







Prediction of the development of islet autoantibodies through integration of environmental, genetic, and metabolic markers

Bobbie-Jo M. Webb-Robertson^{1,2}  | Lisa M. Bramer³  | Bryan A. Stanfill³  | Sarah M. Reehl³ | Ernesto S. Nakayasu¹  | Thomas O. Metz¹  | Brigitte I. Frohnert⁴  | Jill M. Norris²  | Randi K. Johnson²  | Stephen S. Rich⁵  | Marian J. Rewers⁴ 

¹Biological Sciences Division, Pacific Northwest National Laboratory, Richland, Washington

²Colorado School of Public Health, University of Colorado Anschutz Medical Campus, Aurora, California

³Computing and Analytics Division, Pacific Northwest National Laboratory, Richland, Washington

⁴Barbara Davis Center for Diabetes, University of Colorado Anschutz Medical Campus, Aurora, Colorado

⁵Center for Public Health Genomics, University of Virginia, Charlottesville, Virginia

Correspondence

Bobbie-Jo M. Webb-Robertson, Pacific Northwest National Laboratory, PO BOX 999, MSIN J4-18, Richland, WA 99352.
Email: bj@pnnl.gov

Funding information

National Center for Advancing Translational Sciences, Grant/Award Numbers: UC4 DK100238, ULI TR000064, ULI TR001082

Abstract

Background: The Environmental Determinants of the Diabetes in the Young (TEDDY) study has prospectively followed, from birth, children at increased genetic risk of type 1 diabetes. TEDDY has collected heterogeneous data longitudinally to gain insights into the environmental and biological mechanisms driving the progression to persistent islet autoantibodies.

Methods: We developed a machine learning model to predict imminent transition to the development of persistent islet autoantibodies based on time-varying metabolomics data integrated with time-invariant risk factors (eg, gestational age). The machine learning was initiated with 221 potential features (85 genetic, 5 environmental, 131 metabolomic) and an ensemble-based feature evaluation was utilized to identify a small set of predictive features that can be interrogated to better understand the pathogenesis leading up to persistent islet autoimmunity.

Results: The final integrative machine learning model included 42 disparate features, returning a cross-validated receiver operating characteristic area under the curve (AUC) of 0.74 and an AUC of ~0.65 on an independent validation dataset. The model identified a principal set of 20 time-invariant markers, including 18 genetic markers (16 single nucleotide polymorphisms [SNPs] and two HLA-DR genotypes) and two demographic markers (gestational age and exposure to a prebiotic formula). Integration with the metabolome identified 22 supplemental metabolites and lipids, including adipic acid and ceramide d42:0, that predicted development of islet autoantibodies.

Conclusions: The majority (86%) of metabolites that predicted development of islet autoantibodies belonged to three pathways: lipid oxidation, phospholipase A2 signaling, and pentose phosphate, suggesting that these metabolic processes may play a role in triggering islet autoimmunity.

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2020 The Authors. *Journal of Diabetes* published by Ruijin Hospital, Shanghai Jiaotong University School of Medicine and John Wiley & Sons Australia, Ltd.

KEYWORDS

autoimmunity, genetics, machine learning, metabolomics

Highlights

- A machine learning model can significantly predict imminent development of islet autoimmunity based on environmental, genetic, and metabolic features.
- The machine learning algorithm feature selection identified type 1 diabetes-associated single nucleotide polymorphisms from The Environmental Determinants of the Diabetes in the Young (TEDDY) analysis are correlated to related data from the Diabetes Autoimmunity Study in the Young (DAISY) study.
- Most of the metabolic features predicting the development of islet autoantibodies belonged to three pathways; lipid oxidation, phospholipase A2 signaling, and pentose phosphate.

1 | INTRODUCTION

The risk of type 1 diabetes (T1D) involves both genetic and nongenetic factors. Our understanding of the role of human leukocyte antigen (HLA) and other genes in development of islet autoimmunity and subsequent progression to T1D is continually expanding.^{1–3} Recent work has focused on understanding how these genetic factors interact with environmental factors and biomarkers of T1D risk.^{4–7} Better prediction of the risk of T1D incorporating multiple predictive factors to offer new strategies for early diagnosis and treatment is one of the core goals of birth cohort studies studying genetically susceptible children, such as The Environmental Determinants of Diabetes in the Young (TEDDY) and Diabetes Auto Immunity Study in the Young (DAISY). Herein we focus on prediction of imminent progression to the development of persistent islet autoantibodies to insulin (IAA), glutamic acid decarboxylase (GADA), or insulinoma antigen-2 (IA-2A) in TEDDY participants by integrating data such as metabolomics that are measured within 6 months prior to the diagnosis of persistent autoimmunity with associated risk factors, such as genetics and environment.

Identification of robust molecular markers from large and complex data has been noted as one of the major challenges of personalized medicine.^{8,9} One strategy is to utilize a knowledge-based approach, where known risk factors for a disease are combined in a machine learning framework to make individualized predictions. Alternatively, a data-driven approach can be taken that allows potential markers, such as metabolite characterization from high-throughput 'omic studies, to be incorporated into the model and further interrogated to identify an optimal subset of risk factors to make statistical-based predictions of interest. In recent studies we took the latter

approach to evaluate a small cohort from DAISY at various time points prior to the development of persistent autoantibodies Islet autoimmunity (IA) to both predict phases of development and also to identify the core features of importance.¹⁰ In this study, we expand on this approach to perform machine learning-based ensemble feature selection on 314 children who had metabolomic data across time in the TEDDY nested case-control study.¹¹ We utilize a probability-based machine learning integration strategy combined with an optimization-based feature selection process to identify a core set of predictive markers. We further examine the core features that drive the machine learning model and assess the mechanistic changes associated with these features. We present results in the context of cross-validation and an independent holdout set.

2 | METHODS

2.1 | Participants selection and data generation

The TEDDY study includes 8676 participants with increased T1D risk HLA-DR/DQ genotypes, recruited before the age of 4.5 months from four countries; United States, Germany, Sweden, and Finland.¹² The children were evaluated for the development of persistent islet autoantibodies every 3 months until either the development of T1D or 48 months. At this point for those with autoantibody seroconversion visits continue every 3 months, and for the rest, visits proceed at 6-month intervals. All study participants had written informed consent from a parent or primary caretaker. The TEDDY study was approved by the local institutional review

boards where the data were collected and is monitored by an external evaluation committee formed by the National Institutes of Health (NIH).

2.2 | Islet autoantibody measurements

IAA, GADA, or IA-2A were measured in two laboratories by radiobinding assays as previously described.^{13,14} In the United States, all sera were assayed at the Barbara Davis Center for Diabetes at the University of Colorado Denver; in Europe, all sera were assayed at the University of Bristol, United Kingdom. Both laboratories have previously shown high sensitivity and specificity¹⁴ as well as concordance. To optimize concordance, harmonized assays for GADA and IA-2A (12) replaced previous assays, in January 2010. Based on a receiver-operator curve (ROC) analysis, prior samples that needed to be reanalyzed with the harmonized assays, included Denver GADA between -0.015 and 0.042 ; Bristol GADA between 10.69 and 36.72 ; Denver IA-2A between -0.004 and 0.016 ; and Bristol IA-2A between 6.69 and 10.58 . All positive islet autoantibodies and 5% of negative samples were retested in the other reference laboratory and deemed confirmed if concordant.

2.3 | Outcomes

Persistent confirmed islet autoimmunity (IA) was defined as two consecutive visits positive for a specific islet autoantibody confirmed in a second laboratory. Date of IA was the draw date of the first sample of the two consecutive samples that deemed the child persistent confirmed positive for any islet autoantibody.

2.4 | Nested case-control study

Because of the large cohort size, it is cost and time prohibitive to perform many 'omics-based analyses, such as microbiome^{15,16} and metabolomics,¹⁷ for all the samples. The case-control study design can also help to reduce the batch effects associated with large-scale 'omics studies. The TEDDY Data Coordinating Center generated nested case-control pairs using a design that is based on the time point at which a child is positive for an event. The event is either the presentation of persistent islet autoantibodies or clinical diagnosis of T1D. Once a case is defined, an associated control is selected from all event-free participants at that same time point for the case. The best control for the case is selected based on matching factors of clinical center, sex, and family history of

T1D.^{11,18,19} In the analyses presented here, we included 157 cases who developed persistent islet autoantibodies and 157 matched controls, a 1:1 nested case-control design described in detail by Lee et al.¹¹

2.5 | Data sources

The predictive model developed here was based on three sources of data: (a) participant risk factors (RF) previously identified in the study population,^{6,11} (b) participant genetic risk (T1D-associated single nucleotide polymorphisms [SNPs]³), and (c) participant metabolomic risk (metabolites and lipids). The goal of this modeling effort is the prediction of progression to persistent autoimmunity, thus all data used is either independent to or collected prior to the observance of autoantibodies.

The first data source was a combination of participant risk factors, including genetic features, associated with T1D risk (RF-SNP). The patient risk factors were selected based on their availability in the case-control data and included (a) gestational age in weeks, (b) exposure to cow's milk formula, (c) exposure to prebiotic formula, (d) ethnicity/race (white, unknown, multiracial), and (e) HLA risk genotypes (DR3/4, DR4/4, DR4/8, DR4/1, DR4/13, DR3/3) previously described in.⁵ Except for gestational age, each of these were represented as binary variables and thus in the dataset is represented as 12 variables. The T1D-associated SNP data were generated using a custom genotyping array (Illumina ImmunoChip) containing 186 000 SNPs associated with autoimmune disease. The data were collected for all TEDDY individuals with genotyping conducted by the TEDDY Genetics Laboratory at the University of Virginia Center for Public Health Genomics. A total of 85 SNPs significantly associated with T1D were used in this analysis.^{3,20}

The second source of data was time-varying metabolite and lipid measurements (MET-LIP) from participants at time points prior to the development of persistent autoimmunity. Untargeted metabolomics and lipidomics data were generated for all cases and controls in the TEDDY nested case-control study. Primary metabolites and lipids were quantified from citrated plasma using gas chromatography-time-of-flight mass spectrometry (GC-TOF MS) and charged surface hybrid liquid chromatography coupled to quadrupole TOF MS (CSH-QTOF MS),²¹ respectively, at the NIH West Coast Metabolomics Center at the University of California, Davis, California. The GC-TOF MS metabolomics data acquisition followed previously described protocols²² followed by data processing and compound identification using the BinBase algorithm²³ and normalization using the sum approach. There were 156 identified metabolites quantified for

analysis. Metabolites with less than 10% missing values were processed with random forest imputation²⁴ and those with 10% or more missing data were removed. For complex lipids, samples were extracted by methyl-tert-butyl ether/methanol/water and analyzed using CSH-QTOF MS in both positive and negative electrospray ionization (ESI). Lipid chromatogram peak detection and alignment used Mass Profiler Professional (Agilent, Santa Clara, CA). Peaks detected in at least 30% of samples were identified and quantification back-filled using the Fiehn laboratory's LipidBlast spectral library.²⁵ Locally weighted scatter plot smoother (LOESS)-based normalization was corrected for batch effects by adjusting individual samples to intermittent quality control samples.²⁶ There were 652 lipids identified across both positive and negative ESI modes.

2.6 | Integrative machine learning

We built machine learning models to predict cases vs controls at three time horizons to the development autoantibodies: (a) 0 months (the time at which positive autoantibodies were first detected for cases), (b) 3 months prior, and (c) 6 months prior. Data were assembled by identifying the age of the case at the time of being classified as autoantibody positive and selecting the sampling time point for the control that matched the age of the case as close as possible. In addition, there are situations in which a child is identified as both a control and a case because of the risk set sampling used for the nested case control study design from the longitudinal study. These situations were removed from the dataset because they cause issues with the independence assumptions between subjects of the machine learning models. Most children are sampled approximately every 3 months and, thus, we used data that represent the case and control pair at the two prior sampling time points based on the age of the case at confirmation of autoimmunity. The point of seroconversion (0 months) is included to evaluate if the prediction of imminent progression can be achieved with similar accuracy once autoantibodies are detectable.

Of all case-control pairs available for the autoimmunity endpoint, there were 314 children (157 pairs) that contained the RF-SNP and MET-LIP data at all three time points. We segregated a validation set prior to machine learning consisting of 25% of the pairs (78 total children) with the remaining 236 children utilized as the training set. Descriptive statistics of the training and validation sets are shown in Supplemental Table S1. The 39 pairs in the validation set were selected at random but have similar overall characteristics as the training set (last column of Table S1).

The workflow of the analysis performed is shown in Figure 1. The first step separated the validation data from the training data to ensure that data quality and filtering is independent of the validation set and will not bias the downstream machine learning evaluation. Once the validation set was segregated, the data were preprocessed and the model was developed. As is common for machine learning,²⁷ the metabolites and lipids were preprocessed with a paired *t* test (comparing cases to controls) in the training set at each time point to reduce the dimensionality before machine learning. A conservative minimum significance threshold of $P = 0.1$ across the time points was selected; this yielded 131 markers consisting of 45 metabolites and 86 lipids. The metabolites and lipids in the validation set were reduced to match the training set but were not utilized in generating the statistics for the down-selection criteria.

Multiple machine learning algorithms, including logistic regression, K-nearest neighbors, linear discriminant analysis, Naïve Bayes classifier, random forest, and a support vector machine (linear kernel), were completed on the base level RF-SNP features and MET-LIP features of with five-fold cross-validation (CV) to evaluate which machine learning algorithm best modeled the underlying structure of the data. A random forest approach was selected for the MET-LIB data (131 metabolome features) and a Naïve Bayes classifier for the RF-SNP (90 SNP/environmental features). The models were merged as the product of the posterior probability from each machine learning algorithm to attain a single probabilistic score for each child.^{28,29} We validated that the merged model was equal to or superior to combining the two sources of data into a single Naïve Bayes or random forest model. At seroconversion the merged model returned significantly larger ROC area under the curve (AUC)s than either a single Naïve Bayes or random forest model based on 100 repetitions of fivefold CV at a paired *t* test *P*-value threshold of 0.05.

Repeated optimization for feature interpretation (ROFI) generated the features that optimize the ability to separate those that will transition to IA positivity in the three defined time windows.^{10,30} ROFI performs an optimization-based feature selection algorithm repeated 500 times and the importance of a feature is defined as the percentage of times it was selected for inclusion in the model over the 500 independent optimization runs based on five-fold CV. Given the paired nature of the nested case-control study, the CV process ensured that pairs were placed together in a training or a holdout set in the CV in order to reduce bias from potential pairwise correlation. Once the feature importance metrics were acquired, the values were sorted and the features to be included in the final model development were selected. The final model was generated on the full training data of 236 TEDDY children based on the features selected

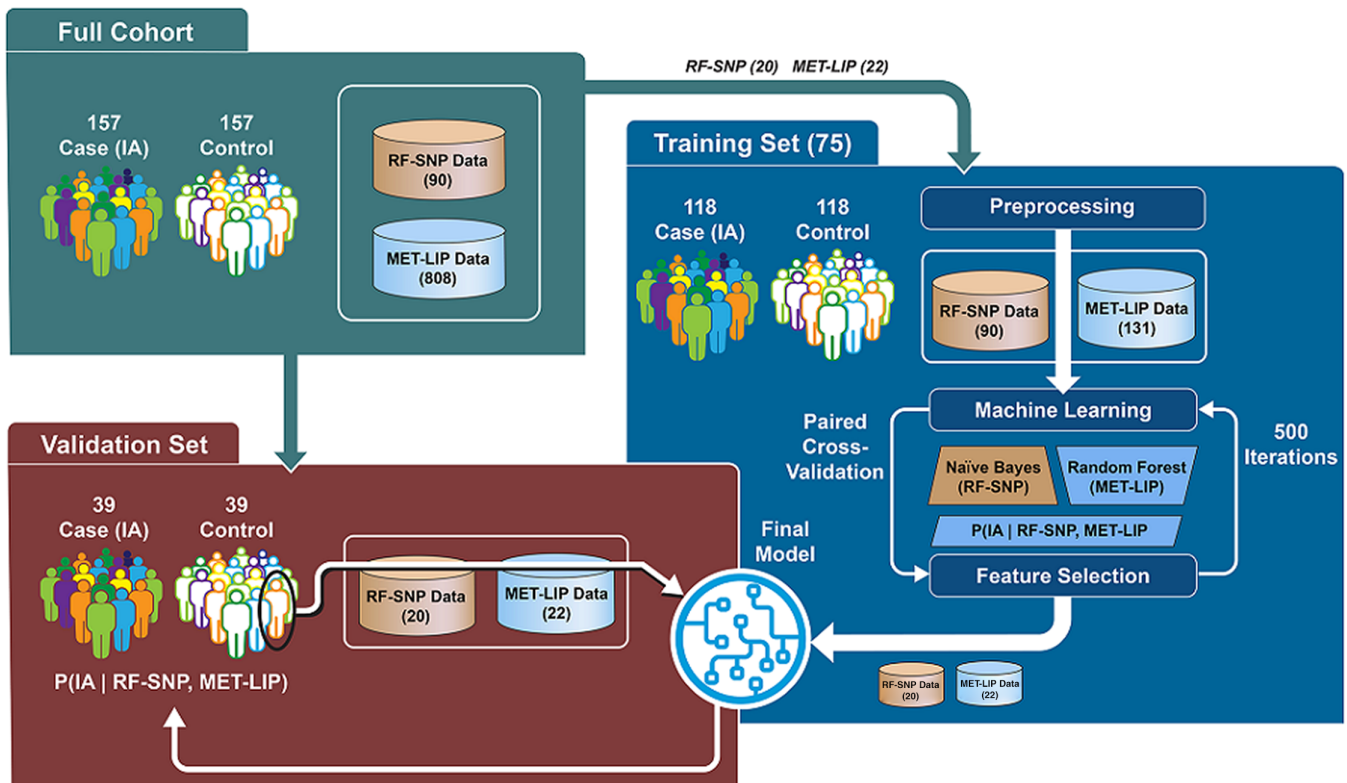


FIGURE 1 The Environmental Determinants of the Diabetes in the Young (TEDDY) data processing and machine learning workflow for a single time point first separated the data into a training and validation set. The training data is used to perform prefiltering and ensemble-based feature selection. The final repeated optimization for feature interpretation (ROFI) generated model was created from the full training set and applied to the validation set to benchmark performance. MET-LIP, metabolite and lipid measurements; RF-SNP, risk factors-single nucleotide polymorphisms

from ROFI; the machine learning model was applied to the validation set to generate the likelihood that each of the 78 children to develop persistent autoimmunity with the defined time frame, with the full process repeated for each of the three time points (Figure 1). The machine learning algorithms and feature selection method are available at <https://github.com/pmartR/peppuR>.

3 | RESULTS

The data here were derived from the nested case-control design of TEDDY and, therefore, features of family history, sex, clinical center, and age are not included in the model because they are utilized as matching criteria. The machine learning model generates the probability that a TEDDY child will transition from a control state to the development of persistent autoantibodies within the next x months ($x = 0, 3, 6$) based on the defined risk factors, genetic profile, and metabolome. The average age of the children that developed autoantibodies, “cases,” was ~ 2 years old (range from 0.72 to 4 years); thus,

predictions for children not in this age range may not be applicable.

3.1 | Feature selection and performance

We performed ROFI-based feature selection on 131 metabolome features, 85 T1D-associated SNPs features, in addition to five features representing HLA category, gestational age, cow’s milk formula exposure, prebiotic formula exposure, and ethnicity/race. The RF-SNP data was the same at each time point, but the metabolome is represented as three datasets of quantitative metabolite and lipid values at 0, -3 , and -6 m from seroconversion. The feature selection process identified 42 features as optimal utilizing an order-based statistic to define the selection threshold.³¹ As seen in Figure 2 there is a dramatic improvement for both the cross-validated training data (boxplots) and the associated accuracy of the validation set (point estimate) at all time points for the feature selection model vs using all 221 features. Of these 42 features 20 were associated

with the risk factors and genetic markers, and the remaining 22 were from the metabolome (11 metabolites and 11 lipids). Figure 3 displays the importance, defined as the percentage of solutions including the feature, of each these 42 candidates identified by ROFI. Figure 2 also demonstrates that the prediction of imminent development can roughly be predicted with the same accuracy 3 and 6 months prior to the time point when the autoantibodies are observed.

3.2 | Development of IA: Machine learning

Gestational age is one of potential risk factors for T1D reported previously.³² For the training set, the average gestational age of cases was 0.48 weeks longer than that

of the controls ($P \sim 0.008$) that decreases to a statistically insignificant difference of 0.09 weeks in the validation set ($P \sim 0.811$), Figure 4A. It is not clear whether the result in the validation set is due to low statistical power or random fluctuation in the training/validation process. Figures 4B-D show the other three important risk factors - prebiotic exposure, HLA-DR3/4, and HLA-DR3/3 genotypes - follow expected patterns^{4,33} and are more strongly correlated between the training and validation sets. The metabolite with the strong feature importance score (adipic acid) and one of the top lipids (PC[40:5]) separate patterns across time for the full cohort, training, and validation sets for the T1D case-control matched pairs (Figure 4E, Figure 4F). These markers have been associated with diabetes³⁴⁻³⁶ and patterns of increased abundance of these metabolites appear to predict the

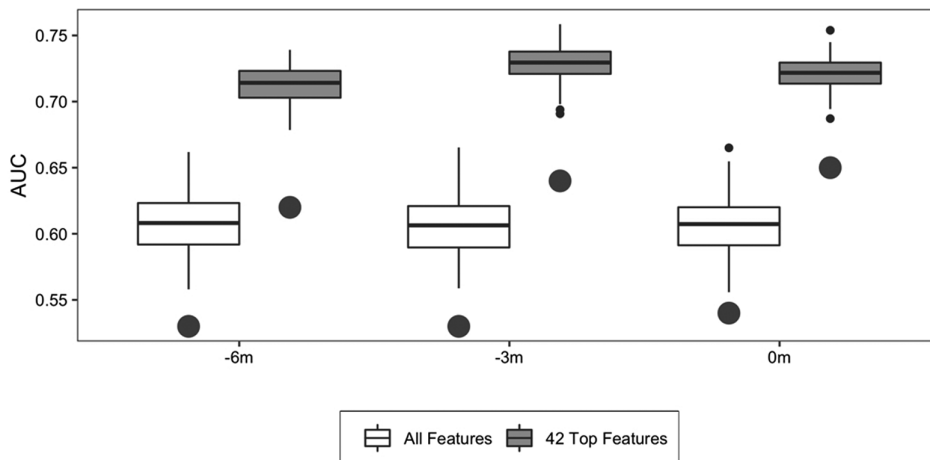


FIGURE 2 Overall accuracy and variability of the training data (boxplots) based on cross-validation and the associated accuracy of the model when applied to the validation set (large dots) for both all features and the 42 selected features. The boxes of the training data results represent the 25th and 75th percentiles and the line indicates the median accuracy with extreme values represented by the small dots. AUC, area under the curve

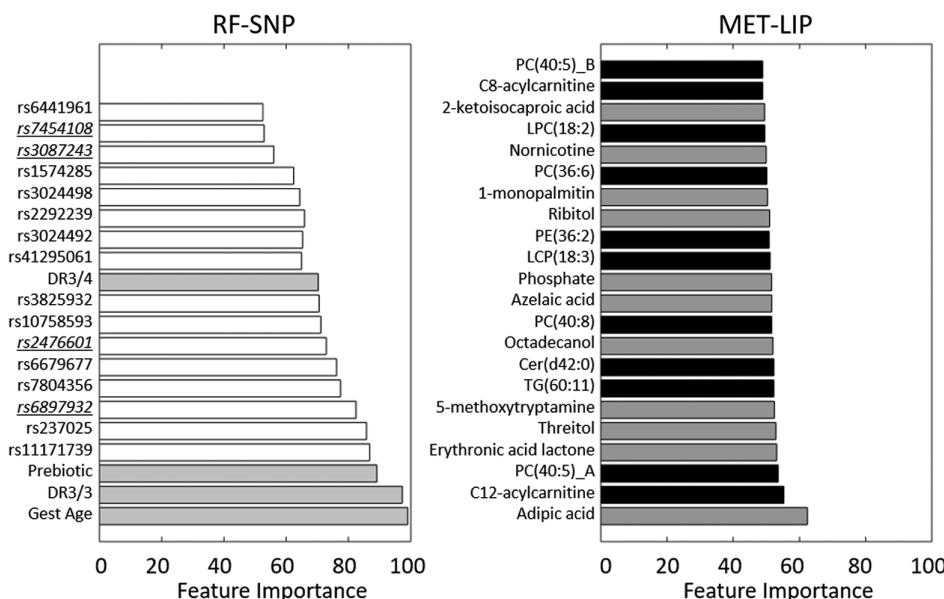


FIGURE 3 Bar graphs showing the feature importance of each of the selected 42 candidates where time-invariant markers are on the left with single nucleotide polymorphisms (SNPs) in white and other risk factors in gray and metabolomics data on the right where lipids are in black and metabolites in gray. The underlined SNPs are those highlighted in Figure 5. MET-LIP, metabolite and lipid measurements; RF-SNP, risk factors-single nucleotide polymorphisms

development of persistent autoantibodies in the TEDDY population.

There were 16 SNPs that were identified by the machine learning approach that were able to distinguish the cases from controls for the development of autoantibodies. Because DAISY is a distinct cohort with similar goals to TEDDY as well as having genetic data on a subset of participants, we evaluated whether the SNPs identified in this analysis from TEDDY have a similar pattern in DAISY. In particular, DAISY is a longitudinal, observational birth cohort study of 2547 high-risk children followed to development of autoimmunity and T1D.³⁷⁻³⁹ For DAISY, genetic information was collected on 25 children that either ended the study disease free or with confirmed persistent autoantibodies. There were four SNPs that overlapped the two studies (underlined in Figure 3), for which Figure 5 shows that the minor allele frequency (MAF) of the SNPs is similar between TEDDY

and DAISY (eg, rs246601 MAF in DAISY is 0.12 and MAF in TEDDY is 0.19). We estimated the correlation within each group (“case” or control) within each SNP and demonstrated that the control samples had an average Pearson correlation of 0.921 and the case samples of 0.802 (average 0.862), a confirmation that these core T1D-associated SNPs appear to be robustly associated with outcome across studies.

To investigate the mechanistic indicators of the feature selection, we evaluated the SNP-defined putative T1D target genes and metabolic pathways.⁴⁰ Twelve out of the 16 SNPs (75%) had putative target genes related to the immune system (Figure 6A). Metabolites were mainly grouped in three pathways: lipid oxidation, phospholipase signaling, and pentose phosphate. Triacylglycerol TG(60:11), 1-monopalmitin, C8-acylcarnitine, C12-acylcarnitine, adipic acid, and azelaic acid are metabolites of the lipid storage and oxidation pathway (Figure 6B). The degradation of triacylglycerols have

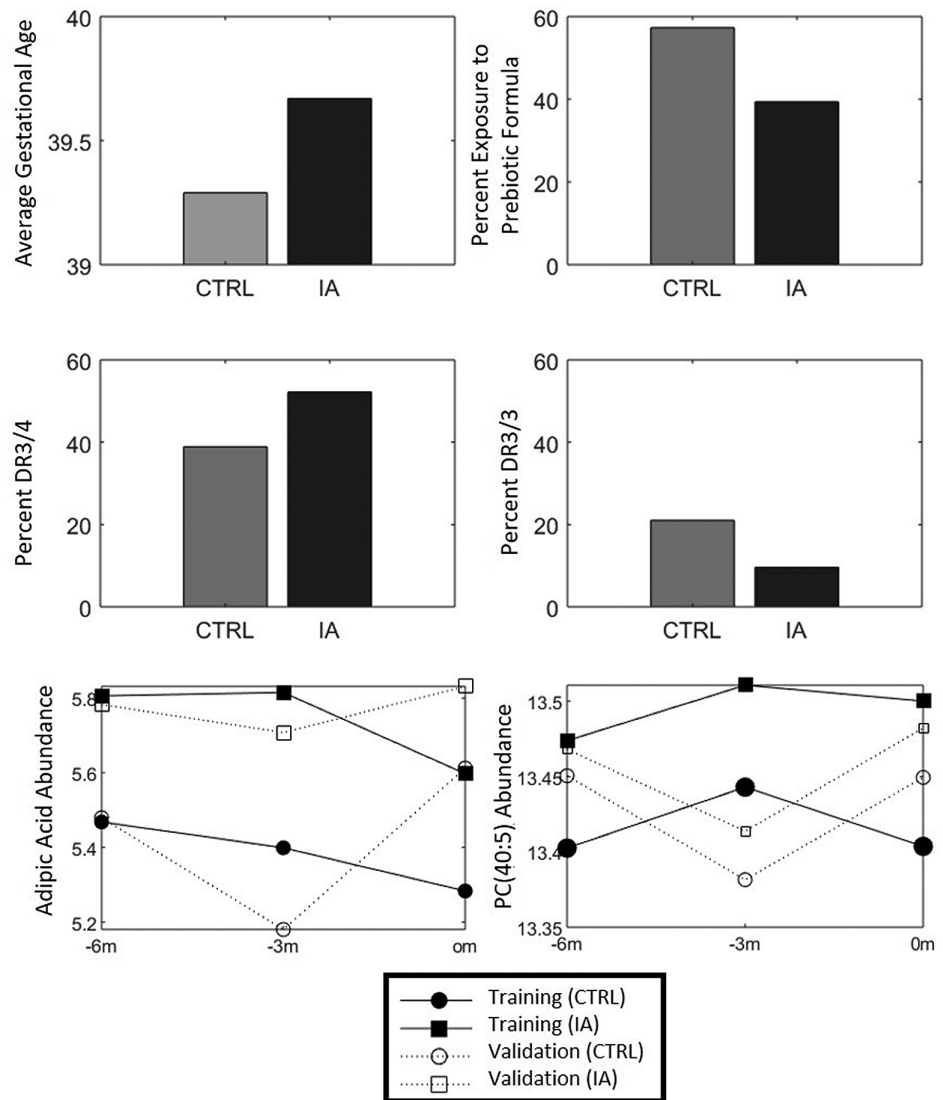


FIGURE 4 Data graphs showing the directional changes of (A) gestational age, (B) exposure to prebiotic formula, (C) DR3/4 and (D) DR3/3, as well as the temporal changes for (E) adipic acid and (F) phosphatidylcholine PC(40:5) for the IA and associated control samples (CTRL)

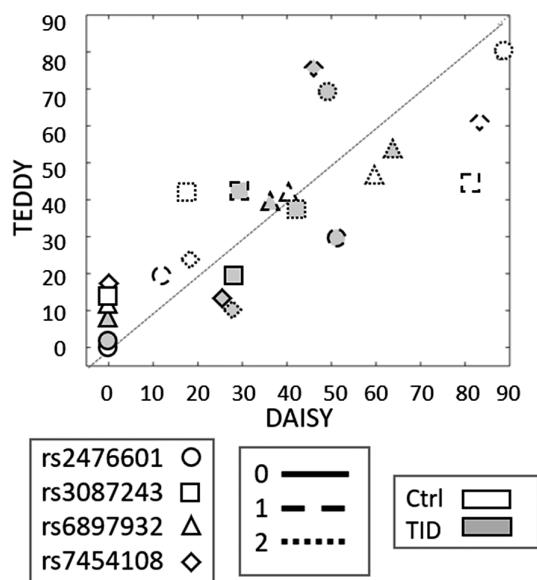


FIGURE 5 Scatter plot of percentage of children in each genotype for four of the top single nucleotide polymorphisms (SNPs) in terms of feature importance that overlap with SNPs currently being studied in the DAISY cohort. DAISY, Diabetes Auto Immunity Study in the Young; TEDDY, The Environmental Determinants of the Diabetes in the Young; T1D, type 1 diabetes

been shown to occur in inflammation, fueling the high energy demands required for this process.⁴¹ The second pathway with many metabolic features predictive of autoimmunity was the phospholipase A2 signaling pathway, including phosphatidylcholine PC(40:5), lysophosphatidylcholines LPC(18:2) and LPC(18:3), and ceramide cer(d42:0) (Figure 6C). It has been reported that a phospholipase A2 from human islets is activated by cytokines and endoplasmic reticulum (ER) stress, leading to beta-cell death.³⁵ Phospholipase activation leads to the degradation of phosphatidylcholines (PC) into lysophosphatidylcholines (LPC) and free fatty acids, which in turn activates a neutral sphingomyelinase that cleaves sphingomyelins (SM) into phosphocholines and ceramides (Cer(d42:0)) (Figure 6C). The accumulation of ceramides triggers the apoptotic cascade resulting in the death of beta cells.^{34,35} The last pathway containing several metabolic features predictive of autoimmunity is the pentose metabolism (Figure 6D). Recently, the pentose phosphate pathway was shown to be regulated in peripheral blood mononuclear cells in children IA.⁴² The pentose phosphate pathway is usually repressed in immune cells to prevent damage by toxic reactive oxygen species, but it is upregulated during autoimmune responses to supply the high metabolic demands of active leukocytes.⁴³ Overall, the identified metabolic features reflect processes that are regulated during an autoimmune response.

4 | DISCUSSION

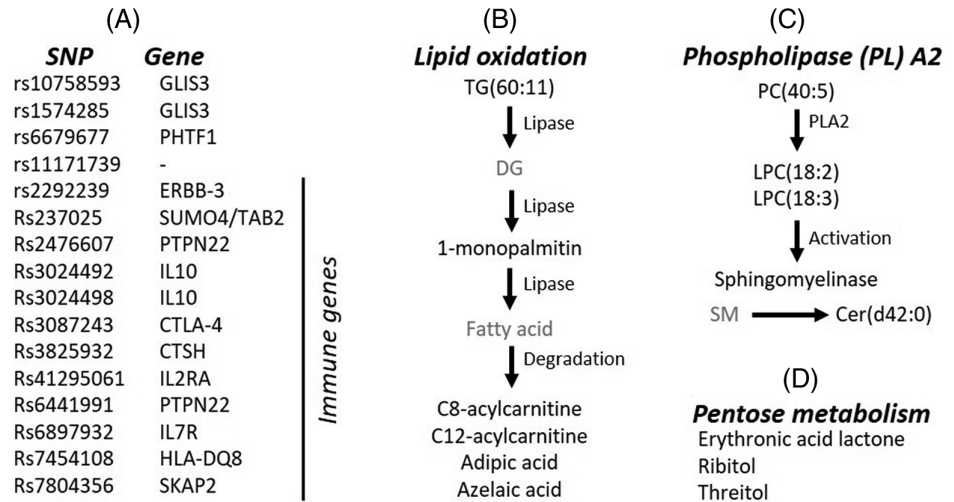
To date, the most common approaches to predicting outcomes of diabetes have focused on regression or small variable subsets selected via expert knowledge. For example, one recent study determined the likelihood that a child will progress to T1D by the age of 6 years if they have presented with persistent autoantibodies at the age of 3 years.²⁷ The logistic regression model to make this prediction utilized five predictors that were selected based on known associations, such as IA-2 antibody positivity status and hemoglobin A1c (HbA1c) level, achieving a sensitivity of 0.91 at a specificity of 0.59, yielding a ROC curve with an AUC of 0.80. The conclusion of this work was that continued developments of such models are necessary to better understand the complexity of the disease and address long-term goals of precision diabetes.^{44,45} Biomarkers hold vast potential for clinical utility both in terms of diagnosis and prognosis of disease but also in drug discovery.^{46,47}

An alternate strategy to regression modeling is to take a data-driven approach to allow a large collections of potential risk factors and molecular markers to be integrated into the predictive modeling. These machine learning models are amenable to the extraction of features that can provide improved prediction.^{30,48,49} Frohnert et al used a machine learning approach to mine large multi-omic predictors (genomics, proteomics, metabolomics, and demographics) of seroconversion and T1D and returned high accuracy, an AUC of 0.91 for the prediction of seroconversion based on cross-validation of 25 case and control subjects.¹⁰ In this study we also undertook the evaluation of seroconversion with several core differences. The population from TEDDY is a much larger and more heterogeneous. The point of seroconversion is more tightly controlled to approximately 3 and 6 months prior. Finally, the size of TEDDY allowed for independent training and validation data to evaluate the robustness of the machine learning model. One caveat to this current analysis is the predefined state of cases and controls based on matching criteria exclude potentially important factors, such as gender, clinical site, and family history, from being included in the prediction model.

In this report, we identified 42 feature candidates (SNPs, traditional risk factors, metabolites, and lipids) that, in combination, predict development of autoimmunity in increased genetic risk TEDDY participants. When interrogated, these features are associated with three biological pathways: lipid oxidation, phospholipase A2 signaling, and pentose phosphate pathway - suggesting that these processes might serve as key predictive processes during development of IA. These pathways reflect processes that are regulated during an autoimmune

FIGURE 6 Overview of the selected features and functions.

(A) List of single nucleotide polymorphisms (SNPs) and their respective genes. (B-D) Features belonging to the lipid oxidation pathway (B), phospholipase A2 (C) and pentose metabolism (D). The metabolites in gray were not selected as features but were added to the figures merely to complete the pathways



response. These markers were identified via a data-driven predictive model of imminent development of IA, evaluating time-invariant risk factors in combination with time-varying metabolic features. Models such as these may lead the field closer to the goals of precision medicine and improved understanding of the underlying biological mechanisms driving T1D.²⁷ Improved understanding of the interactions between genetic factors and diet or metabolism, on the development of autoimmunity could inform new interventions to prevent or delay the onset of T1D.

ACKNOWLEDGEMENTS

We thank the TEDDY study for providing data and all TEDDY children and their families for their continued participation in this study. We also like to thank PNNL graphic designer Michael Perkins for assistance in preparing Figures. B.M.W is the guarantor of this work and, as such, had full access to all the data in the study and takes responsibility for the integrity of the data and accuracy of the data analysis.

This research was performed by an external analytic partner under the auspices of the TEDDY Study Group, which is funded by U01 DK63829, U01 DK63861, U01 DK63821, U01 DK63865, U01 DK63863, U01 DK63836, U01 DK63790, UC4 DK63829, UC4 DK63861, UC4 DK63821, UC4 DK63865, UC4 DK63863, UC4 DK63836, UC4 DK95300, and UC4 DK100238 and by Contract no. HHSN267200700014C from the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK), National Institute of Allergy and Infectious Diseases (NIAID), National Institute of Child Health and Human Development (NICHD), National Institute of Environmental Health Sciences (NIEHS), Centers for Disease Control and Prevention (CDC), and JDRF. TEDDY is supported in part by the National Institutes of Health/

National Center for Advancing Translational Sciences Clinical and Translational Science Awards UL1 TR000064 (University of Florida) and the University of Colorado (UL1 TR001082), and TEDDY grant UC4 DK100238. Machine learning work was performed at the Pacific Northwest National Laboratory; Battelle operates PNNL for the DOE under contract DE-AC05-76RLO01830.

DISCLOSURE


All authors disclose no conflicts of interest.

ORCID

Bobbie-Jo M. Webb-Robertson  <https://orcid.org/0000-0002-4744-2397>

Lisa M. Bramer  <https://orcid.org/0000-0002-8384-1926>

Bryan A. Stanfill  <https://orcid.org/0000-0003-0612-5333>

Ernesto S. Nakayasu  <https://orcid.org/0000-0002-4056-2695>


Thomas O. Metz  <https://orcid.org/0000-0001-6049-3968>

Brigitte I. Frohnert  <https://orcid.org/0000-0002-6636-4048>

Jill M. Norris  <https://orcid.org/0000-0001-8674-2598>

Randi K. Johnson  <https://orcid.org/0000-0001-9345-4439>

Stephen S. Rich  <https://orcid.org/0000-0003-3872-7793>

Marian J. Rewers  <https://orcid.org/0000-0003-3829-9207>

REFERENCES

- Regnell SE, Lernmark A. Early prediction of autoimmune (type 1) diabetes. *Diabetologia*. 2017;60(8):1370-1381.
- Rich SS, Concannon P. Role of type 1 diabetes-associated SNPs on autoantibody positivity in the type 1 diabetes genetics consortium: overview. *Diabetes Care*. 2015;38(Suppl 2):S1-S3.

3. Torn C, Hadley D, Lee HS, et al. Role of type 1 diabetes-associated SNPs on risk of autoantibody positivity in the TEDDY study. *Diabetes*. 2015;64(5):1818-1829.
4. Frederiksen BN, Kroehl M, Fingerlin TE, et al. Association between vitamin D metabolism gene polymorphisms and risk of islet autoimmunity and progression to type 1 diabetes: the diabetes autoimmunity study in the young (DAISY). *J Clin Endocrinol Metab*. 2013;98(11):E1845-E1851.
5. Krischer JP, Lynch KF, Lernmark A, et al. Genetic and environmental interactions modify the risk of diabetes-related autoimmunity by 6 years of age: the TEDDY study. *Diabetes Care*. 2017;40(9):1194-1202.
6. Uusitalo U, Lee HS, Andren Aronsson C, et al. Early infant diet and islet autoimmunity in the TEDDY study. *Diabetes Care*. 2018;41(3):522-530.
7. Mishra SP, Wang S, Nagpal R, et al. Probiotics and prebiotics for the amelioration of type 1 diabetes: present and future perspectives. *Microorganisms*. 2019;7(3):67.
8. Lin Y, Qian F, Shen L, Chen F, Chen J, Shen B. Computer-aided biomarker discovery for precision medicine: data resources, models and applications. *Brief Bioinform*. 2017;20(3):952-975.
9. Vafae F, Diakos C, Kirschner MB, et al. A data-driven, knowledge-based approach to biomarker discovery: application to circulating microRNA markers of colorectal cancer prognosis. *NPJ Syst Biol Appl*. 2018;4:20.
10. Frohnert BI, Webb-Robertson BJ, Bramer LM, et al. Predictive modeling of type 1 diabetes stages using disparate data sources. *Diabetes*. 2020;69(2):238-248.
11. Lee HS, Burkhardt BR, McLeod W, et al. Biomarker discovery study design for type 1 diabetes in the environmental determinants of diabetes in the young (TEDDY) study. *Diabetes Metab Res Rev*. 2014;30(5):424-434.
12. Rewers M, Hyoty H, Lernmark A, et al. The environmental determinants of diabetes in the young (TEDDY) study: 2018 update. *Curr Diab Rep*. 2018;18(12):136.
13. Babaya N, Yu L, Miao D, et al. Comparison of insulin autoantibody: polyethylene glycol and micro-IAA 1-day and 7-day assays. *Diabetes Metab Res Rev*. 2009;25(7):665-670.
14. Bonifacio E, Beyerlein A, Hippich M, et al. Genetic scores to stratify risk of developing multiple islet autoantibodies and type 1 diabetes: a prospective study in children. *PLoS Med*. 2018;15(4):e1002548.
15. Stewart CJ, Ajami NJ, O'Brien JL, et al. Temporal development of the gut microbiome in early childhood from the TEDDY study. *Nature*. 2018;562(7728):583-588.
16. Vatanen T, Franzosa EA, Schwager R, et al. The human gut microbiome in early-onset type 1 diabetes from the TEDDY study. *Nature*. 2018;562(7728):589-594.
17. Fan S, Kind T, Cajka T, et al. Systematic error removal using random Forest for normalizing large-scale untargeted Lipidomics data. *Anal Chem*. 2019;91(5):3590-3596.
18. Lee HS, Lynch KF, Krischer JP, Group, T. S. Nested case-control data analysis using weighted conditional logistic regression in the environmental determinants of diabetes in the young (TEDDY) study: a novel approach. *Diabetes Metab Res Rev*. 2020;36(1):e3204.
19. Stanfill B, Reehl S, Bramer L, et al. Extending classification algorithms to case-control studies. *Biomed Eng Comput Biol*. 2019;10:1179597219858954.
20. Cariaso M, Lennon G. SNPedia: a wiki supporting personal genome annotation, interpretation and analysis. *Nucleic Acids Res*. 2012;40(Database issue):D1308-D1312.
21. Cajka T, Smilowitz JT, Fiehn O. Validating quantitative untargeted Lipidomics across nine liquid chromatography-high-resolution mass spectrometry platforms. *Anal Chem*. 2017;89(22):12360-12368.
22. Fiehn O. Metabolomics by gas chromatography-mass spectrometry: combined targeted and untargeted profiling. *Curr Protoc Mol Biol*. 2016;114:30.4.1-30.4.32.
23. Kind T, Wohlgemuth G, Lee DY, et al. FiehnLib: mass spectral and retention index libraries for metabolomics based on quadrupole and time-of-flight gas chromatography/mass spectrometry. *Anal Chem*. 2009;81(24):10038-10048.
24. Stekhoven DJ, Buhlmann P. MissForest-non-parametric missing value imputation for mixed-type data. *Bioinformatics*. 2012;28(1):112-118.
25. Kind T, Liu KH, Lee DY, DeFelice B, Meissen JK, Fiehn O. LipidBlast in silico tandem mass spectrometry database for lipid identification. *Nat Methods*. 2013;10(8):755-758.
26. Dunn WB, Broadhurst D, Begley P, et al. Procedures for large-scale metabolic profiling of serum and plasma using gas chromatography and liquid chromatography coupled to mass spectrometry. *Nat Protoc*. 2011;6(7):1060-1083.
27. Jacobsen LM, Larsson HE, Tamura RN, et al. Predicting progression to type 1 diabetes from ages 3 to 6 in islet autoantibody positive TEDDY children. *Pediatr Diabetes*. 2019;20:263-270.
28. Beagley N, Stratton KG, Webb-Robertson BJM. VIBE 2.0: visual integration for Bayesian evaluation. *Bioinformatics*. 2010;26(2):280-282.
29. Webb-Robertson BJ, McCue LA, Beagley N, et al. A Bayesian integration model of high-throughput proteomics and metabolomics data for improved early detection of microbial infections. *Pac Symp Biocomput*. 2009;451-463. <https://pubmed.ncbi.nlm.nih.gov/19209722/>.
30. Webb-Robertson, B.-J. M.; Bramer, L. M.; Reehl, S. M.; Metz, T. O.; Zhang, Q.; Rewers, M.; Frohnert, B. I., *ROFI - The Use of Repeated Optimization for Feature Interpretation*. 2016 *International Conference on Computational Science and Computational Intelligence (CSCI) 2016*, 29–33.
31. Zhu M, Chipman HA. Darwinian evolution in parallel universes: a parallel genetic algorithm for variable selection. *Dent Tech*. 2006;48(4):491-502.
32. Khashan AS, Kenny LC, Lundholm C, et al. Gestational age and birth weight and the risk of childhood type 1 diabetes: a population-based cohort and sibling design study. *Diabetes Care*. 2015;38(12):2308-2315.
33. Krischer JP, Liu X, Lernmark A, et al. The influence of type 1 diabetes genetic susceptibility regions, age, sex, and family history on the progression from multiple autoantibodies to type 1 diabetes: a TEDDY study report. *Diabetes*. 2017;66(12):3122-3129.
34. Lei X, Bone RN, Ali T, et al. Evidence of contribution of iPLA2beta-mediated events during islet beta-cell apoptosis due to proinflammatory cytokines suggests a role for iPLA2beta in T1D development. *Endocrinology*. 2014;155(9):3352-3364.
35. Lei X, Zhang S, Bohrer A, Barbour SE, Ramanadham S. Role of calcium-independent phospholipase a(2)beta in human

- pancreatic islet beta-cell apoptosis. *Am J Physiol Endocrinol Metab.* 2012;303(11):E1386-E1395.
36. Wang TJ, Ngo D, Psychogios N, et al. 2-Aminoadipic acid is a biomarker for diabetes risk. *J Clin Invest.* 2013;123(10):4309-4317.
37. Frederiksen B, Kroehl M, Lamb MM, et al. Infant exposures and development of type 1 diabetes mellitus: the diabetes autoimmunity study in the young (DAISY). *JAMA Pediatr.* 2013;167(9):808-815.
38. Frohnert BI, Ide L, Dong F, et al. Late-onset islet autoimmunity in childhood: the diabetes autoimmunity study in the young (DAISY). *Diabetologia.* 2017;60(6):998-1006.
39. Rewers M, Bugawan TL, Norris JM, et al. Newborn screening for HLA markers associated with IDDM: diabetes autoimmunity study in the young (DAISY). *Diabetologia.* 1996;39(7):807-812.
40. Huang DW, Sherman BT, Tan Q, et al. The DAVID gene functional classification tool: a novel biological module-centric algorithm to functionally analyze large gene lists. *Genome Biol.* 2007;8(9):R183.
41. McGillicuddy FC, Chiquoine EH, Hinkle CC, et al. Interferon gamma attenuates insulin signaling, lipid storage, and differentiation in human adipocytes via activation of the JAK/STAT pathway. *J Biol Chem.* 2009;284(46):31936-31944.
42. Laimighofer M, Lickert R, Fuerst R, et al. Common patterns of gene regulation associated with cesarean section and the development of islet autoimmunity - indications of immune cell activation. *Sci Rep.* 2019;9(1):6250.
43. Muschen M. Metabolic gatekeepers to safeguard against autoimmunity and oncogenic B cell transformation. *Nat Rev Immunol.* 2019;19(5):337-348.
44. Marshall SM. Precision diabetes: a realistic outlook on a promising approach. *Diabetologia.* 2017;60(5):766-768.
45. Mohan V, Radha V. Precision diabetes is slowly becoming a reality. *Med Princ Pract.* 2019;28:1-9.
46. Sanhueza C, Kohli M. Clinical and novel biomarkers in the Management of Prostate Cancer. *Curr Treat Options Oncol.* 2018;19(2):8.
47. Wishart DS. Emerging applications of metabolomics in drug discovery and precision medicine. *Nat Rev Drug Discov.* 2016;15(7):473-484.
48. Bahado-Singh RO, Yilmaz A, Bisgin H, et al. Artificial intelligence and the analysis of multi-platform metabolomics data for the detection of intrauterine growth restriction. *PLoS One.* 2019;14(4):e0214121.
49. Chen Z, Zhao P, Li F, et al. iLearn: an integrated platform and meta-learner for feature engineering, machine-learning analysis and modeling of DNA, RNA and protein sequence data. *Brief Bioinform.* 2020;21(3):1047-1057.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.

How to cite this article: Webb-Robertson B-JM, Bramer LM, Stanfill BA, et al. Prediction of the development of islet autoantibodies through integration of environmental, genetic, and metabolic markers. *Journal of Diabetes.* 2021;13:143-153. <https://doi.org/10.1111/1753-0407.13093>