Supplementary File 15: Extended Materials/Subjects and Methods:

Autism Spectrum Disorder Diagnosis Verification

The Mobile Autism Risk Assessment (MARA), a brief level-2 screening measure for ASD, was collected from ASD participants [28]. The MARA is a behavioural questionnaire designed to screen children who are at high risk for ASD. Parents filled out the questionnaire about their children's behavior electronically and their answers were used in a classifier that outputs an ASD severity score [28]**.** Additionally, parents were asked to submit a short video of their child with and child without ASD enrolled in the study via encrypted file share. These videos were 2-6 minutes in length and were of the child during an independent or paired play session. These videos were then rated by minimally trained raters approved by the Stanford Institutional Review Board to score the presence and severity of a set of 30 behavioral features. Behavioral feature data collection methods are further described by Tariq et al [60].

The median of the scores across multiple raters were fed to previously published Machine Learning classifiers to predict ASD risk scores. [27,29]. By combining these risk scores with the parent-report screening tool (MARA), as well as parent-reported physician diagnosis, we confirmed diagnosis using majority rules consensus. We excluded (3) children and their TD siblings for whom the consensus did not agree with original parent diagnosis. We decided to include two children whose classifications did not reach majority consensus after realizing that the selected video-based classifiers did not include sensory dysregulation feature items. When viewing the sensory dysregulation item scores for these children, they scored in the ASD range, agreeing with their parent-reported diagnosis and their MARA scores.

Stool Collection and Storage

Stool samples were collected by the parents using a preservative buffer (Norgen Biotek, ON, Canada). The sampling consisted of 8 collection tubes. At the initial timepoint, we collected two samples per child: one sample was preserved at room temperature in a preservative buffer, and the second one was collected in a tube without a preservative buffer but immediately frozen at home at -20. This frozen sample was shipped back overnight with two ice packs provided to the participants. Other samples were shipped back every two weeks at room temperature in the preservative buffer. All shipping material was provided and pre-paid to the participants. Toilet collection containers were provided in the kit for collecting stool (sitting over the rear portion of the toilet). Participants were in contact with our clinical coordinator (by email or by phone). Once received, stool samples were stored at -80C until processing.

Sequence Processing, Filtering, and Taxonomic Annotation

Raw sequence reads were processed with DADA2 applying default settings for filtering, learning errors, dereplication, ASV inference, and chimera removal [62]. Truncation quality (truncQ) was set to 2. Ten nucleotides were then trimmed from each terminus of each read, both forward and reverse. An average of 156, 246 reads per sample library remained after processing the raw reads. For strain level ASV assignment, ASVs were mapped to an in-house strain database (StrainSelect,

https://www.secondgenome.com/platform/data-analysis-tools/strainselect, version 2019 (SS19)) using USEARCH (usearch_global) in the same manner as a recent study by Shah [63]. StrainSelect is a repository of strain identifiers obtained from gene sequencing, genome sequencing, draft genomes, and metagenomic assemblies, and assigns taxonomy using taxonomic annotations adapted from the Genome Taxonomy Database (GTDB). All sequences matching a unique strain at an identity ≥ 99% were assigned a strain-level annotation. To ensure specificity of these strain matches, a non-zero difference between the identity of the best match and the second best match was required *e.g.,* 99.75 vs. 99.5). Reads of the ASVs matched to the same strain were summed to represent the reads of the strain. If a unique strain match was not achieved, then species level and higher taxonomic placement was estimated with sintax (-cutoff 0.80) [64].

Normalization and Taxa Filtration

Before filtration, we attained a median read depth of 1.3 * 10^5 reads and a mean depth of 1.6*10^5 reads. Taxa not present in at least 3% of the samples were removed. Taxa abundances were normalized using DESeq2 or Cumulative Sum Scaling (CSS) depending on the contrast analysis performed. DESeq2 uses a negative binomial model and normalization by performing variance stabilization on taxa counts, then fitting a generalized linear model with log links to the normalized counts [65]. CSS normalizes counts by removing biases from taxa that are preferentially amplified in a sample-specific manner [66]. Due to how DESeq2 normalized minimized intra-group variance within families more so than CSS (ee Supplementary Information File 16), DESeq2 was used as the primary normalization in our gut-microbial community analysis.

In addition, taxa that significantly vary over time within the same individual were removed to increase the chance of identifying taxa directly related to core phenotype characteristics, rather than changes due to diet or season. A Friedman test was used to model ASV abundance as dependent on timepoint for each individual, and ASVs that were significantly related to timepoint ($p < .1$) were removed. 64 ASVs were removed from DESeq normalized data, 78 ASVs were removed from CSS normalized data, and 72 ASVs were removed from unnormalized data.

Analysis of Differentially Abundant Taxa between Phenotypes

ASV counts between the two phenotypes were compared using three common methods of differential analysis for amplicon data (DESeq2, MetagenomeSeq, and ANCOM2.1) [65, 66, 67]. Multiple analysis methods were employed to assess consistency of results and minimize spurious associations. In each of the above tests, we included the timepoint in the design matrix to account for differences between timepoints when estimating the effect of phenotype on any given taxa abundance.

Metadata Comparisons Between Cohorts

For each categorical variable from the metadata, a chi-square test was performed between the two cohorts. Wilcox-ranked sum tests were performed on non-longitudinal numerical metadata (age,

general dietary habits, etc.) Two-way mixed repeated measures anovas were used on numerical metadata that were collected longitudinally.

Identifying Confounding Factors impacting Microbial Structure

In order to address the potentially confounding variables, variables identified as driving factors in the PERMANOVA test and had significance in either a chi-square, wilcoxon-ranked sum test, or mixed two-way repeated measure ANOVA were also tested to find taxa that significantly differed based on these variables. MetagenomeSeq was used to identify specific ASVs associated with these confounding variables (Table 1). In addition, associations between beta-diversity and specific behavioral characteristics (MARA) were made only within the ASD cohort since TD individuals did not complete this questionnaire.

Identifying Driving Factors in Gut-Microbial Community Structure

A PERMANOVA test consisting of the adonis and betadisper functions from the vegan package were used to assess whether or not metadata variables collected in this study via questionnaires had a significant association with the gut-microbial community structure of our samples. Permutations were performed with the strata option, in order to constraint permutation according to our repeated measure study design. This test was performed using Bray-Curtis distances and DESeq2 normalized counts. Variables with a significant adonis p-value (<0.05) and insignificant betadisper p-values were identified to be significant driving factors.

Random Forest Model for Phenotype Classification

All models were trained and tested using 7 fold cross-validation, where siblings were placed in the same fold together. Samples from the same individual (over 3 timepoints) were treated as independent samples and also placed in a fold together. Optimal tree depth was chosen using 3 fold repeated cross validation on the training set for each of the 7 original folds. Area under the ROC curve was reported as a performance metric, and variable importance averaged across 7 folds was calculated as mean Gini index decrease. ASV counts were normalized using DESeq2 before being used in the model

Taxa Correlations with Anxiety Changes

Anxiety in the last 2 weeks was reported by caretakers on a scale of 1 to 3 where 1 meant low and 3 high anxiety. This metric was used to measure changes in anxiety within the same individual across time. We calculated the change in anxiety from timepoints 1 to 2, timepoints 2 to 3, and timepoints 1 to 3 for each individual. Samples without associated anxiety scores were removed. Respective changes in ASV abundance were calculated as log2 fold changes. We then correlated the log2 fold changes with changes in anxiety using a spearman correlation, and applied a multiple hypothesis correction. Positive values indicated an increase in anxiety or ASV abundance across timepoints for the same individual.