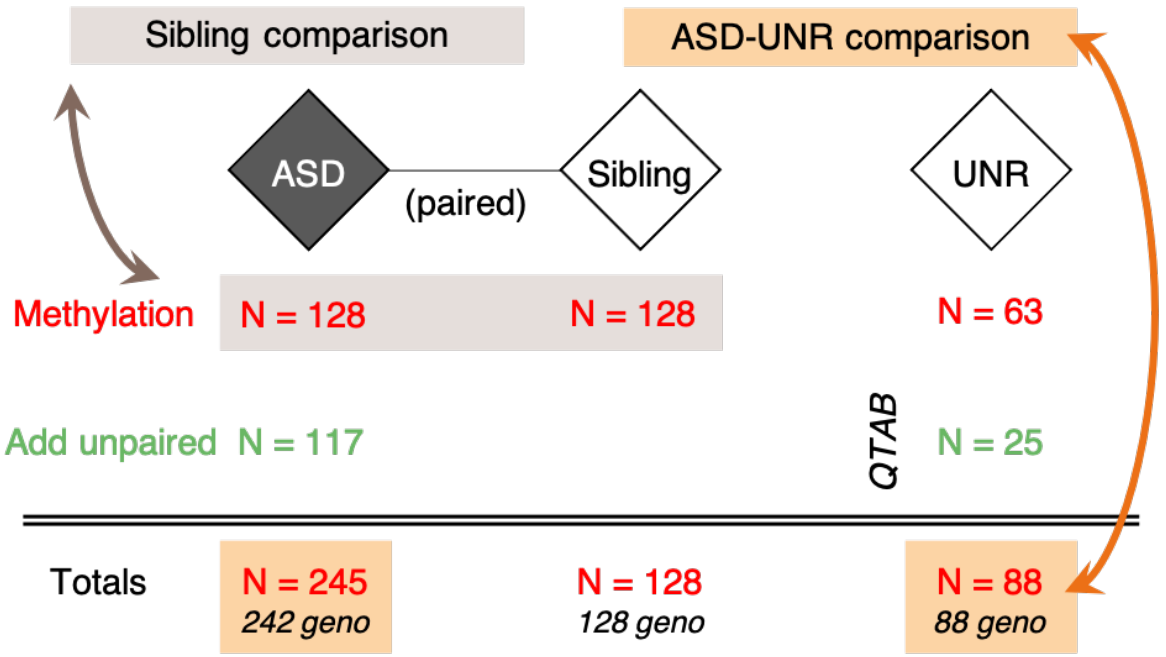


Figure 2.5.1.1 Schematic of the DNA methylation study design.

2.5.2 Cell-type proportions

The DNA methylation (DNAm) data is generated from whole blood which essentially is a mixture of several different white blood cell types. Each cell type has a specific DNA methylation profile at specific locations in the genome. It is important to know if any differences detected between test groups is driven by differences in these mixtures (proportions) of different cell-types or not. Using the cleaned DNA methylation data, we deconvolved blood cell-type proportions (CTPs; neutrophils, CD4+ T-cells, CD8+ T-cells, B-cells, NK cells, monocytes, eosinophils) in each individual, using a human immune cell reference dataset (GSE35069). We looked for association of CTPs with participant group and included CTPs as covariates in association analyses.

2.5.3 Variance components analysis

We estimated the proportion of variance in autism diagnosis that was associated with all DNA methylation probes combined, using the Omics-data-based restricted maximum likelihood (OREML) method implemented in the OSCA [30] software. We additionally performed analyses of autism polygenic score, Short Sensory Profile (SSP_sensory) raw sensory score, Children's Sleep Habits Questionnaire (CSHQ) raw score, a composite score for IQ (composite scores from the NIH Toolbox age-adjusted questionnaire) and DQ (Mullen's Scales of Early Learning non-verbal composite score), hereafter referred to as "IQ-DQ", and age, and we adjusted for covariates, including sex, genotype PCs, cell-type proportions and age (with the exception of analyses in which the response variable was age). The purpose of these variance component analyses was to quantify the upper limit of the ability of DNA methylation data to predict autism diagnostic status.

2.5.4 Methylome-wide association study (MWAS) and meta-analysis

We used a mixed-linear model approach called MOA (MLM-based omic association) implemented in the package OSCA[30] to generate MWAS summary statistics. We chose MOA as it has been shown to increase power to detect associations while adequately controlling for false positives[30]. We did not include covariates in this analysis, as the variance-covariance matrix used in MOA automatically detects these within a random-effects framework. We used the METAL[37] software to perform an MWAS meta-analysis of AAB-QTAB and the MINERvA[38] study from Denmark, analysing only those methylation probes shared across all three studies.

2.6 Lipidomics

2.6.1 Data and quality control

We engaged with The Baker Institute in Melbourne to generate lipidomics data in non-fasting blood plasma samples from n=765 children and adolescents recruited into the Australian Autism Biobank (AAB)[35] and Queensland Twin Adolescent Brain (QTAB) study. The sample included 485 participants with a diagnosis of autism, 160 undiagnosed siblings (SIB), and 120 unrelated undiagnosed (UNR) children, 96 from the AAB and 24 from QTAB. Participant groups were well matched for age, whereas there was a male bias for the ASD group, as is expected given the known sex bias for autism (Figure 2.6.1.1).

The Metabolomics Profiling Facility at The Baker Institute generated semi-quantitative data on a total of 825 lipid species in 41 lipid classes (Appendix A), using an Agilent liquid chromatography–mass spectrometry (LC/MS) platform comprising a 1290 series HPLC combined with a 6495C triple quadrapole mass spectrometer. Raw mass spectrometry data was processed using the MassHunter Quant (B08) software from Agilent. Concentrations for each lipid species were obtained from the ratio of the peak area to the corresponding internal standard, and values for lipid classes were calculated by summing the concentrations of the individual species within each class.

Established and validated in-house pipelines were used for quality control and filtering of lipid species. First, technical variation between batches was removed by performing a median centring procedure using the pooled plasma quality control (PQC) samples included in the extractions. For each lipid, the median concentration for all PQCs in each batch was calculated. The lipid concentrations in each batch are then multiplied by the ratio of the overall PQC median across all batches (i.e., the median of all PQC medians) over the batchwise PQC median. Second, missing values (i.e., lipids below the detection threshold) were imputed to half the minimum observed concentration. Third, sample outliers were then detected on the basis of a Z-score >3 standard deviations from the mean, or distance from the origin in a principal component analysis was in the 99th percentile of the distribution. On the basis of these criteria, seven outlier samples were identified and removed. We also excluded TG [NL] from classes and TG [SIM] from species, in addition to a further n=10 lipid species for which >10% of the variance in lipid concentration was explained by batch and injection order, and we confirmed that these excluded lipids had negligible association with autism diagnosis in a model of lipid concentration ~ autism diagnosis (all explaining <=1% of variance). The final dataset comprised n=781 lipid species in 39 classes in a total of n=758 individuals (479 ASD, 160 SIB, 119 UNR)

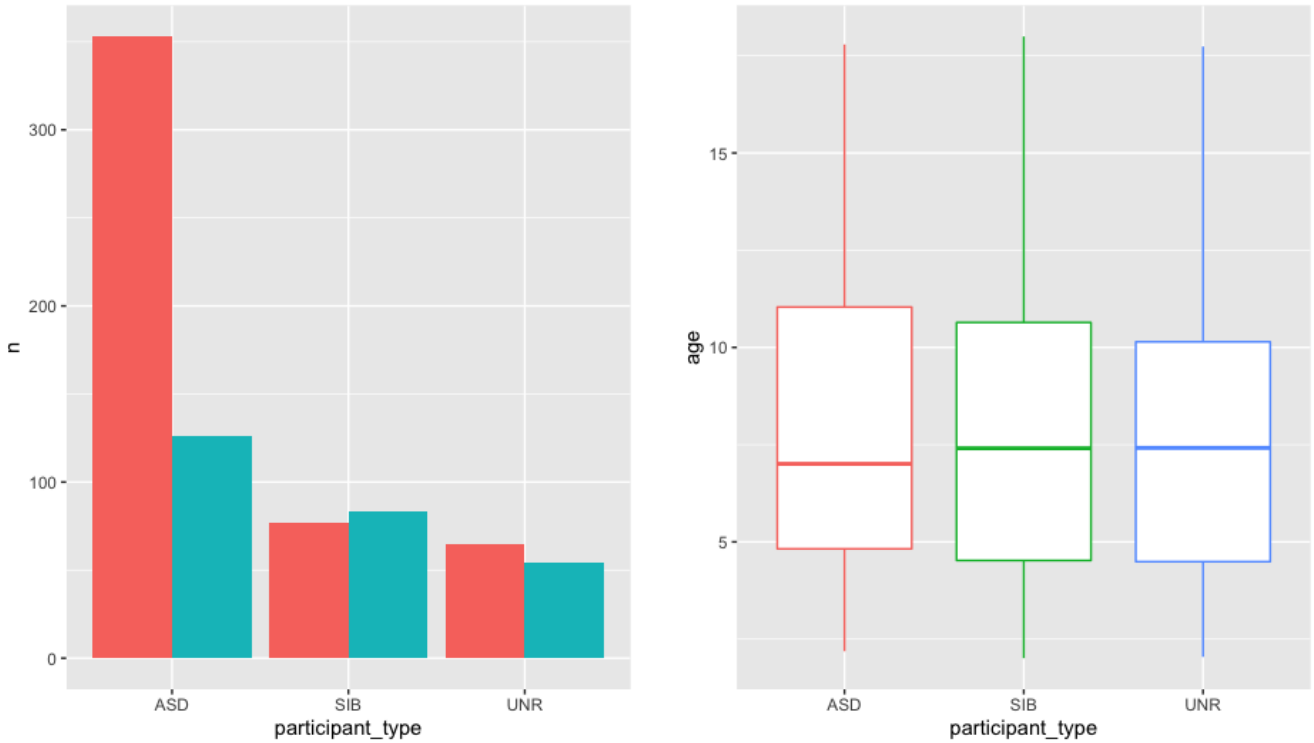


Figure 2.6.1.1 Breakdown of participant groups by sex (left: red=male, blue=female) and age (right).

2.6.2 Variance components analysis

We estimated the proportion of variance in our sample attributable to Autism Spectrum Disorder diagnosis and each of six other traits (age, Tanner stage, ID_DQ, sleep problems, gross motor skills, Bristol Stool Chart) that was associated with all lipid species combined, using the Omics-data-based restricted maximum likelihood (OREML) method implemented in the OSCA software. We also performed OREML analyses for each of five dietary phenotypes, including total cholesterol, protein, fats, sugars and carbohydrates. We applied an inverse normal transformation to the lipid species data prior to building the Omics-data-based Relationship Matrix (ORM), and we performed analyses with and without covariates, which included age (except where the trait was age or Tanner stage), sex, batch and storage time. We additionally adjusted for time of day as a sensitivity analysis, in the subset of individuals for which this data was available, and for dietary variables (except where the trait was a dietary phenotype), in the subset of n=260 participants for whom the Australian Eating Survey was available. For dietary phenotypes, we adjusted for demographic and batch covariates (as described above), in addition to energy intake.

2.6.3 Lipidome wide association analysis (LWAS)

We tested for association between individual lipids and Austim Spectrum Disorder diagnosis, ID_DQ, sleep problems (CSHQ), age and Tanner developmental stage. We used linear or logistic regression, with and without covariates, which included sex, batch, injection order and storage time for all analyses, and additionally age in all but the age and Tanner stage analyses. We applied backwards stepwise regression (using the R package MASS, which optimises the model based on the AIC) to the lipids passing Bonferroni correction (threshold calculated by dividing p=0.05 by the number of lipids included within each association analysis) to account for correlation between lipid species and classes. We applied an inverse-normal transformation to the lipidomics data and excluded the n=7 outliers. Sensitivity analyses for the effect of storage time outliers (n=64) showed minimal difference for all traits other than Autism Spectrum Disorder diagnosis, so we retained all n=758 participants for analyses of traits other than Autism, for which the sample size was n=694.

2.7 Systems genomics-based prediction

2.7.1 Rationale

An important question in autism research is whether genomic information can assist in early and accurate diagnosis of autism. Here, we present a preliminary assessment of systems genomics predictors of autism in the AAB. The rationale of the study design – which was to combine genetic and genomic predictors – is that the latter may capture variation due to autism-related environmental exposures, and so may improve the accuracy of a genetic predictor, which can never be a perfect predictor, because the genetic contribution (i.e., heritability) of autism, although high, is less than 1. Genome-wide data on single nucleotide polymorphisms (SNPs), blood-derived DNA methylation, gut metagenomics and lipidomics in the AAB – generated by Project 1.042RC – were available for this analysis.

Genomic prediction typically involves the estimation of effect sizes for predictors in discovery datasets that are independent of the target sample in which the predictor is applied and evaluated. This is important to avoid the so-called “winners’ curse”, whereby the effect sizes of the most strongly associated genomic variables *within* a cohort-specific analysis are inflated, and because confounding of batch effects with diagnosis within a single cohort can bias results. Alternatively, if independent datasets are not available, K-folds cross validation can be used within a single cohort, whereby subsets of the data (typically ~20%) are held-out for evaluation of predictors estimated in the remainder of the data, averaging across K (e.g., 5) data partitions. It is important to note that

this approach may not circumvent the issues noted above, and it is only possible if the dataset is sufficiently large to enable partitioning of the data into multiple discovery and target samples.

In our analyses of the AAB data, the former strategy was possible for predictors based on genome-wide SNP and DNA methylation data, for which independent, publicly available datasets exist (see below). On the other hand, no independent gut metagenomics or lipidomics data for autism is available, and the AAB datasets are insufficient in size for K-folds cross validation (see below). For this reason, this preliminary report on genomic prediction in the AAB is based on genetic (i.e., SNP) and DNA methylation-based predictors.

We were unable to evaluate an integrated diagnostic protocol combining systems genomics predictors with variables from behavioural assessments, because no assessments were available in all of the study groups: children diagnosed with autism, siblings, and unrelated children without a diagnosis.

2.7.2 Data quality control

Genotyping data: We used association test summary statistics from the largest published genome-wide association study (GWAS) for Autism Spectrum Disorder from the Psychiatric Genomics Consortium[7], which reported five independent ($p<5e-8$, $r^2<0.1$) genome-wide significant associations from analysis of n=46,350 participants (18,381 diagnosed, 27,969 undiagnosed).

We generated polygenic score (PGS) weights using SBayesR – a Bayesian method that takes GWAS summary statistics as input[39]. This method shrinks SNP effect sizes while still maximising variance explained by "binning" SNPs into a mixture of normally distributed priors, accounting for linkage disequilibrium (i.e., correlations between SNPs) and adapting to the genetic architecture specific to the trait. SBayesR has been shown to outperform other PGS methods regardless of the underlying genetic architecture of the trait[10,39]. SBayesR requires two inputs: 1) GWAS summary statistics from which HapMap3 SNPs with imputation INFO filter>0.8 were extracted, retaining only those SNPs that passed QC in both AAB and the UK Biobank (UKB), and 2) linkage disequilibrium matrices built using HapMap3 SNPs from a subset of 50,000 unrelated Europeans from the UKB. SBayesR was run with the default inputs: --pi 0.95, 0.02, 0.02, 0.01; gamma 0, 0.01, 0.1, 1; chain-length 10000; burn-in 2000; out-freq 10. We excluded the MHC region on chromosome 6, due to the complex patterns of linkage disequilibrium in this region, using the –exclude-mhc flag.

To generate PGS for Autism Spectrum Disorder, we multiplied the best guess genotypes in the target sample (i.e. AAB individuals and UKB controls) by the per SNP effect sizes (reweighted by SBayesR[39]), summing across all SNPs, using the PLINK –score function. We restricted analyses of

the target dataset to the subset of participants of inferred European ancestry. The PGS was standardised by subtracting the mean and dividing by the standard deviation of the UKB controls.

Methylation data: We used summary statistics from the MINERvA DNA methylation study of Danish infant blood spot samples[38]. We note that this is a valuable dataset as samples were obtained from infants before Autism Spectrum Disorder diagnosis (versus other datasets including the AAB which involved ascertainment based on prior diagnosis). We applied a p-value threshold for association of methylation probes autism in the MINERvA dataset of $p<1x10^{-3}$, leaving 415 probes. Of these, 397 overlapped with probes in the AAB methylation dataset, which had undergone filtering to exclude probes with standard deviation <0.2, and MASK probes which are often subject to technical issues such as cross-hybridisation[40]. Subsequently, we generated methylation genomic scores (MGS) from these 397 probes, by multiplying their weights from the reference MINERvA study by the normalised methylation beta values in the AAB methylation dataset and summing across all 397 probes. We then standardised these MGS scores by subtracting the mean and dividing by the standard deviation of the individuals included within the AAB dataset.

Justification of exclusion of metagenomics and lipidomics data: We did not include the metagenomics or lipidomics data within this predictor for a few reasons: first, there are no external datasets that we could use as independent discovery data to identify predictors for evaluation in the AAB as a target cohort. Second, there was negligible association of both the gut microbiome and lipidome with autism diagnosis, and thus there is limited benefit from using a K-folds cross-validation approach with this data.

3. Findings

3.1 Biospecimen receipt and processing

Biospecimens collected from Australian Autism Biobank ("Biobank") participants recruited at each of the four clinical sites (Perth, Melbourne, Sydney, Brisbane) were receipted and processed by The University of Queensland Human Studies Unit (HSU).

Over the duration of the active recruitment phase of the Biobank (ending 30th June 2018), a total of 3,707 biological samples were receipted by HSU, including blood samples from >2,000 Biobank participants, stool samples from >400 child participants, urine samples from nearly 600 child participants and hair samples from 663 child participants (Table 3.1.1).

Table 3.1.1: Autism CRC biospecimens receipted and processed at the UQ HSU Laboratory by participant group.

Sample type	Proband	Mother	Father	Sibling	Control	Total participants
Blood	702	588	399	220	117	2026
Stool	220	0	0	98	83	401
Urine	318	0	0	144	125	587
Hair	357	0	0	169	137	663
Saliva	13	4	2	5	6	30

Receipted biospecimens were processed to generate the sample fractions described in Table 2.2.1. Each (intact) blood collection (comprising EDTA and SST tubes) yielded whole blood (x1), plasma (x2), red blood cells (x1), serum (x2), and buffy coat (nucleated white blood cells), from which genomic DNA was isolated. Intact stool and urine collections from autistic children, non-autistic siblings and controls yielded 6 and 8 aliquots, respectively. Not all receipted blood, stool and urine collections yielded all fractions, due primarily to lower-than-expected volumes.

In total, across all receipted biospecimens, in excess of 24,000 high-quality sample fractions were generated by HSU and on-shipped (with the exception of the buffy coat and saliva samples) to the Biobank at Wesley Medical Research (Table 3.1.2).

Table 3.1.2: Biospecimen fractions generated by the HSU Laboratory for on-shipping to the Biobank.

Sample type	Proband	Mother	Father	Sibling	Control	Total participants	Total fractions
Whole Blood	657	584	392	213	115	1961	1961
Plasma A#	677	587	396	217	117	1994	1994^
Plasma B#	661	587	395	216	116	1975	1975
Red Blood Cells	676	587	395	217	117	1992	1992
Serum A#	696	586	394	219	49	1944	1944
Serum B#	683	585	394	218	45	1925	1925
Paxgene	149	131	90	56	9	435	435
Stool A#	219	0	0	98	81	398	1196^
Stool B#	213	0	0	96	80	389	1159
Urine aliquot	318	0	0	144	125	587	4735
Hair	357	0	0	169	137	663	663
Saliva	13	4	2	5	6	30	30
DNA Stock*	707	638	423	220	118	2106	2106
DNA Variables*	725	631	422	241	122	2141	2141

* 'DNA Variables' provides the total number of DNA tubes and 'DNA Stock' details the total number of samples for which there is extracted DNA.

^ A number of the samples/fractions may have been used for various genomic assays.

Plasma and serum samples were divided into A and B samples for biobanking. Stool samples were collected in duplicate (A and B), with each then divided into three fractions, for a total of six per participant.

In addition to receipt and processing of primary biospecimens, HSU also receipted into the Biobank a large collection of blood samples and associated sample fractions, RNA and DNA from a Western Australian family-based autism cohort (192 autistic, 171 mothers, 107 fathers) recruited between 2011 and 2016, on-sent from PathWest (Table 3.1.3).

Table 3.1.3: PathWest blood samples and fractions receipted by HSU.

Sample type	Proband	Mother	Father	Sibling	Control	Unknown	Total participants	Total fractions
Blood	188	170	106	0	0	4	468	N/A
Plasma	187	168	106	0	0	1	462	938
Serum	185	165	105	0	0	1	456	468
Buffy Coat	183	165	102	0	0	1	451	568
RNA	185	119	73	0	0	3	380	749
DNA Stock	1	3	4	0	0	0	8	8
DNA Variables	187	168	107	0	0	0	462	462

The total number of biological samples (blood: 2,490, stool: 398, urine: 587, hair: 663, saliva: 30) and associated fractions (plasma, serum, red blood cells, buffy coat, stool, urine, hair, saliva, RNA, DNA) in the Biobank was 4,168 and >24,000, respectively.

3.2 Genetics

A full description of the analysis of common single nucleotide polymorphisms (SNP) and rare copy number variants (CNVs) in the Biobank dataset is available in Yap *et al.* (2021)[4].

3.2.1 Common genetic variation

A total of 7,068,672 SNPs (6,991,521 autosomal markers and 77,151 on chromosome X) and 2,477 Biobank individuals (including *n*=436 families with both parents and ≥1 affected child) passed quality control, with Europeans the predominant ancestry (*n*=1,964 individuals, *n*=323 families), followed by South Asian (*n* = 248 individuals), East Asian (*n* = 45) and African (*n* = 10).

We used polygenic scoring (PGS) to characterise common genetic variation for autism, IQ and sleep (chronotype) in the Biobank. We included IQ because others have reported a positive genetic correlation with autism[7], which is paradoxical as intellectual disability is a common co-occurring condition. We included chronotype because sleep issues are common among people on the spectrum[41,42], potentially due to shared genetic factors[7].

We first tested for differences in PGS for autism, IQ and chronotype between each of the Biobank groups (ASD, SIB, UNR) and the UKB controls. For Autism, the mean PGS was significantly higher in the Biobank Autism Spectrum Disorder group than the UKB controls ($p = 6.1x10^{-13}$), but the Biobank SIB ($p = 4.9x10^{-3}$) and UNR ($p = 3.0x10^{-3}$) groups also had higher mean ASD PGS than the UKB at a nominal threshold, and there was no difference in mean ASD PGS between the ASD and SIB groups or ASD and UNR controls (Figure 3.2.1.1a). We found no evidence for Biobank group differences for IQ PGS, including in relation to the UKB controls (Figure 3.2.1.1b), or for chronotype PGS. We also found no evidence for over-transmission of autism or IQ PGS to autistic children versus their non-autistic siblings in the Biobank, likely due to the small size of the dataset.

Next, we explored whether the ASD PGS predicted diagnosis of autism in the Biobank, in addition to other Autism-related phenotypes, including ADOS-2 calibrated severity score in the ASD group, Social Responsiveness Scale (SRS) in the SIB and UNR groups, and the Communication Checklist-Adult in parents (chosen in lieu of any autism spectrum questionnaires available across all participant groups). We also determined whether ASD PGS was predictive of indicators for age of first parental concern and age of diagnosis, and other autism-associated phenotypes, including the Short Sensory Profile in the ASD group and cognitive measures in children (WISC-IV composite score or MSEL non-verbal developmental quotient) and parents (WASI matrix reasoning score). We stratified analyses of the cognitive phenotypes in children into ASD and SIB/UNR groups, because IQ is genetically correlated with autism, and so we hypothesised that ASD PGS may exhibit different relationships with IQ or developmental quotient depending on diagnostic status. No associations (after multiple testing correction) were found between ASD PGS and ASD diagnostic status, likely because the Grove *et al.* ASD GWAS remains underpowered for prediction. However, there were nominal associations between ASD PGS and quantitative traits such as MSEL non-verbal developmental quotient ($r = -0.11$, $p = 7.4x10^{-3}$ in the ASD group alone) and marginal association with parental WASI matrix reasoning score ($r = 0.07$, $p = 5.5x10^{-2}$).

We performed similar prediction analyses using IQ PGS, given evidence (as noted above) for a positive genetic correlation with autism, which is counter-intuitive in view of the co-occurrence of ID with autism. IQ PGS showed a significant positive correlation with parental IQ (WASI matrix reasoning domain, $r = 0.17$, $p = 8.0x10^{-7}$; Figure 3.2.1.1c). We also observed a nominally significant correlation between IQ PGS and WISC-IV composite score in the SIB/UNR group ($r = 0.24$, $p = 2.1x10^{-3}$; Figure 3.2.1.1e), whereas there was no evidence for a correlation in the ASD group ($r=0.07$, $p=0.24$), including when the ASD group was stratified by ID (Figure 3.2.1.1d). There were no significant relationships between IQ PGS and MSEL non-verbal developmental quotient, age of diagnosis, age of parental concern or Short Sensory Profile.

Finally, we investigated whether chronotype PGS predicted propensity for sleep conditions in the Biobank, given evidence that sleep issues are common in autism[41,42]. There was no evidence for a significant correlation between chronotype PGS and Children's Sleep Habits Questionnaire (CSHQ) raw score among all children (r = 0.06, p = 7.7x10^{-2}), but we observed a weak (nominally significant) positive correlation in the ASD group (r = 0.13, p = 1.9x10^{-3}; Figure 3.2.1f).

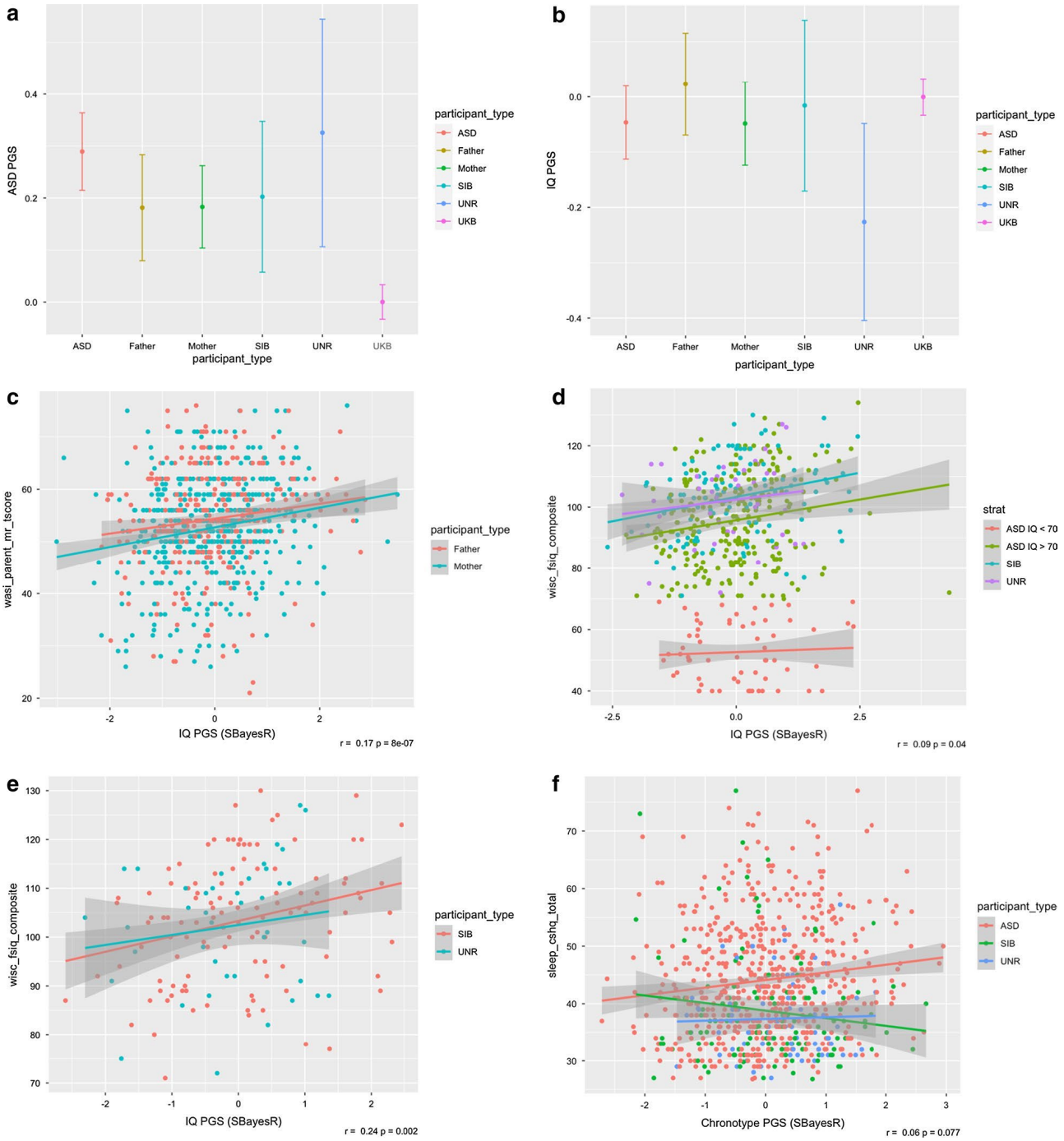


Figure 3.2.1.1 Analysis of polygenic scores (PGS) for ASD, IQ and chronotype in the Australian Autism Biobank. Mean PGS (± 95% CI) for the ASD, SIB and UNR groups for (a) ASD, and (b) IQ. Scatterplots illustrating correlation between the following pairs of traits, with coefficient and p-value in the bottom right corner of each panel for the single combined analysis: (c) IQ PGS and measured IQ (WASI) in parents with fathers in red and mothers in green (overall

r=0.17, 8.0x10^{-7}); (d) IQ PGS and measured IQ (WISC fsiq composite) in the AAB ASD group with IQ < 70 (red), ASD group with IQ ≥ 70 (green), SIB group (blue) and UNR group (purple), (overall r = 0.1, p = 4.0x10^{-2}); (e) IQ PGS and measured IQ (WISC fsiq composite) in the SIB (red) and UNR (green) groups (overall r = 0.24, p = 2.1x10^{-3}); (f) chronotype PGS and Children's Sleep Habits Questionnaire in the ASD (red), SIB (green) and UNR (blue) groups (overall r = 0.06, p = 7.7x10^{-2}). The correlation coefficient for the ASD subset in red is r = 0.13 (p = 1.9x10^{-3}).

3.2.2 Rare copy number variation

We obtained consensus calls for 885 rare exonic CNVs from 723 individuals in the Biobank cohort (Table 3.2.1), after quality control (see Methods). As a first step, we determined whether our pipeline validated previously reported CNVs, finding that 7 of 8 clinical genetic diagnoses were identified, except for an individual with Phelan-McDermid syndrome (caused by a 22q13 deletion of chr22: 51,159,408–51,166,043 which was smaller than the lower detection size in our pipeline).

We identified 13 individuals with CNVs that overlapped high-confidence autism- or ID-associated CNVs from the ClinGen (https://dosage.clinicalgenome.org/pathogenic_region.shtml) and/or DECIPHER (https://decipher.sanger.ac.uk/disorders/syndromes/list)[43] databases (Table 3.2.2), four of which had been reported in the Biobank phenotype dataset. Ten of the 13 were from the ASD group, and three from mothers (including one transmitted from mother to child). The phenotypic data of the participants with overlapping autism/ID-associated CNVs was inspected, finding that many of these participants had reported one or more clinical features (e.g., developmental delay, ID, seizures, macrocephaly and/or sleep disturbances) consistent with their genetic diagnosis.

Table 3.2.2.1 Summary of group CNV statistics in the Australian Autism Biobank

Group	n_all	n_ind	n CNV	% with CNV	no. median	no. mean	ASD/ID CNV del	ASD/ID CNV dup	ASD/ID/DD genes del	ASD/ID/DD genes dup
ASD	885	263*	330	29.72	0	0.37	11	4	5	3
FTR	504	134	167	26.59	0	0.33	0	0	1	0
MTR	752	232	280	30.85	0	0.37	3	0	3	0
SIB	218	57	68	26.15	0	0.31	0	0	0	0
UNR	116	37	40	31.90	0	0.34	0	0	0	0

n_all: number of individuals within the entire group; n_ind: number of individuals with a rare exonic CNV in the group; n_CNV: number of rare exonic CNVs in that group; % with CNV: % of individuals with a rare exonic CNV. no. median: median number of rare exonic CNVs; no. mean: mean number of rare exonic CNVs; ASD/ID CNV del: number of rare exonic deletion CNVs overlapping with the ClinGen + DECIPHER deletion CNV set; ASD/ID CNV dup: number of rare

exonic duplication CNVs overlapping with the ClinGen + DECIPHER duplication CNV set; ASD/ID/DD genes del: number of rare exonic deletion CNVs overlapping with the Satterstrom et al. + DDD gene set; ASD/ID/DD genes dup: number of rare exonic duplication CNVs overlapping with the Satterstrom et al. + DDD gene set (100% overlap required).

A further 12 individuals (8 autistic, 4 parents) were found to carry CNVs overlapping exons of protein coding genes with prior robust evidence for association with autism in the most recent autism whole-exome sequencing study[44] or developmental delay (DD), as reported by the Deciphering Developmental Disorders (DDD) study[45] (Tables 3.2.3). We identified three instances in which the CNV appeared to have been transmitted from mother to child. The status of the remaining CNVs (i.e., transmitted or *de novo*) was unclear, because we did not have complete parental information for those children.

We also identified large (>1Mb in length) CNVs in 37 Biobank participants (19 ASD, 1 SIB, 3 UNR, 11 mothers, 3 fathers) that did not overlap any known autism or ID-associated CNVs or gene, a subset of which were shared by relatives, including a 1.3 Mb 16p23.1 deletion occurring in a father-child (ASD) pair; a 1.9 Mb 2q37.3 deletion occurring in a mother–child (ASD) pair; a 2.4 Mb 4q35.2 duplication occurring in a mother and two of her children (both in the UNR group); and a 2.3 Mb 1p34.2 deletion in an identical twin pair (both in the ASD group).

Table 3.2.2.2 ASD/ID-associated CNVs detected in the Australian Autism Biobank dataset.

Diagnosed children						
CNVs with a critical gene						
Group	**Sex**	**Type**	**CNV coordinates**	**ASD/ID-associated CNV**	**Reference coordinates**	**Critical gene coordinates**
ASD	M	Dup	chr17:1196088-1326656	17p13.3 (Miller-Dieker syndrome) region (includes YWHAE and PAFAH1B1)^	chr17:1247833-2588909	YWHAE: chr17:1247569-1268350; PAFAH1B1: chr17:2541583-2585096
ASD	M	Dup	chr17:29111368-30343735	17q11.2 recurrent region (includes NF1)	chr17:29097069-30264027	chr17:29422328-29701173
CNVs without a critical gene						
Group	**Sex**	**Type**	**CNV coordinates**	**ASD/ID-associated CNV**	**Reference coordinates**	**Overlap % of reference**
ASD	F	Dup	chr15:22321690-32515100	15q11q13 recurrent (PWS/AS) region (BP1-BP3, Class 1)	chr15:22832519-28379874	100.00

ASD	M	Del	chr15:29079105-32515100	15q13.3 recurrent region (D-CHRNA7 to BP5) (includes CHRNA7 and OTUD7A)	chr15:32019621-32445405	100.00
ASD	M	Del	chr15:31007901-32515100	15q13.3 recurrent region (D-CHRNA7 to BP5) (includes CHRNA7 and OTUD7A)	chr15:32019621-32445405	100.00
ASD	M	Del	chr15:31115226-32515100	15q13.3 recurrent region (D-CHRNA7 to BP5) (includes CHRNA7 and OTUD7A)	chr15:32019621-32445405	100.00
ASD*	M	Del	chr16:21973913-22414463	Recurrent 16p12.1 microdeletion (neurodevelopmental susceptibility locus)	chr16:21946524-22467284	84.60
ASD	M	Del	chr16:28832565-29044745	16p11.2 recurrent region (distal, BP2-BP3) (includes SH2B1)	chr16:28822635-29046499	94.78
ASD	F	Del	chr22:18877787-21461607	22q11.2 recurrent (DGS/VCFS) region (proximal, A-D) (includes TBX1)	chr22:15912231-21465672	99.84
ASD	F	Del	chrX:6488784-8135053	Xp22.31 recurrent region (includes STS)	chrX:6455812-8133195	98.02
Parents						
MTR*	M	Del	chr16:21956457-22414463	Recurrent 16p12.1 microdeletion (neurodevelopmental susceptibility locus)	chr16:21946524-22467284	87.95
MTR	F	Del	chr22:19036154-20244259	22q11.2 recurrent (DGS/VCFS) region (proximal, A-B) (includes TBX1)	chr22:18912231-20287208	87.86
MTR	F	Del	chrX:6456940-8135053	Xp22.31 recurrent region (includes STS)	chrX:6455812-8133195	99.93

*ASD/ID-associated CNVs were taken from ClinGen and DECIPHER datasets, filtering for ASD/ID-associated loci. For reference CNVs with a critical gene, AAB CNVs were annotated where there was any overlap with the critical gene, with the critical gene coordinates provided in the "Other information" column. For reference CNVs without a critical gene, the AAB CNV was called as overlapping with an ASD/ID-associated CNV based on ≥ 80% overlap with the reference coordinates, with percentage overlap provided in the "Other information" column. Genome coordinates are hg19. Biobank Sample IDs have been anonymized. *Refers to parent–child pairs between which ASD/ID-associated CNVs appeared to be inherited in this dataset.*

Table 3.2.2.3 CNVs identified in the Australian Autism Biobank overlapping ASD- and ID/DD-associated genes.

Group	Sex	CNV coordinates	Gene	Overlap % of gene	Cytoband
Diagnosed children					
ASD*	F	chr2:148730454-148883419	MBD5	21	2q23.1
ASD+	M	chr10:27978030-28041669	MKX	78	10p12.1
ASD	M	chr19:10609319-12464434	ELAVL3	100	19p13.2
ASD^	M	chr20:61824507-62321517	KCNQ2	100	20q13.33
ASD^	M	chr20:61802599-62268955	KCNQ2	100	20q13.33
ASD	M	chr2:32277654-32818823	SPAST	100	2p22.3
ASD	M	chr4:6104865-7415038	KIAA0232	100	4p16.1
ASD	F	chr15:22321690-32515100	GABRB3	100	15q12
Parents transmitting CNVs					
Mother*	F	chr2:148730454-148883419	MBD5	21	2q23.1
Mother+	F	chr10:27978030-28041669	MKX	78	10p12.1
Mother^	F	chr20:61802599-62321517	KCNQ2	100	20q13.33

**, +, ^: denote CNVs shared by individuals within the same family (either inherited from parents, or shared between siblings), suggesting inheritance. CNV: copy number variant.*

A density plot of CNV deletions and duplications across all individuals, in relation to chromosome, cytoband, ASD/ID-associated CNVs and ASD/ID/DD-associated genes is shown in Figure 3.2.2.1.

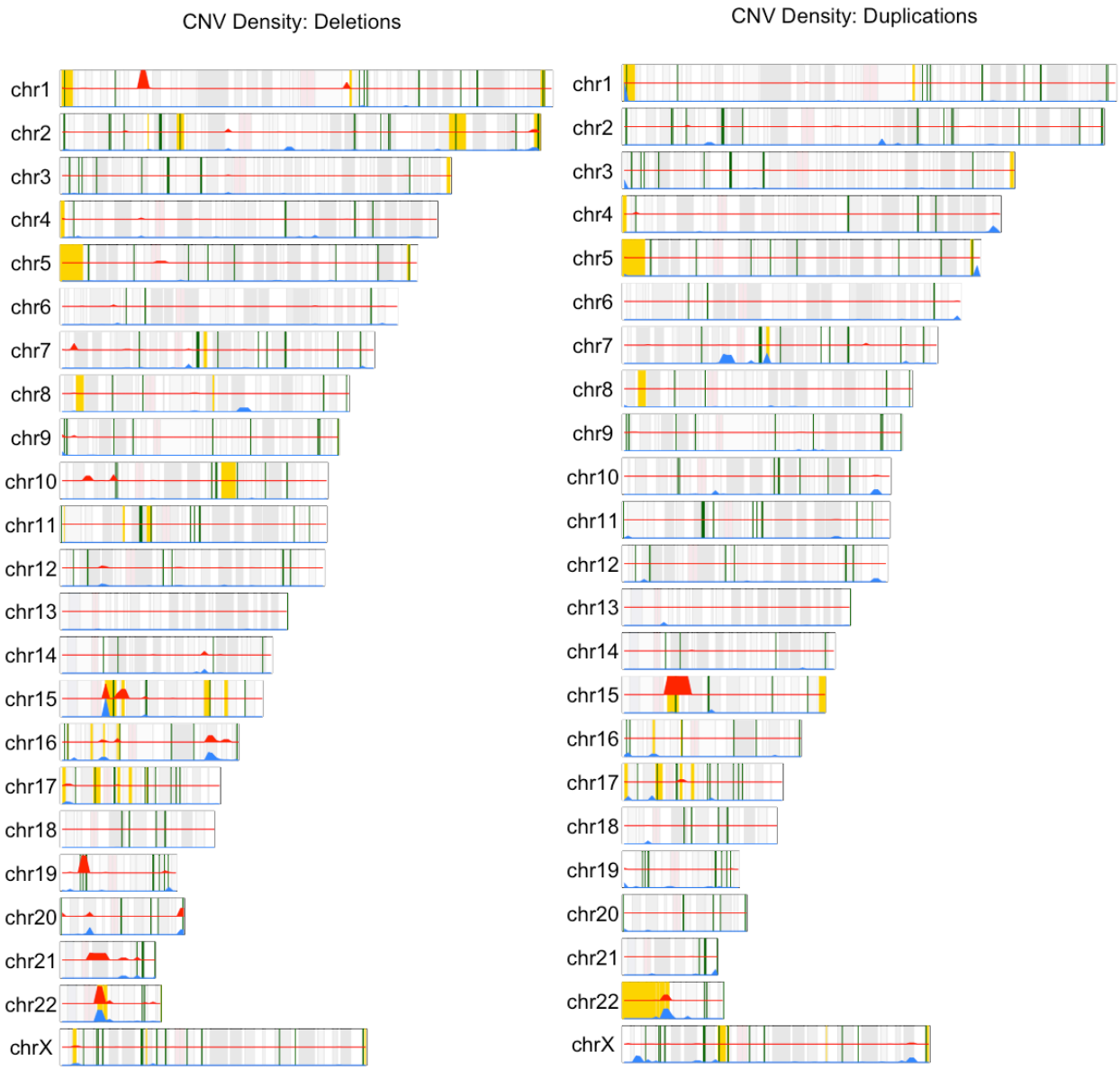


Figure 3.2.2.1 Karyograms showing location and density of deletion and duplication CNVs identified in the Australian Autism Biobank. Red density track represents CNVs detected in the ASD subset. Blue density track represents CNVs detected in the non-ASD subset (undiagnosed siblings, unrelated undiagnosed children, parents), noting that there are instances in which parents have CNVs overlapping ASD/ID-associated regions. Yellow regions depict ASD/ID-associated CNVs from ClinGen and DECIPHER. Green regions denote ASD/ID/DD-associated genes reported by Satterstrom et al. and the DDD study.

3.3 Gut microbiome

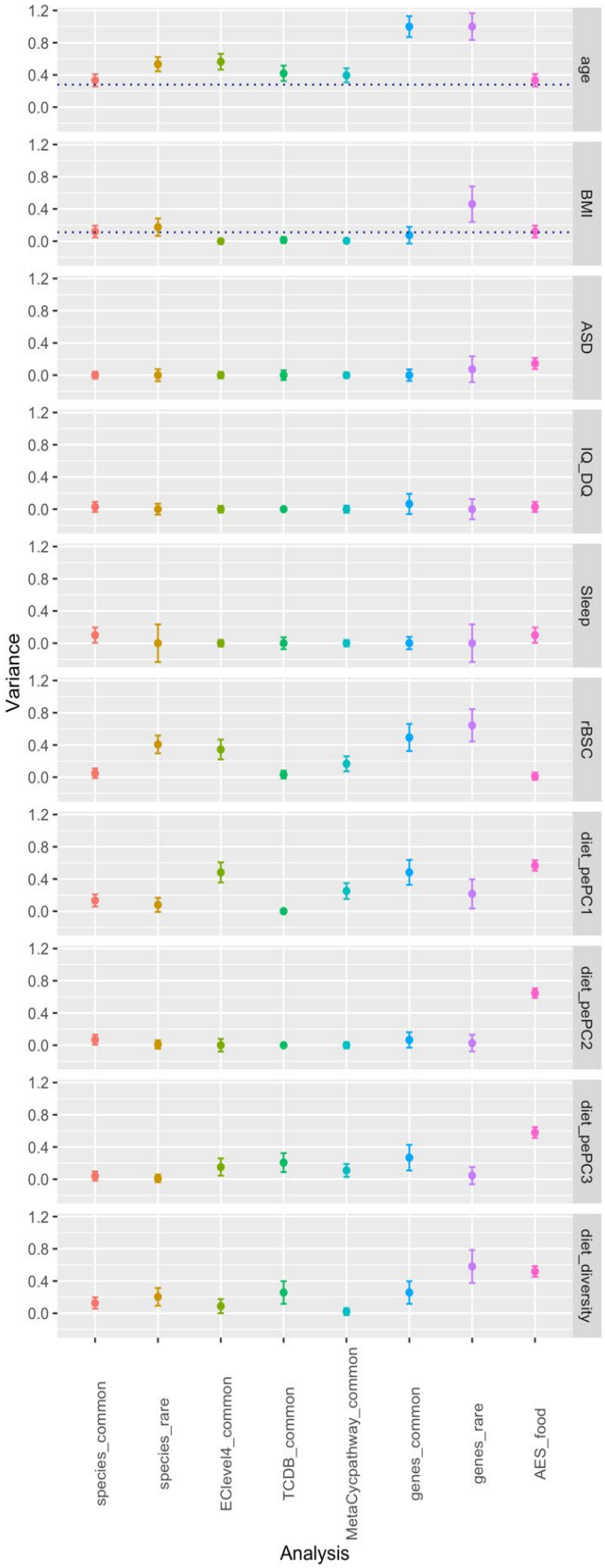
3.3.1 Data summary

Stool-derived metagenomic profiles were generated by Microba from a total of 199 age and sex-matched Biobank child participants and 49 unrelated controls from the Queensland Twin Adolescent Brain (QTAB) study. Between 4.8 and 24.8 million reads passed quality control per sample. Down sampling to a standard depth of 7 million read pairs was performed to alleviate the effect of differential sensitivity in association analyses, with all QC-ed reads retained for the n=19 samples with <7M reads. Data on n=247 individuals with matching dietary data (Australian Eating Survey [26,27]) were retained for analyses. A total of 1,758 species were identified in the dataset, of which 607 remained for analysis after removing low prevalence species (<10 non-zero values).

3.3.2 Variance in autism diagnostic status and other traits associated with the microbiome

First, we estimated the proportion of phenotypic variance associated with the gut microbiome for autism diagnosis and a range of other traits, including neurodevelopmental traits (IQ-DQ, Children's Sleep Habits Questionnaire (CHSQ) raw score), phenotypes with intuitive relationships with the microbiome (dietary pe_PCs, dietary diversity, stool consistency; see Methods), and CD4 T-cell proportion. This variance estimate – the microbiome-association-index (or "b^2") – is analogous to heritability (h^2) from genetic analyses and provides an upper limit for predictive ability. Whereas h^2 reflects causality, b^2 may reflect cause or consequence of trait variation. This analysis involves calculating an "omics relatedness matrix" (ORM) from microbiome features between each pair of individuals and regressing against the trait in a linear mixed model framework[46]. We created ORMs from several measures of microbiome composition, at the level of species and genes, and stratifying into common (those with median count > 0) versus rare features (median count=0, but with >=10 non-zero counts in the full sample).

As a benchmarking exercise, we first analysed age and BMI. Our results based on common species (n=96) were consistent with previous species-level b^2 estimates for these traits in adults (Rothschild *et al* 2020) (Figure 3.3.3.1). Notably, gene-level ORMs were associated with greater variance for both age and BMI (Figure 3.3.3.1), suggesting that taxonomic and functional microbiome measures capture different elements of phenotypic variance. These results were robust to antibiotics usage.



In notable contrast to the results for age, species- and gene-level b^2 estimates for autism diagnosis were weak and non-significant (maximum: rare genes b^2=7%, SE=16%, p=0.33; covariates: age, sex, pe_PC1-3), as were those for IQ-DQ (common species b^2=6%, SE=13%, p=0.39; covariates: age, sex), CSHQ raw score (common species b^2=10%, SE=9%, p=0.17; covariates: age, sex) (Figure 3.3.2.1) and clr-transformed CD4 T-cell proportion (results not shown).

Unlike these neurodevelopmental and immune traits, we observed strong FDR-significant b^2 estimates for both stool consistency (rBSC) (covariates: age, sex, group, pe_PC1-3; rare species b^2=41%, SE=11%, p=8.7x10^{-6}; rare genes b^2=64%, SE=20%, p=2.5x10^{-5}) and dietary pe_PC1 (rare genes: b^2=48%, SE=15%, p=3.8x10^{-4}; covariates: age, sex, participant group) (Figure 3.3.2.1).

Figure 3.3.2.1 Proportion of phenotypic variance associated with microbiome composition. Rows denote phenotypes, including benchmarking traits (age and BMI; with dotted lines showing results from the Rothschild et al. (2020) species-level analysis), neuropsychiatric traits (ASD, IQ-DQ composite score, CSHQ raw score) and microbiome-related traits (stool consistency measured as regrouped Bristol Stool Chart (rBSC), pe_PC1-3, dietary diversity calculated using Shannon index). The y-axis shows the proportion of total phenotypic variance associated with the relevant measure of microbiome composition. Columns denote dataset used to generate the ORM. "*common" indicates ORMs calculated using variables with median>0; "*rare" indicates ORMs calculated using variables with median=0; all datasets are from the metagenomics dataset except "AES_food" which is based on the Australian Eating Survey food-level data.

3.3.3 Differentially abundant species and genes

Next, we looked for specific microbial taxa and genes associated with autism diagnosis, testing for differential abundance of 607 species, 297 genera, 38 orders, 15 phyla. We used ANCOMv2.1[31], a robust, non-parametric method that accounts for multiple testing and adequately controls the false positive rate[32].

In a comparison of ASD and the combined SIB+UNR groups, with age, sex and pe_PC1-3 as covariates, only the species *Romboutsia timonensis* was significantly differentially abundant (lower abundance in ASD) at the conventional detection threshold >0.7 (Figure 3.3.3.1a). We were unable to account for familial relatedness due to model singularities, so performed a sensitivity analysis, comparing the ASD and UNR groups (minus SIBs), again finding *R. timonensis*, and reduced *Erysipelatoclostridium sp003024675* at detection threshold >0.6 (Figure 3.3.3.1b).

The results were largely consistent when excluding participants (n=10) reporting antibiotic use at sample collection. In differential-presence testing (Fisher's exact test, ASD vs UNR), the same two taxa (*R. timonensis* $p=3.9x10^{-4}$, 56/99 ASD vs 78/97 UNR; *E. sp003024675* $p=1.5x10^{-4}$, 6/99 ASD vs 25/97 UNR) were also the top ranked species, though neither survived FDR-correction. In permutation testing (n=1000 random shuffles of diagnostic labels for each sample), these taxa had $p \leq 0.001$ when compared to the empirical distributions for both ANCOM and Fisher's exact tests.

Notably, we failed to replicate previously-reported autism-gut microbiome associations[47] with the genus *Prevotella*, phylum Firmicutes, Clostridiales clusters and species of *Bifidobacterium*. However, we note that *R. timonensis* (family Peptostreptococcaceae, order Clostridiales, class Clostridia, phylum Firmicutes A) and *E. sp003024675* (family Erysipelatoclostridiaceae, order Erysipelotrichales, class Bacilli, phylum Firmicutes) are members of these phylogenetic groups. Poor replication may be related to: 1) prior microbiome studies being underpowered and prone to sampling biases; 2) technical differences between metagenomics and 16S rRNA sequencing – the former providing more detailed taxonomic resolution, which is relevant as *R. timonensis* was only recently isolated in human gut in the setting of malnutrition[48]; 3) different statistical methods, with variable use of adequate adjustment for multiple-testing and/or confounders.

Unlike 16S sequencing studies (which have dominated the autism microbiome literature), metagenomics sequencing permits functional insights. Hence, we looked for microbial genes associated with autism diagnosis. We focused on relative abundances of metagenome annotations to MetaCyc groups, MetaCyc pathways and EC gene families (as opposed to the entire gene set due to computational and multiple-testing burden), in order of increasing resolution, finding no significant associations.

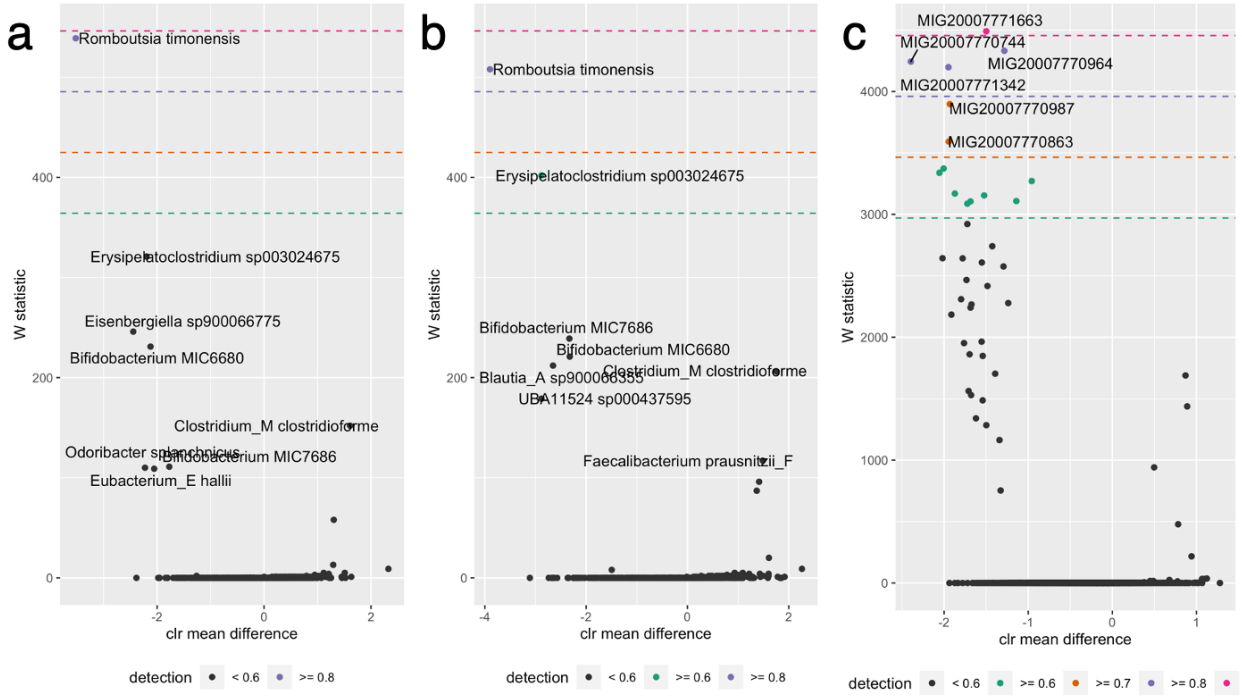


Figure 3.3.3.1 Differential abundance testing in ASD at the level of species (a, b) and genes (c) using ANCOM. Plots show differentially abundant species for (a) the full ASD analysis (n=246), (b) ASD versus UNR excluding SIB (n=196) and differentially abundant (c) gene-level results for ASD vs SIB+UNR analysis, focusing on n=4,950 genes (with >10 non-zero counts) mapping to the R. timonensis genome. The W-statistic (y-axis) represents, for feature_i of n total features, the count of Bonferroni-Hochberg-significant p-values from regressing the additive-log-ratio-transformation (= log(feature/feature_i) against ASD and covariates age, sex and pe-PC1-3). Dotted lines show W-statistic quantile detection thresholds ≥0.6 (green), ≥0.7 (orange), ≥0.8 (purple) and ≥0.9 (pink) respectively; features exceeding detection threshold >0.7 are considered significantly differentially abundant. The x-axis ("clr mean difference") shows the coefficient for regressing the clr-transformed feature against the variable of interest (in this case, Autism diagnosis).

Next, we sought to identify specific genes or pathways from *R. timonensis* underlying the autism-associated signal. We performed ASD vs SIB+UNR differential abundance testing for MetaCyc groups, MetaCyc pathways, EC gene families, and \n=4,950 specific genes with >10 non-zero values across samples) directly mapping to *R. timonensis* in our dataset (covariates: age, sex, pe_PC1-3). We identified six genes with detection threshold >0.7 (Figure 3.3.3.1c), one of which overlapped with the EC gene set, whereas there were no associations with MetaCyc groups or pathways. Consistent with the species-level direction of effect, all significantly differentially abundant genes were reduced in the ASD group. Their functions included metabolism of amino acids (L-glutamine, L-lysine, L-methionine and L-threonine), purines and pyrimidines, carbohydrates (galactose), as well as bacterial spore germination and dsDNA digestion. We note

that these results represent potential transcription, and that metatranscriptomics data would be needed to evaluate actual expression.

We performed species-level differential abundance analysis for IQ-DQ composite score, finding that lower *Bifidobacterium sp002742445* passed detection threshold >0.7 in the ASD vs SIB+UNR comparison, but only passed detection threshold >0.6 in the ASD vs UNR sensitivity analysis. There were no virome associations in the ASD vs SIB+UNR analysis (n=200 taxa).

3.3.4 Restricted diet mediates autism-microbiome associations

Whereas autism-associated signals were scarce in the metagenomics data, there were consistent associations with diet: in the variance component analysis, the dietary ORM was strongly associated with autism (R^2=14%, SE=7%, p=2.2x10^{-5}) (Figure 3.3.4.1), and we observed significantly lower dietary pe_PC3 in autism (suggesting reduced meat intake) compared to SIB and UNR, after adjusting for age and sex. Given that children on the spectrum favour restricted diets[49-53], we explored the effect of dietary restrictedness on the microbiome and stool consistency.

The ASD group had significantly less-diverse diet (estimated using dietary alpha-diversity from the Shannon index of 123 food-level variables) than both SIB and UNR (one-way ANOVA p=1.3x10^{-7}), even after adjusting for age and sex (one-way ANOVA p=2.2x10^{-6}) (Figure 3.3.4.1a). Investigating this link between autism and dietary diversity further, we hypothesised, on the basis that there was no evidence for direct relationships between autism and microbiome alpha- or beta-diversity, that microbiome diversity may be affected downstream of diet. Consistent with this hypothesis, there was a significant positive correlation between dietary and taxonomic diversity (r=0.25, p=6.3x10^{-5}) (Figure 3.3.4.1c), and in reciprocal regression analyses, dietary and taxonomic diversity were significant predictors of each other (Figure 3.3.4.1d-e). Furthermore, the largest effects in the dietary diversity regression were from group (UNR: b=0.035, p=3.0x10^{-6}; SIB: b=0.021, p=1.4x10^{-2}), whereas taxonomic diversity was not associated with group (Figure 3.3.4.1e). This suggests that autism-associated dietary restrictedness (but not diagnosis itself) is associated with reduced microbiome diversity. These effects were robust in sensitivity analyses accounting for antibiotic and probiotic use.

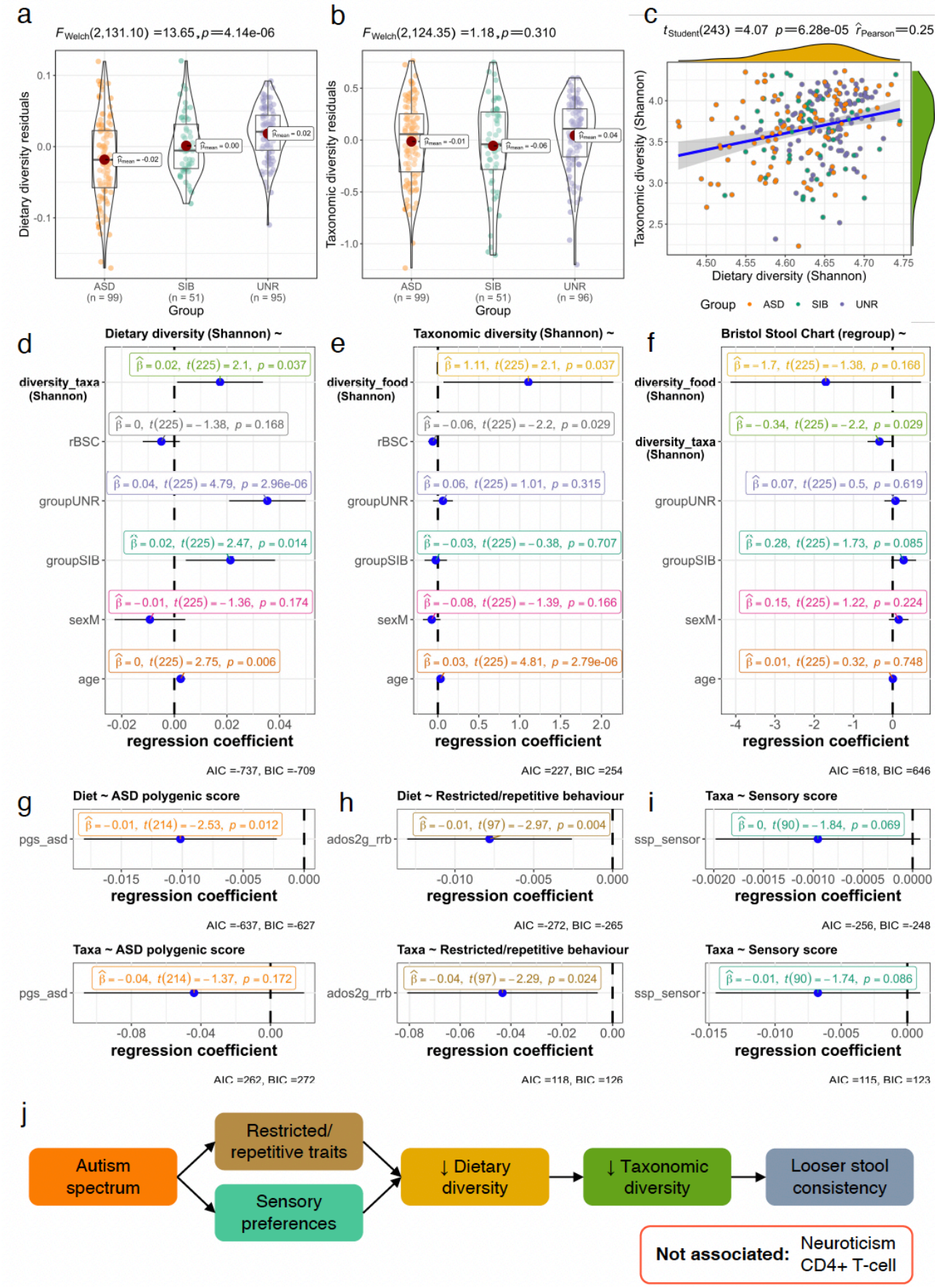

a
$F_{Welch}(2,131.10) = 13.65, p = 4.14e-06$
Dietary diversity residuals
ASD (n = 99)
SIB (n = 51)
UNR (n = 95)
Group
b
$F_{Welch}(2,124.35) = 1.18, p = 0.310$
Taxonomic diversity residuals
ASD (n = 99)
SIB (n = 51)
UNR (n = 96)
Group
c
$t_{Student}(243) = 4.07$ $p = 6.28e-05$ $\hat{r}_{Pearson} = 0.25$
Taxonomic diversity (Shannon)
Dietary diversity (Shannon)
Group ASD SIB UNR
d
Dietary diversity (Shannon) ~
diversity_taxa (Shannon)
rBSC
groupUNR
groupSIB
sexM
age
regression coefficient
AIC =-737, BIC =-709
e
Taxonomic diversity (Shannon) ~
diversity_food (Shannon)
rBSC
groupUNR
groupSIB
sexM
age
regression coefficient
AIC =227, BIC =254
f
Bristol Stool Chart (regroup) ~
diversity_food (Shannon)
diversity_taxa (Shannon)
groupUNR
groupSIB
sexM
age
regression coefficient
AIC =618, BIC =646
g
Diet ~ ASD polygenic score
pgs_asd
regression coefficient
AIC =-637, BIC =-627
Taxa ~ ASD polygenic score
pgs_asd
regression coefficient
AIC =262, BIC =272
h
Diet ~ Restricted/repetitive behaviour
ados2g_rrb
regression coefficient
AIC =-272, BIC =-265
Taxa ~ Restricted/repetitive behaviour
ados2g_rrb
regression coefficient
AIC =118, BIC =126
i
Taxa ~ Sensory score
ssp_sensor
regression coefficient
AIC =-256, BIC =-248
Taxa ~ Sensory score
ssp_sensor
regression coefficient
AIC =115, BIC =123
j
Autism spectrum
Restricted/ repetitive traits
Sensory preferences
↓ Dietary diversity
↓ Taxonomic diversity
Looser stool consistency
Not associated: Neuroticism CD4+ T-cell

Figure 3.3.4.1 Relationships between dietary and taxonomic diversity (measured using Shannon Index) and ASD-related phenotypes. a) Boxplot of dietary diversity (residuals after regressing out age and sex) between participant groups (ANOVA $p=2.1x10^{-6}$). b) Boxplot of taxonomic diversity (residuals after regressing out age and sex) between participant groups (ANOVA p=0.15). c) Correlation between dietary and taxonomic diversity (r=0.25, $p=6.3x10^{-5}$). d-i) Linear model plots[54], showing effect sizes (+/-95%CI), test statistics, degrees of freedom and p-values for each term: d) Linear model coefficients taking dietary diversity as the dependent variable. e) Linear model coefficients taking taxonomic diversity as the dependent variable. f) Linear model coefficients taking rBSC as the dependent variable. g) Linear model coefficients regressing dietary (upper) and taxonomic (lower) diversity against ASD PGS. h) Linear model coefficients regressing dietary (upper) and taxonomic (lower) diversity against ADOS2/G RRB score. i) Linear model coefficients regressing dietary (upper) and taxonomic (lower) diversity against ADHD-ASD-TS PGS. j) Proposed synthesis of relationships between autism spectrum measures, restricted and repetitive interests, sensory preferences, dietary diversity, taxonomic diversity and stool consistency.

Next, we explored relationships between dietary and taxonomic diversity and stool consistency (rBSC). We replicated a previously reported[55-57] inverse relationship between rBSC and taxonomic diversity (b=-0.41 $p=3.7x10^{-3}$ without covariates) that was robust to covariates. We also identified nominally significant association of rBSC with dietary diversity (b=-2.34, $p=3.7x10^{-2}$ in the model without covariates), although this did not survive covariate adjustment. Notably, the taxonomic diversity model (without covariates) explained greater variance (R^2=3.2%) than the dietary diversity analysis (R^2=1.5%), and in models of rBSC with both taxonomic and dietary diversity fitted as explanatory variables, only taxonomic diversity was significant (Figure 3.3.4.1f). Overall, this suggests that looser stool consistency (higher rBSC) is proximally related to reduced taxonomic diversity, which is downstream of reduced dietary diversity. This mechanism may explain findings of increased gastrointestinal issues with increased repetitive behaviours[58].

3.3.5 Upstream drivers of dietary and taxonomic diversity

We investigated whether behavioural factors diagnostic of autism are upstream drivers of restricted diet and reduced dietary and taxonomic diversity. To achieve this, we leveraged both psychometric measures and polygenic scores (for autism and other phenotypes).

First, we confirmed the autism-dietary diversity association through analysis of continuous autism-spectrum measures. We identified an inverse association between autism polygenic score (Yap et al., 2021) and dietary diversity (b=-1.0e2, $p=1.2x10^{-2}$), but not taxonomic diversity (b=-4.4e-2, p=0.17) (Figure 3.3.4.1g). Phenotypically, we observed a negative, marginal association between dietary diversity and Social Responsiveness Scale t-score ($b=-6.4x10^{-4}$, $p=7.8x10^{-2}$; n=97 AAB children: 10 ASD and 87 SIB/UNR).

Second, we hypothesised that repetitive-restrictive behaviours may underlie a restricted diet upstream of microbiome changes. Phenotypically, we observed FDR-significant negative association between higher combined[59] ADOS2/G restricted and repetitive behaviour (RRB) scores and taxonomic diversity (b=-4.3 $x10^{-2}$, p=3.8 $x10^{-3}$, with covariates age and sex: b=-6.4 $x10^{-3}$, p=1.8 $x10^{-2}$), and nominal negative association with dietary diversity (b=-4.3 $x10^{-2}$, p=2.4 $x10^{-2}$) (n=97 ASD group only, Figure 3.3.4.1h). We then leveraged GWAS summary statistics from a cross-trait analysis[33] of autism, attention-deficit hyperactivity disorder, and Tourette's Syndrome (hereafter called ASD-ADHD-TS) to generate PGS for restrictive-repetitive behaviours. We confirmed that ASD-ADHD-TS PGS correlated with ADOS-2/G RRB scores in the full AAB (r=0.10, p=3.2$x10^{-3}$, n=868), and so represents a genetic proxy for RRBs. We found marginal negative association between ASD-ADHD-TS PGS and dietary diversity (b=-7.2$x10^{-3}$, p=0.10), but not with taxonomic diversity (b=-4.7$x10^{-2}$, p=0.18; Figure 3.3.3i).

Third, on the basis that sensory sensitivity may also underlie restricted dietary preferences[60], we explored relationships with Short Sensory Profile raw Sensory score in a small autism-only subset of the data for which this instrument was completed and found marginal associations with both dietary (b=-9.5$x10^{-4}$, p=6.9$x10^{-2}$) and taxonomic diversity (b=-6.8$x10^{-2}$, p=8.6$x10^{-2}$), (n=91).

In contrast, we found no evidence for hypothesized links between both dietary and taxonomic diversity measures and autism-associated traits such as anxiety, neuroticism[61,62] or functional gastrointestinal disorders (e.g., irritable bowel syndrome).

Overall, these data suggest that autism-associated preferences and behaviours drive reduced dietary diversity, which mediates weak autism-microbiome relationships (Figure 3.3.4.1j). However, we cannot rule out the possibility that these downstream microbiome effects could feed-back and influence behaviour, given that there was a stronger association between ADOS-2/G RRB score and taxonomic diversity than dietary diversity (Figure 3.3.4.1h).

A manuscript describing analysis of this dataset has been provisionally accepted for publication in *Cell*: Yap, C.X., et al. "Restricted diet mediates autism-gut microbiome associations."

3.4 DNA methylation

Blood-derived genome-wide DNA methylation data were generated for a total of 468 Biobank participants, including 255 autistic children, 125 non-autistic siblings and 88 unrelated controls using the Illumina EPIC Human Methylation array.

3.4.1 Variance explained by the methylome

We used OREML to quantify the proportion of variance for various traits associated with variation in genome-wide DNA methylation. The traits analysed included autism diagnostic status, autism polygenic score, Short Sensory Profile (SSP_sensory) raw sensory score, Children's Sleep Habits Questionnaire (CSHQ) raw score, a composite score for IQ (composite scores from the NIH Toolbox age-adjusted questionnaire) and DQ (Mullen's Scales of Early Learning non-verbal composite score), hereafter referred to as "IQ-DQ", and age. We also examined the effect of including covariates (age, sex, genotype PCs and cell-type proportions). Within the AAB-QTAB dataset, we found no evidence for an association between DNA methylation and any neurodevelopmental trait, whereas we identified a very strong association for age.

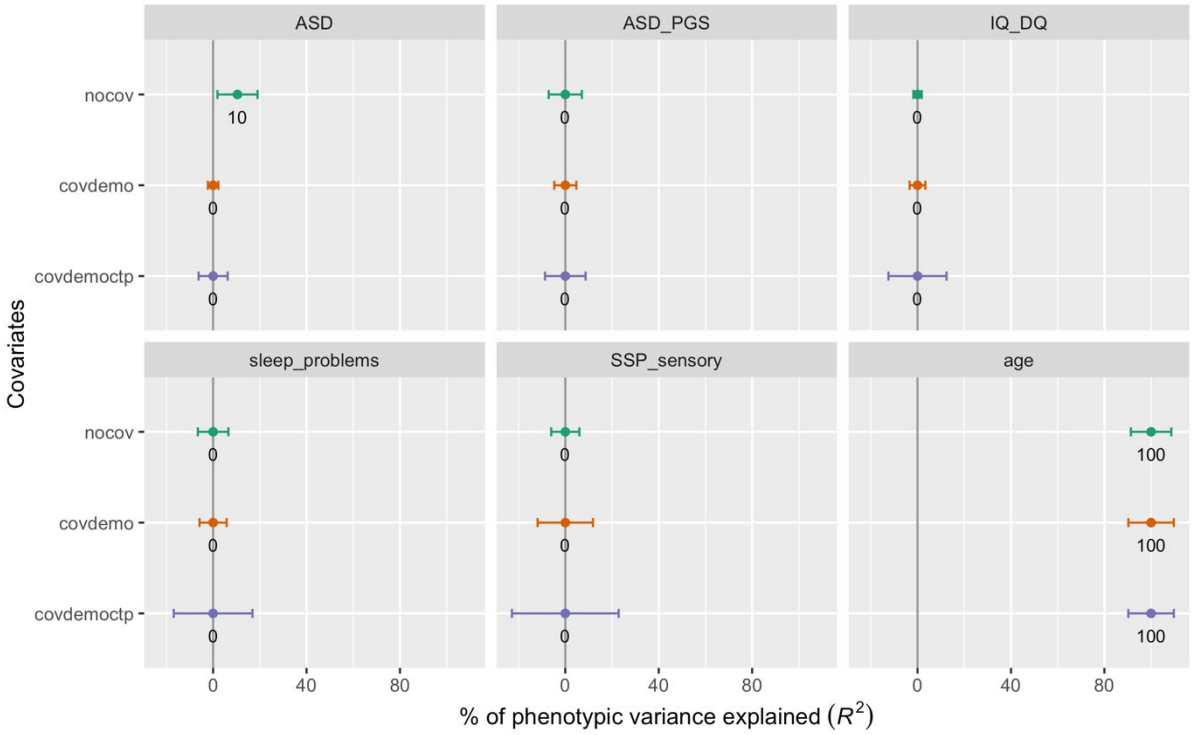


Figure 3.4.1.1: Variance component analysis for traits (grey box text) using a relatedness matrix calculated using methylation data from the AAB-QTAB study. Colours denote different combinations of covariates: green denotes no covariates; orange denotes demographic covariates (sex, batch, genotyping PCs, and age for all phenotypes except for age); purple denotes demographic variables and principal components capturing >95% of variance in cell-type proportions. Text denotes the percentage of variance in the trait associated with the methylation relatedness matrix. Error bars correspond to standard error for the estimate. ASD: autism spectrum disorder diagnosis; ASD_PGS: ASD polygenic score; IQ_DQ: composite intelligence quotient and developmental quotient score; SSP_sensory: raw sensory score on the Short Sensory Profile-2.

3.4.2 Methylome-wide association study (MWAS) and meta-analysis

We tested for association among 301,589 methylation probes and autism diagnostic status, finding no associations within the AAB-QTAB cohort. There were also no significant associations detected from meta-analysing MWAS summary statistics from AAB-QTAB (n=468) and MINERvA (n=1,263) cohorts (total n=1,731 participants) (Figure 3.4.2.1).

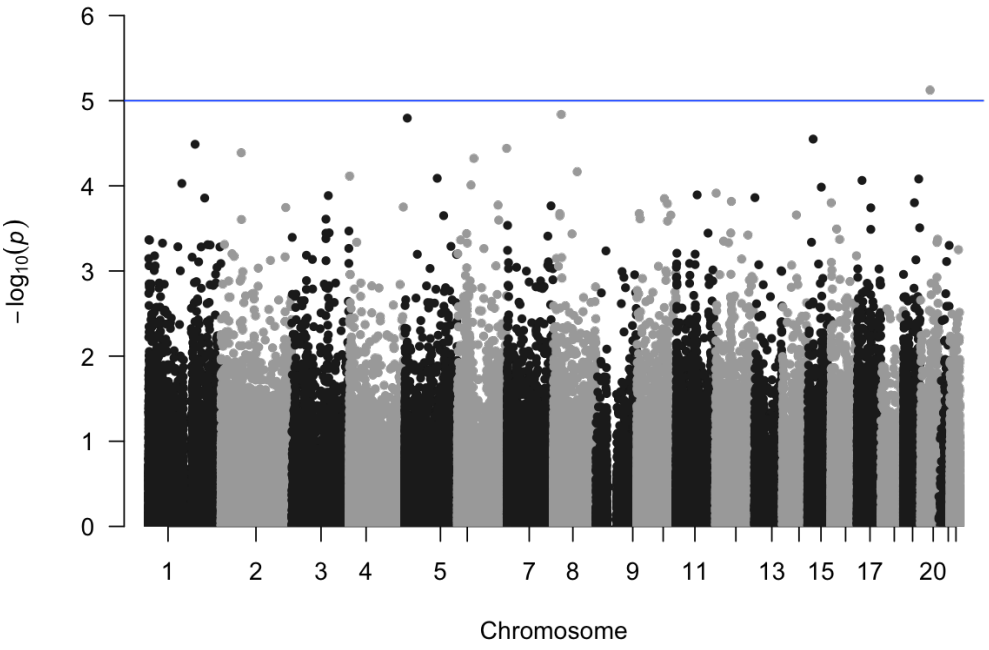


Figure 3.4.2.1: Manhattan plot for the AAB-QTAB and MINERvA MWAS meta-analysis.

The most significantly-associated probe in the AAB-QTAB and MINERvA meta-analysis was cg01122366 (Z=4.79, p=1.67x10^{-6}) – a probe that was only in the AAB-QTAB dataset as it was only available on the EPIC array and not the 450K array. This probe is within the gene *GALNT2*, a gene for which it has been reported that loss-of-function variants are associated with a genetic syndrome which includes neurodevelopmental delay and autistic features[63]. The positive direction effect (i.e., increased methylation is associated with increased propensity for autism) is also consistent with the clinical syndrome being characterised by loss-of-function variants.

The next-most significant hit corresponded to the methylation probe cg12699865 (p=7.51x10^{-6}, Z-score=-4.48), on chromosome 20 within the 5'UTR of the *RALY* gene. This probe was the lead hit in the MINERvA analysis (p=7.63e-7)[38], but had negligible signal in the AAB-QTAB cohort (p=0.62), suggesting that this association was driven by the larger MINERvA cohort. Reassuringly, however, both studies demonstrated a negative direction of effect for the association between DNA methylation status and autism diagnosis.

3.4.3 Conclusions

We found minimal evidence for association between DNA methylation and diagnosis of autism, including in variance components analyses considering all DNA methylations jointly, and in methylome-wide association studies of individual probes. We cannot rule out a small effect of DNA methylation, and so larger studies will be required to fully assess the relationship between DNA methylation and autism. It is also possible that DNA methylation profiles in blood – the main focus of this study due to our interest in identifying genomic biomarkers in an accessible tissue – do not capture DNA methylation changes in specific brain regions and/or cell types relevant to autism, since DNA methylation is known to tissue and cell-type specific. A further possibility is that putative DNA methylation changes associated with autism occur during embryonic development, or the early post-natal period, prior to autism diagnosis. Overall, further study will be required to establish the association of DNA methylation with diagnosis of autism.

3.5 Lipidomics

Data was available on 781 lipid species in 39 classes for a total of 758 (479 ASD, 160 SIB, 119 UNR) AAB and QTAB participants, after quality control procedures. A detailed description of outlier samples is provided on page 50.

Variance components analyses: We used the OREML method to assess the association of all lipid species (inverse normal transformed) jointly with autism, IQ_DQ, sleep problems, age and Tanner developmental stage. We identified strong associations with age and Tanner that were robust to inclusion of demographic, batch and dietary covariates (Figure 3.5.1). In comparison, a significant association with autism diagnosis in the analysis without covariates was absent in the analysis with demographic and batch covariates and was found to be explained by the presence of n=64 storage time outliers, all in the ASD group. These storage time outliers belong to the PathWest sample, which was collected prior to other subsets of the AAB. Interestingly, we observed significant and robust associations with IQ_DQ, and weaker but nonetheless robust associations with sleep problems. No association was observed for gross motor skills or Bristol Stool Chart.

In the OREML analysis of dietary phenotypes, we observed strong and significant associations with total cholesterol and total protein that were robust to demographic and batch covariates and also energy intake (Figure 3.5.2). In comparison, whereas we observed comparable associations with total fats, sugars and carbohydrates in analyses without covariates and demographic and

batch covariates, these did not survive adjustment for energy intake. The association of lipids with cholesterol and protein (in particular) intake, but not fats, is notable.

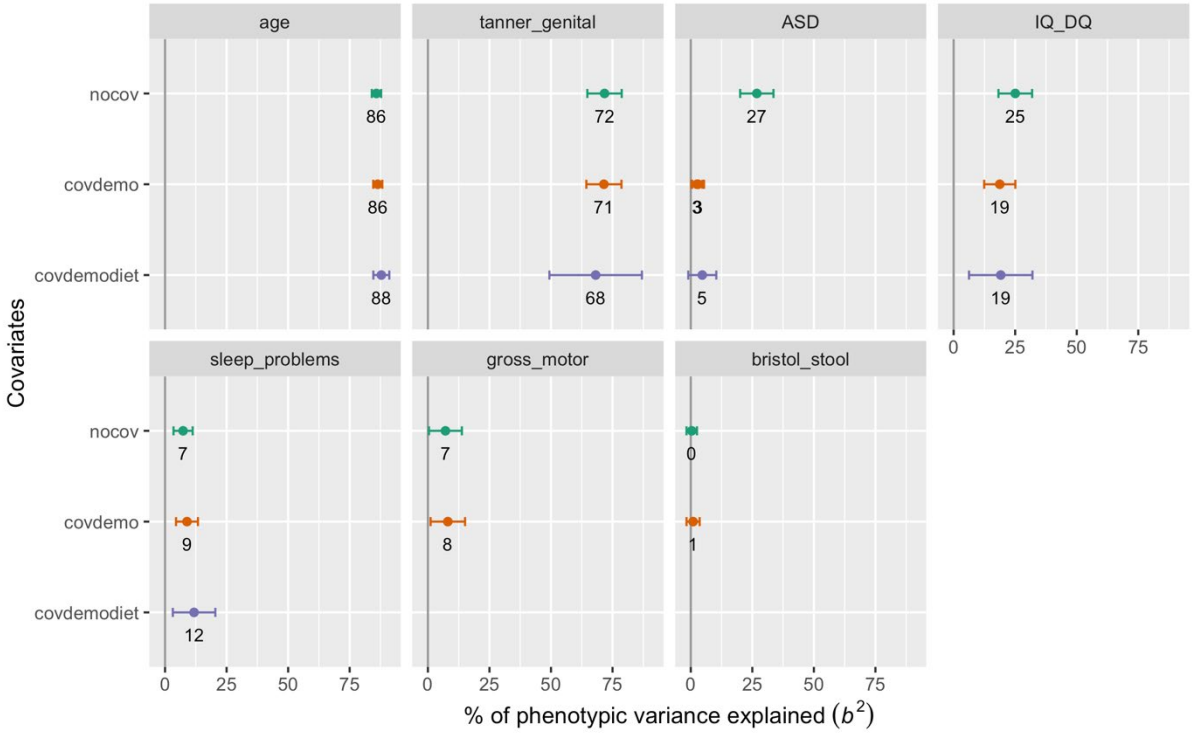


Figure 3.5.0.1 Variance components analyses for age-related and ASD-related traits (grey box text) using a relatedness matrix calculated using inverse normal transformed lipid species data from the AAB-QTAB study. Colours denote different combinations of covariates: green denotes no covariates ("nocov"); orange refers to the analysis including age (except for age and Tanner phenotypes), sex, batch, injection order and storage time covariates ("covdemo"); purple refers to the analysis on the subset of participants for whom dietary data was available (n=260; "covdemodiet"). Text denotes the percentage of variance in the trait associated with the lipidomics relatedness matrix. Error bars correspond to standard error for the estimate. ASD: autism spectrum disorder diagnosis; IQ_DQ: composite intelligence quotient and developmental quotient score.

Lipidome-wide association studies (LWAS): We tested for association between individual lipids and autism diagnosis, ID/DQ, sleep problems, age and Tanner developmental stage. We performed analyses at the level of lipid classes and lipid species. For autism, we found one significantly associated lipid class (Total PC(O), OR=0.73, SE=0.09, $p=7.1x10^{-4}$), but no lipid species (Figure 3.5.3). For IQ-DQ, we identified two lipid species (PC(P-35:2) (b) and LPC(O-22:0); Figure 3.5.4), and for sleep problems (measured using the CSHQ), two lipid classes (total PE(O), total LPE(P)) and five lipid species (PC(15:0_22:6), PC(P-18:0/22:6), PE(P-19:0/20:4) (b), PE(P-20:1/22:6), dimethyl-CE(22:6; Figure 3.5.5) were significant in the final models. Subsequent analyses accounting for diet in a sub-set of 260 individuals with data from the Australian Eating Survey (AES), indicated that these associations with autism, ID/DQ and sleep were largely

explained by diet, although, in the case of autism this is known to be confounded with diagnosis. In contrast to these neurodevelopmental and behavioural traits, we observed much stronger associations with age (n=7 lipid classes and n=78 remaining in the final model after backwards stepwise regression; Figure 3.5.6) and Tanner developmental stage (n=3 lipid classes and n=14 species in the final model; Figure 3.5.7). See Appendix B for complete details of all associated lipid classes and species.

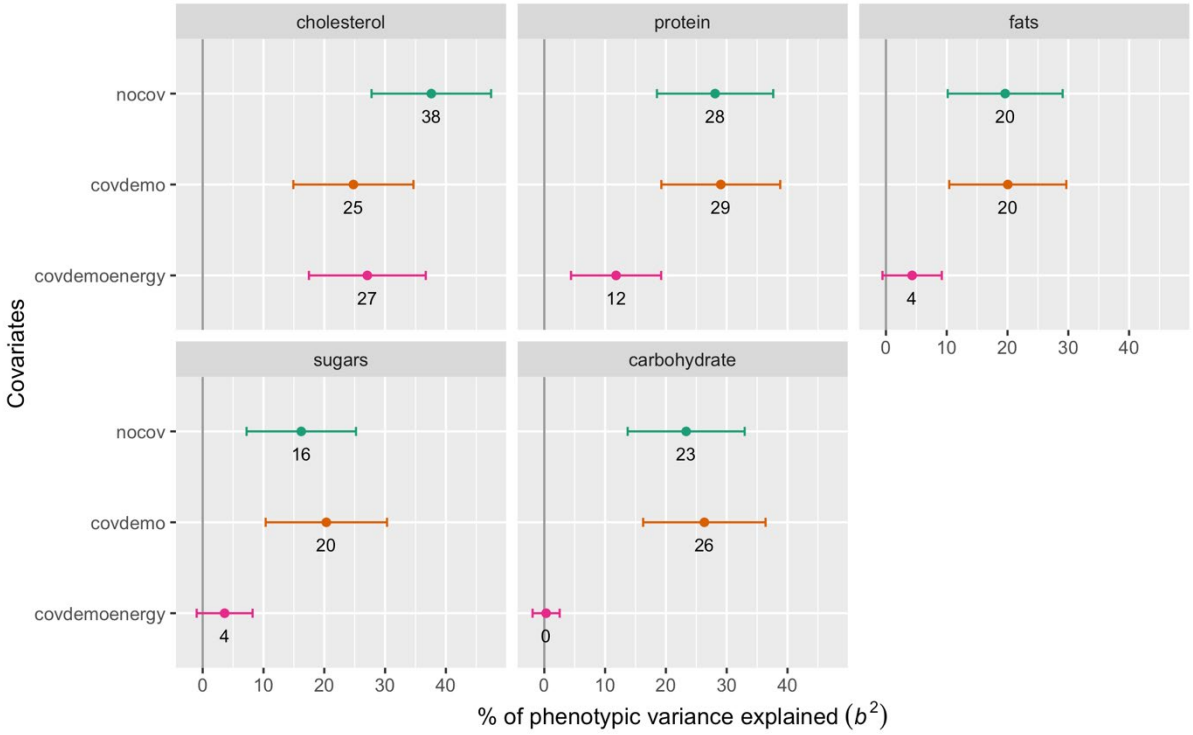


Figure 3.5.0.2 Variance components analyses for dietary phenotypes (grey box text) using a relatedness matrix calculated using inverse normal transformed lipid species data from the AAB-QTAB study. Colours denote different combinations of covariates: green denotes no covariates ("nocov"); orange refers to the analysis including age, sex, batch, injection order and storage time covariates ("covdemo"); purple refers to the analysis also adjusted for energy intake (n=260; "covdemoenergy"). Text denotes the percentage of variance in the trait associated with the lipidomics relatedness matrix. Error bars correspond to standard error for the estimate.

Dyslipidaemia: The existence of a subtype of autism characterised by dyslipidaemia has been proposed on the basis of large-scale analysis of electronic health record data, including clinical measures of cholesterol and triglycerides[64]. We sought to replicate these findings, by applying clinical thresholds from adults to identify individuals in our AAB+QTAB dataset with dyslipidaemia (cholesterol >5.5 mmol/L, triglycerides >2.0 mmol/L). We found no over-representation of autistic participants or participants with intellectual disability (defined here as IQ<70) with respect to high cholesterol or triglyceride levels (Figure 3.5.8). To account for other variables that may affect lipid levels, we took the residuals from a regression of lipid concentrations on covariates (age, sex,

batch, injection order and storage time), and defined high cholesterol and triglyceride groups as those participants in the top decile. In this analysis, there was again no relationship with autism diagnosis, but there was a nominal association between intellectual disability and higher triglycerides (OR=1.94, p=3.4e-2).

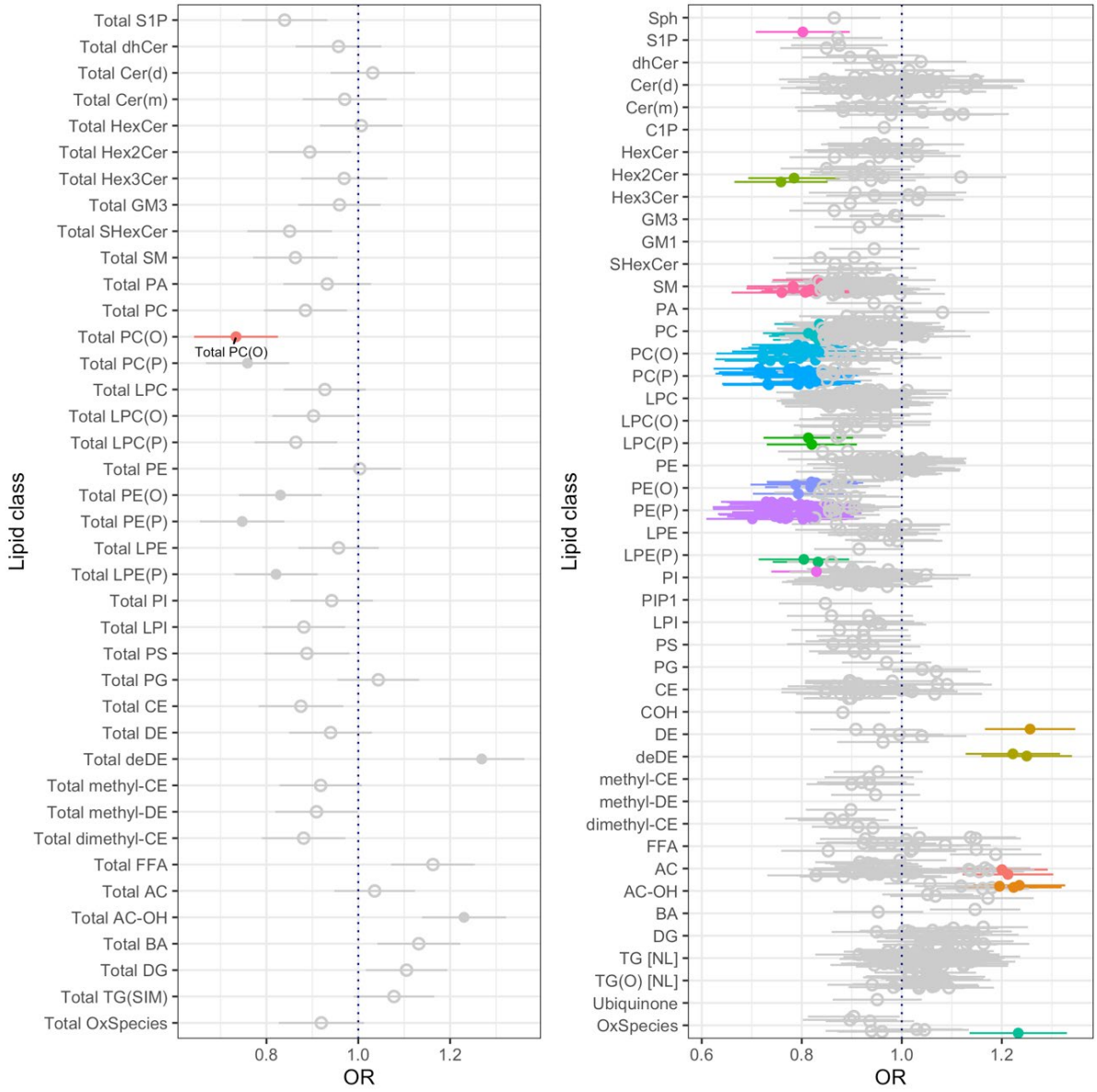


Figure 3.5.0.3: Association of lipid classes (left) and species (right) with ASD. Coloured lipids (classes, species) are nominally significant; named lipids are significant after multiple testing correction.

Outlier samples: We investigated the n=7 participants identified as outliers on the basis of quality control criteria (see Section 2.6). These outliers did not appear to be an artefact of batch processing, or haemolysis, but interestingly, three overlapped with a set of n=12 samples recorded at sample processing as being visually “fatty”, and two of these three were siblings, suggesting a

shared effect (whether it be genetic or environmental). A fourth outlier sample belonged to a participant with a CNV deletion encompassing the *LDLR* gene (encoding the LDL receptor), which is an important gene involved in cholesterol regulation. Rare genetic variation in *LDLR* causing reduction or loss in function is associated with familial hypercholesterolaemia, and there is also a common genetic signal, as captured using GWAS. Others have also found that a five-exon cluster of the LDLR gene carries autism-segregating variation[65].

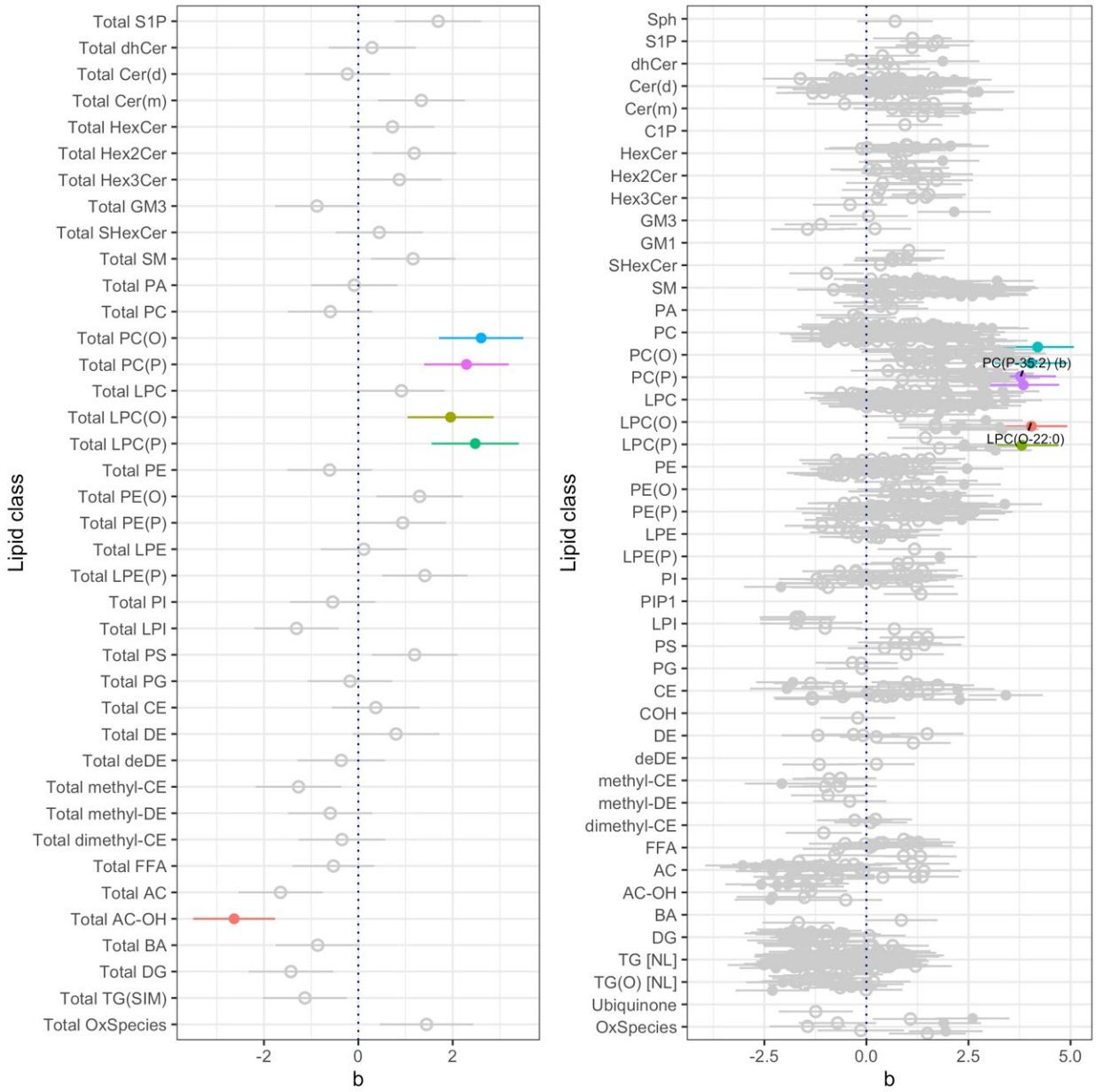


Figure 3.5.0.4: Association of lipid classes (left) and species (right) with IQ-DQ. Coloured lipids (classes, species) are nominally significant; named lipids are significant after multiple testing correction.

Notably, the participant with the *LDLR* deletion had the highest plasma concentrations of the following lipid classes: dihydroceramides (dhCer), dihexosylceramides (Hex2Cer), sulfatide (SHexCer), cholesteryl ester (CE), dehydrocholesterol ester (DE), and among the top-five highest plasma concentrations for free cholesterol (COH), ceramide (Cer(d)), GM3 ganglioside (GM3), sphingomyelin (SM), PC, LPC, PE(O), PI, and particularly high plasma concentrations of cholesterol esters, cholesterol and sphingomyelin. Overall, this suggests that a potential biological explanation could be identified for four of seven outlier samples.

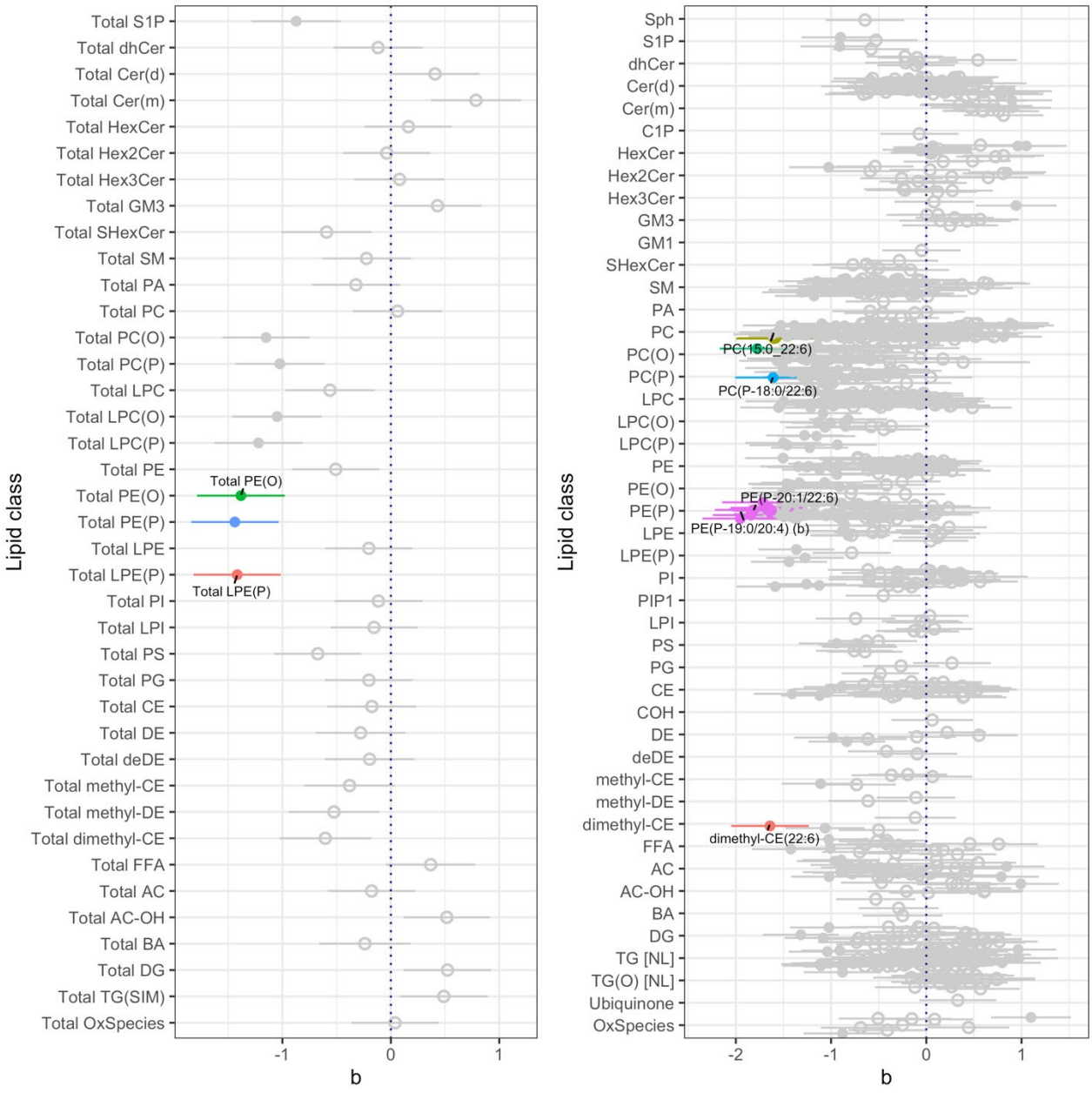


Figure 3.5.0.5: Association of lipid classes (left) and species (right) with sleep problems. Coloured lipids (classes, species) are nominally significant; named lipids are significant after multiple testing correction.

Considering the 12 visually “fatty” plasma samples, it was notable that seven had among the ten-highest concentrations of diacylglycerols, and five were among the ten-highest for triacylglycerols (TG [NL] / TG [SIM]) and akyldiacylglycerols (TG(O) [NL] / TG(O) [SIM]), suggesting that these lipid classes may be related to “fatty” plasma appearance. Another potential explanation for outlier samples may include participants having a fatty meal prior to blood sample collection, as fasting status was not recorded. Unfortunately, dietary data was only available for n=2 of the adverse event samples.

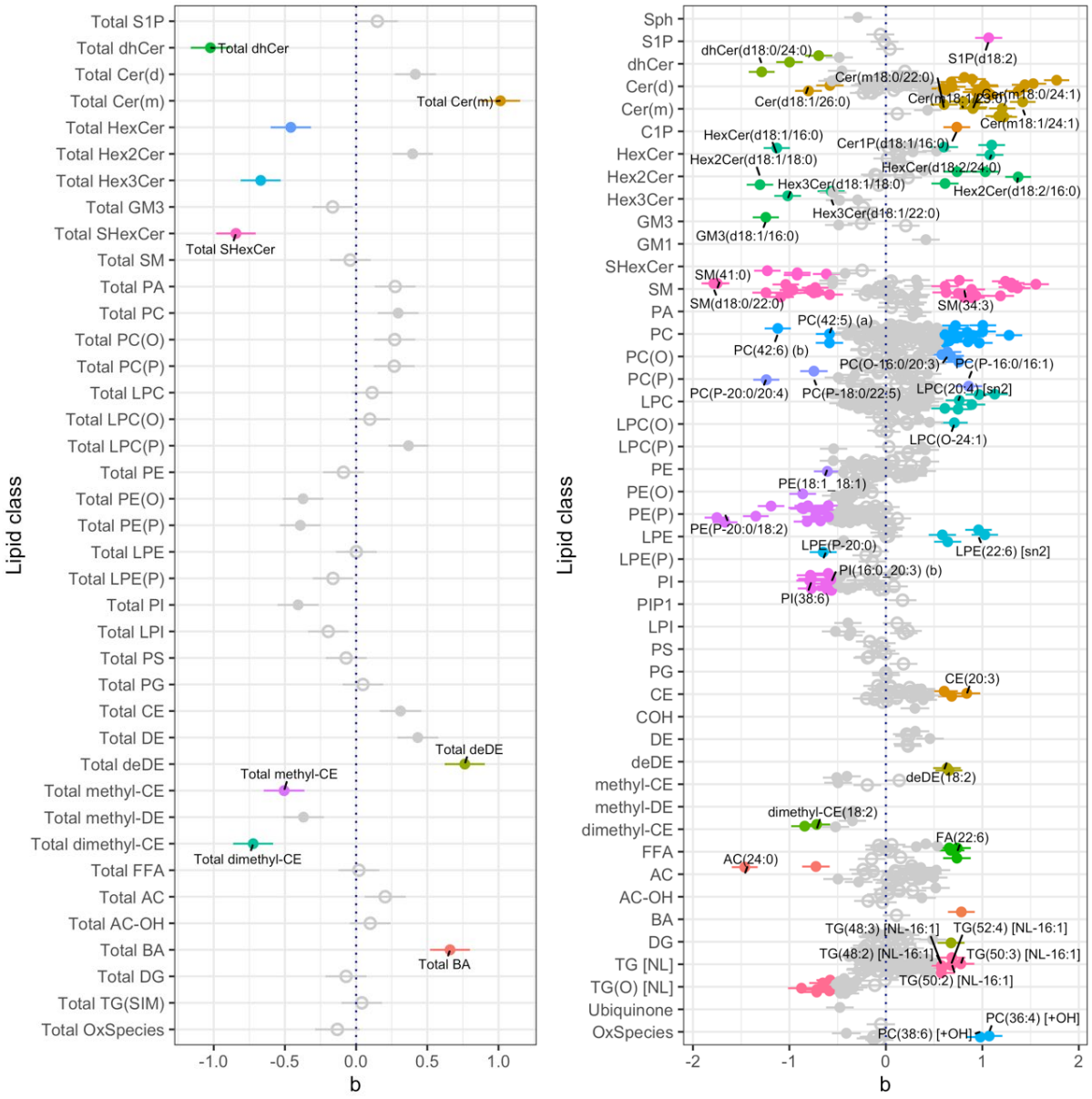


Figure 3.5.0.6: Association of lipid classes (left) and species (right) with age. Coloured lipids (classes, species) are nominally significant; named lipids are significant after multiple testing correction.

Heterogeneity analyses: On the basis that some outlier samples have biologically meaningful explanations, we investigated if variance in lipid concentrations differed between autistic and non-autistic groups, in addition to ID vs non-ID groups, using Levene's test. For these analyses, we included the 7 outlier samples identified by the Baker Institute, but excluded the 64 storage time outliers. We identified a single lipid class (phosphatidic acid, PA) – but no lipid species – with significantly greater variance in autistic children compared to siblings and unrelated, undiagnosed children, after Bonferroni correction (Figure 3.5.9).

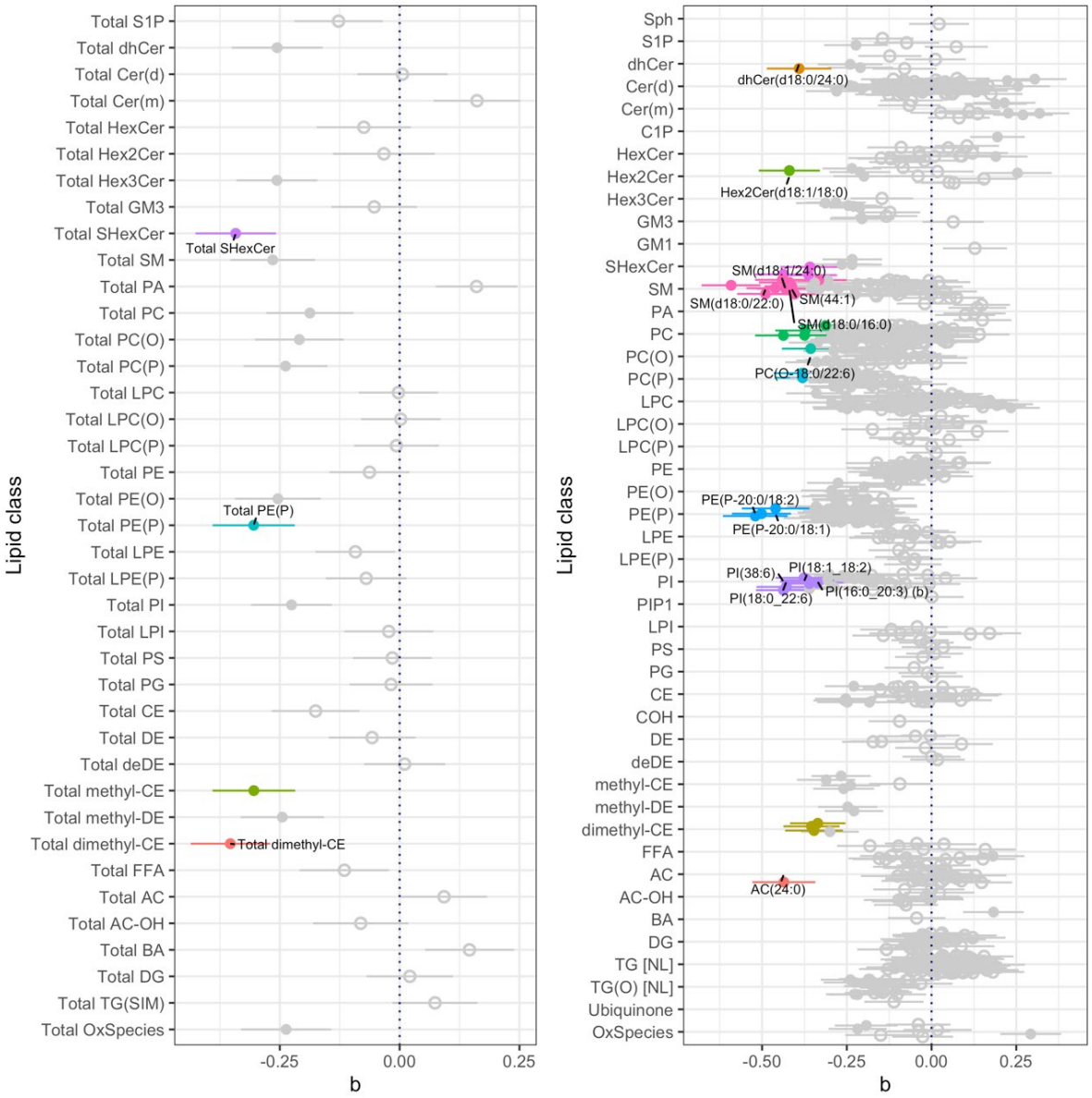


Figure 3.5.0.7: Association of lipid classes (left) and species (right) with Tanner developmental stage. Coloured lipids (classes, species) are nominally significant; named lipids are significant after multiple testing correction.

3.5.1 Conclusions

Our analyses revealed large effects of age and developmental stage (Tanner) on the lipidome, and more modest associations with neurodevelopmental and behavioural traits including diagnosis of autism, IQ_DQ and sleep disturbances, although only IQ_DQ, and to a lesser extent sleep showed robust associations in both the variance components and LWAS analyses after adjusting for covariates. The associations with autism, IQ_DQ and sleep all implicate the plasmalogen pathway, although sensitivity analyses in children for whom dietary data was available suggested that diet may influence this association. Further analyses will be needed to tease apart the relationship between diet, autism diagnosis, ID diagnosis, sleep and lipids. Another notable association was for increased phosphatidic acid (PA) variance in autism and ID, given that PA is essential for stability of the mTOR signalling pathway, which is known to harbour a number of well characterised genes for neurodevelopmental disorders[66]. A limitation of our study is that we did not have access to fasting bloods, and consequently it is possible that some of the variance in lipid concentrations may be explained by proximity (and type) of meal preceding blood sample collection. We are also unable to attempt replication of our findings, since independent lipidomics data in autism does not yet exist.

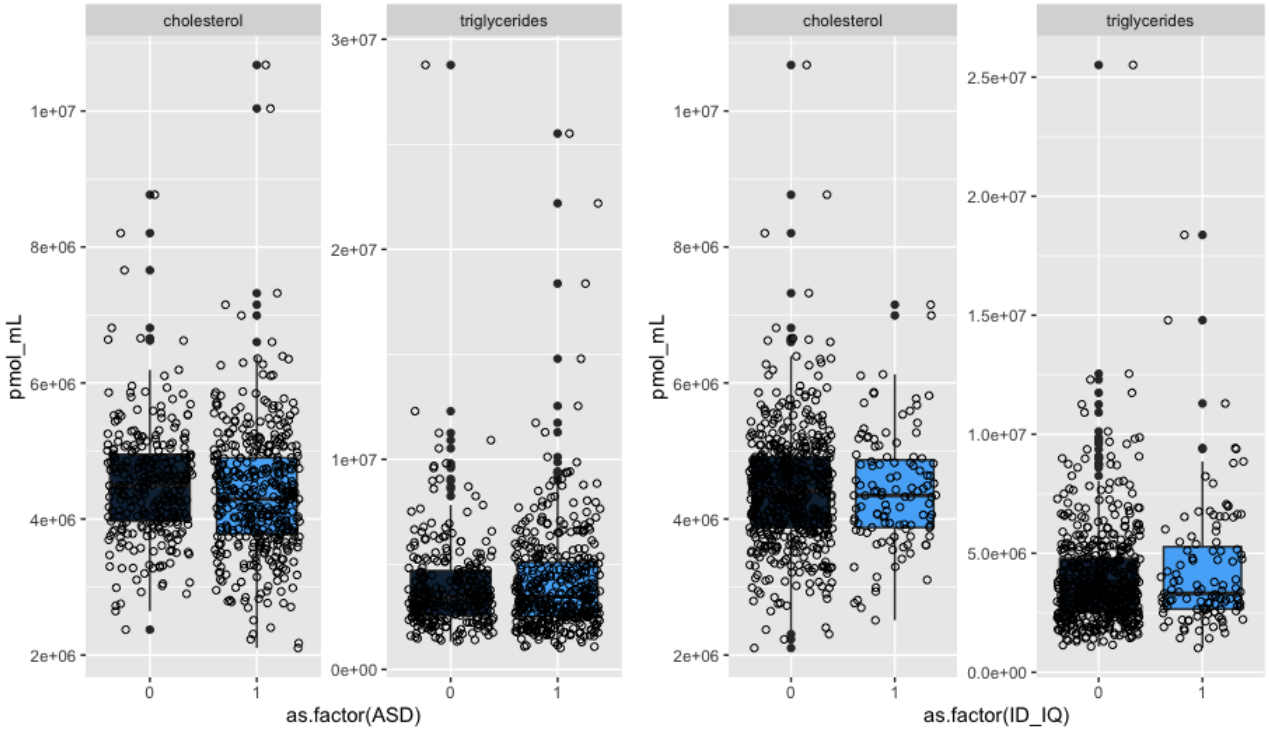


Figure 3.5.1.1: Distribution of total cholesterol and total triglycerides in ASD versus non-ASD (left panels) and ID-DQ vs non-ID-DQ (right panels). Individuals with a diagnosis are designated by "1". Filled dots indicate outliers.

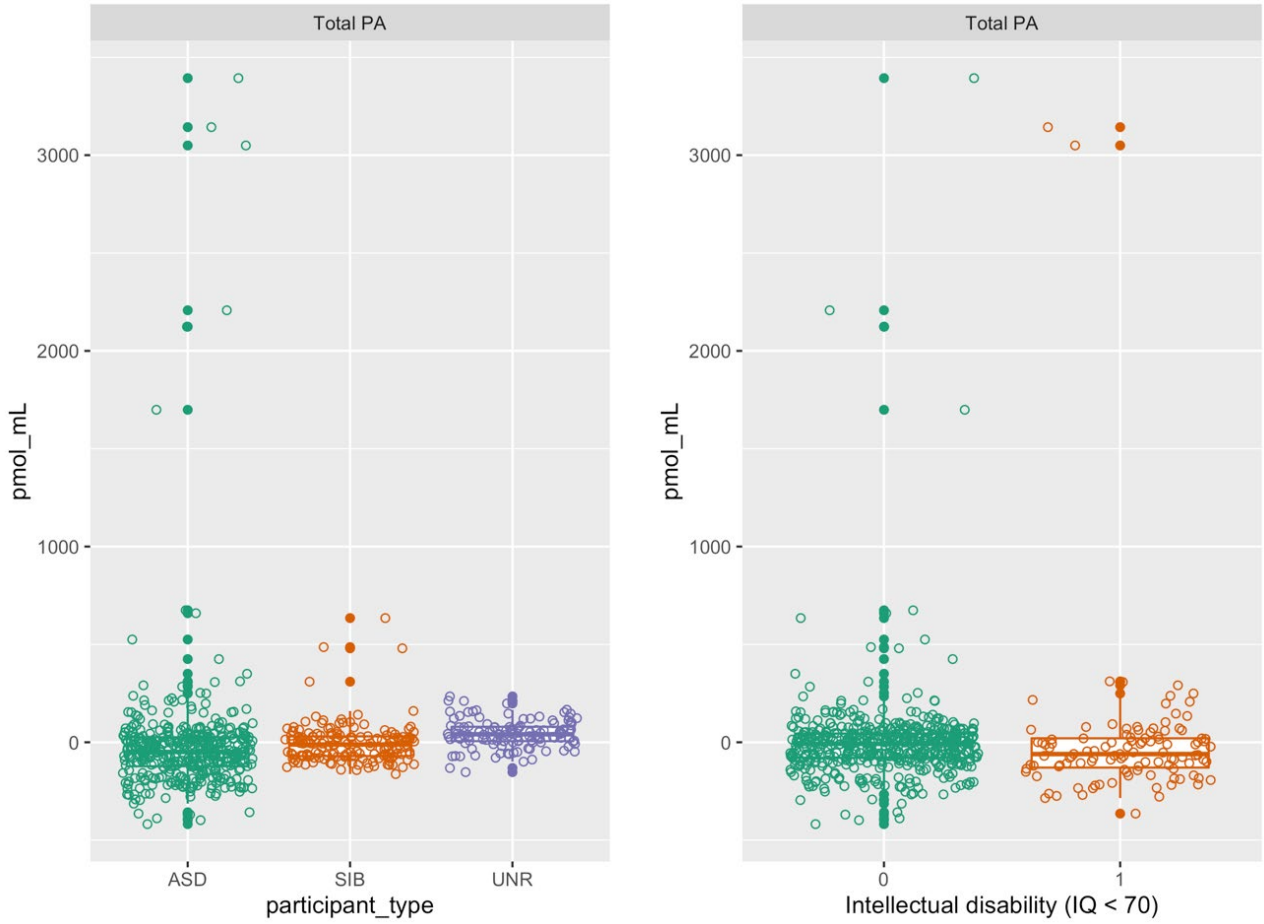


Figure 3.5.1.2: Distribution of total PA in ASD compared to siblings (SIB) and undiagnosed, unrelated (UNR) (left) and in ID ("1") compared to non-ID ("0") (right). Filled dots indicate outliers.

3.6 Systems genomics-based prediction

We evaluated the predictive potential of genetic (PGS), DNA methylation (MGS) and combined genetic and DNA methylation (PGS and MGS) predictors in the AAB using logistic regression. Using data on a total of n=448 AAB participants for whom both PGS and MGS were available, we calculated the difference in Nagelkerke's pseudo-R^2 between a baseline model, in which Autism diagnosis (encoded as a 0/1 binary variable, with diagnosed children encoded as a 1, and their siblings and unrelated undiagnosed children encoded as 0) was regressed against age and sex, and models in which PGS, MGS and both PGS and MGS were included as additional independent variables.

The Nagelkerke's pseudo-R^2 for the baseline model was 8.4% (p=5.5e-7), reflecting strong associations of age and sex with autism diagnosis in the AAB cohort (i.e. due to ascertainment), which is consistent with epidemiological observations that Autism diagnosis is more common

among males than females. The addition of PGS to the baseline model increased Nagelkerke's pseudo-R^2 to 8.6% (p=1.5e-6), a gain of 0.2% (p=0.33), and adding MGS to the baseline model returned a Nagelkerke's pseudo-R^2 of 8.7% (p=1.3e-6), a gain of 0.3% (p=0.26). In comparison, adding both PGS and MGS to the baseline model increased Nagelkerke's pseudo-R^2 to 9.1% (p=2.7e-6), a gain of 0.7% (p=0.30). Reassuringly, the effect sizes were in the expected directions, with both PGS and MGS being associated with increased likelihood of Autism diagnosis. Overall, these results suggest that integrating genetic and methylation data is likely to improve prediction, although current predictors are weak, and reflects the fact that the available reference datasets remain under-powered.

3.6.1 Caveats and limitations

We note that there are a number of significant limitations in our analysis, and we strongly caution against the use of these results beyond this proof-of-principle demonstration.

1. We acknowledge that the predictive accuracy is weak, which is likely attributable to lack of power in the (external) discovery genotyping and methylation datasets, which is in part due to the complexity and heterogeneity of autism. Indeed, this low accuracy was anticipated from our first discussions with the Autism CRC.

2. We are unable to penetrate this heterogeneity as we lacked clinical datasets that were available across all groups (diagnosed, siblings, and undiagnosed unrelated children).

3. This is a very small dataset, meaning that the results are not widely generalisable.

4. We note that our dataset was ascertained for participants for whom a diagnosis already existed. This contrasts with the design of the MINERvA methylation dataset, for which infant blood spots were retrieved after having identified people with a subsequent diagnosis in the Danish health registry. A strength of MINERvA is that the DNA methylation differences are identified prior to diagnosis; however, we do not know enough about temporal DNA methylation trajectories to know if we would expect them to be present in the age-group of our AAB samples.

4. Limitations

This study had a number of significant limitations. First, the sample size of the Biobank is insufficient for statistically powered analyses of genomic variation, as we noted from the outset of the study. Additionally, external datasets required for building genetic and genomic predictors for autism, including from independent genome-wide and methylome-wide association studies are few (and also underpowered). For some data types, including gut metagenomics and metabolomics, independent data in autism cohorts is currently lacking (demonstrating the novelty of the study), which limits opportunities for inclusion of metagenomics and metabolomics data in a systems genomics predictor for autism that can be evaluated in the Biobank.

A second limitation is that the cross-sectional study design, comprising once-only collection of data and biospecimens following diagnosis is not amenable to assessment of genomic predictors in early childhood prior to diagnosis. For this reason, although we find negligible association of autism with, for example, the gut microbiome, we cannot rule out the possibility that such data collected in infancy may have predictive value.

Third, integrating systems genomics predictors with a clinical diagnostic instrument based on behavioural surveillance is not feasible in the Biobank because data from diagnostic instruments in autistic children are largely not available in non-autistic siblings and controls.

5. Implications for research and practice

The primary research implication from this project is that, as anticipated, much larger genomic datasets will be needed to fully evaluate genomics predictors in autism. Achieving this goal will require coordinated international collaboration to achieve data harmonisation and the necessary sample sizes.

A number of potentially important clinical implications emerge from our analysis of gut microbiome data in the Biobank, which supports a “top-down” model of causality, whereby restricted diet in autism influences the gut microbiome and stool consistency, contrary to evidence from animal models suggesting a causal effect (i.e. “bottom-up”) of the microbiome on autism-related behaviours. First, our findings indicate that microbiome-based interventions in autism (e.g. faecal microbiota transplantation) may have no beneficial effect, but may carry risks. Second, management of an adequate diet in autism should be a priority, given the known relationship between reduced microbiome diversity and poorer health outcomes.

6. Recommendations

The project team's key recommendations from this study relate to the need for data sharing. We strongly recommend that Biobank biospecimens, genetic data and phenotype data be contributed to international autism genetics consortia, including the Psychiatric Genomics Consortium and Autism Sequencing Consortium. These Consortia have the largest sample sizes globally, and thus participation in these studies represents the best opportunity for Biobank participants to contribute to new understanding of autism through genetic discoveries.

In the same vein, we strongly support efforts to establish international consortia for analysis of DNA methylation, gut metagenomics and metabolomics data in autism, given the need for sample sizes that are beyond the scope of any single research group or study cohort. We strongly encourage open-access data sharing to increase the rate of progress in the field.

A second recommendation is to actively explore options for prospective recruitment and collection of biospecimens and clinical and lifestyle data from children before and after diagnosis.

7. References

1. Alvares GA, Dawson PA, Dissanayake C, et al. "Study protocol for the Australian autism biobank: an international resource to advance autism discovery research." *BMC Pediatr. 2018;18(1):284.*
2. McCarthy S, Das S, Kretzschmar W, et al. "A reference panel of 64,976 haplotypes for genotype imputation." *Nat Genet. 2016;48(10):1279-1283.*
3. Loh P-R, Danecek P, Palamara PF, et al. "Reference-based phasing using the Haplotype Reference Consortium panel." *Nat Genet. 2016;48:1443.*
4. Yap CX, Alvares GA, Henders AK, et al. "Analysis of common genetic variation and rare CNVs in the Australian Autism Biobank." *Molecular autism. 2021;12(1):12.*
5. Yang J, Benyamin B, McEvoy BP, et al. "Common SNPs explain a large proportion of the heritability for human height." *Nat Genet. 2010;42(7):565-569.*
6. Yang J, Lee SH, Goddard ME, Visscher PM. "GCTA: a tool for genome-wide complex trait analysis." *Am J Hum Genet. 2011;88(1):76-82.*
7. Grove J, Ripke S, Als TD, et al. "Identification of common genetic risk variants for autism spectrum disorder." *Nat Genet. 2019;51(3):431-444.*
8. Sullivan PF, Geschwind DH. "Defining the Genetic, Genomic, Cellular, and Diagnostic Architectures of Psychiatric Disorders." *Cell. 2019;177(1):162-183.*
9. Lloyd-Jones LR, Zeng J, Sidorenko J, et al. "Improved polygenic prediction by Bayesian multiple regression on summary statistics." *Nat Commun. 2019;10(1):5086.*
10. Ni G, Zeng J, Revez JR, et al. "A comprehensive evaluation of polygenic score methods across cohorts in psychiatric disorders." *medRxiv. 2020:2020.2009.2010.20192310.*
11. Grove J, Ripke S, Als TD, et al. "Identification of common genetic risk variants for autism spectrum disorder." *Nat Genet. 2019;51(3):431-444.*
12. Savage JE, Jansen PR, Stringer S, et al. "Genome-wide association meta-analysis in 269,867 individuals identifies new genetic and functional links to intelligence." *Nat Genet. 2018;50(7):912-919.*
13. Jones SE, Lane JM, Wood AR, et al. "Genome-wide association analyses of chronotype in 697,828 individuals provides insights into circadian rhythms." *Nat Commun. 2019;10(1):343.*
14. Chen W, Wu Y, Zheng Z, et al. "Improved analyses of GWAS summary statistics by reducing data heterogeneity and errors." *bioRxiv. 2020:2020.2007.2009.196535.*
15. Weiner DJ, Wigdor EM, Ripke S, et al. "Polygenic transmission disequilibrium confirms that common and rare variation act additively to create risk for autism spectrum disorders." *Nat Genet. 2017;49(7):978-985.*
16. Marshall CR, Howrigan DP, Merico D, et al. "Contribution of copy number variants to schizophrenia from a genome-wide study of 41,321 subjects." *Nat Genet. 2017;49(1):27-35.*
17. Wang K, Li M, Hadley D, et al. "PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data." *Genome Res. 2007;17(11):1665-1674.*
18. Pinto D, Pagnamenta AT, Klei L, et al. "Functional impact of global rare copy number variation in autism spectrum disorders." *Nature. 2010;466(7304):368-372.*
19. Durinck S, Spellman PT, Birney E, Huber W. "Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt." *Nat Protoc. 2009;4(8):1184-1191.*
20. ClinGen. "Pathogenic CNV regions." *In: ClinGen, ed2020.*
21. DECIPHER. "CNV Syndromes." *In: DECIPHER, ed2020.*
22. Firth HV, Richards SM, Bevan AP, et al. "DECIPHER: Database of Chromosomal Imbalance and Phenotype in Humans Using Ensembl Resources." *Am J Hum Genet. 2009;84(4):524-533.*

23. Satterstrom FK, Kosmicki JA, Wang J, et al. "Large-Scale Exome Sequencing Study Implicates Both Developmental and Functional Changes in the Neurobiology of Autism." *Cell. 2020.*
24. McRae JF, Clayton S, Fitzgerald TW, et al. "Prevalence and architecture of de novo mutations in developmental disorders." *Nature. 2017;542(7642):433-438.*
25. Parks DH, Rigato F, Vera-Wolf P, et al. "Evaluation of the Microba Community Profiler for Taxonomic Profiling of Metagenomic Datasets From the Human Gut Microbiome." *Front Microbiol. 2021;12:643682.*
26. Watson JF, Collins CE, Sibbritt DW, Dibley MJ, Garg ML. "Reproducibility and comparative validity of a food frequency questionnaire for Australian children and adolescents." *International Journal of Behavioral Nutrition and Physical Activity. 2009;6(1):62.*
27. Collins CE, Burrows TL, Truby H, et al. "Comparison of Energy Intake in Toddlers Assessed by Food Frequency Questionnaire and Total Energy Expenditure Measured by the Doubly Labeled Water Method." *Journal of the Academy of Nutrition and Dietetics. 2013;113(3):459-463.*
28. Rothschild D, Weissbrod O, Barkan E, et al. "Environment dominates over host genetics in shaping human gut microbiota." *Nature. 2018;555(7695):210-215.*
29. Bokulich NA, Chung J, Battaglia T, et al. "Antibiotics, birth mode, and diet shape microbiome maturation during early life." *Science Translational Medicine. 2016;8(343):343ra382.*
30. Zhang F, Chen W, Zhu Z, et al. "OSCA: a tool for omic-data-based complex trait analysis." *Genome Biology. 2019;20(1):107.*
31. Mandal S, Van Treuren W, White RA, Eggesbø M, Knight R, Peddada SD. "Analysis of composition of microbiomes: a novel method for studying microbial composition." *Microbial Ecology in Health and Disease. 2015;26(1):27663.*
32. Weiss S, Xu ZZ, Peddada S, et al. "Normalization and microbial differential abundance strategies depend upon data characteristics." *Microbiome. 2017;5(1):27.*
33. Yang Z, Wu H, Lee PH, et al. "Cross-disorder GWAS meta-analysis for Attention Deficit/Hyperactivity Disorder, Autism Spectrum Disorder, Obsessive Compulsive Disorder, and Tourette Syndrome." *bioRxiv. 2019:770222.*
34. Nagel M, Jansen PR, Stringer S, et al. "Meta-analysis of genome-wide association studies for neuroticism in 449,484 individuals identifies novel genetic loci and pathways." *Nature Genetics. 2018;50(7):920-927.*
35. Alvares GA, Dawson PA, Dissanayake C, et al. "Study protocol for the Australian autism biobank: an international resource to advance autism discovery research." *BMC Pediatr. 2018;18(1):284.*
36. Min JL, Hemani G, Davey Smith G, Relton C, Suderman M. "Meffil: efficient normalization and analysis of very large DNA methylation datasets." *Bioinformatics (Oxford, England). 2018;34(23):3983-3989.*
37. Willer CJ, Li Y, Abecasis GR. "METAL: fast and efficient meta-analysis of genomewide association scans." *Bioinformatics. 2010;26(17):2190-2191.*
38. Hannon E, Schendel D, Ladd-Acosta C, et al. "Elevated polygenic burden for autism is associated with differential DNA methylation at birth." *Genome Medicine. 2018;10(1):19.*
39. Lloyd-Jones LR, Zeng J, Sidorenko J, et al. "Improved polygenic prediction by Bayesian multiple regression on summary statistics." *Nature Communications. 2019;10(1):5086.*
40. Zhou W, Laird PW, Shen H. "Comprehensive characterization, annotation and innovative use of Infinium DNA methylation BeadChip probes." *Nucleic Acids Res. 2017;45(4):e22-e22.*
41. Mazurek MO, Dovgan K, Neumeyer AM, Malow BA. "Course and Predictors of Sleep and Co-occurring Problems in Children with Autism Spectrum Disorder." *J Autism Dev Disord. 2019;49(5):2101-2115.*
42. Reynolds AM, Soke GN, Sabourin KR, et al. "Sleep Problems in 2- to 5-Year-Olds With Autism Spectrum Disorder and Other Developmental Delays." *Pediatrics. 2019;143(3).*

43. Firth HV, Richards SM, Bevan AP, et al. "DECIPHER: Database of Chromosomal Imbalance and Phenotype in Humans Using Ensembl Resources." *Am J Hum Genet. 2009;84(4):524-533.*
44. Satterstrom FK, Kosmicki JA, Wang J, et al. "Large-Scale Exome Sequencing Study Implicates Both Developmental and Functional Changes in the Neurobiology of Autism." *Cell. 2020;180(3):568-584 e523.*
45. *"*Deciphering Developmental Disorders S. Prevalence and architecture of de novo mutations in developmental disorders." *Nature. 2017;542(7642):433-438.*
46. Zhang F, Chen W, Zhu Z, et al. "OSCA: a tool for omic-data-based complex trait analysis." *Genome Biol. 2019;20(1):107.*
47. Ho LKH, Tong VJW, Syn N, et al. "Gut microbiota changes in children with autism spectrum disorder: a systematic review." *Gut Pathogens. 2020;12(1):6.*
48. Ricaboni D, Mailhe M, Khelaifia S, Raoult D, Million M. "Romboutsia timonensis, a new species isolated from human gut." *New Microbes New Infect. 2016;12:6-7.*
49. Bandini LG, Anderson SE, Curtin C, et al. "Food Selectivity in Children with Autism Spectrum Disorders and Typically Developing Children." *The Journal of Pediatrics. 2010;157(2):259-264.*
50. Sharp WG, Postorino V, McCracken CE, et al. "Dietary Intake, Nutrient Status, and Growth Parameters in Children with Autism Spectrum Disorder and Severe Food Selectivity: An Electronic Medical Record Review." *Journal of the Academy of Nutrition and Dietetics. 2018;118(10):1943-1950.*
51. Marí-Bauset S, Zazpe I, Mari-Sanchis A, Llopis-González A, Morales-Suárez-Varela M. "Food Selectivity in Autism Spectrum Disorders: A Systematic Review." *Journal of Child Neurology. 2013;29(11):1554-1561.*
52. Schreck KA, Williams K. "Food preferences and factors influencing food selectivity for children with autism spectrum disorders." *Res Dev Disabil. 2006;27(4):353-363.*
53. Panossian C, Lyons-Wall P, Whitehouse A, et al. "Young Adults with High Autistic-Like Traits Displayed Lower Food Variety and Diet Quality in Childhood." *Journal of Autism and Developmental Disorders. 2020.*
54. Patil I. "{ggstatsplot}: 'ggplot2' Based Plots with Statistical Details." *CRAN. 2018.*
55. Vandeputte D, Falony G, Vieira-Silva S, Tito RY, Joossens M, Raes J. "Stool consistency is strongly associated with gut microbiota richness and composition, enterotypes and bacterial growth rates." *Gut. 2016;65(1):57-62.*
56. Hadizadeh F, Walter S, Belheouane M, et al. "Stool frequency is associated with gut microbiota composition." *Gut. 2017;66(3):559.*
57. Zhernakova A, Kurilshikov A, Bonder MJ, et al. "Population-based metagenomics analysis reveals markers for gut microbiome composition and diversity." *Science. 2016;352(6285):565.*
58. Chakraborty P, Carpenter KLH, Major S, et al. "Gastrointestinal problems are associated with increased repetitive behaviors but not social communication difficulties in young children with autism spectrum disorders." *Autism. 2020:1362361320959503.*
59. Lord C, Rutter, M., DiLavore, P. C., Risi, S., Gotham, K., & Bishop, S. L. "Autism diagnostic observation schedule, 2nd ed. (ADOS-2)." *Torrance, CA: Western Psychological Services; 2012.*
60. Cermak SA, Curtin C, Bandini LG. "Food selectivity and sensory sensitivity in children with autism spectrum disorders." *J Am Diet Assoc. 2010;110(2):238-246.*
61. Yang B, Wei J, Ju P, Chen J. "Effects of regulating intestinal microbiota on anxiety symptoms: A systematic review." *Gen Psychiatr. 2019;32(2):e100056-e100056.*
62. Kim H-N, Yun Y, Ryu S, et al. "Correlation between gut microbiota and personality in adults: A cross-sectional study." *Brain, Behavior, and Immunity. 2018;69:374-385.*
63. Zilmer M, Edmondson AC, Khetarpal SA, et al. "Novel congenital disorder of O-linked glycosylation caused by GALNT2 loss of function." *Brain : a journal of neurology. 2020;143(4):1114-1126.*

64. Luo Y, Eran A, Palmer N, et al. "A multidimensional precision medicine approach identifies an autism subtype characterized by dyslipidemia." *Nature medicine. 2020;26(9):1375-1379.*
65. Luo Y, Eran A, Palmer N, et al. "A multidimensional precision medicine approach identifies an autism subtype characterized by dyslipidemia." *Nature Medicine. 2020;26(9):1375-1379.*
66. Crino PB. "The mTOR signalling cascade: paving new roads to cure neurological disease." *Nat Rev Neurol. 2016;12(7):379-392.*

Appendix A

Metabolite Class	N	Metabolite species
Acylcarnitines (AC)	27	AC(10:0), AC(12:0), AC(12:1), AC(13:0), AC(14:0), AC(14:1), AC(14:2), AC(15:0)(a), AC(15:0)(b), AC(16:0), AC(16:1), AC(17:0), AC(18:0), AC(18:1), AC(18:2), AC(18:3), AC(20:3)(a), AC(20:3)(b), AC(20:4), AC(20:5), AC(22:5), AC(22:6), AC(24:0), AC(24:1)(a), AC(24:1)(b), AC(26:0), AC(26:1)
AC-OH	10	AC(14:0)-OH, AC(14:1)-OH, AC(16:0)-OH, AC(16:1)-OH, AC(18:0)-OH, AC(18:1)-OH, AC(20:3)-OH, AC(22:5)-OH, AC(24:0)-OH, AC(24:1)-OH
Bile Acids (BA)	2	CA, dxCA
C1P	1	Cer1P(d18:1/16:0)
CE	27	CE(14:0), CE(15:0), CE(16:0), CE(16:1), CE(16:2), CE(17:0), CE(17:1), CE(18:0), CE(18:1), CE(18:2), CE(18:3), CE(20:0), CE(20:1), CE(20:2), CE(20:3), CE(20:4), CE(20:5), CE(22:0), CE(22:1), CE(22:4), CE(22:5), CE(22:6), CE(24:0), CE(24:1), CE(24:4), CE(24:5), CE(24:6)
Cer(d)	48	Cer(d16:1/16:0), Cer(d16:1/18:0), Cer(d16:1/20:0), Cer(d16:1/22:0), Cer(d16:1/23:0), Cer(d16:1/24:0), Cer(d16:1/24:1), Cer(d17:1/16:0), Cer(d17:1/18:0), Cer(d17:1/20:0), Cer(d17:1/22:0), Cer(d17:1/23:0), Cer(d17:1/24:0), Cer(d17:1/24:1), Cer(d18:1/14:0), Cer(d18:1/16:0), Cer(d18:1/18:0), Cer(d18:1/19:0), Cer(d18:1/20:0), Cer(d18:1/21:0), Cer(d18:1/22:0), Cer(d18:1/23:0), Cer(d18:1/24:0), Cer(d18:1/24:1), Cer(d18:1/26:0), Cer(d18:2/14:0), Cer(d18:2/16:0), Cer(d18:2/18:0), Cer(d18:2/20:0), Cer(d18:2/21:0), Cer(d18:2/22:0), Cer(d18:2/23:0), Cer(d18:2/24:0), Cer(d18:2/24:1), Cer(d18:2/26:0), Cer(d19:1/16:0), Cer(d19:1/18:0), Cer(d19:1/20:0), Cer(d19:1/22:0), Cer(d19:1/23:0), Cer(d19:1/24:0), Cer(d19:1/24:1), Cer(d19:1/26:0), Cer(d20:1/22:0), Cer(d20:1/23:0), Cer(d20:1/24:0), Cer(d20:1/24:1), Cer(d20:1/26:0)
Cer(m)	11	Cer(m18:0/20:0), Cer(m18:0/22:0), Cer(m18:0/23:0), Cer(m18:0/24:0), Cer(m18:0/24:1), Cer(m18:1/18:0), Cer(m18:1/20:0), Cer(m18:1/22:0), Cer(m18:1/23:0), Cer(m18:1/24:0), Cer(m18:1/24:1)
COH	1	COH
DE	6	DE(16:0), DE(18:1), DE(18:2), DE(20:4), DE(20:5), DE(22:6)
deDE	2	deDE(18:2), deDE(20:4)
DG	25	DG(14:0_16:0), DG(14:0_18:2), DG(16:0_16:0), DG(16:0_16:1), DG(16:0_18:1), DG(16:0_18:2), DG(16:0_20:4), DG(16:0_22:5), DG(16:0_22:6), DG(16:1_18:1), DG(18:0_18:1), DG(18:0_18:2), DG(18:0_20:4), DG(18:0_22:6), DG(18:1_18:1), DG(18:1_18:2), DG(18:1_18:3), DG(18:1_20:3), DG(18:1_20:4), DG(18:1_20:5), DG(18:1_22:5), DG(18:1_22:6), DG(18:2_18:2), DG(18:2_20:4), DG(18:2_22:6)
dhCer	6	dhCer(d18:0/16:0), dhCer(d18:0/18:0), dhCer(d18:0/20:0), dhCer(d18:0/22:0), dhCer(d18:0/24:0), dhCer(d18:0/24:1)
Dimethyl-CE	4	dimethyl-CE(18:1), dimethyl-CE(18:2), dimethyl-CE(20:4), dimethyl-CE(22:6)
FFA	16	FA(14:0), FA(15:0), FA(16:0), FA(16:1), FA(17:1), FA(18:0), FA(18:1), FA(18:2), FA(18:3), FA(20:2), FA(20:3), FA(20:4), FA(20:5), FA(22:4), FA(22:5), FA(22:6)
GM1	1	GM1(d18:1/16:0)
GM3	7	GM3(d18:1/16:0), GM3(d18:1/18:0), GM3(d18:1/20:0), GM3(d18:1/22:0), GM3(d18:1/24:0), GM3(d18:1/24:1), GM3(d18:2/24:1)

Hex2Cer	10	Hex2Cer(d16:1/16:0), Hex2Cer(d16:1/24:1), Hex2Cer(d18:1/16:0), Hex2Cer(d18:1/18:0), Hex2Cer(d18:1/20:0), Hex2Cer(d18:1/22:0), Hex2Cer(d18:1/24:0), Hex2Cer(d18:1/24:1), Hex2Cer(d18:2/16:0), Hex2Cer(d18:2/18:0), Hex2Cer(d18:2/24:1)
Hex3Cer	6	Hex3Cer(d18:1/16:0), Hex3Cer(d18:1/18:0), Hex3Cer(d18:1/20:0), Hex3Cer(d18:1/22:0), Hex3Cer(d18:1/24:0), Hex3Cer(d18:1/24:1)
HexCer	14	HexCer(d16:1/18:0), HexCer(d16:1/20:0), HexCer(d16:1/22:0), HexCer(d16:1/24:0), HexCer(d18:1/16:0), HexCer(d18:1/18:0), HexCer(d18:1/20:0), HexCer(d18:1/22:0), HexCer(d18:1/24:0), HexCer(d18:1/24:1), HexCer(d18:2/18:0), HexCer(d18:2/20:0), HexCer(d18:2/22:0), HexCer(d18:2/24:0)
LPC	61	LPC(14:0) [sn1], LPC(14:0) [sn2], LPC(15-MHDA) [sn1] / LPC(17:0) [sn2], LPC(15-MHDA) [sn1] [104_sn1], LPC(15-MHDA) [sn2], LPC(15:0) [sn1], LPC(15:0) [sn2], LPC(16:0) [sn1], LPC(16:0) [sn2], LPC(16:1) [sn1], LPC(16:1) [sn2], LPC(17:0) [sn1], LPC(17:1) (a) [sn1] [104_sn1], LPC(17:1) [sn1] (a) / LPC(17:1) [sn2] (b), LPC(17:1) [sn1] (b), LPC(17:1) [sn2] (a), LPC(18:0) [sn1], LPC(18:0) [sn2], LPC(18:1) [sn1], LPC(18:1) [sn2], LPC(18:2) [sn1], LPC(18:2) [sn2], LPC(18:3) (a) [sn1] [104_sn1], LPC(18:3) [sn1] (a)/LPC(18:3) [sn2] (b), LPC(18:3) [sn1] (b), LPC(18:3) [sn2] (a), LPC(19:0) (a) [sn1] [104_sn1], LPC(19:0) [sn1] (a) / LPC(19:0) [sn2] (b), LPC(19:0) [sn1] (b), LPC(19:0) [sn2] (a), LPC(19:1) (a), LPC(19:1) (b), LPC(19:1) (c), LPC(20:0) [sn1], LPC(20:0) [sn2], LPC(20:1) [sn1], LPC(20:1) [sn2], LPC(20:2) [sn1], LPC(20:2) [sn2], LPC(20:3) [sn1], LPC(20:3) [sn2], LPC(20:4) [sn1], LPC(20:4) [sn2], LPC(20:5) [sn1], LPC(20:5) [sn2], LPC(22:0) [sn1], LPC(22:0) [sn2], LPC(22:1) [sn1], LPC(22:1) [sn2], LPC(22:4) [sn1], LPC(22:4) [sn2], LPC(22:5) (n3) [sn1] [104_sn1], LPC(22:5) [sn1] (n3)/LPC(22:5) [sn2] (n6), LPC(22:5) [sn1] (n6), LPC(22:5) [sn2] (n3), LPC(22:6) [sn1], LPC(22:6) [sn2], LPC(24:0) [sn1], LPC(24:0) [sn2], LPC(26:0) [sn1], LPC(26:0) [sn2]
LPC(O)	10	LPC(O-16:0), LPC(O-18:0), LPC(O-18:1), LPC(O-20:0), LPC(O-20:1), LPC(O-22:0), LPC(O-22:1), LPC(O-24:0), LPC(O-24:1), LPC(O-24:2)
LPC(P)	6	LPC(P-16:0), LPC(P-17:0)(a), LPC(P-17:0)(b), LPC(P-18:0), LPC(P-18:1), LPC(P-20:0)
LPE	12	LPE(16:0) [sn1], LPE(16:0) [sn2], LPE(18:0) [sn1], LPE(18:0) [sn2], LPE(18:1) [sn1], LPE(18:1) [sn2], LPE(18:2) [sn1], LPE(18:2) [sn2], LPE(20:4) [sn1], LPE(20:4) [sn2], LPE(22:6) [sn1], LPE(22:6) [sn2]
LPE(P)	4	LPE(P-16:0), LPE(P-18:0), LPE(P-18:1), LPE(P-20:0)
LPI	7	LPI(18:0), LPI(18:1) [sn1], LPI(18:1) [sn2], LPI(18:2) [sn1], LPI(18:2) [sn2], LPI(20:4) [sn1], LPI(20:4) [sn2]
Methyl-CE	5	methyl-CE(18:0), methyl-CE(18:1), methyl-CE(18:2), methyl-CE(20:4), methyl-CE(22:6)
Methyl-DE	2	methyl-DE(18:1), methyl-DE(18:2)
OxSpecies	9	CE(18:2) [+OH], CE(20:4) [+OH], CE(22:6) [+OH], LPC(18:2) [+OH], LPC(20:4) [+OH], LPC(22:6) [+OH], PC(34:2) [+OH], PC(36:4) [+OH], PC(38:6) [+OH]
PA	6	PA(34:1), PA(36:1), PA(36:2), PA(36:3), PA(36:4), PA(40:6)

PC	83	PC(14:0_16:0), PC(14:0_20:4), PC(14:0_22:6), PC(15-MHDA_18:1), PC(15-MHDA_18:2), PC(15-MHDA_20:4), PC(15-MHDA_22:6), PC(15:0_20:3), PC(15:0_20:4), PC(15:0_22:6), PC(16:0_16:0), PC(16:0_18:0), PC(16:0_18:1), PC(16:0_18:2), PC(16:0_18:3) (a), PC(16:0_18:3) (b), PC(16:0_20:3) (a), PC(16:0_20:3) (b), PC(16:0_20:4), PC(16:0_20:5), PC(16:0_22:6), PC(16:1_18:2), PC(16:1_20:4), PC(16:1_22:6), PC(17:0_18:1), PC(17:0_18:2), PC(17:0_20:4), PC(17:0_22:6), PC(17:1_18:2), PC(18:0_18:1), PC(18:0_18:2), PC(18:0_20:3), PC(18:0_20:4), PC(18:0_22:4), PC(18:0_22:5) (n3)/PC(20:1_20:4), PC(18:0_22:5) (n6), PC(18:0_22:6), PC(18:1_18:1), PC(18:1_18:2), PC(18:1_20:3), PC(18:1_22:6) (a), PC(18:1_22:6) (b), PC(18:2_18:2), PC(18:2_20:5), PC(20:0_20:4), PC(28:0), PC(31:0) (a), PC(31:0) (b), PC(31:1), PC(32:1), PC(32:2), PC(33:0) (a), PC(33:0) (b), PC(33:1), PC(33:2), PC(34:5), PC(35:5), PC(36:6) (a), PC(38:2), PC(38:4) (b), PC(38:5) (a), PC(38:5) (b), PC(38:6) (a), PC(38:7) (c), PC(39:5) (a), PC(39:5) (b), PC(40:7) (a), PC(40:8), PC(42:10), PC(42:2), PC(42:3), PC(42:4), PC(42:5) (a), PC(42:5) (b), PC(42:6) (a), PC(42:6) (b), PC(42:7), PC(42:8), PC(42:9), PC(44:12), PC(44:4) (a), PC(44:4) (b), PC(44:5)
PC(O)	33	PC(O-16:0/16:0), PC(O-16:0/20:3), PC(O-16:0/20:4), PC(O-16:0/22:6), PC(O-18:0/18:1), PC(O-18:0/18:2), PC(O-18:0/20:4), PC(O-18:0/22:6), PC(O-18:1/18:1), PC(O-18:1/18:2), PC(O-32:1), PC(O-32:2), PC(O-34:1), PC(O-34:2), PC(O-34:4), PC(O-35:4), PC(O-36:0), PC(O-36:5), PC(O-38:5), PC(O-40:5), PC(O-40:7), PC(O-42:4) (a), PC(O-42:4) (b), PC(O-42:5) (a), PC(O-42:5) (b), PC(O-42:6), PC(O-42:7), PC(O-42:8), PC(O-44:6), PC(O-44:7), PC(O-46:7) (a), PC(O-46:7) (b), PC(O-46:8)
PC(P)	29	PC(P-15:0/20:4) (a), PC(P-15:0/20:4) (b), PC(P-16:0/14:0), PC(P-16:0/16:0), PC(P-16:0/16:1), PC(P-16:0/18:0), PC(P-16:0/18:1), PC(P-16:0/18:2), PC(P-16:0/18:3), PC(P-16:0/20:4), PC(P-16:0/20:5), PC(P-16:0/22:6), PC(P-17:0/20:4) (a), PC(P-17:0/20:4) (b), PC(P-18:0/18:2), PC(P-18:0/20:4), PC(P-18:0/22:5), PC(P-18:0/22:6), PC(P-18:1/18:1), PC(P-18:1/22:6), PC(P-20:0/20:4), PC(P-35:2) (a), PC(P-35:2) (b), PC(P-36:3), PC(P-38:5) (a), PC(P-38:5) (b), PC(P-42:5), PC(P-44:5), PC(P-46:8)
PE	37	PE(15-MHDA_18:1), PE(15-MHDA_18:2), PE(15-MHDA_20:4), PE(15-MHDA_22:6), PE(16:0_16:0), PE(16:0_16:1), PE(16:0_18:1), PE(16:0_18:2), PE(16:0_18:3) (a), PE(16:0_18:3) (b), PE(16:0_20:3), PE(16:0_20:4), PE(16:0_20:5), PE(16:0_22:6), PE(16:1_18:2), PE(16:1_20:4), PE(17:0_18:1), PE(17:0_18:2), PE(17:0_20:4), PE(17:0_22:6), PE(18:0_18:1), PE(18:0_18:2), PE(18:0_20:3) (a), PE(18:0_20:3) (b), PE(18:0_20:4), PE(18:0_22:4), PE(18:0_22:5) (n3), PE(18:0_22:5) (n6), PE(18:0_22:6), PE(18:1_18:1), PE(18:1_18:2), PE(18:1_22:6) (a), PE(18:1_22:6) (b), PE(20:0_20:4), PE(36:0), PE(38:5)(a), PE(38:5) b)
PE(O)	14	PE(O-16:0/18:2), PE(O-16:0/20:3), PE(O-16:0/20:4), PE(O-16:0/22:4), PE(O-16:0/22:6), PE(O-18:0/20:4), PE(O-18:0/22:5), PE(O-18:0/22:6), PE(O-18:1/18:2), PE(O-18:1/22:6), PE(O-34:1), PE(O-36:5), PE(O-38:5) (a), PE(O-38:5) (b)

PE(P)	54	PE(P-15:0/20:4) (a), PE(P-15:0/20:4) (b), PE(P-15:0/22:6) (a), PE(P-15:0/22:6) (b), PE(P-16:0/18:1), PE(P-16:0/18:2), PE(P-16:0/18:3), PE(P-16:0/20:3) (a), PE(P-16:0/20:3) (b), PE(P-16:0/20:4), PE(P-16:0/20:5), PE(P-16:0/22:4), PE(P-16:0/22:5) (n3), PE(P-16:0/22:5) (n6), PE(P-16:0/22:6), PE(P-17:0/20:4) (a), PE(P-17:0/20:4) (b), PE(P-17:0/22:6) (a), PE(P-17:0/22:6) (b), PE(P-18:0/18:1), PE(P-18:0/18:2), PE(P-18:0/18:3), PE(P-18:0/20:3) (a), PE(P-18:0/20:3) (b), PE(P-18:0/20:4), PE(P-18:0/20:5), PE(P-18:0/22:4), PE(P-18:0/22:5) (n3), PE(P-18:0/22:5) (n6), PE(P-18:0/22:6), PE(P-18:1/18:1) (a), PE(P-18:1/18:1) (b), PE(P-18:1/18:2) (a), PE(P-18:1/18:2) (b), PE(P-18:1/18:3), PE(P-18:1/20:3) (a), PE(P-18:1/20:3) (b), PE(P-18:1/20:4) (a), PE(P-18:1/20:4) (b), PE(P-18:1/20:5) (a), PE(P-18:1/20:5) (b), PE(P-18:1/22:4), PE(P-18:1/22:5) (a), PE(P-18:1/22:5) (b), PE(P-18:1/22:6) (a), PE(P-18:1/22:6) (b), PE(P-19:0/20:4) (a), PE(P-19:0/20:4) (b), PE(P-20:0/18:1), PE(P-20:0/18:2), PE(P-20:0/20:4), PE(P-20:0/22:6), PE(P-20:1/20:4), PE(P-20:1/22:6)
PG	3	PG(34:1), PG(36:1), PG(36:2)
PI	27	PI (38:5) (b), PI(15-MHDA_18:1)/PI(17:0_18:1), PI(15-MHDA_18:2)/PI(17:0_18:2), PI(15-MHDA_20:4)\PI(17:0_20:4), PI(16:0/16:0), PI(16:0_16:1), PI(16:0_20:3) (a), PI(16:0_20:3) (b), PI(16:0_20:4), PI(18:0_18:1), PI(18:0_20:2), PI(18:0_20:3) (a), PI(18:0_20:3) (b), PI(18:0_20:4), PI(18:0_22:4), PI(18:0_22:5) (n3), PI(18:0_22:5) (n6), PI(18:0_22:6), PI(18:1_18:2), PI(20:0_20:4), PI(34:0), PI(34:1), PI(36:2), PI(37:6), PI(38:5) (a), PI(38:6), PI(39:6)
PIP1	1	PIP1(38:4)
PS	7	PS(36:1), PS(36:2), PS(38:3), PS(38:4), PS(38:5), PS(40:5), PS(40:6)
S1P	4	S1P(d16:1), S1P(d18:0), S1P(d18:1), S1P(d18:2)
SHexCer	6	SHexCer(d18:1/16:0(OH)), SHexCer(d18:1/16:0), SHexCer(d18:1/24:0(OH)), SHexCer(d18:1/24:0), SHexCer(d18:1/24:1(OH)), SHexCer(d18:1/24:1)
SM	47	SM(34:3), SM(35:2) (b), SM(37:1), SM(37:2), SM(38:3) (a), SM(38:3) (b), SM(40:3) (a), SM(40:3) (b), SM(40:4), SM(41:0), SM(41:1) (a), SM(42:3), SM(42:4), SM(43:1), SM(43:2) (b), SM(43:2) (c), SM(44:1), SM(44:2), SM(44:3), SM(d16:1/19:0), SM(d16:1/23:0)/SM(d17:1/22:0), SM(d16:1/24:1), SM(d17:1/14:0), SM(d17:1/16:0), SM(d17:1/24:1), SM(d18:0/14:0), SM(d18:0/16:0), SM(d18:0/22:0), SM(d18:1/14:0)/SM(d16:1/16:0), SM(d18:1/16:0), SM(d18:1/17:0)/SM(d17:1/18:0), SM(d18:1/18:0)/SM(d16:1/20:0), SM(d18:1/20:0)/SM(d16:1/22:0), SM(d18:1/22:0)/SM(d16:1/24:0), SM(d18:1/23:0)/SM(d17:1/24:0), SM(d18:1/24:0), SM(d18:1/24:1), SM(d18:2/14:0), SM(d18:2/16:0), SM(d18:2/17:0), SM(d18:2/18:0), SM(d18:2/18:1), SM(d18:2/20:0), SM(d18:2/22:0), SM(d18:2/23:0), SM(d18:2/24:0), SM(d19:1/24:1)
Sph	1	Sph(d18:1)

TG [NL]	77	TG(48:0) [NL-16:0], TG(48:0) [NL-18:0], TG(48:1) [NL-16:1], TG(48:1) [NL-18:1], TG(48:2) [NL-14:0], TG(48:2) [NL-14:1], TG(48:2) [NL-16:1], TG(48:2) [NL-18:2], TG(48:3) [NL-14:0], TG(48:3) [NL-16:1], TG(48:3) [NL-18:3], TG(49:1) [NL-16:1], TG(49:1) [NL-17:1], TG(50:0) [NL-18:0], TG(50:1) [NL-14:0], TG(50:1) [NL-16:0], TG(50:1) [NL-18:1], TG(50:2) [NL-14:0], TG(50:2) [NL-16:1], TG(50:2) [NL-18:1], TG(50:2) [NL-18:2], TG(50:3) [NL-14:0], TG(50:3) [NL-14:1], TG(50:3) [NL-16:1], TG(50:3) [NL-18:2], TG(50:3) [NL-18:3], TG(50:4) [NL-14:0], TG(50:4) [NL-18:3], TG(50:4) [NL-20:4], TG(51:0) [NL-16:0], TG(51:1) [NL-17:0], TG(51:2) [NL-15:0], TG(51:2) [NL-17:0], TG(51:2) [NL-17:1], TG(52:1) [NL-18:0], TG(52:1) [NL-18:1], TG(52:2) [NL-16:0], TG(52:2) [NL-18:2], TG(52:3) [NL-16:1], TG(52:3) [NL-18:2], TG(52:4) [NL-16:1], TG(52:4) [NL-18:2], TG(52:4) [NL-18:3], TG(52:5) [NL-18:3], TG(52:5) [NL-20:4], TG(52:5) [NL-20:5], TG(53:2) [NL-17:1], TG(53:2) [NL-18:1], TG(54:0) [NL-18:0], TG(54:1) [NL-18:1], TG(54:2) [NL-18:0], TG(54:2) [NL-20:1], TG(54:3) [NL-18:1], TG(54:3) [NL-18:2], TG(54:4) [NL-18:2], TG(54:4) [NL-20:3], TG(54:5) [NL-18:3], TG(54:5) [NL-20:4], TG(54:6) [NL-18:3], TG(54:6) [NL-20:4], TG(54:6) [NL-20:5], TG(54:6) [NL-22:6], TG(54:7) [NL-20:5], TG(54:7) [NL-22:6], TG(56:6) [NL-20:4], TG(56:6) [NL-22:5], TG(56:7) [NL-20:4], TG(56:7) [NL-20:5], TG(56:7) [NL-22:5], TG(56:7) [NL-22:6], TG(56:8) [NL-20:4], TG(56:8) [NL-20:5], TG(56:8) [NL-22:6], TG(56:9) [NL-22:6], TG(58:10) [NL-22:6], TG(58:8) [NL-22:6], TG(58:9) [NL-22:6]
TG(O) [NL]	20	TG(O-50:1) [NL-15:0], TG(O-50:1) [NL-16:0], TG(O-50:1) [NL-17:1], TG(O-50:1) [NL-18:1], TG(O-50:2) [NL-16:1], TG(O-50:2) [NL-18:1], TG(O-50:2) [NL-18:2], TG(O-50:3) [NL-18:2], TG(O-52:0) [NL-16:0], TG(O-52:1) [NL-16:0], TG(O-52:1) [NL-18:1], TG(O-52:2) [NL-16:0], TG(O-52:2) [NL-17:1], TG(O-52:2) [NL-18:1], TG(O-54:2) [NL-17:1], TG(O-54:2) [NL-18:1], TG(O-54:3) [NL-17:1], TG(O-54:3) [NL-18:1], TG(O-54:4) [NL-17:1], TG(O-54:4) [NL-18:2]
TG(SIM)	34	TG(48:0) [SIM], TG(48:1) [SIM], TG(48:2) [SIM], TG(48:3) [SIM], TG(49:1) [SIM], TG(50:0) [SIM], TG(50:1) [SIM], TG(50:2) [SIM], TG(50:3) [SIM], TG(50:4) [SIM], TG(51:0) [SIM], TG(51:1) [SIM], TG(51:2) [SIM], TG(52:1) [SIM], TG(52:2) [SIM], TG(52:3) [SIM], TG(52:4) [SIM], TG(52:5) [SIM], TG(53:2) [SIM], TG(54:0) [SIM], TG(54:1) [SIM], TG(54:2) [SIM], TG(54:3) [SIM], TG(54:4) [SIM], TG(54:5) [SIM], TG(54:6) [SIM], TG(54:7) [SIM], TG(56:6) [SIM], TG(56:7) [SIM], TG(56:8) [SIM], TG(56:9) [SIM], TG(58:10) [SIM], TG(58:8) [SIM], TG(58:9) [SIM]
Ubiquinone	1	Ubiquinone

Appendix B

Summary statistics from linear model analyses of lipid classes and species that remained Bonferroni significant after backwards stepwise regression. “Pheno”: phenotype (dependent variable) in the association test; “Data”: dataset – either lipid classes (“lipids_class”) or lipid species (“lipids”); “Lipid name”: name of the lipid; “Effect”: effect size for the association (OR for ASD, beta for all other traits); “SE”: standard error; “P”: p-value for association; “N”: number of participants included in that analysis.

Phenotype	Data	Lipid.name	Effect	SE	P	N
ASD	lipids_class	Total PC(O)	0.73	0.09	7.10E-04	694
IQ_DQ	lipids	PC(P-35:2) (b)	3.77	0.87	1.48E-05	642
IQ_DQ	lipids	LPC(O-22:0)	4.04	0.88	4.38E-06	642
sleep_problems	lipids_class	Total PE(O)	-1.38	0.40	6.43E-04	611
sleep_problems	lipids_class	Total LPE(P)	-1.42	0.40	4.22E-04	611
sleep_problems	lipids	PC(15:0_22:6)	-1.60	0.40	6.20E-05	611
sleep_problems	lipids	PC(P-18:0/22:6)	-1.61	0.40	5.51E-05	611
sleep_problems	lipids	PE(P-19:0/20:4) (b)	-1.95	0.40	8.70E-07	611
sleep_problems	lipids	PE(P-20:1/22:6)	-1.82	0.41	8.32E-06	611
sleep_problems	lipids	dimethyl-CE(22:6)	-1.64	0.41	5.40E-05	611
age	lipids_class	Total dhCer	-1.02	0.14	1.31E-13	758
age	lipids_class	Total Cer(m)	1.01	0.14	2.64E-13	758
age	lipids_class	Total SHexCer	-0.84	0.14	1.14E-09	758
age	lipids_class	Total deDE	0.76	0.14	6.50E-08	758
age	lipids_class	Total methyl-CE	-0.51	0.14	4.04E-04	758
age	lipids_class	Total dimethyl-CE	-0.72	0.14	2.72E-07	758
age	lipids_class	Total BA	0.66	0.14	2.48E-06	758
age	lipids	S1P(d18:2)	1.07	0.14	3.50E-14	758
age	lipids	dhCer(d18:0/24:0)	-1.29	0.13	1.04E-21	758
age	lipids	Cer(d17:1/18:0)	0.69	0.14	1.19E-06	758
age	lipids	Cer(d17:1/24:1)	0.94	0.14	3.49E-11	758
age	lipids	Cer(d18:1/20:0)	0.62	0.15	2.09E-05	758
age	lipids	Cer(d18:1/26:0)	-0.80	0.14	8.59E-09	758
age	lipids	Cer(d18:2/24:0)	1.03	0.14	2.17E-13	758
age	lipids	Cer(m18:0/22:0)	0.60	0.14	2.00E-05	758
age	lipids	Cer(m18:0/24:1)	0.90	0.14	8.53E-11	758
age	lipids	Cer(m18:1/23:0)	0.80	0.14	1.09E-08	758
age	lipids	Cer(m18:1/24:1)	1.42	0.13	2.38E-26	758
age	lipids	Cer1P(d18:1/16:0)	0.74	0.14	1.31E-07	758

age	lipids	HexCer(d18:1/16:0)	-1.13	0.14	1.15E-16	758
age	lipids	HexCer(d18:2/24:0)	1.10	0.14	2.92E-15	758
age	lipids	Hex2Cer(d18:1/18:0)	-1.31	0.14	2.75E-21	758
age	lipids	Hex2Cer(d18:2/16:0)	1.37	0.13	1.92E-24	758
age	lipids	Hex3Cer(d18:1/18:0)	-1.02	0.14	2.66E-13	758
age	lipids	Hex3Cer(d18:1/22:0)	-0.57	0.14	6.11E-05	758
age	lipids	GM3(d18:1/16:0)	-1.24	0.14	5.22E-20	758
age	lipids	SHexCer(d18:1/16:0)	-0.92	0.14	2.73E-11	758
age	lipids	SHexCer(d18:1/24:0)	-0.92	0.14	3.15E-11	758
age	lipids	SHexCer(d18:1/24:1)	-1.23	0.14	3.88E-19	758
age	lipids	SM(34:3)	0.81	0.14	8.74E-09	758
age	lipids	SM(40:3) (a)	0.93	0.14	2.78E-11	758
age	lipids	SM(41:0)	-1.75	0.13	8.57E-43	758
age	lipids	SM(41:1) (a)	-0.58	0.14	3.28E-05	758
age	lipids	SM(42:4)	1.31	0.14	1.14E-21	758
age	lipids	SM(43:1)	-1.01	0.14	1.37E-13	758
age	lipids	SM(44:1)	-1.09	0.14	1.56E-15	758
age	lipids	SM(d18:0/14:0)	-0.71	0.14	3.79E-07	758
age	lipids	SM(d18:0/16:0)	-1.24	0.14	1.47E-18	758
age	lipids	SM(d18:0/22:0)	-1.78	0.13	1.25E-44	758
age	lipids	SM(d18:1/23:0)/SM(d17:1/24:0)	-1.04	0.14	4.08E-14	758
age	lipids	SM(d18:1/24:0)	-0.96	0.14	3.13E-12	758
age	lipids	SM(d18:2/14:0)	1.30	0.14	2.65E-21	758
age	lipids	SM(d18:2/16:0)	1.56	0.14	1.51E-30	758
age	lipids	SM(d18:2/18:0)	1.19	0.14	3.55E-17	758
age	lipids	SM(d18:2/20:0)	1.24	0.14	6.35E-19	758
age	lipids	SM(d18:2/22:0)	0.76	0.14	9.93E-08	758
age	lipids	PC(14:0_22:6)	0.60	0.14	2.11E-05	758
age	lipids	PC(16:1_18:2)	0.97	0.14	2.94E-12	758
age	lipids	PC(16:1_20:4)	0.88	0.14	2.73E-10	758
age	lipids	PC(18:0_20:3)	0.85	0.14	1.42E-09	758
age	lipids	PC(32:2)	0.65	0.14	3.75E-06	758
age	lipids	PC(38:6) (a)	0.75	0.14	1.22E-07	758
age	lipids	PC(40:8)	0.84	0.14	1.53E-09	758
age	lipids	PC(42:5) (a)	-0.58	0.14	3.94E-05	758
age	lipids	PC(42:6) (b)	-1.12	0.14	1.80E-16	758
age	lipids	PC(42:8)	0.90	0.14	1.87E-10	758
age	lipids	PC(O-16:0/20:3)	0.64	0.14	1.08E-05	758
age	lipids	PC(P-16:0/16:1)	0.86	0.14	9.87E-10	758
age	lipids	PC(P-18:0/22:5)	-0.75	0.14	1.64E-07	758

age	lipids	PC(P-20:0/20:4)	-1.24	0.14	4.87E-20	758
age	lipids	LPC(20:4) [sn2]	0.75	0.14	1.05E-07	758
age	lipids	LPC(O-24:1)	0.71	0.14	3.93E-07	758
age	lipids	PE(18:1_18:1)	-0.61	0.14	1.53E-05	758
age	lipids	PE(P-15:0/20:4) (a)	-0.85	0.14	6.45E-10	758
age	lipids	PE(P-16:0/18:1)	-0.81	0.14	7.41E-09	758
age	lipids	PE(P-18:0/18:1)	-0.87	0.14	3.06E-10	758
age	lipids	PE(P-18:0/18:2)	-0.60	0.14	2.44E-05	758
age	lipids	PE(P-18:0/22:5) (n6)	-0.73	0.14	2.06E-07	758
age	lipids	PE(P-18:1/18:1) (a)	-0.82	0.14	4.80E-09	758
age	lipids	PE(P-20:0/18:2)	-1.67	0.13	2.45E-38	758
age	lipids	LPE(22:6) [sn2]	0.96	0.14	3.59E-12	758
age	lipids	LPE(P-20:0)	-0.65	0.14	3.77E-06	758
age	lipids	PI(16:0_20:3) (b)	-0.56	0.14	6.27E-05	758
age	lipids	PI(38:6)	-0.77	0.14	3.00E-08	758
age	lipids	CE(20:3)	0.84	0.14	2.57E-09	758
age	lipids	deDE(18:2)	0.63	0.14	1.03E-05	758
age	lipids	dimethyl-CE(18:2)	-0.72	0.14	3.42E-07	758
age	lipids	FA(22:6)	0.74	0.14	1.49E-07	758
age	lipids	AC(24:0)	-1.46	0.13	1.28E-27	758
age	lipids	TG(48:2) [NL-16:1]	0.57	0.14	4.62E-05	758
age	lipids	TG(48:3) [NL-16:1]	0.58	0.14	4.49E-05	758
age	lipids	TG(50:2) [NL-16:1]	0.68	0.14	1.19E-06	758
age	lipids	TG(50:3) [NL-16:1]	0.78	0.14	3.09E-08	758
age	lipids	TG(52:4) [NL-16:1]	0.68	0.14	1.54E-06	758
age	lipids	PC(36:4) [+OH]	1.07	0.14	3.93E-15	758
age	lipids	PC(38:6) [+OH]	0.98	0.14	1.38E-11	758
tanner_genital	lipids_class	Total SHexCer	-0.34	0.08	4.65E-05	224
tanner_genital	lipids_class	Total PE(P)	-0.30	0.09	3.94E-04	224
tanner_genital	lipids_class	Total dimethyl-CE	-0.35	0.08	1.96E-05	224
tanner_genital	lipids	dhCer(d18:0/24:0)	-0.39	0.10	4.08E-05	224
tanner_genital	lipids	Hex2Cer(d18:1/18:0)	-0.42	0.09	3.35E-06	224
tanner_genital	lipids	SM(44:1)	-0.41	0.08	3.26E-07	224
tanner_genital	lipids	SM(d18:0/16:0)	-0.42	0.09	2.55E-06	224
tanner_genital	lipids	SM(d18:0/22:0)	-0.49	0.08	5.81E-09	224
tanner_genital	lipids	SM(d18:1/24:0)	-0.43	0.08	2.93E-07	224
tanner_genital	lipids	PC(O-18:0/22:6)	-0.36	0.08	2.52E-05	224
tanner_genital	lipids	PE(P-20:0/18:1)	-0.46	0.10	4.22E-06	224
tanner_genital	lipids	PE(P-20:0/18:2)	-0.52	0.10	5.23E-08	224
tanner_genital	lipids	PI(16:0_20:3) (b)	-0.34	0.08	4.44E-05	224

tanner_genital	lipids	PI(18:0_22:6)	-0.43	0.09	1.18E-06	224
tanner_genital	lipids	PI(18:1_18:2)	-0.38	0.08	7.91E-06	224
tanner_genital	lipids	PI(38:6)	-0.44	0.08	6.42E-08	224
tanner_genital	lipids	AC(24:0)	-0.44	0.09	2.65E-06	224

AutismCRC

Autism CRC
The University of Queensland
Long Pocket Precinct
Level 3, Foxtail Building
80 Meiers Road
Indooroopilly Qld 4068
T +61 7 3377 0600
E info@autismcrc.com.au
W autismcrc.com.au

@autismcrc

Our values

Inclusion

Working together with those with the lived experience of autism in all we do

Innovation

New solutions for long term challenges

Evidence

Guided by evidence-based research and peer review

Independence

Maintaining autonomy and integrity

Cooperation

Bringing benefits to our partners; capturing opportunities they cannot capture alone

Australian Government
Department of Industry, Science, Energy and Resources

AusIndustry
Cooperative Research Centres Program