

MR. JAEHYUN PARK (Orcid ID : 0000-0003-2900-8340)

PROF. SOO-JONG HONG (Orcid ID : 0000-0003-4760-9262)

PROF. BONG-SOO KIM (Orcid ID : 0000-0003-1243-8280)

Article type : Letter to the Editor

Manuscript Acceptance Date: 28-Mar-2021

Multi-omics analyses implicate *EARS2* in the pathogenesis of atopic dermatitis

To the Editor:

In recent years, omics data for studies on atopic dermatitis (AD) have become more available. Most studies, however, have focused on a single type of omics data, owing to which, interplays between biological aspects may be missed or spurious correlations may be concluded. Here, using genetic and microbial profiles obtained from Cohort for Childhood Origin of Asthma and Allergic Diseases (COCOA) study¹, we built a prediction model of AD in 6-month-old infants using three omics datasets: microarray transcriptome data for host gene expression, and 16S rRNA microbiome and metagenome shotgun data for compositional and functional characteristics of the intestinal microbiota, respectively. Performance was evaluated using area under the curve (AUC) values

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the [Version of Record](#). Please cite this article as [doi: 10.1111/all.14837](#)

This article is protected by copyright. All rights reserved

corresponding to receiver operating characteristic (ROC) curves, and the relative importance of each omics data was investigated using McFadden's R-square².

We used 84 fecal samples with complete information of all three omics datasets. Among the subjects, 38 had AD while 46 were unaffected. In addition, sex, delivery method, feeding method, and family history of AD were included as clinical variables. The prediction model was built using the LASSO regression model. A 10-fold cross-validation was used to evaluate the performance of the estimated prediction model. Feature selection was conducted beforehand with a sparse DIABLO with a single component in the R package mixOmics^{3,4}. The details of data acquisition and preprocessing, variable imputation, feature selection, model building, and evaluation are provided in the Supplement.

According to the results of cross-validation, host expression of *EARS2* had a unique 100% selection probability (Table 1). AUC of the multi-omics prediction model was 0.715, which was slightly lower than that of the model with only transcriptome data and clinical variables (AUC 0.793, DeLong test p-value 0.128). The result indicated that incorporation of microbial information into the prediction model of AD does not provide an additional gain in accuracy.

The final prediction model was built with all 84 samples with three omics datasets and clinical variables. Figure 1 shows the 10 model coefficients with the largest absolute values in the model. Variables with higher selection probabilities in cross-validation tended to have larger effect size coefficients. *EARS2* showed the largest coefficient of 3.38, which meant the odds of having AD was $e^{3.38} \approx 29.5$ times higher if the normalized expression of *EARS2* increased by 1. The *EARS2* gene encodes a mitochondrial glutamyl-tRNA synthetase⁵. Thus, our results suggested that depletion of glutamate via mitochondrial tRNA synthesis might be important in the pathogenesis, similar to the previous report that insufficient glutamine due to *CARD11* mutation could be a cause of AD⁶. However, the actual mechanism should be studied by downstream metabolite analyses.

The level of gene expression explained most of the variances in the prediction model (McFadden's R² 32.66%), followed by metagenome shotgun data (4.91%) and microbiota (1%); clinical variables did not explain any variance in the prediction model. The unexplained part of variance was 61.47%, which probably indicated the requirement of other information in predicting the disease, including the

Accepted Article

profile of genetic mutations, metabolites, or interactions across the omics data. We noted that the small number of analyzed samples has a risk of selection bias, and generalization of the results to the entire Korean population should be made carefully.

In conclusion, we used various omics profiles to construct a prediction model of AD, and found several potentially crucial factors for the prediction, including the host expression of *EARS2*. Future studies, focusing on both revealing the effect of glutamyl-tRNA synthase on atopic eczema and verification of the results with a larger population, would be recommended.

References

1. Yang HJ, Lee SY, Suh DI, et al. The Cohort for Childhood Origin of Asthma and allergic diseases (COCOA) study: design, rationale and methods. *BMC pulmonary medicine*. 2014;14(1):109.
2. McFadden D Conditional logit analysis of qualitative choice behavior. 1973.
3. Singh A, Gautier B, Shannon CP, et al. Diablo - an integrative, multi-omics, multivariate method for multi-group classification. *bioRxiv*. 2016.
4. Rohart F, Gautier B, Singh A, Lê Cao KA. mixOmics: an R package for 'omics feature selection and multiple data integration. *PLOS computational biology*. 2017;13(11):e1005752.
5. Nagao A, Suzuki T, Katoh T, Sakaguchi Y, Suzuki T. Biogenesis of glutaminyl-mt tRNA^{Gln} in human mitochondria. *Proceedings of the National Academy of Sciences of the United States of America*. *Proceedings of the National Academy of Sciences*. 2009;106(38):09076021.
6. Ma CA, Stinson JR, Zhang Y, et al. Germline hypomorphic CARD11 mutations in severe atopic disease. *Nature genetics*. 2017;49(8):1192.

Multi-omics analyses implicate *EARS2* in the pathogenesis of atopic dermatitis

Jaehyun Park¹, Sharon M. Lutz², Seungil Choi³, Sanghun Lee⁴, Sangcheol Park⁵, Kangjin Kim⁶, Hosik Choi⁷, Hansoo Park^{8,9}, So Yeon Lee¹⁰, Scott T. Weiss¹¹, Soo-Jong Hong¹⁰, Bong-Soo Kim^{12*}, Sungho Won^{1, 5, 6, 13*}

¹Interdisciplinary Program in Bioinformatics, College of Natural Sciences, Seoul National University, Seoul, Republic of Korea

²Department of Population Medicine, Harvard Medical School and Harvard Pilgrim Health Care Institute, Boston, MA, United States

³Department of Statistics, College of Natural Sciences, Seoul National University, Seoul, Republic of Korea

⁴Department of Medical Consilience, Graduate School, Dankook University, Yongin-Si, Republic of Korea

⁵Seoul National University Institute of Health and Environment, Seoul, Republic of Korea

⁶Department of Public Health Science, Graduate School of Public Health, Seoul National University, Seoul, Republic of Korea

⁷Graduate School, Department of Urban Big Data Convergence, University of Seoul, Republic of Korea

⁸Department of Biomedical Science and Engineering, Gwangju Institute of Science and Technology (GIST), Gwangju, Republic of Korea

⁹Genome & Company, Seoungnam-si, Gyeonggi-do, Republic of Korea

¹⁰Department of Pediatrics, Childhood Asthma Atopy Center, Environmental Health Center, Asan Medical Center, University of Ulsan College of Medicine, Seoul, Republic of Korea

¹¹Channing Division of Network Medicine, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA, United States

¹²Department of Life Science, Multidisciplinary Genome Institute, Hallym University, Chuncheon, Gangwon-do, Republic of Korea

¹³RexSoft Corps, Seoul, Republic of Korea

Present addresses:

Seungil Choi: Department of Biostatistics, Graduate School of Public Health, University of Pittsburgh, Pittsburgh, PA, United States

Sangcheol Park: Department of Neurosurgery, Brigham and Women's Hospital, Boston, MA, United States

Correspondence

Sungho Won
sunghow@gmail.com

Bong-Soo Kim
bkim79@hallym.ac.kr

Keywords

atopic dermatitis (AD), *EARS2*, multi-omics analysis, prediction model

Acknowledgement

This study was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (NRF-2020R1F1A01072033); the Bio & Medical Technology Development Program of the National Research Foundation (NRF) funded by the Ministry of Science & ICT (MSIT) (2017M3A9F3043834); and the Korea Centers for Disease Control and Prevention (2008E3303000, 2009E3303300, 2011E3302100, 2012E3301200, 2013E5100300, 2014E5100400, 2014E5100401, 2014E5100402, 2017E6700200, 2017E6700201, 2017E6700202, and 2020E670200).

We would like to thank Editage (www.editage.co.kr) for English language editing.

Conflict of Interest

All of the authors have nothing to disclose. The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

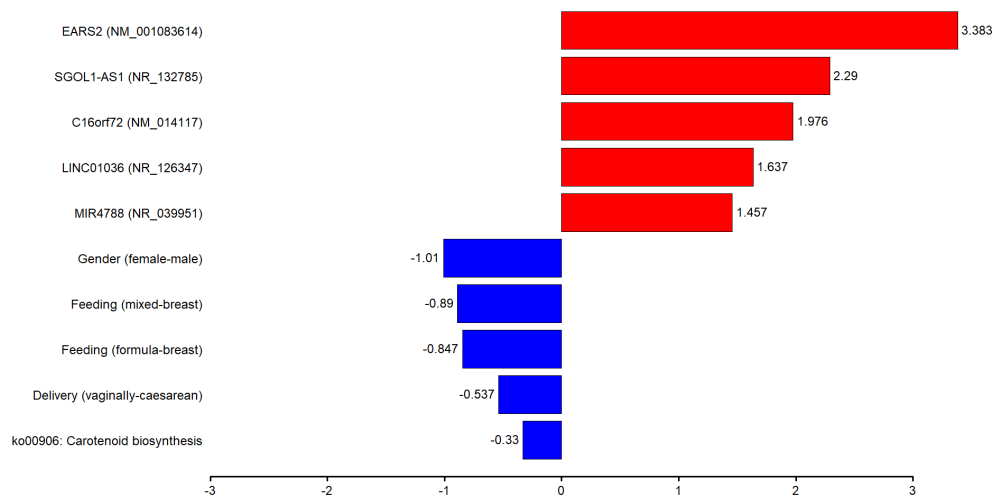
Author Contributions

JP contributed to statistical analyses, writing of the manuscript, and preparation of the tables. SC, SP, KK, HC, and SW contributed to statistical analyses. SL, SYL, and BSK contributed to the medical and biological interpretation of the results. BSK and SW contributed to the review and revision of the analyses and the manuscript. All authors reviewed the manuscript.

Table 1. Cross-validation results for multi-omics analyses. Cross-validation of the prediction model, which used all three multi-omics data, was conducted. Selection probability for each covariate indicates the proportion of folds where it was included in the prediction. Covariates with their selected proportions larger than or equal to 60% are listed. For genes from microarray, the mRNA accession numbers are also shown. Average coefficients indicate the average of coefficients from the prediction model in each fold.

Omics	Variable	Selection probability (%)	Average coefficient	OR
transcriptome	<i>EARS2</i> (NM_001083614)	100	7.061	1.166×10^3
transcriptome	<i>SGOL1-AS1</i> (NR_132785)	80	2.810	16.610
transcriptome	<i>LINC01036</i> (NR_126347)	60	3.474	32.266
transcriptome	<i>C16orf72</i> (NM_014117)	60	4.589	98.396
metagenome	ko00906: Carotenoid biosynthesis	60	-0.422	0.656

Figure 1. Top 10 biggest regression coefficients in the final prediction model. Variables are ordered by the absolute values of the coefficients. For genes from microarray, mRNA accession numbers are shown. For feeding types, considering three types, 2 dummy variables were used.



all_14837_f1.png

Supplementary materials for “Multi-omics analyses implicate *EARS2* in the pathogenesis of atopic dermatitis”

Supplementary material: full and detailed description of the methods.

COCOA Study

For the prediction of AD, we used 199 subjects from the Cohort for Childhood Origin of Asthma and Allergic Diseases (COCOA) study¹. The latter is a prospective birth cohort study in Korea conducted from September 2007, representing a Korean inner-city population. The COCOA data collection schedule included 11 routine standardized follow-up assessments of all children at 6 months and every year thereafter until 10 years of age, regardless of allergic disease development. From the cohort, we selected 95 patients with AD and sampled the non-patients using propensity scores calculated with age, sex, and feeding type. We decided to select 6-month-old infants, since we focused on the interaction between gut microbiota and the host in early childhood as an important factor in AD; further, microbial variation was expected to be minimal at that age, since the gut microbiota is known to be affected by diet, which is relatively simple at 6 months.

Microarray Transcriptome Data

Of the 199 subjects, microarray transcriptome data were generated for fecal samples were collected from infants who did not take antibiotics at 6 months of age and immediately frozen at -80 °C before being processed for RNA extraction. Total RNA was extracted from colonocytes and subjected to gene expression profiling on Affymetrix GeneChip Human Gene 2.0 ST Arrays (Thermo Fisher Scientific, Inc., Waltham, MA, USA) by Macrogen, Inc (Geumcheon-gu, Seoul, Republic of Korea). Signal values for 44,625 probes were normalized using RMA. Probes without their corresponding gene symbol or duplicated data were removed, and 30,980 probes were utilized for statistical analyses without any further preprocessing.

Microbiome Data Based on 16S rRNA Gene

We generated microbiota compositional profiles using 16S rRNA amplicons. DNA extracted from 149 and 48 stool samples from 6-month-old infants was sequenced using 454 pyrosequencing and Illumina MiSeq platform, and read counts for 898 and 205 genera were obtained, respectively. DNA extraction and sample processing were performed as described by Lee *et al.*². A total of 76 genera were commonly observed on the two sequencing platforms. Since quality control requires actual read counts, and we cannot directly compare the different platforms, we focused on the common genera.

Genera for each sequencing data were separately filtered based on the criteria detailed by Li *et al.*³ and Bokulich *et al.*⁴, whereas the actual cutoff values were set empirically. Genera were excluded if the relative ratios of read counts summed for all subjects to the total read counts were less than 0.05%, or if the proportion of subjects with at least one or more read counts was less than a quarter. The genera remaining after quality control in both datasets were retained. For each subject, we transformed the data to counts per million (CPM) values using the R package edgeR⁵, and log-transformed CPMs (log-CPMs) were used for statistical analyses. To adjust the batch effect caused by the sequencing method, log-CPMs were separately centered and scaled for each sequencing data, and then pooled.

Metagenome Shotgun Data

The profiling of microbial functions, using stool samples of infants, with whole metagenome sequencing was conducted using the method described by Lee *et al.*². Briefly, the Illumina adapter sequences and sequences with short read lengths (< 150 bp in each pair) and low Q scores (< 20) were

removed from raw sequences using the Nsoni clip tool within the Nsoni high-throughput sequencing data analysis toolset (ver. 0.127). Contaminating human genes were removed using BBMap by mapping to the reference human genome. Functional profiles of the filtered sequences were obtained using HUMAnN2. The KEGG Orthology (KO) ID was obtained using ‘humann2_regroup_table’ within HUMAnN2 utility scripts.

Two datasets were separately generated by sequencing microbial communities from 58 and 40 subjects, and reads per kilobase (RPK) values for each KEGG pathway were obtained. Three hundred and sixty-one pathways were commonly observed for both of the sequenced datasets. We excluded any pathway if the proportion of subjects with one or more reads was less than a quarter. Thereafter, log-CPM transformation was conducted using edgeR⁵. Log-CPM values were separately centered and scaled for each dataset before pooling to cope with the batch effect.

Due to the pipeline, which aligns sequenced reads to the bacterial genome, some genes in the pathways may not be mapped to the reference genome and marked as “unclassified.” We noted that pathways that were not mapped to the bacterial genome were unnecessary; therefore, they were excluded from the analysis.

Clinical Covariates for Prediction Models

Sex, delivery method, feeding method, and family history of atopic dermatitis were included in the prediction model as clinical variables. For these variables, missing values were imputed with the missForest^{6,7}. Supplementary Table 1 describes the descriptive statistics of those variables and they were compared between AD and normal groups using the chi-squared test performed by R package coin 1.2-2⁸. According to Supplementary Table 1, a family history of AD was significantly more prevalent in patients with AD (chi-squared test p-value 9.32×10^{-5}) than in normal individuals. No significant difference was observed for the other variables ($p > 0.2$).

Subjects for Statistical Analyses

For statistical analyses, we considered 84 subjects with all three omics profiles. Among them, metagenome sequencing data for 37 and 47 were from 454 pyrosequencing and Illumina MiSeq, respectively, and that for 56 and 28 were from separate datasets of the metagenome shotgun. The samples did not show statistically significant differences in sex, delivery mode, and family history (chi-squared p-value > 0.2) compared to 115 samples that were not included in the analyses. However, the analyzed samples showed significantly different proportions of feeding methods (p-value 7.69×10^{-5}), with a high proportion of formula feedings (15 out of 84 compared to 1 out of 115).

The imbalance in family history of AD was weakened in 84 subjects (Supplementary Table 1; chi-squared test p-value 0.340). For the other variables, there were still no significant differences between patients with AD and normal individuals ($p > 0.05$).

Feature Selection

It has been repeatedly reported that prediction accuracy can be improved with feature selection if the number of covariates is much larger than the sample size^{9,10}. Each omics data set consists of many different components. Thus, for each 10-fold cross-validation, important features were chosen with a training set, and then used to build a prediction model. Feature selection was conducted with a sparse DIABLO with a single component in the R package mixOmics^{11,12}. The design matrix was set as suggested by the original mixOmics paper¹², and the number of selected features was chosen with an internal tuning function in mixOmics with default parameters. The number of selected variables ranged from 5 to 50.

Prediction Model Building and Evaluation

Prediction models were built with clinical variables and all omics datasets: microarray transcriptome data and microbial community data. For microbial data, we used both the compositional data with 16S rRNA sequencing and functional data with metagenome sequencing, since the former focuses on the compositional aspects and can help develop the prediction at a lower cost, whereas the latter focuses on the functional aspects of microbial community.

A much larger number of variables than sample size was available in the omics data, and prediction models were built with penalizing coefficients for selected features from multi-omics data with a LASSO regression model¹³. We chose the LASSO model because it offers a stronger degree of feature selection, which may be important in building the prediction model. This may be advantageous to our data compared to other models, such as ridge regression. Each variable was standardized before building each of the models, although the resulting coefficients were returned to the original scale. Clinical variables were excluded from penalization. Tuning parameters for penalization were chosen with an additional 5-fold cross validation within the training set. Each training data were split into five subgroups, and prediction models with tuning parameters specified *a priori* were estimated with four subgroups. Thereafter, areas under the curves (AUCs) were estimated with the remaining subgroup, and tuning parameters that generated the largest AUC were chosen. Prediction models with the best choice of tuning parameters were built with the training set and applied to the test set, and the probabilities of each participant affected were estimated. The probabilities were used to draw ROC curves, and AUCs were used to evaluate the performance of each model. Data management and analyses were performed with R version 3.4.3¹⁴, and penalized regression was built with functions in the R package glmnet¹⁵.

The ridge regression model was also used; however, the performance (AUC 0.712) was not significantly different from the result of LASSO.

We evaluated the relative importance of each omics data point, using the estimated prediction model, by calculating the McFadden's R-square value¹⁶. If we denote the predicted and true phenotypes by $\hat{\mu}_i$ and Y_i for subjects in a test set, respectively, the deviance for subject i would be defined as

$$Y_i \log \frac{Y_i}{\hat{\mu}_i} + (1 - Y_i) \log \frac{1 - Y_i}{1 - \hat{\mu}_i} \quad (1)$$

The deviances were evaluated by summing those of all subjects. If we let Δ_F and Δ_0 be the deviance of the full model and reduced model, respectively, variability explained by the full model (Δ_F) using McFadden's R-square can be defined by

$$\left(1 - \frac{\Delta_F}{\Delta_0}\right) \times 100$$

To evaluate the importance of each omics data, we excluded single omics data from the full model. If the reduced model is denoted by R , relative importance of the excluded omics data may be obtained by

$$\left(1 - \frac{\Delta_F}{\Delta_0}\right) \times 100 - \left(\left(1 - \frac{\Delta_R}{\Delta_0}\right) \times 100\right) = \frac{\Delta_R - \Delta_F}{\Delta_0} \times 100$$

Selection Probability

The selection probability and average regression coefficient for each covariate were calculated using 10-fold cross-validation. The selection probability of a covariate indicates the relative proportion of the prediction model including it.

Supplementary Table 1. Descriptive statistics of clinical variables after variable imputation for all 199 (left) and 84 subjects (right) used in the prediction models. P-values were obtained using chi-squared tests with the R package coin.

Variable	Category	Control (n = 104)	AD (n = 95)	p-value	Control in the model (n = 46)	AD in the model (n = 38)	p-value
Sex	Male	56	57	0.381	23	27	0.0504
	Female	48	38		23	11	
Delivery mode	Caesarean	30	23	0.546	15	11	0.718
	Vaginal	74	72		31	27	
Feeding type*	Breast	33	41	0.222	13	15	0.552
	Mixed	61	48		24	17	
	Formula	10	6		9	6	
Family history*	0	61	28	9.32×10^{-5}	21	13	0.340
	1	42	67		24	25	
	2	1	0		1	0	

References

1. Yang HJ, Lee SY, Suh DI, et al. The Cohort for Childhood Origin of Asthma and allergic diseases (COCO) study: design, rationale and methods. *BMC Pulmonary Medicine*. 2014;14(1):109.
2. Lee MJ, Kang MJ, Lee SY, et al. Perturbations of gut microbiome genes in infants with atopic dermatitis according to feeding type. *Journal of Allergy & Clinical Immunology*. 2018;141(4):1310-1319.
3. Li K, Bihan M, Methé BA. Analyses of the stability and core taxonomic memberships of the human microbiome. *PLOS ONE*. 2013;8(5):e63139.
4. Bokulich NA, Subramanian S, Faith JJ, et al. Quality-filtering vastly improves diversity estimates from Illumina amplicon sequencing. *Nature Methods*. 2013;10(1):57-59.
5. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*. 2010;26(1):139-140.
6. Stekhoven DJ. missForest: nonparametric missing value imputation using random forest. <http://cran.r-project.org/web/packages/missForest/index.html>. Published; 2013. Accessed.
7. Stekhoven DJ, Bühlmann P. MissForest—non-parametric missing value imputation for mixed-type data. *Bioinformatics*. 2012;28(1):112-118.
8. Hothorn T, Hornik K, Van De Wiel MA, Zeileis A. A Lego system for conditional inference. *The American Statistician*. 2006;60(3):257-263.
9. Dash M, Liu H. Feature selection for classification. *Intelligent Data Analysis*. 1997;1(3):131-156.
10. Chandrashekar G, Sahin F. A survey on feature selection methods. *Computers & Electrical Engineering*. 2014;40(1):16-28.
11. Singh A, Gautier B, Shannon CP, et al. Diabolo - an integrative, multi-omics, multivariate method for multi-group classification. *bioRxiv*. 2016.
12. Rohart F, Gautier B, Singh A, Lê Cao KA. mixOmics: an R package for 'omics feature selection and multiple data integration. *PLOS Computational Biology*. 2017;13(11):e1005752.
13. Tibshirani R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B*. 1996;58(1):267-288.
14. R: A language and environment for statistical computing [computer program]. version 3.4.3. Vienna, Austria: R Foundation for Statistical Computing; 2017.
15. Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*. 2010;33(1):1-22.
16. McFadden D. Conditional logit analysis of qualitative choice behavior. In: Zarembka P, ed. *Frontiers in econometrics*. New York: Academic Press; 1974:105-142.