Infant Gut Microbiome Associated With Cognitive Development

Supplement 1

Supplementary Methods and Materials

DNA Isolation

Participating families were provided with a sample collection kit that included 2 tubes (one for backup) each containing 1 ml Allprotect reagent (Valencia, CA). Parents were instructed to collect approximately 200mg of feces from a single soiled diaper, immediately place it in a tube completely submerged in reagent, and return through overnight mail (samples submerged in Allprotect can be stored up to 7 days at 15-25ºC). Once received, the tubes were stored at -80°C until analysis. At the completion of the study, stool samples (200 mg) were transferred to sterile 2 ml tubes containing 200 mg of 212-300 μm glass beads (Sigma, St. Louis, MO) and 1.4 ml of Qiagen ASL buffer (Valencia, CA). Bead-beating was then carried out for 5 minutes in 1 minute intervals in a Qiagen TissueLyser II at 30Hz. Subsequently, samples were incubated at 95°C for 5 minutes and centrifuged at 21000 x g for 5 minutes. To remove PCR inhibitors, supernatants were transferred to new 2 ml-tubes containing InhibiEx inhibitor adsorption tablets (Qiagen) and vortexed vigorously. After a brief centrifugation, supernatants were aspirated and transferred to a new tube with Qiagen AL buffer containing Proteinase K (600IU/ μl). Samples were then incubated at 70°C for 10 minutes. DNA was purified using a standard on-column purification method with Qiagen buffers AW1 and AW2 as washing agents, and eluted in 10mM Tris (pH 8.0).

16S rRNA Amplicon Sequencing

12.5 ng of total DNA was amplified using a combination (4:1) of Universal and *Bifidobacterium*-specific primers targeting the V1-V2 region of the bacterial 16S rRNA gene (1; 2); primer sequences contained overhang adapters appended to the 5' end of each primer for compatibility with Illumina sequencing platform. The complete sequences of the primers were:

8F - 5'
TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGAGAGTTTGATCCTGGCTCAG3'

BifidoF- 5'
TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGAGGGTTCGATTCTGGCTCAG3'

338R - 5'
GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGGCTGCCTCCCGTAGGAGT3'

Master mixes contained 12.5 ng of total DNA, 0.2 µM of each primer and 2x KAPA HiFi HotStart ReadyMix (KAPA Biosystems, Wilmington, MA). The thermal profile for the amplification of each sample had an initial denaturing step at 95°C for 3 minutes, followed by a cycling of denaturing of 95°C for 30 seconds, annealing at 55°C for 30 seconds and a 30 second extension at 72°C (25 cycles), a 5-minute extension at 72°C and a final hold at 4°C. Each 16S amplicon was purified using the AMPure XP reagent (Beckman Coulter, Indianapolis, IN). Next, each sample was amplified using a limited cycle PCR program, adding Illumina sequencing adapters and dual-index barcodes (index 1(i7) and index 2(i5)) (Illumina, San Diego, CA) to the amplicon target. The thermal profile for the amplification of each sample had an initial denaturing step at 95°C for 3 minutes, followed by a denaturing cycle of 95°C for 30 seconds, annealing at 55°C for 30 seconds and a 30 second extension at 72°C (8 cycles), a 5-minute extension at 72°C and a final hold at 4°C. The final libraries were again purified using the AMPure XP reagent (Beckman Coulter), quantified and normalized prior to pooling. The DNA library pool was then denatured with NaOH, diluted with hybridization buffer and heat denatured before loading on the MiSeq reagent cartridge (Illumina) and on the MiSeq instrument (Illumina). Automated cluster generation and paired–end sequencing with dual reads were performed according to the manufacturer's instructions.

Sequencing Data Analysis

Multiplexed paired-end fastq files were produced from the sequencing results of the Illumina MiSeq using the Illumina software configureBclToFastq. The paired-end fastqs were joined into a single multiplexed, single-end fastq using the software tool fastq-join. Demultiplexing and quality filtering was performed on the joined results. Quality analysis reports were produced using the FastQC software. Bioinformatics analysis of bacterial 16S amplicon sequencing data was conducted using the Quantitative Insights Into Microbial Ecology (QIIME) software (3). OTU picking was performed on the quality filtered results using pick_de_novo_otus.py. Chimeric sequences were detected and removed using ChimeraSlayer. Alpha and beta diversity analysis were performed on the data set using the QIIME routines: alpha_rarefaction.py and beta_diversity_through_plots.py (4; 5), respectively. Summary reports of taxonomic assignment by sample and all categories were produced using QIIME summarize_taxa_through_plots.py and summarize_otu_by_cat.py.

Microbiota Clustering

In order to identify groups of subjects with similar microbial communities, we applied various distance metrics including Jensen-Shannon distance (JSD), square root of JSD (rJSD), unweighted UniFrac distance, weighted UniFrac distance and Bray-Curtis distance (BC) to the relative genus abundance in each individual. Both UniFracs were computed using QIIME, JSD, rJSD and BC were computed using R or R package *phyloseq*. Partitioning Around Medoids (PAM) was used as the clustering algorithm. Results were assessed for the optimal number of clusters using the Calinski-Harabasz (CH) Index, silhouette index (SI), and prediction strength (PS). We also evaluated the performance of the clustering by using between-class analysis through R package *ade4* using JSD as input. The between-class analysis is a particular case of principal components analysis (PCA) with an instrumental variable. The between-class analysis enables us first to find the principal components based on the center of gravity of each group in a way to highlight differences between groups and then to link each sample with its group. The visualization plot of between-class analysis indicates the quality of clustering and the homogeneity of the samples in each cluster.

Genera Analysis and Co-occurrence Networks

A modified Kruskal Wallis test for zero-inflated data (ZIKW) (http://www-stat.wharton.upenn.edu/~wanjiew/Testzero.pdf) was applied to identify significant differences in relative genera abundance between clusters. This method is designed to adjust for excessive zeros in the data and has higher statistical power compared to standard rank-based tests making it ideal for microbiome analysis. We corrected for multiple comparisons using Benjamini-Hochberg false discovery rate (FDR < 5%). Due to the relatively small sample size, a permuted method was used to obtain more accurate p-values. To better understand the unique community dynamics of the 1 year microbiome, we used one representative genera from each cluster as a seed to generate co-occurrence networks. To be selected as a seed, the following criteria had to be met: 1) the genus must show a significant difference between clusters after FDR correction, 2) the genus must have a median relative abundance value above 0.01 in at least 1 cluster, and 3) the median value for that genus must be higher in the cluster it represents than in the other 2 clusters. Where multiple genera met these criteria, the one with the highest test statistic was selected. Spearman correlations were computed between *Faecalibacterium* (representing Cluster 1), *Bacteroides* (representing Cluster 2), and an unnamed genus of Ruminococcaceae (representing Cluster 3) and all other genera. Absolute correlations above 0.3 transformed into connections between two genera in the co-occurrence genera network. Genera with less than 1% abundance were removed from the networks. The Cytoscape (6) package was used to generate the network figures with a spring-based layout.

Quality Control for Cognitive Testing

Given the potential challenges of cognitively assessing 1 and 2 year olds, quality control is very important. Cognitive testing staff members are well-versed in testing materials and highly skilled in handling and engaging infants and toddlers for sessions that run from 30 to 60 minutes. The goal is to obtain data that are representative of the child's potential. Parents are asked about the representativeness of their child's behavior during the assessment, and testing staff record their impressions of the child's state following the assessment to note if there are concerns about the representativeness of scores. Staff and research interns review the videos of these sessions and systematically score 8 behavioral dimensions on a 4-point scale for each of the 5 Mullen Scales. The dimensions include levels of overall alertness, focused attention, freedom from apprehension, freedom from negative affect, cooperativeness, persistence, flexibility, and enthusiasm. These ratings are then used to make decisions about what performance data may be compromised, necessitating exclusion from data analyses.

Predictor Covariate Identification

Relevant variables were created based on literature review of important factors that may influence early gut microbiome development. In order to identify environmental variables that could act as confounders due to their association with cluster membership or alpha diversity, we generated contingency tables with the Fisher Exact Test for categorical variables and performed linear mixed effect model (LMM) for continuous variables (see Table 1 and Supplementary Table S1). Binary categorical variables were filtered by requiring at least a 5% frequency in the study population. The p-values of both the Fisher's Exact Test and LMM were combined and corrected for multiple comparisons using FDR. Variables with q-values less than 0.05 were used as covariates in the subsequent predictive models for cluster or alpha diversity analyses. The following is a list of variables assessed and how they were created: income (subjects were stratified into low (<200% federal poverty level), middle (200%-400% federal poverty level), high (>400% federal poverty level), or not told groups based on the reported household income and federal poverty level for the household size at the year of visit), maternal & paternal psychiatric history (binary variable created for self-report or medical record positive for psychiatric history of schizophrenia spectrum disorder, bipolar disorder, depressive disorder, anxiety disorder, obsessive-compulsive disorder, attention-deficit/hyperactivity disorder, Tourette's syndrome, or autism spectrum disorders), cesarean section (medical record review), single or twin gestation (medical record review), sex (medical record review), >24hr stay in neonatal intensive care (medical record review), maternal ethnicity (parental report), paternal ethnicity (parental report), surgical anesthesia (medical record review), older siblings (parental report and medical record review), currently breastfeeding (parental report), ever given formula (parental report), currently given formula (parental report), given milk other than breastmilk or formula (parental report), type of other milk (such as almond, soy, coconut from parental report), symptoms of illness in previous week (parental report), gastrointestinal symptoms in previous week (parental report), antibiotics within last year (binary variable created from parental report and medical record review where oral antibiotics count as a positive finding), antibiotics during pregnancy (medical record review), gestational age at birth (medical record review), birth weight (medical record review), Apgar at 5 minutes (medical record review), maternal age at birth (parental report), maternal education (parental report), paternal age at birth (parental report), paternal education (parental report).

Covariate Identification for Outcome Variables

As part of the larger parent studies, a moment-based method (7; 8) to select fixed effects in linear mixed effects models was used to identify important covariates for global brain volumes (1 year n=526, 2 year n=375), regional brain volumes (1 year n=526, 2 year n=375), Mullen scores (1 year n=820, 2 year n=681), and longitudinal Mullen scores from 1 to 2 years (n=618). A list of potential predictors was considered that included sex, birth weight, twin status, delivery method, Apgar score at 5 minutes, gestational age at birth, stay in neonatal intensive care unit for more than 24 hours, Chiari I malformation, mild ventriculomegaly, major infection from birth to 1 year, major infection from 1 to 2 years, surgical anesthesia from birth to 1 year, surgical anesthesia from 1 to 2 years, age at 1 year MRI/2 year MRI/1 year Mullen/2 year Mullen, maternal education, paternal education, maternal age, paternal age, maternal ethnicity, paternal ethnicity, maternal psychiatric history, paternal psychiatric history, maternal smoking during pregnancy, total household income. For fixed effects selection, an adaptive Lasso penalty using the feasible generalized least squares estimator as an initial was applied. The BIC statistic was used to select the tuning parameter of the adaptive Lasso. Twins were treated as repeated measures. Before applying our variable selection method, all covariates were standardized and the response variable was centered. Bootstrap methods were applied 1000 times to assess the stability of the results. Variables were included as covariates for global volumes if they were selected more than 800 times for at least one volume. Variables were included as covariates for regional volumes if they were selected more than 800 times for at least 5% of ROIs. Variables were included as covariates for single time point Mullen cognitive outcomes and longitudinal Mullen outcomes from 1 to 2 years if they were selected more than 800 times for at least 5% of cognitive metrics.

Covariates:

1 Year Global Brain Volumes:
birth weight, sex, age at 1 year MRI, twin status, maternal education, paternal education

2 Year Global Brain Volumes:
age at 2 year MRI, sex, birth weight, paternal education

1 Year Regional Brain Volumes:
Intracranial Volume

2 Year Regional Brain Volumes:
Intracranial Volume

1 Year T-Score Mullen Outcomes:
Age at Mullen Testing

2 Year T-Score Mullen Outcomes:
sex, maternal education, paternal age, paternal ethnicity, twin status, income

1-2 Year Difference Raw Score Mullen Outcomes:
sex, age at Mullen testing, maternal education, paternal education, paternal ethnicity, twin status, total household income, maternal psychiatric history

Cluster:
Cesarean section, paternal ethnicity, currently breastfeeding

Alpha Diversity:
Paternal ethnicity, older siblings

Supplementary Table S1. Modified Kruskal Wallis test for zero-inflated data to test for significant differences in relative genera abundance between clusters.

	cluster1 median	cluster2 median	cluster3 median	statistic	pval	qval[a]
Bacteroides	0.3739	0.5046	0.0014	86.82	0.0000	0.0000
Clostridiales;f_;g	0.0076	0.0008	0.0033	79.54	0.0000	0.0000
Faecalibacterium	0.1555	0.0004	0.0545	78.20	0.0000	0.0000
Ruminococcaceae;Other	0.0015	2.10E-05	0.0063	77.89	0.0000	0.0000
Ruminococcaceae;g	0.0218	0.0018	0.0342	70.24	0.0000	0.0000
Clostridiales;Other;Other	0.0297	0.0029	0.0368	66.33	0.0005	0.0038
Lachnospira	0.0208	0.0002	0.0132	64.27	0.0003	0.0027
Ruminococcus	0.0029	0.0010	0.0024	63.52	0.0009	0.0060
Rikenellaceae;g	0.0015	1.31E-05	2.00E-05	52.01	0.0028	0.0165
Collinsella	0.0006	6.49E-06	2.00E-05	50.93	0.0039	0.0207
Coprococcus	0.0042	4.76E-05	0.0107	47.20	0.0531	0.2147
Dialister	0.0003	3.90E-05	0.0001	41.69	0.0753	0.2661
Clostridiaceae;g	0.0012	0.0005	0.0035	41.11	0.1491	0.4331
Dorea	0.0021	0.0008	0.0078	39.85	0.1716	0.4331
Blautia	0.0152	0.0037	0.0419	38.61	0.1990	0.4794
Lactococcus	0.0001	5.89E-06	5.70E-06	38.55	0.0204	0.0983
Parabacteroides	0.0001	1.18E-05	1.46E-05	33.20	0.1291	0.4025
Prevotella	0.0001	0.0001	0.0004	31.45	0.4951	0.9545
Oscillospira	0.0074	0.0076	0.0085	30.33	0.5223	0.9545
Lachnospiraceae;Other	0.0323	0.0150	0.0621	28.96	0.5952	0.9998
Megamonas	1.83E-05	2.67E-05	1.79E-05	26.94	0.1038	0.3438
Enterococcus	4.16E-05	0.0003	0.0001	25.89	0.3609	0.7970
Streptococcus	0.0055	0.0063	0.0040	23.60	0.8206	0.9998
Fusobacterium	1.67E-05	4.20E-05	3.14E-05	22.92	0.2235	0.5150
Clostridiaceae;Other	0.0002	0.0001	0.0001	22.49	0.6597	0.9998
Sutterella	0.0001	0.0030	0.0102	22.03	0.8865	0.9998
Unassigned;Other;Other;Other;Other;Other	0.0007	0.0010	0.0007	21.26	0.8945	0.9998
Enterobacteriaceae;Other	1.87E-05	0.0001	0	20.95	0.1686	0.4331
Coriobacteriaceae;Other	0.0004	0.0007	0.0004	20.61	0.8617	0.9998
Bifidobacterium	0.0071	0.0160	0.0085	19.65	0.9267	0.9998
Lachnospiraceae;g	0.0513	0.0638	0.1022	19.35	0.9368	0.9998
[Eubacterium]	0.0007	0.0010	0.0042	19.15	0.9490	0.9998
Lactobacillus	9.99E-06	0.0001	1.29E-05	18.58	0.4651	0.9481
Haemophilus	0.0002	0.0005	0.0002	16.13	0.9582	0.9998
Veillonella	0.0053	0.1493	0.0240	15.94	0.9782	0.9998

	cluster1 median	cluster2 median	cluster3 median	statistic	pval	qval[a]
Paraprevotella	0	0	0	15.60	0.0360	0.1590
Enterobacteriaceae;g	0.0012	0.0021	0.0016	15.29	0.9853	0.9998
Erysipelotrichaceae;g	0.0025	0.0033	0.0068	14.96	0.9865	0.9998
Bilophila	0	0	0	14.89	0.1621	0.4331
Akkermansia	1.57E-05	8.6170E-06	6.36E-06	13.13	0.6232	0.9998
Christensenellaceae;g	0	0	0	11.49	0.0567	0.2147
Clostridium	0.0015	0.0042	0.0097	10.21	0.9994	0.9998
[Ruminococcus]	0.0084	0.0305	0.0219	7.18	0.9998	0.9998
Acidaminococcus	0	0	0	6.86	0.6224	0.9998
RF32;f_;g	0	0	0	6.26	0.4412	0.9353
Megasphaera	9.15E-06	3.87E-05	2.00E-05	5.93	0.9708	0.9998
Dysgonomonas	0	0	0	4.21	0.5198	0.9545
Phascolarctobacterium	0	0	0	3.83	0.8577	0.9998
Epulopiscium	0	4.79E-06	1.47E-05	3.16	0.9647	0.9998
S24-7;g	0	0	0	2.98	0.6934	0.9998
Burkholderiales;Other;Other	0	0	0	1.97	0.6076	0.9998
Catenibacterium	0	0	0	0.50	0.9568	0.9998
[Prevotella]	0	0	0	0.40	0.9216	0.9998

[a] Gray cells are significant post FDR correction

Supplementary Table S2. Alpha Diversity Covariate Identification, q-values

	Faith's Phylogenetic Diversity		Shannon Index		Observed Species		Chao1	
	beta	q-value[a]	beta	q-value	beta	q-value	beta	q-value
Income		0.83		0.39		0.42		0.47
High	-0.46		-0.29		-8.64		-13.74	
Mid	-0.27		-0.10		-4.66		-5.90	
Not Told	-0.19		0.02		-5.96		-0.26	
Maternal Psychiatric History	-0.09	0.88	0.08	0.98	1.03	0.81	12.33	0.47
Paternal Psychiatric History	-0.22	0.88	-0.27	0.70	-4.65	0.68	-7.88	0.80
Cesarean Section	0.09	0.88	-0.01	0.98	2.53	0.68	-2.03	0.86
Twin Gestation	0.67	0.42	0.31	0.29	11.52	0.06	10.95	0.47
Sex	0.01	0.97	0.15	0.73	1.13	0.81	1.77	0.86
NICU	0.13	0.88	0.25	0.70	4.53	0.68	-2.09	0.86
Maternal Ethnicity		0.88		0.84		0.34		0.45
Black	0.25		0.13		8.37		13.66	
Asian	0.17		0.14		12.76		24.62	
Native American	0.28		0.21		8.56		11.66	
Paternal Ethnicity		0.60		0.29		0.05		0.04
Black	0.59		0.26		12.26		17.60	
Asian	0.69		0.55		21.83		39.35	
Native American	-0.43		0.41		-4.32		-18.43	
Surgical Anesthesia	-0.46	0.88	-0.15	0.98	-4.78	0.68	-15.02	0.75
Older Siblings	1.12	0.01	0.38	0.14	11.60	0.06	16.60	0.27
Currently Breastfed	-0.79	0.06	-0.26	0.39	-11.05	0.06	-14.45	0.38
Ever Given Formula	0.51	0.80	0.16	0.73	6.84	0.42	9.27	0.72
Currently Given Formula	0.14	0.88	0.00	0.98	-2.82	0.68	-3.88	0.85
Given Milk Other Than Breastmilk or Formula	0.39	0.88	0.08	0.98	5.78	0.68	9.63	0.79
Type of Other Milk	0.16	0.88	0.02	0.98	2.08	0.75	2.34	0.86
Symptoms of Illness in Previous Week	-0.16	0.88	0.01	0.98	-2.10	0.68	-7.03	0.77
Gastrointestinal Symptoms in Previous Week	-0.08	0.88	-0.30	0.39	-9.08	0.24	-6.86	0.80
Antibiotics within Last Year	-0.30	0.80	-0.36	0.14	-5.61	0.39	-9.96	0.47
Antibiotics During Pregnancy	0.41	0.88	-0.05	0.98	4.94	0.68	6.05	0.80
Gestational Age at Birth	-0.01	0.80	0.00	0.73	-0.20	0.42	-0.20	0.79
Birth Weight	0.00	0.88	0.00	0.87	0.00	0.68	0.00	0.80
APGAR5	-0.04	0.88	-0.02	0.98	-1.49	0.68	2.94	0.80
Maternal Age	-0.04	0.83	-0.02	0.73	-0.34	0.68	-0.41	0.80
Maternal Education	0.01	0.95	-0.01	0.98	-0.50	0.68	-0.78	0.80

	Faith's Phylogenetic Diversity		Shannon Index		Observed Species		Chao1	
	beta	q-value[a]	beta	q-value	beta	q-value	beta	q-value
Paternal Age	0.01	0.88	0.00	0.98	0.20	0.68	-0.10	0.87
Paternal Education	-0.03	0.88	-0.01	0.98	-0.89	0.45	-0.31	0.86

[a] Gray cells are significant post FDR correction

Supplementary Table S3. PICRUSt analysis results. 3a. KEGG Ortholog differences between clusters; 3b. KEGG L3 pathway differences between clusters; 3c. KEGG L2 pathway differences between clusters; 3d - KEGG L1 pathway differences between clusters.

Please note: due to its large size, Supplementary Table S3 is provided as a separate, multisheet Excel file.

Supplementary Table S4. Sensitivity analyses controlling for cluster or alpha diversity in reciprocal analyses.

Change in Cluster P/Q-Values with Alpha Diversity as Covariate

	Cluster w/CH1 Covariate	Cluster w/OS Covariate	Cluster w/SI Covariate	Cluster w/FPD Covariate
ELC (1yr)	0.766	0.711	0.841	0.615
GM (1yr)	0.803	0.821	0.858	0.791
FM (1yr)	0.803	0.821	0.858	0.791
VR (1yr)	0.803	0.821	0.858	0.791
RL (1yr)	0.803	0.821	0.858	0.791
EL (1yr)	0.952	0.961	0.979	0.847
ELC (2yr)	0.026	0.051	0.006	0.026
GM (2yr)	1.000	0.642	0.931	0.273
FM (2yr)	0.722	0.536	0.931	0.273
VR (2yr)	1.000	1.000	1.000	0.212
RL (2yr)	0.041	0.071	0.004	0.035
EL (2yr)	0.041	0.071	0.004	0.035

Change in Alpha Diversity P/Q Values with Cluster as Covariate

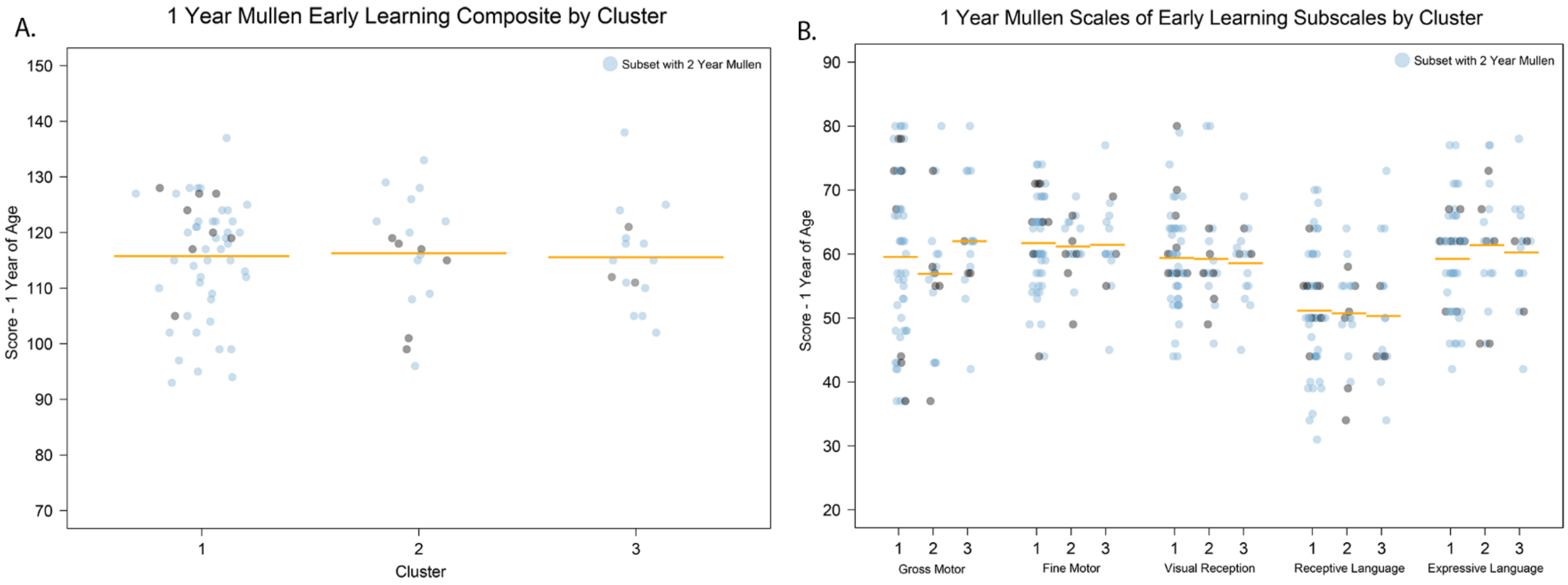
	CH1 w/Cluster Covariate	OS w/Cluster Covariate	SI w/Cluster Covariate	FPD w/Cluster Covariate
ELC (1yr)	0.806	0.359	0.347	0.246
GM (1yr)	0.981	0.768	0.768	0.788
FM (1yr)	0.981	0.868	0.768	0.669
VR (1yr)	0.768	0.981	0.981	0.967
RL (1yr)	0.367	0.177	0.170	0.040
EL (1yr)	0.669	0.669	0.669	0.367
ELC (2yr)	0.072	0.082	0.983	0.155

	CH1 w/Cluster Covariate	OS w/Cluster Covariate	SI w/Cluster Covariate	FPD w/Cluster Covariate
GM (2yr)	0.597	0.597	0.783	0.222
FM (2yr)	0.295	0.155	0.783	0.047
VR (2yr)	0.047	0.047	0.047	0.134
RL (2yr)	0.597	0.783	0.783	0.783
EL (2yr)	0.019	0.047	0.576	0.060

No Change

Loss of significance

Gain of significance



Supplementary Figure S1. Performance on Mullen Scales of Early Learning at 1 year of age does not differ between clusters in the subset (N=69) of infants with 2 year data. A. Individual value plot showing performance on the Mullen Early Learning Composite by cluster. B. Individual value plot showing secondary analysis of the Mullen Scale performance by cluster. Covariates for both analyses: cesarean section, paternal ethnicity, currently breastfeeding, age at Mullen testing.

Supplementary References

1. Edwards U, Rogall T, Blöcker H, Emde M, Böttger EC (1989): Isolation and direct complete nucleotide determination of entire genes. Characterization of a gene coding for 16S ribosomal RNA. *Nucleic Acids Res*. 17: 7843–7853.

2. Fierer N, Hamady M, Lauber CL, Knight R (2008): The influence of sex, handedness, and washing on the diversity of hand surface bacteria. *Proc Natl Acad Sci U S A*. 105: 17994–9.

3. Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, *et al.* (2010): correspondence QIIME allows analysis of high- throughput community sequencing data Intensity normalization improves color calling in SOLiD sequencing. *Nat Publ Gr*. 7: 335–336.

4. Lozupone C, Knight R (2005): UniFrac : a New Phylogenetic Method for Comparing Microbial Communities. *Appl Environ Microbiol*. 71: 8228–8235.

5. Lozupone C, Hamady M, Knight R (2006): UniFrac-an online tool for comparing microbial community diversity in a phylogenetic context. *BMC Bioinformatics*. 7: 371.

6. Christmas, Rowan; Avila-Campillo, Iliana; Bolouri, Hamid; Schwikowski, Benno; Anderson, Mark; Kelley, Ryan; Landys, Nerius; Workman, Chris; Ideker, Trey; Cerami, Ethan; Sheridan, Rob; Bader, Gary D.; Sander C (2005): Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Am Assoc Cancer Res Educ B*. 12–16.

7. Ahn M, Zhang HH, Lu W (2012): Moment-based Method for Random Effects Selection in Linear Mixed Models. *Stat Sin*. 22: 1539–1562.

8. Knickmeyer RC, Xia K, Lu Z, Ahn M, Jha SC, Zou F, *et al.* (2016): Impact of Demographic and Obstetric Factors on Infant Brain Volumes: A Population Neuroscience Study. *Cereb Cortex*. 1–10.

Supplement 2

Click here to download Supplemental Information: Supplement 2_Carlson et al.xlsx

Supplementary Table S3a - KEGG Ortholog differences between clusters

	X.Intercept.	beta C2	beta C3
K01643	9.249E-05	0.00021832	7.46152E-05
K01644	0.000100327	0.000228985	6.34616E-05
K03313	0.000219947	0.000137581	-0.000150025
K01011	9.85976E-05	0.000234738	8.62544E-05
K01673	0.000297879	0.00019026	-0.000140101
K00122	1.38008E-05	8.13581E-05	-1.40668E-06
K08989	9.70681E-06	7.22099E-05	1.18768E-06
K10094	9.35114E-06	6.62544E-05	9.95398E-08
K01057	0.000195749	0.000106136	-0.000153665
K09477	9.53058E-06	6.83577E-05	6.86541E-07
K06726	6.10423E-05	0.000167045	4.34161E-05
K06956	4.33251E-05	0.00011366	3.30742E-05
K07588	0.000239531	0.000154049	-0.000184622
K07326	2.00652E-05	0.000137384	-1.87564E-06
K03404	2.78929E-05	0.000136275	-3.92187E-06
K03322	0.00023117	0.000113603	-0.000163076
K00261	2.23596E-05	7.85088E-05	-5.61378E-06
K03855	1.72502E-05	7.39584E-05	4.65294E-06
K00824	4.87472E-05	0.000109464	2.63476E-05
K03319	4.31982E-05	0.000185498	-1.21897E-05
K11065	0.00021442	0.000104084	-0.000157472
K01646	3.32048E-05	9.95389E-05	1.29578E-05
K09162	1.07519E-05	6.55459E-05	9.4895E-07
K04086	1.60129E-05	7.15148E-05	-5.54314E-06
K01847	0.00042839	0.000181001	-0.000354045
K08590	0.000219338	0.000116308	-0.000151302
K01007	5.32833E-05	0.000118691	5.58672E-05
K07238	0.000619709	-0.000187985	0.000164149
K00772	3.84051E-05	0.0001108	3.55328E-05
K13256	2.1372E-05	7.8137E-05	4.27497E-06
K09771	1.39403E-05	7.62084E-05	-9.80761E-07
K11709	1.41584E-05	7.51706E-05	-1.99953E-06
K03638	1.54248E-05	7.49891E-05	-3.18922E-06
K11707	1.4651E-05	7.47215E-05	-1.74093E-06
K11708	1.46506E-05	7.4721E-05	-1.7408E-06
K08680	1.54872E-05	7.67571E-05	-2.76036E-06
K11710	1.46509E-05	7.47219E-05	-1.74089E-06
K00313	2.59297E-05	8.31572E-05	4.31722E-06
K00370	1.56327E-05	6.88447E-05	-8.5005E-06
K04758	0.001294768	-0.000194984	0.00048376
K00086	1.80685E-05	6.57483E-05	1.13464E-05
K00015	1.21767E-05	7.05901E-05	-2.13814E-06
K00620	0.000976489	-0.000351388	0.000257327
K02548	0.000524321	0.000288298	-0.000322913
K11719	1.44311E-05	7.19497E-05	-1.29223E-06
K06923	0.000453011	-0.000149207	0.000157882
K00366	5.39235E-05	0.000111693	3.30904E-05
K12982	1.8416E-05	7.57201E-05	-1.18564E-06
K01906	1.80977E-05	7.13375E-05	-1.79316E-06
K00036	0.000204723	5.11811E-05	-0.000148901
K04759	0.00118096	-0.000181638	0.000175508
K01601	1.85125E-05	7.10681E-05	-2.16273E-06