Characteristics and Mechanisms of a Sphingolipid-associated Childhood Asthma Endotype

Daniela Rago, Msc, PhD; Casper-Emil T. Pedersen, MSc, PhD; Mengna Huang, PhD; Rachel S. Kelly, PhD; Gözde Gürdeniz, MSc, PhD; Nicklas Brustad, MD; Hanna Knihtil, MD, PhD; Kathleen A. Lee-Sarwar, MD, PhD; Andréanne Morin, MSc, PhD; Morten A. Rasmussen, MSc, PhD; Jakob Stokholm, MD, PhD; Klaus Bønnelykke, MD, PhD; Augusto A. Litonjua, MD, MPH; Craig E. Wheelock, PhD; Scott T. Weiss, MD, MS; Jessica Lasky-Su, ScD; Hans Bisgaard, MD, DMSc; Bo L. Chawes, MD, PhD, DMSc

CONTENTS

Table S2. Correlation analysis results between the age 6 months and 6 years plasma levels of the sphingolipids associated with early-onset asthma development before age 3 years, which were discovered in $COPSAC_{2010}$ and replicated in VDAART.

Table S3. sRAW measurements at age 6 years in relation to plasma phosphosphingolipids at 6 years of age: interaction and stratified models with 17q21 genotype.

Figure S1. QQ-plot of the p-values expected vs. observed from univariate models associating the: (**A**) Number of lower respiratory tract infections in the first 3 years of life, (**B**) Number of wheezing episodes in the first 3 years of life, (**C)** Age at onset of first severe exacerbations in the first 3 years of life, (**D**) Asthma cross-sectional at age 6 years, (**E**) sRAW at age 6 years, (**F**) FEV1 at age 6 years, (**G)** PD20 at age 6 years, (**H**) FEV1/FVC at age 6 years with the 6 months plasma metabolomic profiles.

Figure S2. QQ-plot of the p-values expected vs. observed from univariate models associating the 6 years plasma metabolomic profile with asthma at 6 years of age.

Figure S3 QQ-plot of the p-value expected vs. observed from univariate model associating the: (**A**) sRAW at 6years of age, (**B**) FEV1 at 6years of age, (**C**) PD20 at 6 years of age, (**D)** FEV1/FVC at 6years of age with 6 years metabolomic profiles.

Figure S4. Results from the PLS-DA model validation. (**A**) The histogram represents the distribution of the averaged values (from 100 test-sets) of AUC results from the permutation test, and the red vertical lines represents the averaged value (from 100 test set) from the original class. (**B**) The histogram represents the distribution of the averaged values (from 100 test-sets) of Classification Error (CE) results from the permutation test, and the red vertical lines represents the averaged CE value (from 100 test set) from the original class.

Figure S5. PCA scores plot of the entire dataset comprising the two time points. The samples are color based on the age of the child.

eMethods

Study populations

COPSAC$_{2010}$: It is a population-based mother-child cohort with 738 pregnant women recruited between 22-26 weeks of gestation. The pregnant women participated in a double blind randomized control trial (DB-RCT) in which they during pregnancy week 22-26 were randomly assigned in a 1:1 ratio to receive either 2.4 g per day of n−3 long-chain polyunsaturated fatty acids (LCPUFAs) or placebo till 1 week postpartum. Furthermore, a subgroup of the pregnant women (n=623) were also enrolled in a nested DB-RCT in which they were assigned to 2400 IU of vitamin D per day on top of the recommended pregnancy supplement of 400 IU/d, or placebo.

The offspring were subsequently followed during their first 6 years of life attending 12 scheduled clinical visits at age 1 week, 1, 3, 6, 12, 18, 24, 30, and 36 months after birth, and yearly thereafter at the COPSAC clinical research center. In addition, they had a daily diary filled by the parents prospectively from birth capturing troublesome lung symptoms (cough, wheeze, and dyspnea), anti-asthmatic treatment, skin symptoms and treatment, and infections.

VDAART: It is a selected mother-child cohort with pregnant non-smoking women recruited from three sites across the USA between October 2009 and July 2011 (Boston, San Diego and St Louis) between 10 to 18 weeks of gestation with a history of asthma, eczema, or allergic rhinitis, or women who conceived the child with a man with a history of such diseases. Women participated to a double-blind randomized control trial to either a daily dose of 4,000 IU vitamin D3 or a placebo tablet until delivery. All women additionally received a daily multivitamin containing 400 IU vitamin D3.

Randomization took into account the study-site and race.

The families were followed prospectively from birth by quarterly telephone interviews and yearly in-clinic visits.

The VDAART metabolomics dataset comprised 469 samples with 834 compounds (653 knowns, 181 unknowns). A total of 65 compounds were excluded due to missingness >= 30%, therefore the final dataset used for the replication included 769 compounds (617 knowns, 152 unknowns) and 421 samples (the exclusion was due to not having age 3 asthma/recurrent wheeze status, missing age 1 BMI, missing age 1 vitamin D, asthma/recurrent wheeze diagnosed before age 1). Missing values were imputed with half of the minimum value of each metabolite.

Primary outcomes

Asthma diagnosis

Asthma was solely diagnosed by the COPSAC pediatricians based on a validated symptom algorithm, which includes the following criteria: verified diary recordings of at least five episodes of troublesome lung symptoms within six months, each lasting at least three consecutive days; symptoms typical of asthma; need for intermittent rescue use of inhaled corticosteroids; and response to a three-month course of inhaled corticosteroids and relapse upon ending treatment[15].

Lower respiratory tract infections, wheezing episodes and acute severe exacerbations

Lower respiratory tract infections included bronchiolitis and pneumonia. Bronchiolitis was defined as cough, tachypnea, chest retractions, auscultative widespread crepitation, or rhonchi

in an infant less than 1 year of age. Pneumonia was diagnosed in children with significant cough, tachypnea, fever, and abnormal lung stethoscopy[56].

Wheezing episodes were captured from verified diary recordings and defined as episodes of troublesome lung symptoms lasting at least three consecutive days. Number of episodes at age 0-3 years was the endpoint.

An acute severe wheeze exacerbation was defined as acute asthma-like symptoms requiring hospitalization, oral prednisolone (1-2 mg/kg for 3-7 days) or high-dose inhaled corticosteroid treatment (at least 1600 mcg budesonide per day for 14 days). In case of hospitalization without involvement of the COPSAC research unit, hospital records were retrieved and reviewed to confirm that symptom history and treatment fulfilled the above criteria[57].

Lung function measurements at 6 years of age

Spirometry was performed using a MasterScope Pneumoscreen spirometer (Erich Jäeger, Würzburg, Germany). A minimum of three tests with a within-test difference in forced expiratory volume in the first second (FEV1) of maximum 10% was completed. The best FEV1 was used in the analysis after calibrating the values for age, sex and heigh[58].

Whole-body plethysmography was done using a MasterScope sealed bodybox (Erich Jäeger, Würzburg, Germany). A minimum of two assessments of specific airway resistance (sRAW) with a within-test difference of maximum 0.3 kPa/s were obtained, using the mean of the two measurements in the analysis after calibrating the values for age, sex and height[59].

Bronchial responsiveness was assessed by measuring FEV1 after a saline inhalation and after subsequent inhalations of methacholine in dose steps from 10 to 2560 μg in concentrations of either 3.834 mg/ml (low) or 30.68 mg/ml (high) (APS Pro, CareFusion, 234 GmbH, Germany).

The provocative dose of methacholine producing a 20% fall in FEV1 (PD20)[60] was estimated from the dose-response curve fitted with a logistic function[61].

Covariates

Breastfeeding duration for the 6 months metabolomics data

Information on exclusive breastfeeding duration was acquired during the scheduled visits and used to adjust the analyses using the 6 months metabolomics data.

BMI

Information about BMI was collected at the scheduled visits. For the analysis the BMI measurements were WHO z-scored and used to adjust analyses of both 6 months and 6 years metabolomics.

Prenatal n-3 LCPUFA and vitamin D Interventions

Treatment allocation to both DB-RCTs in the COPSAC2010 cohort was used to adjust the Cox regression models.

Blood sample collection and storage for metabolomics

Blood samples were drawn from the children at the scheduled visits at 6 months and at 6 years of age. The blood samples were collected in EDTA tubes and left at room temperature for 30 min and thereafter spun down for 10 min at 4000 rpm. The supernatants were collected and stored at -80 C until further analysis.

Liquid Chromatography-Mass Spectrometry (LC-MS) metabolomics analysis

Sample preparation

The sample preparation was done using an automated system MicroLab STAR® system from Hamilton Company. Samples were fortified with recovery standards for quality control (QC) purposes, then extracted with methanol under vigorous shaking for 2 min (Glen Mills GenoGrinder 2000) and finally centrifuged. The resulting extracts were divided into four aliquots, then placed on a TurboVap® (Zymark) to remove the organic solvent and thereafter stored overnight under nitrogen before preparation for LC-MS/MS analysis.

LC-MS/MS analysis

The analytical analysis was carried out using ACQUITY Ultra-Performance Liquid Chromatography (UPLC) (Waters, Milford, USA) with Q Exactive™ Hybrid Quadrupole-Orbitrap™ mass spectrometer interfaced with heated electrospray ionization (HESI-II) source (ThermoFisher Scientific, Waltham, Massachusetts, USA). The sample extracts were reconstituted in solvents compatible to each of the four LC-MS methods utilized: two separated reverse phase UPLC-ESI(+)MS/MS methods optimized for hydrophilic and hydrophobic compounds; one reverse phase UPLC-(-)MS/MS using basic optimized conditions; and one HILIC/UPLC-(-)MS/MS. The MS analysis alternated between full scan MS and data-dependent MS^n scans using dynamic exclusion. The scan range for both ionization modes was 70-1000 *m/z*.

Throughout this study, unknown metabolites have w level 4 annotation and they are reported with the corresponding m/z value, whereas "**" indicates level 3 annotation (putatively

characterized compound class), "*" is level 2 annotation (putatively annotated compound using spectral libraries), and no label indicates a level 1 annotation (identified compound with a chemical reference standard).

Quality control and Data pre-processing

Metabolomics dataset

The raw data was extracted, and the area-under-the-curve was used for semi-quantification of each peak. The peak identification was based on three matching criteria: retention time/index (RI) range, mass accuracy (+/- 10 ppm) and MS/MS spectra of authenticated standards or recurrent unknown entities present in the spectra library.

The instrumental performances were assessed by a well-characterized human plasma (as a technical replicate), extracted water samples (as processed blanks), and a mixture of QC standards (not interfering with the measurement of endogenous compounds) spiked in every analyzed sample prior to injection, which were all run through the batch analyses. The instrument variability was determined by calculating the median relative standard deviation (RSD) for the QC standards spiked in, which was around 7%. The total process variability was determined by calculating the median RSD for all endogenous metabolites (i.e., non-instrument standards) present in the technical replicates, which was within 10%.

Data normalization was applied to all the datasets from the four analytical platforms to compensate for the inter-day batch variation. Each variable was corrected in run-day by registering the medians to equal one (1.00) and normalizing each variable, accordingly. The final set of variables from all the four platforms was imported into Matlab (Version 9.3, the Mathworks, Inc, MA, USA) and R (Version 3.60) for statistical analysis. Principal component analysis (PCA) on the entire dataset, comprising samples from 6 months and 6 years and 1,138

unique metabolites, showed a separation based on the child's age (using the first two principal components). Samples from children aged 4-8 months were more similar compared to samples collected from children at older ages (0.9-4 years of age) and therefore we only include samples from age 4-8 months in the 6 month time-point data analysis. For the 6 year time-point data analysis we included all samples (5-7.3 years) since samples all cluster together (**Figure S10**). The metabolomics data set was afterwards split into two subsets based on the 6 month and 6 year time-points. A further quality control step was performed independently in the two data sets, removing samples with ≥ 30% missing values as well as features with ≥ 95% missing values.

17q21 genotyping dataset

Genotypes were called with Illumina Genome Studio software. We excluded individuals with individual genotyping call rate <0.95, gender mismatch, genetic duplicates or outlying heterozygosity >0.27 and <0.037.

Nasal brushing transcriptomic dataset

Samples with missing genotype data or a depth below 8M exon mapped reads were removed as well as samples with >3-fold absolute deviation from the median count per million. Furthermore, contaminated samples were detected using VerifyBamID.[1,2]. Genes with <1 count per million (CPM) in >85% of the samples or located on X and Y chromosomes were removed. Data were log-transformed and normalized using trimmed mean of M-value [3] methods and variance modeling (voom) [4].

Reference

1. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics 29, 15–21 (2013).

2. Jun, G. et al. Detecting and estimating contamination of human DNA samples in sequencing and array-based genotype data. Am. J. Hum. Genet. 91, 839–848 (2012).

3. Robinson, M. D. & Oshlack, A. A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biol. 11, R25 (2010).

4. Law, C. W., Chen, Y., Shi, W. & Smyth, G. K. voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. Genome Biol. 15, R29 (2014).

Data analysis

Univariate regression analysis

Linear regression analysis was performed using R “stats” package to relate the metabolite levels at 6 months and at 6 years of age with the lung function measurements (FEV1, sRAW and PD20) at 6 years of age. Furthermore, it was used to investigate the relationship between the gene expression with sRAW and sphingolipids measurements. Prior to using gene expression data, we regressed out the first ten principal components of the entire gene expression dataset.

Cox proportional hazards regression models were employed to relate metabolite levels at 6 months of age with asthma development from infancy until 3 years of age, and development of acute severe exacerbations using the R package “Survival”. The Proportional hazard

assumption was tested with the Schoenfeld residuals, using the function cox.zph in R package "Survival".

The replication in VDAART of the nominally significant sphingolipids was done using their age 1 year metabolomics data with Cox proportional hazards regression models using asthma development from 1 to 3 years of age using a FDR<0.05 significance level threshold.

General linear regression model with a quasi-Poisson distribution was used to associate the number of wheezing episodes or the number of lower respiratory tract infections with plasma metabolome at 6 months of age. Logistic regression analysis was employed for the cross-sectional asthma analyses.

All models were adjusted for exclusively breastfeeding duration, z-scored child BMI, LCPUFA and vitamin D intervention for the 6 months metabolome dataset and z-scored BMI for the 6 years dataset.

Multivariate partial least squares discriminant analysis (PLS-DA)

Supervised multivariate analysis, specifically PLS-DA was employed to explore the association between the metabolite levels and sRAW measurements. The sRAW variable expressed as a continuous measurement was split into quartiles and only the samples in the lower (25th percentile) and the upper quartile (75th percentile) were utilized for the analysis as a two-class problem.

PLS_Toolbox (Version 8.61, Eigenvector Research, Inc., MA, USA) was employed for the PLS-DA. 100 subsets were randomly drawn from the entire dataset. Each of the subsets was successively randomly split into a training set (90% of the data set) and validation set (10% of the data set) making sure that each class was equally represented in the validation test. For each iteration, variable importance for projection (VIP)-based PLS-DA was applied using 10-fold

cross-validation to calculate the optimal number of components based on the lowest misclassification error, and 10% of the lowest VIP metabolites were iteratively excluded until the model reached the best performance in terms of classification error. In each iteration, the predictive performance of the obtained model was assessed using the validation set in terms of classification error (CE) and the area under the receiver operator characteristic curve (AUROC). The global model performance was expressed in terms of mean CE and AUROC of the 100 discriminant models, which were compared with CE and AUROC from 100 random class-permuted PLS-DA models.

The final set of metabolites was chosen based on the common variables present at least in 90% of the generated models and used to build a new final PLS-DA model using the entire original dataset. Furthermore, a permutation test using 100x iterations was applied to assess the classification performance.

eFigures and eTables

Table S1. Characteristics of the children in the age 6 months and 6 years plasma metabolomics datasets.

Characteristics	6 months Children (N=577)	6 years Children (N=513)
Fish Oil Supplementation		
Placebo	289 (50.1%)	256 (49.9%)
N3-LCPUFA	288 (49.9%)	257 (50.1%)
Vitamin D Supplementation		
Not in the trial	80 (13.9%)	80 (15.6%)
Placebo	242 (41.9%)	212 (41.3%)
Vitamin D	255 (44.2%)	221 (43.1%)
Gender		
Female	283 (49.0%)	240 (46.8%)
Male	294 (51.0%)	273 (53.2%)
Asthma 0-3 years of age		
No	473 (82.0%)	-
Yes	104 (18.0%)	-
Asthma at 5 years of age		
No	493 (85.4%)	-
Yes	54 (9.4%)	-
Missing	30 (5.2%)	-
Asthma at 6 years of age		
No	496 (86.0%)	467 (91.0%)
Yes	38 (6.6%)	38 (7.4%)
Missing	43 (7.5%)	8 (1.6%)
Number of Lower Respiratory Tract Infections		
Mean (SD)	0.648 (1.18)	
Median [Min, Max]	0 [0, 13.0]	
Missing	9 (1.6%)	
Number of wheezy episodes		
Mean (SD)	6.70 (6.63)	
Median [Min, Max]	4.00 [0, 37.0]	
Missing	21 (3.6%)	

Wheezy exacerbation 0-3 years of age		
No	530 (91.9%)	
Yes	47 (8.1%)	
zBMI		
Mean (SD)	0.04 (0.91)	0.01 (0.81)
Median [Min, Max]	-0.02 [-2.31, 2.78]	-0.01 [-2.18, 2.96]
Missing	7 (1.2%)	1 (0.2%)
Exclusively Breastfeeding Duration (Days)		
Mean (SD)	106 (58.4)	-
Median [Min, Max]	122 [0, 239]	-
Missing	2 (0.3%)	-
sRAW		
Mean (SD)	1.11 (0.242)	1.10 (0.239)
Median [Min, Max]	1.07 [0.59, 2.01]	1.07 [0.58, 2.01]
Missing	78 (13.5%)	17 (3.3%)
FEV1		
Mean (SD)	1.32 (0.16)	1.32 (0.16)
Median [Min, Max]	1.32 [0.81, 1.76]	1.32 [0.77, 1.75]
Missing	105 (18.2%)	50 (9.7%)
PD20		
Mean (SD)	2.60 (4.84)	2.68 (5.01)
Median [Min, Max]	0.81 [0, 37.8]	0.84 [0, 37.8]
Missing	156 (27.0%)	91 (17.7%)
FEV1/FVC		
Mean (SD)	0.922 (0.0590)	0.922 (0.0591)
Median [Min, Max]	0.935 [0.616, 1.04]	0.934 [0.616, 1.04]
Missing	105 (18.2%)	50 (9.7%)
Eosinophils count at 18 months		
Mean (SD)	0.22 (0.189)	-
Median [Min, Max]	0.17 [0.01, 1.36]	-
Missing	244 (42.3%)	-
Neutrophils count at 18 months		
Mean (SD)	2.86 (1.46)	-
Median [Min, Max]	2.58 [0.716, 9.51]	-
Missing	244 (42.3%)	-
rs12936231		
0	138 (23.9%)	116 (22.6%)
1	273 (47.3%)	230 (44.8%)
2	115 (19.9%)	109 (21.2%)

Missing	51 (8.8%)	58 (11.3%)
rs7216389		
0	133 (23.1%)	112 (21.8%)
1	271 (47.0%)	230 (44.8%)
2	121 (21.0%)	111 (21.6%)
Missing	52 (9.0%)	60 (11.7%)
rs2305480		
0	104 (18.0%)	85 (16.6%)
1	259 (44.9%)	217 (42.3%)
2	162 (28.1%)	151 (29.4%)
Missing	52 (9.0%)	60 (11.7%)
rs4065275		
0	126 (21.8%)	108 (21.1%)
1	275 (47.7%)	233 (45.4%)
2	124 (21.5%)	112 (21.8%)
Missing	52 (9.0%)	60 (11.7%)

Table S2. Correlation analysis results between the age 6 months and 6 years plasma levels of the sphingolipids associated with early-onset asthma development before age 3 years, which were discovered in $COPSAC_{2010}$ and replicated in VDAART.

Metabolite	Class and Metabolic Sub-pathway	Correlation Coefficient
glycosyl-N-stearoyl-sphingosine (d18:1/18:0)	Ceramides	0.22
palmitoyl sphingomyelin (d18:1/16:0)	Sphingomyelins	0.27
sphingomyelin (d18:0/18:0, d19:0/17:0)*		0.16
sphingomyelin (d18:1/17:0, d17:1/18:0, d19:1/16:0)		0.21
sphingomyelin (d18:1/18:1, d18:2/18:0)		0.10
sphingomyelin (d18:2/18:1)*		0.13
sphingomyelin (d18:2/23:1)*		0.29
stearoyl sphingomyelin (d18:1/18:0)		0.16

Table S3. sRAW measurements at age 6 years in relation to plasma phosphosphingolipids at 6 years of age: interaction and stratified models with 17q21 genotype.

Metabolite	SNP (rs number)	Model	β-estimate	CI	Stratum (copies of the risk allele)	p-value
sphinganine-1-phosphate	rs12936231 (risk allele: C)	Metabolite*SNP	-0.032	-0.062 -0.001		0.043
sphingosine 1-phosphate		Metabolite*SNP	-0.028	-0.06 0.004		0.08
sphinganine-1-phosphate			-0.016	-0.052 0.019	0	0.26
sphinganine-1-phosphate			-0.038	-0.071 -0.004	1	0.02
sphinganine-1-phosphate			-0.080	-0.128 -0.033	2	0.003
sphinganine-1-phosphate	rs2305480 (risk allele: T)	Metabolite*SNP	-0.021	-0.055 0.013		0.23
sphingosine 1-phosphate		Metabolite*SNP	-0.005	-0.038 0.029		0.78
sphinganine-1-phosphate	rs4065275 (risk allele: T)	Metabolite*SNP	-0.022	-0.056 0.012		0.21
sphingosine 1-phosphate		Metabolite*SNP	-0.019	-0.053 0.014		0.26
sphinganine-1-phosphate	rs7216389 (risk allele: T)	Metabolite*SNP	-0.028	-0.06 0.004		0.09
sphingosine 1-phosphate		Metabolite*SNP	-0.019	-0.051 0.014		0.27

‡ : significant p-value ≦ 0.05

Figure S1. QQ-plot of the p-values expected vs. observed from univariate models associating the: (**A**) Number of lower respiratory tract infections in the first 3 years of life, (**B**) Number of wheezing episodes in the first 3 years of life, (**C)** Age at onset of first severe exacerbations in the first 3 years of life, (**D**) Asthma cross-sectional at age 6 years, (**E**) sRAW at age 6 years, (**F**) FEV1 at age 6 years, (**G)** PD20 at age 6 years, (**H**) FEV1/FVC at age 6 years with the 6 months plasma metabolomic profiles.

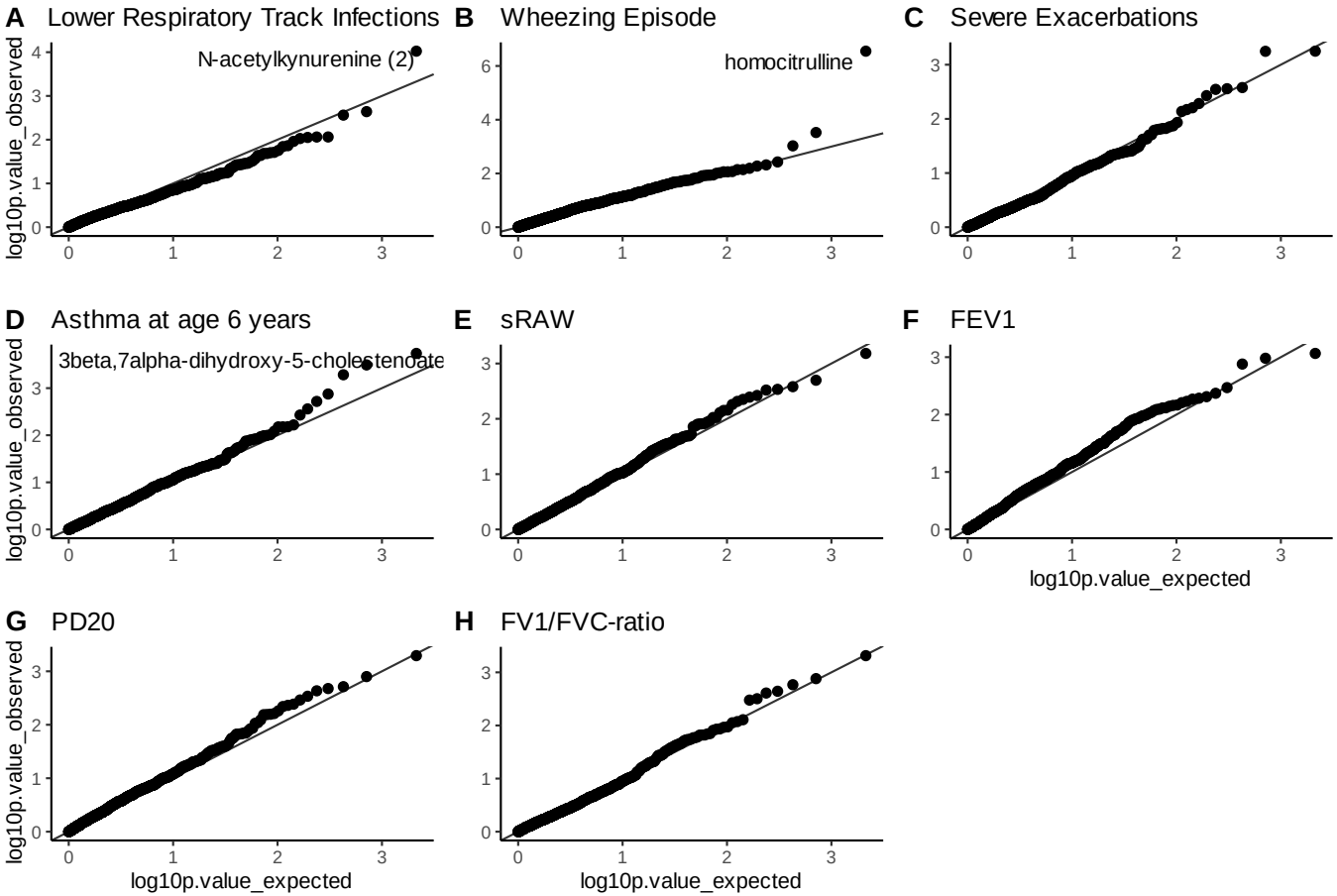

Figure S2. QQ-plot of the p-values expected vs. observed from univariate models associating the 6 years plasma metabolomic profile with asthma at 6 years of age.

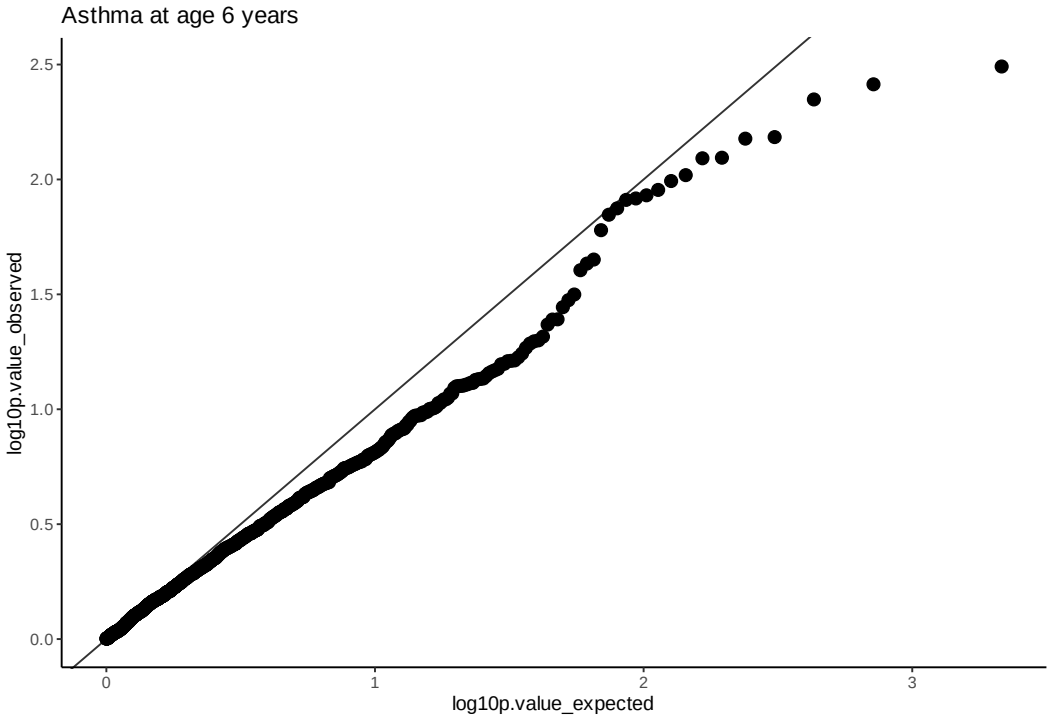

Figure S3 QQ-plot of the p-value expected vs. observed from univariate model associating the: (**A**) sRAW at 6years of age, (**B**) FEV1 at 6years of age, (**C**) PD20 at 6 years of age, (**D**) FEV1/FVC at 6years of age with 6 years metabolomic profiles.

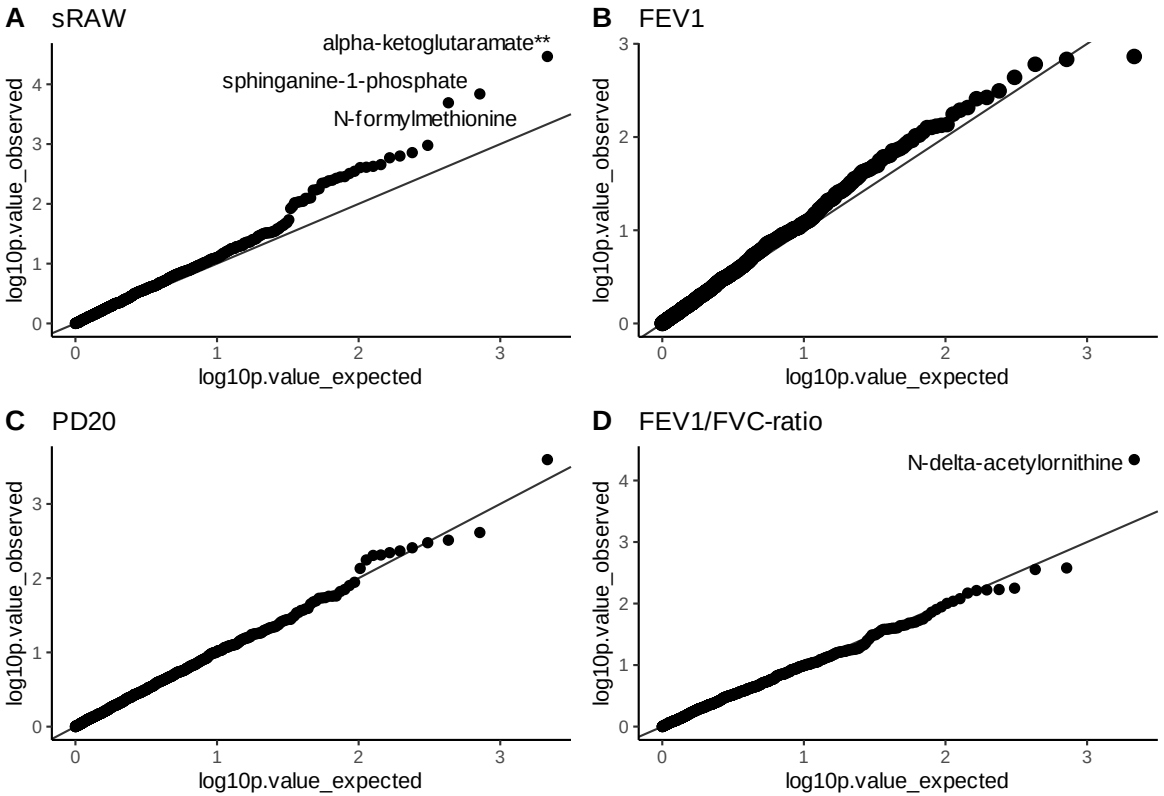

Figure S4. Results from the PLS-DA model validation. (**A**) The histogram represents the distribution of the averaged values (from 100 test-sets) of AUC results from the permutation test, and the red vertical lines represents the averaged value (from 100 test set) from the original class. (**B**) The histogram represents the distribution of the averaged values (from 100 test-sets) of Classification Error (CE) results from the permutation test, and the red vertical lines represents the averaged CE value (from 100 test set) from the original class.

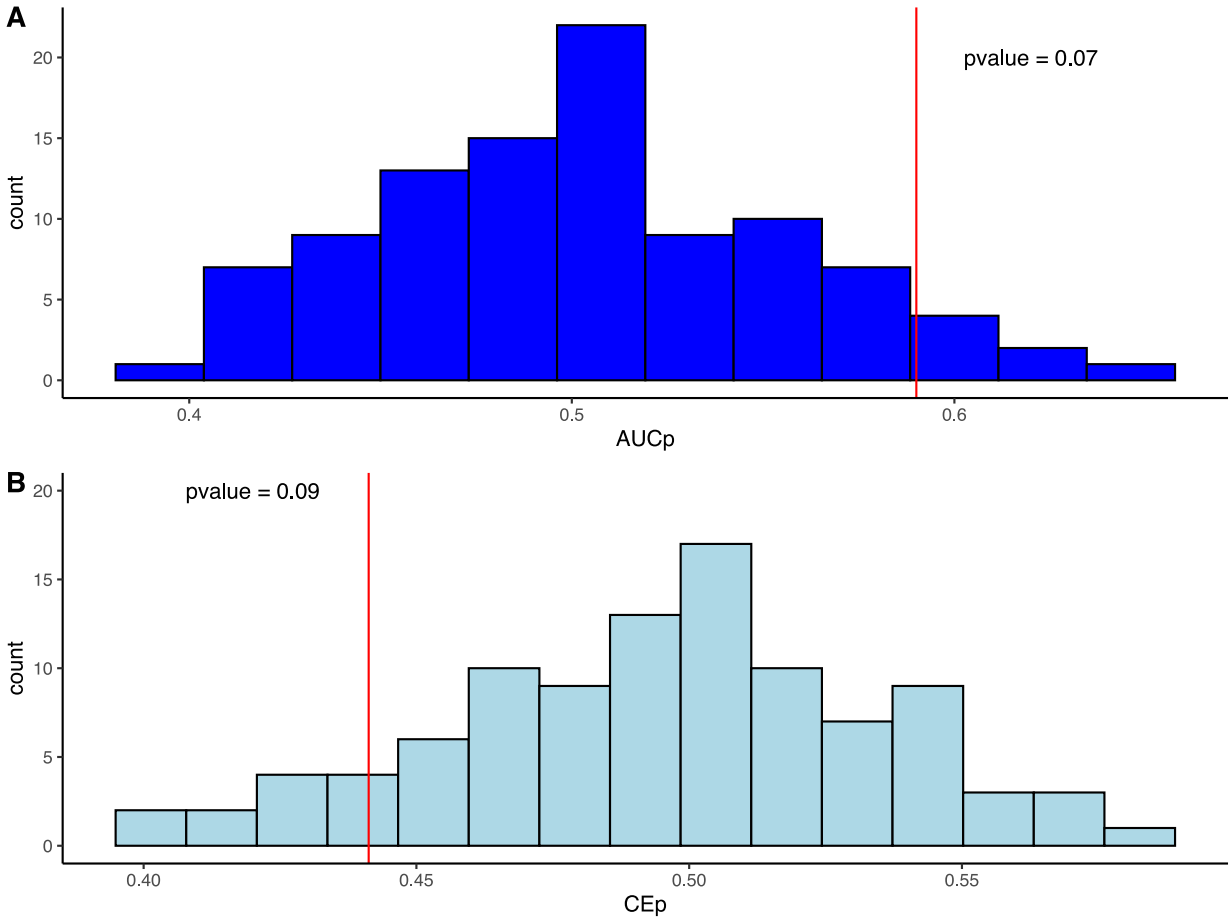

Figure S5. PCA scores plot of the entire dataset comprising the two time points. The samples are color based on the age of the child.