Journal of the American Medical Informatics Association, 24(6), 2017, 1116–1126
doi: 10.1093/jamia/ocx069
Advance Access Publication Date: 22 July 2017
Research and Applications

Research and Applications

A genome-by-environment interaction classifier for precision medicine: personal transcriptome response to rhinovirus identifies children prone to asthma exacerbations

Vincent Gardeux,[1,2,3,†] Joanne Berghout,[1,2,3,†] Ikbel Achour,[1,2,3,†] A Grant Schissler,[1,2,3,4,†] Qike Li,[1,2,3,4] Colleen Kenost,[3] Jianrong Li,[3] Yuan Shang,[1,2,3,5] Anthony Bosco,[6] Donald Saner,[1,3,11] Marilyn J Halonen,[8] Daniel J Jackson,[9] Haiquan Li,[1,3] Fernando D Martinez,[2,7] and Yves A Lussier[1,2,3,10]

[1]Department of Medicine, University of Arizona, Tucson, AZ, USA, [2]BIO5 Institute, University of Arizona, Tucson, AZ, USA, [3]Center for Biomedical Informatics and Biostatistics, University of Arizona, Tucson, AZ, USA, [4]Interdisciplinary Program in Statistics, University of Arizona, Tucson, AZ, USA, [5]Center for Innovation in Brain Science, University of Arizona, Tucson, AZ, USA, [6]Telethon Institute for Child Health Research, Perth, Australia, [7]Department of Pediatrics, University of Arizona, Tucson, AZ, USA, [8]Department of Pharmacology, University of Arizona, Tucson, AZ, USA, [9]Department of Pediatrics, School of Medicine and Public Health, University of Wisconsin, WI, USA, [10]UA Cancer Center, University of Arizona, Tucson, AZ, USA and [11]Banner Health, Phoenix, AZ, USA

[†]These authors contributed equally to the work.

Corresponding Authors: Dr Yves A Lussier and Dr Fernando D Martinez, BIO5 Institute, University of Arizona, 1657 E Helen Street, 251, PO Box 210240, Tucson, AZ 85721, USA. E-mail: yves@email.arizona.edu; fdmartin@email.arizona.edu. Dr Haiquan Li, Department of Medicine, University of Arizona, 1657 E Helen Street, PO Box 210240, Tucson, AZ 85721, USA. E-mail: haiquan@email.arizona.edu.

Received 18 July 2016; Revised 1 May 2017; Accepted 29 June 2017

ABSTRACT

Objective: To introduce a disease prognosis framework enabled by a robust classification scheme derived from patient-specific transcriptomic response to stimulation.

Materials and Methods: Within an illustrative case study to predict asthma exacerbation, we designed a stimulation assay that reveals individualized transcriptomic response to human rhinovirus. Gene expression from peripheral blood mononuclear cells was quantified from 23 pediatric asthmatic patients and stimulated in vitro with human rhinovirus. Responses were obtained via the single-subject gene set testing methodology "N-of-1-pathways." The classifier was trained on a related independent training dataset ($n = 19$). Novel visualizations of personal transcriptomic responses are provided.

Results: Of the 23 pediatric asthmatic patients, 12 experienced recurrent exacerbations. Our classifier, using individualized responses and trained on an independent dataset, obtained 74% accuracy (area under the receiver operating curve of 71%; 2-sided $P = .039$). Conventional classifiers using messenger RNA (mRNA) expression within the viral-exposed samples were unsuccessful (all patients predicted to have recurrent exacerbations; accuracy of 52%).

Discussion: Prognosis based on single time point, static mRNA expression alone neglects the importance of dynamic genome-by-environment interplay in phenotypic presentation. Individualized transcriptomic response quantified at the pathway (gene sets) level reveals interpretable signals related to clinical outcomes.

Conclusion: The proposed framework provides an innovative approach to precision medicine. We show that quantifying personal pathway–level transcriptomic response to a disease-relevant environmental challenge predicts disease progression. This genome-by-environment interaction assay offers a noninvasive opportunity to translate omics data to clinical practice by improving the ability to predict disease exacerbation and increasing the potential to produce more effective treatment decisions.

© The Author 2017. Published by Oxford University Press on behalf of the American Medical Informatics Association.
This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs licence (http://creativecommons.org/licenses/by-nc-nd/4.0/), which permits non-commercial reproduction and distribution of the work, in any medium, provided the original work is not altered or transformed in any way, and that the work properly cited. For commercial re-use, please contact journals.permissions@oup.com

Downloaded from https://academic.oup.com/jamia/article-abstract/24/6/1116/4004728/A-genome-by-environment-interaction-classifier-for
by University of Liverpool user
on 19 October 2017

Key words: precision medicine, gene-by-environment, asthma, prognostic, genomic classifier, personal transcriptome, dynamic expression, virogram, HRV stimulation, PBMC stimulated, pathways

BACKGROUND AND SIGNIFICANCE

The advent of increasingly rapid and cost-effective genome-wide measures (omics methods)[1,2] is transforming our understanding of human pathophysiology and has stimulated numerous initiatives in precision medicine.[3] In particular, our understanding of single-gene disorders is advancing rapidly. However, genes alone do not determine the onset or progression of complex diseases – environmental factors, including lifestyle and specific stimulus exposure, also contribute substantially.[4] Already, a number of stimulation-based bioassays have been developed for diagnosis and treatment in clinical medicine. These include in vivo bronchial challenge tests (eg, inhaled methacholine challenge for asthma) and in vitro antibiotic-sensitivity bacteriograms, though available tests rely on available pathognomonic (diagnostic) biomarkers and have not yet leveraged the power of big-data genomics. Tissue-stimulation assays coupled with omics measurements provide a minimally invasive opportunity to understand an individual's personal genome-by-environment response to a stimulus relevant to disease state or risk of progression. While cohort statistics can identify the most common stimulus-provoked molecular responses between clinical populations,[5] there are no unbiased tools for assessing a whole transcriptome response that can be scaled down to a single patient without requiring inference from reference sets based on well-powered, cross-patient comparisons. Tailoring patient treatment according to the results of a given in vivo or in vitro stimulation thus remains an unmet challenge.[6] Nonetheless, diseases with complex inheritance that are precipitated by acute environmental challenge are conspicuous candidates for assay development to provide clinical insight into using this strategy. Focusing on dynamic gene expression changes within a subject additionally gets to the heart of the matter more effectively than conventional static classifiers derived from population-averaged expression.

We have used an approach that assesses changes to transcript expression between 2 paired samples from a single subject in response to stimulus exposure. This is followed by analyses of biological process involvement through gene set enrichment. Expression changes are determined using a suite of single subject–oriented N-of-1-pathways analytical tools uniquely designed for paired data analysis without use of a reference set.[7–11] In the accompanying experimental assay focused on asthma, live peripheral blood mononuclear cells (PBMCs) were drawn from patients, aliquoted, and exposed in vitro to either human rhinovirus (HRV) stimulus (stimulated) or vehicle control (unstimulated). We have called the provoked transcriptional response a *viral regulome* specific to a personal genome, or a "virogram."[9]

Epidemiological studies have identified viral infection as the most prominent environmental risk factor for severe asthma exacerbation episodes at all ages, with HRV infection (aka the common cold) specifically contributing to 48%–65% of pediatric cases.[12–16,17] These severe asthmatic exacerbations are a major cause of morbidity, require acute hospitalizations, and in rare cases can lead to death.[12] Though the role of viral infection is well established, not all children with respiratory infections become asthmatic or experience significant exacerbations later in life. This supports an additional significant role for non-Mendelian genetic predisposition interacting with viral exposure.[18] Identifying individuals at significantly increased risk for exacerbation before a clinically severe episode occurs would provide benefit at the individual case management and larger population health levels.[19]

We hypothesize that clinically useful pathway classifiers can be derived from dissecting the interaction of an individual patient's aggregate genomic risk with relevant environmental or laboratory assay exposure (genome × environment classifier; G × E classifier). In this study, we designed a robust *dynamic* transcriptional signature that can identify asthmatic patients at risk of exacerbation with a single blood test.

MATERIALS AND METHODS

The framework of our study is shown in Figure 1. We used a published microarray dataset measuring mRNA expression from an in vivo HRV exposure study[20] in non-asthmatic volunteers (training set; see Table 1). These data were used to build classifiers based on differentiating responses associated with symptomatic and asymptomatic infection. The G × E single-subject classifier built from our analysis of individualized within-subject responses was compared to cohort-based classifiers constructed using more standard methods. In our validation set (see Table 1),[9] PBMCs were isolated from asthmatic patients and stimulated in vitro with the same virus to provoke transcriptional changes. RNA was harvested and transcript expression quantified by microarray from paired unstimulated and stimulated cells. Single subject–derived and cohort-derived viral response classifiers were then explored for their ability to predict exacerbation episodes in the asthmatic validation set.

Gene set analysis for classifier construction was performed using Gene Ontology (GO)[21] Biological Process annotations.[22,23] To provide additional clinical translational value, we also introduced a novel visualization tool using star plots to help physicians make treatment decisions from the mechanism-anchored metrics revealed for each patient.

Description of training set to derive responsive features for development of an asthmatic exacerbation classifier

For our training set, we downloaded published microarray data from the Gene Expression Omnibus (GEO; GSE17156)[20] and Table 1. These data were derived from PBMC samples collected from 19 healthy, non-asthmatic volunteers exposed in vivo to HRV via intranasal inoculation. Each volunteer donated 2 blood samples, the first drawn at baseline prior to challenge, the second drawn 48 h after viral exposure. Of these 19 volunteers, 10 individuals developed symptoms of acute respiratory infection after exposure to HRV (symptomatic), while the other 9 had no symptoms (asymptomatic). Note: though the original study included 20 subjects, only 19 pairs of matched mRNA expression datasets were deposited in GEO.

Asthmatic subjects accrued for validation set and clinical phenotype determination

All human studies were carried out according to protocols approved by the Institutional Review Boards of the University of Arizona and the University of Wisconsin (IRB 0600000587 for the TREXA clinical trial and IRB06035604 for the BADGER clinical trial). Twenty-three pediatric asthma patients who previously participated in the

Downloaded from https://academic.oup.com/jamia/article-abstract/24/6/1116/4004728/A-genome-by-environment-interaction-classifier-for
by University of Liverpool user
on 19 October 2017

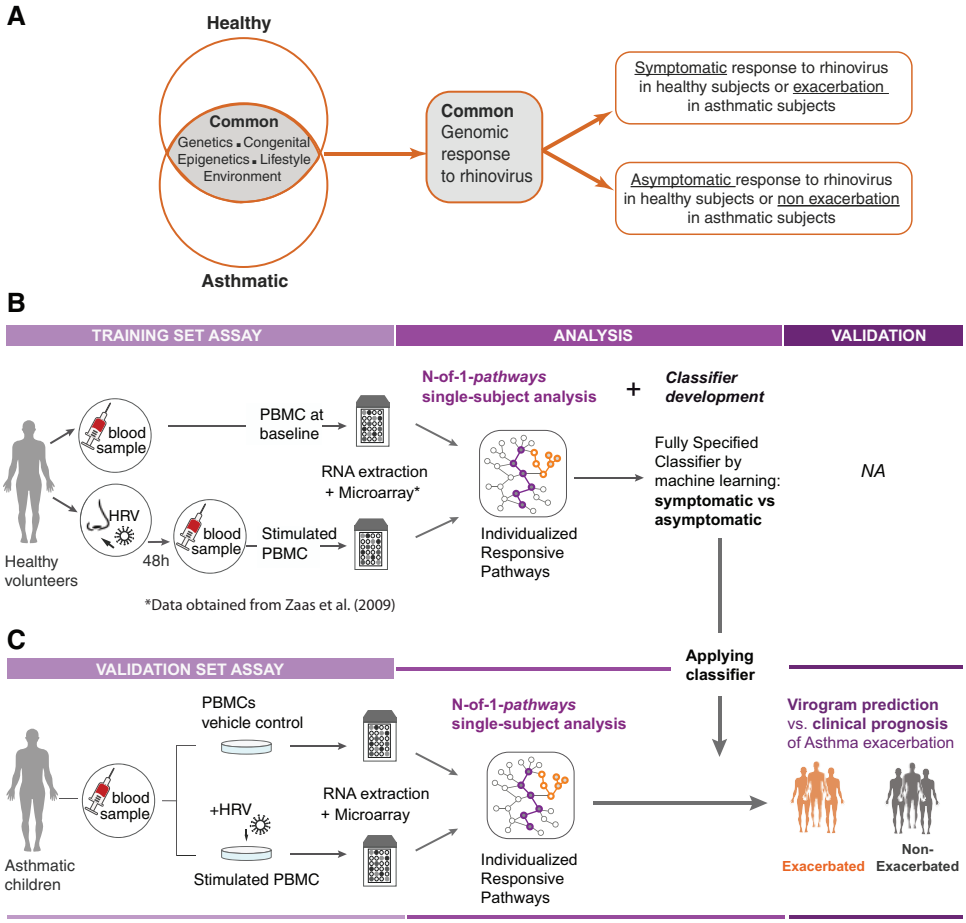


Figure 1. Overview of the experimental and analytic design of the study. We hypothesize that a classifier based on the pathway-level transcriptional responses that differ between symptomatic and asymptomatic responses to HRV infection in healthy patients can predict which asthmatic patients will have exacerbations over a 1-year follow-up period, based on those patients' transcriptomic responses to an in vitro HRV stimulation assay. Panel A illustrates the development of a classifier using innovative features. Both shared and unique genetic and nongenetic variables influence transcript expression and healthy vs disease state within an individual. Exposure to a relevant stimulus (here, HRV infection) reveals relevant pathway gene sets whose genome and environmentally informed responses (G × E response pathways) can be used to predict individual prognosis. Panel B describes the development of classifier from the training set using data from PBMCs of healthy volunteers exposed to HRV in vivo. For each patient, paired microarrays analyzing gene expression before and after HRV exposure were compared using N-of-1-pathways analysis to identify significant Gene Ontology biological process (GO-BP) features describing each response. Responses in asymptomatic patients were then compared to responses in symptomatic patients to develop the classifier. Panel C describes the laboratory stimulation assay of asthmatic patients' PBMCs in this study. As in the training set, paired microarrays for each patient were used to determine the response using N-of-1-pathways. The classifier developed in panel B was then applied to individual responses and used to predict recurrent exacerbation.

Table 1. Descriptions of datasets

Dataset		Training set	Validation set
Purpose		Learn classifier	Predict outcome
Source	Authors	Zaas et al. *Cell Host and Microbe* 2009	Present study
	Data Source	GSE17156 (downloaded 9/17/2014)	GSE68479
Platform	Microarray	Affy. Human Gene U133A 2.0	Affy. Human Gene 1.0ST33297
	Probe	22277	33297
Protocol		Inhaled HRV PBMC samples drawn before and 48 h after HRV inoculation	PBMCs isolated, then incubated in vitro with HRV (stimulated) or vehicle (unstimulated)
Subjects	Total	19 healthy adult volunteers	23 pediatric asthmatic patients
		☐ 10 symptomatic for common cold	☐ 12 recurrent exacerbations of asthma (hospitalizations and/or emergency room visits)
		☐ 9 asymptomatic	☐ 11 no exacerbation of asthma
Samples	Samples:	38 RNA microarrays	46 RNA microarrays
		☐ 19 PBMC drawn prior to infection	☐ 23 PBMC unstimulated
		☐ 19 PBMC drawn after HRV inhalation	☐ 23 PBMC HRV-stimulated in vitro

Downloaded from https://academic.oup.com/jamia/article-abstract/24/6/1116/4004728/A-genome-by-environment-interaction-classifier-for
by University of Liverpool user
on 19 October 2017

Treating Children to Prevent Exacerbations of Asthma (TREXA)[24] and Best Add-on Therapy Giving Effective Responses (BADGER)[25] clinical trials were asked to participate in a yearlong extension study 2 years after the clinical trials ended. Based on exacerbation status, the pediatric asthma patients were classified in 2 groups during the clinical trial: (1) 11 patients who did not receive oral corticosteroids during the year of the extension study were coded as no exacerbation, and (2) 12 patients who received 2 or more courses of oral corticosteroids during the year of the study were coded as RE. (For more details on patient characteristics, see Supplementary Table S1.) Importantly, none of the 23 asthma patients received steroids during the month prior to the extension study when blood samples were collected. We statistically tested for obvious confounding factors between the 2 phenotypes, such as age, gender, family history, etc., and found no evidence of group-specific demographic or clinical differences (Supplementary Table S1).

PBMC sample extraction, HRV stimulation, and RNA expression quantification in the in vitro study to develop the validation set

PBMCs were isolated from blood samples collected from the 23 pediatric asthma patients enrolled in the extension study. Each individual's PBMCs were divided into 2 cultures, nonstimulated and stimulated with human rhinovirus RV-16 provided by James E Gern, University of Wisconsin-Madison. Specifically, 100 µl of the working stock RV-16 was added to 2 mL of PBMC suspension, yielding a final concentration of 2.5×10^6 plaque-forming units/mL. These samples were then incubated at 37°C and 5% CO_2 for 24 h (HRV-stimulated). In parallel, the second aliquot of patient PBMCs was incubated under the same culture conditions, but exposed only to vehicle (unstimulated). Following this, RNA was isolated from each sample, quality was assessed, and then RNA was amplified, tagged, and hybridized on Affymetrix Human Gene 1.0 ST microarrays according to standard operating procedures. Data were deposited in GEO (GSE68479).

Processing of microarray datasets

Within both the training and validation sets, each pair of matched samples (HRV-stimulated and unstimulated) was normalized using Robust Multiple-array Average[26] (2 paired samples at a time to avoid bias in single-subject analyses) from Affymetrix Power Tools.[27] For the training set, each pair of samples derived from each patient was obtained in a $\log_2$-normalized format and further corrected for batch effects using ComBat.[28] In all, 12 157 genes were found in common across both datasets.

Pathway (gene set) annotations using gene ontology biological processes

To obtain mechanistic profiles of responsive pathways for each patient, virus-responsive genes were aggregated into gene sets according to GO annotations of biological processes (GO-BPs)[21,29] using the *org.Hs.eg.db* package[30] (*Homo sapiens*) of Bioconductor,[31] available for R statistical software.[32] Hierarchical GO terms were retrieved using the *org.Hs.egGO2ALLEGS* database (downloaded on May 15, 2013), which contains a list of GO-BP terms along with all of their child nodes. Gene sets comprising a minimum of 15 and a maximum of 500 genes were included, as documented in our previous studies.[7,10,22,23,33]

Transcriptome and pathway (gene set) analysis of gene-by-environment interactions using N-of-1-pathways framework

To determine personally responsive gene set involvement for each patient in each dataset (training and validation sets, Table 1), the N-of-1-pathways Wilcoxon framework was applied using each patient's gene expression microarrays derived from paired samples (HRV-stimulated and nonexposed control PBMCs). Briefly, N-of-1-pathways is designed to create personal profiles of responsive pathways (gene sets) and is based on 3 principles: (1) the individual patient is the sole source/unit of observation, and any statistical descriptions and inferences are intended to relate only to that patient; (2) significance and interpretation are derived from gene sets; and (3) gene set–level information is used to answer questions of clinical importance. Principle 1 allows for detection of individual signals that traditional cohort-level studies may overlook. Principle 2 anchors the results in mechanism, and this affords dimension reduction and interpretation. Principle 3 provides quantitative and qualitative measures to address questions related to patient care.[7]

As described previously,[7,9] each patient's paired gene expression was transformed via N-of-1-pathways Wilcoxon into pathway-level metrics (annotated using GO-BP). Specifically, for each pathway gene set, a Wilcoxon signed-rank test was conducted between the gene expression measurements. This test produces a *P*-value that quantifies whether the central gene expression is different between the pair of samples (higher or lower expressed). In effect, observations from a patient in both datasets (training and validation) were transformed from paired, whole-transcriptome measurements to a simple list of pathway-associated *P*-values for each patient.

The *P*-values obtained from the training set were further manipulated to serve as inputs for the classifier algorithms. The output of N-of-1-pathways was transformed into a matrix of *P*-values representing all patients and pathways (23 patients by 3055 pathways). This matrix was further transformed into a ternary matrix by considering a pathway gene set as responsive if its nominal *P*-value was $<5\%$ and signed to indicate the direction of response. Specifically, the ternary matrix contains the values -1 (significantly lower in HRV-stimulated), 0 (not significant responsive), and 1 (significantly higher).

Principal component analysis

The principal component analyses (PCAs) were computed using the *FactoMineR* package[34] in R (with default parameters) based on the ternary matrix derived from the training set. A projection over the first component was used to assess the statistical significance of the separation shown by the PCA between 2 phenotypes (eg, symptomatic vs asymptomatic). In particular, a Mann-Whitney U test was used to assess the degree of separation of the 2 phenotypes. Projections over the first and third principal components were used to visualize the phenotypic separation.

Classifier modeling using gene-by-environment response pathways as features

The N-of-1-pathways ternary matrix (pathways × patients) that pertains to the analysis of the training set was used for training the classifier through machine learning procedures with feature selection. The training set was transformed into an .arff file using Weka v.3.6.11,[35] and then fed into Weka algorithms for further analysis. Briefly, feature selection was performed on the training set (clinical assay ternary matrix) using a chi-squared test. Using the training

Downloaded from https://academic.oup.com/jamia/article-abstract/24/6/1116/4004728/A-genome-by-environment-interaction-classifier-for
by University of Liverpool user
on 19 October 2017

data alone, chi-squared test of association was performed using the pathway-associated ternary values and the phenotype (symptomatic vs asymptomatic). The pathways were then ranked in ascending order of *P*-values. The pathway gene sets with the smallest *P*-values from the chi-squared test were selected for further modeling based on the best prediction performance by varying the number of features in the classifiers listed below (assessed via leave-one-out cross-validation; Supplementary Table S2). Five different classifier models were assessed for performance on the training set using the default parameters of Java software: random forest classifier, naïve Bayes, decision tree, support vector machine, and nearest neighbor. The random forest classifier model was further assessed for superior performance on the training set using receiver operating characteristic (ROC) curve based on prediction scores extracted from Weka. The area under the curve (AUC) and significance tests were computed with the default parameters of the *verification*[36] package in R. The fully specified classifier derived from the training set was then applied once on the validation set, and the accuracy of predictions was measured from the ROC curve.[37,38]

Visualization of personal gene-by-environment interaction via pathways gene set responsive to stimulation using star plots

Star plots were computed in R using the *stars* function in the default *graphics* package. Each patient has an individual star plot, where each spike represents the response of a specific pathway (GO-BP; method: N-of-1-pathways framework). The spikes within each star plot represents the 20 gene set features of the classifier described in Table 4. Each individual star plot corresponds to a single validation set individual's N-of-1-pathways scores (n = 23 patients; Table 1). An individual pathway is considered upregulated when the N-of-1-pathways test computes a concordant increase in expression of the genes within the pathway in the HRV-stimulated sample as compared to the control. The median radius of the star plot corresponds to no changes in the pathway, while upregulated pathways are assigned values in the white range and downregulated pathways are assigned values in the gray range. Direction-adjusted (signed) *P*-values were transformed into 4 states: (1) significantly upregulated (reaching the border of the star plot), (2) significantly downregulated (center of the star plot), (3) upregulated not reaching statistical significance (midrange radius of the white zone of the star plot), and (4) downregulated not reaching statistical significance (midrange radius of the gray zone of the star plot). For grouping of related pathways in the star plot, similarities between GO-BP terms were determined using Jiang's information theoretic similarity, which ranges from 0 (no similarity) to 1 (perfect match).[39]

RESULTS

Generation of the predictive classifier for rhinovirus stimulation response

As an exploratory analysis, we first computed the principal components resulting from the transformed ternary representation of all N-of-1-pathways scores from the training set (see Materials and Methods section; Principal Component Analysis). The idea was to investigate whether these scores carry information to discriminate asymptomatic from symptomatic individuals prior to fitting a complex classifier. Indeed, the first component from the PCA based on the single-subject dynamic change of gene expression within pathways gene set enabled a rough classification of the 2 clinical

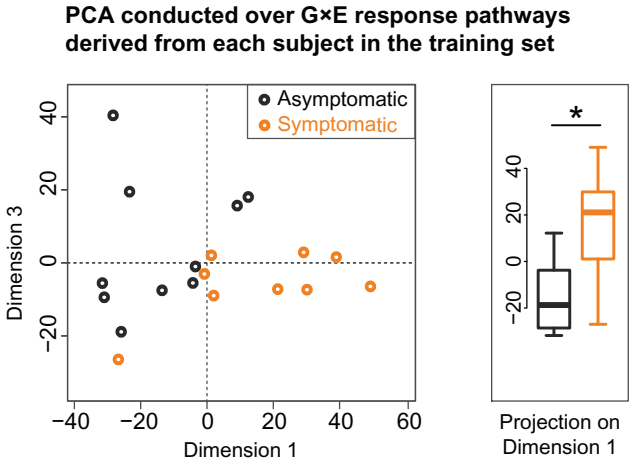


Figure 2. Metrics derived from responsive pathways discriminate asymptomatic from symptomatic subjects in the training set of in vivo HRV-stimulation data. In panel A, principal component analysis was conducted using responsive gene sets derived from each subject in the training set (Figure 1B and Materials and Methods section). The scatter plot on the left illustrates the bivariate relationship of the first and third principal components for each patient's ternary-represented N-of-1-pathways scores derived from paired samples of PBMCs collected at baseline and after HRV exposure. Each point represents a subject in the training set through a linear combination of pathway gene set–level scores that explain the maximal variation in the data (see Materials and Methods section for details on the ternary representation and PCA construction). The first and third principal components show 2 clusters emerging that separate asymptomatic from symptomatic individuals. Thus, N-of-1-pathways scores are associated with the phenotype of interest. On the right, side-by-side box plots display the first principal component scores among asymptomatic and symptomatic subjects, and this component alone significantly dichotomizes the 2 phenotypes (Mann-Whitney U test, $P = .0069$). Panel B lists the 20 Gene Ontology biological processes used as features in our classifier, organized according to broad biological function categories.

phenotypes (asymptomatic vs symptomatic; Figure 2, right). Moreover, the bivariate relationship between the first and third components reveals a striking pattern of phenotypic separation (Figure 2, left). In light of these results, the single-subject scores of the 20 most relevant pathway dynamic expressions produced by the chi-squared test based on feature selection were used to train a classifier (Materials and Methods section). The random forest classifier achieved the best AUC of the ROC curve (86.7%) using leave-one-out cross-validation of the training set (see Materials and Methods section; classifier modeling; Supplementary Table S2).

Classifier derived from pathways responsive to gene-by-environment interactions, under HRV stimulation, enables asthmatic clustering and a clinical risk assessment representation of asthma exacerbation for each subject

We applied the random forest classifier to the dynamic expression change of pathways stimulated in vitro on PBMC samples of the validation set, which showed >70% accuracy in identifying children with and without recurrent asthma exacerbations (Figure 3A). We also tried 4 other classification methods to assess the robustness of the signal regardless of the type of classifier (Table 2). An unsupervised principal component analysis of these pathway scores also showed a natural split between these 2 asthmatic phenotypes (Supplementary Figure S1). To further explore the dynamic change in

Downloaded from https://academic.oup.com/jamia/article-abstract/24/6/1116/4004728/A-genome-by-environment-interaction-classifier-for
by University of Liverpool user
on 19 October 2017

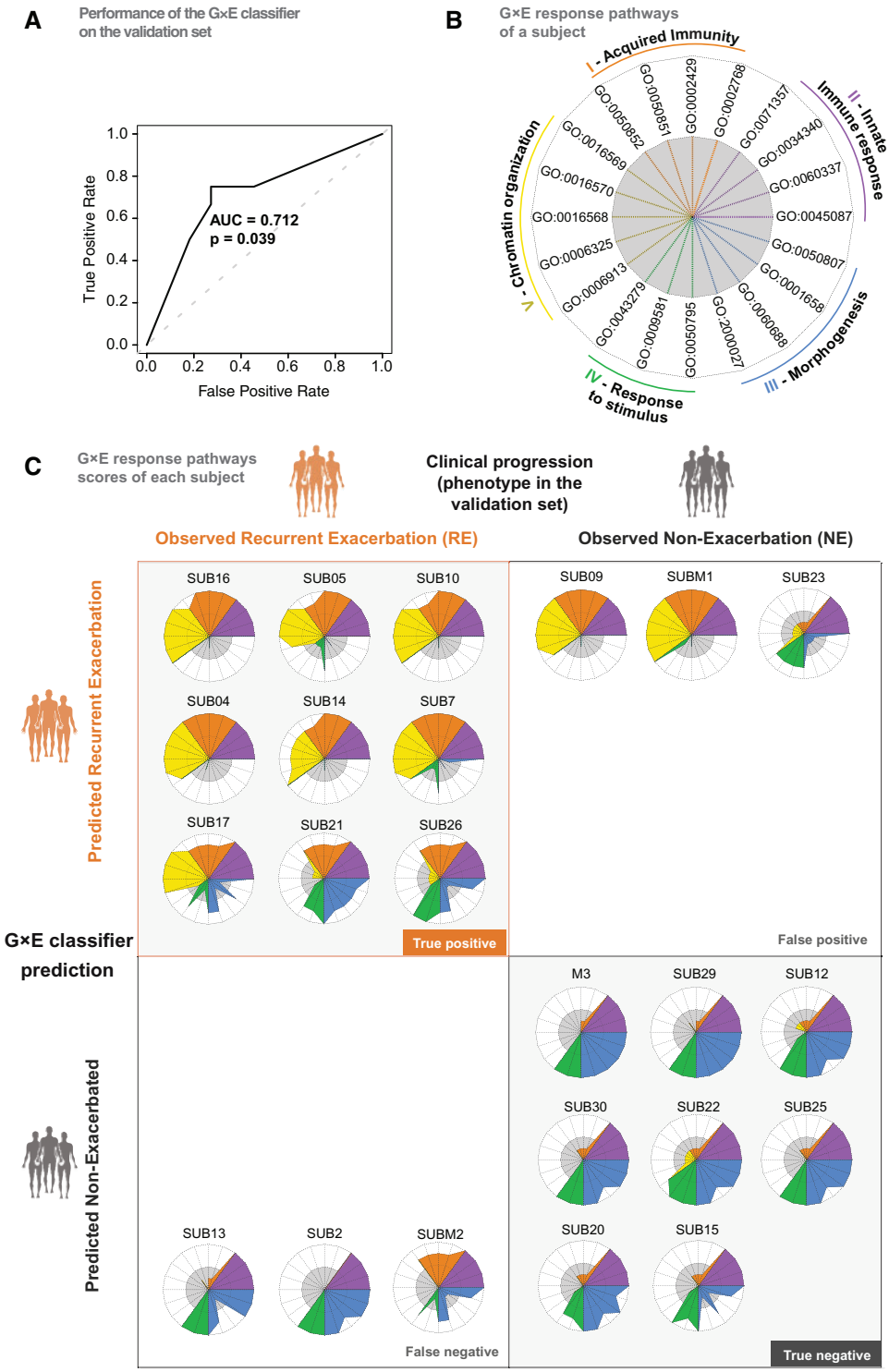


Figure 3. Fully specified classifier derived from responsive pathways in training set performs accurately in independent asthmatic validation/virogram set to predict exacerbation status. Panel A shows receiver operating curve (ROC) of the G × E classifier conducted in the validation set (overall accuracy, 74%; sensitivity, 75%; specificity, 73%). In panel B, the star plots illustrate the level of response to HRV stimulation for each pathway in that patient. The classifier is designed using 20 pathways, with each radial line representing the score of a pathway. The area above and below the gray zone represents upregulation and downregulation, respectively, of any given pathway (see Materials and Methods section for complete details). In panel C, each star plot represents a single subject, with label appearing above (eg, Subject 16 = SUB16). The star plot is located in the quadrant of the contingency table that represents the performance of the G × E classifier on predicting the clinical progression from a specific asthmatic patient's data. This classifier prediction applied to the HRV-stimulation assay recapitulated the clinical progression in 17 out of the 23 asthmatic patients; 6 were misclassified: 3 false positives (SUBM1, SUB09, SUB23) and 3 false negatives (SUB2, SUB23, and SUBM2). One can straightforwardly identify that innate immunity (mauve) is upregulated in every asthmatic patient and does not contribute to the classification in the validation set. On the other hand, observed upregulation in acquired immunity (orange) pathways could be used to correctly classify 19 subjects.

Downloaded from https://academic.oup.com/jamia/article-abstract/24/6/1116/4004728/A-genome-by-environment-interaction-classifier-for
by University of Liverpool user
on 19 October 2017

Table 2. Performance of 5 different pathway response classifiers built on N-of-1-pathway scores in the training set (in vivo) and validated without bias on N-of-1-pathway scores in the validation set (in vitro). The Random Forest (bold) classifier showed the highest scores and was chosen for all other analyses.

Classifier	#Features (GO-BP terms)	Classification performance in the validation set								
		Accuracy	Sensitivity	Specificity	Precision	AUC	TP	FP	FN	TN
Random Forest	**20**	**73.9**	**75**	**72.7**	**75**	**71.2**	**9**	**3**	**3**	**8**
Naïve Bayes	20	**73.9**	75	72.7	75	67.4	9	3	3	8
Decision Tree	20	**65.2**	75	54.5	64.3	64.8	9	5	3	6
Support Vector Machine	20	**69.6**	75	63.6	69.2	69.3	9	4	3	7
Nearest Neighbor	20	**69.6**	66.7	72.7	72.7	69.7	8	3	4	8

The random forest and naïve Bayes classifiers showed the highest metrics on the validation set, and random forest was selected for further optimization.
TP: true positive; FP: false positive; FN: false negative; TN: true negative

Table 3. Our classifier performed optimally across all metrics using 20 gene-set (GO-BP term) features

Classifier	#Features	Classification performance in the validation set								
		Accur	Sens.	Spec.	Prec.	AUC	TP	FP	FN	TN
Random Forest	10	47.8	0	100	NA	68.2	0	0	12	11
Random Forest	20	73.9	75.0	72.7	75.0	71.2	9	3	3	8
Random Forest	30	69.5	58.3	81.8	77.8	65.5	7	2	5	9

TP: true positive; FP: false positive; FN: false negative; TN: true negative

expression within the validation set, we used a heatmap displaying an unsupervised clustering of individuals to visualize transformed N-of-1-pathways scores (Supplementary Figure S2). Further, to assess whether the classifier was dependent on HRV stimulation or due to a general transcriptomic viral response, we developed 2 additional classifiers using (1) influenza A and (2) respiratory syncytial virus infection data (GSE17156; Zaas et al.[20]) as our training sets. We employed the same methodology as described to identify classifier features and attempted to use these to classify our validation set of HRV-stimulated PBMCs from asthmatic patients. The best classifier for either alternative virus was ineffective with every validation set patient-predicted as nonrecurrent, thus yielding no predictive stratification (52% accuracy and 61% AUC for influenza; 48% accuracy, 85% AUC for respiratory syncytial virus; see details in Supplementary Table S3). Moreover, our pathway-level metrics of HRV response outperformed prediction in the validation set compared to conventional mRNA expression–derived classifiers (Supplementary Table S4). Indeed, gene-level classifiers, built in a similar fashion from the HRV-stimulated training set, achieved a best accuracy of only 52% and AUC of 41% on the validation set. Again, this poor AUC is explained by a lack of prediction stratification, as all patients in the validation set were predicted to have REs.

We assessed the optimal number of pathway features to be included in our random forest classifier using the chi-squared method for feature selection. This showed robust performance by 20 to 30 features (Table 3). Guided by parsimony heuristics, we selected 20 features.

Focusing on each patient's PBMC response to HRV stimulation, we developed a star plot representing personal profiles of these 20 pathways (Figure 3B) to report results as a visually interpretable risk assessment of asthma exacerbation. Comparing clinical records of individuals who experienced at least 2 exacerbations in a year to their retrospective molecular assessment of asthma exacerbation, 6 of the 23 individuals were misclassified (SUBM1, SUB09, SUB13, and SUB2, SUB23, and SUBM2; Figure 3C). Acquired immunity (orange), morphogenesis (blue), response to stimulus (green), and chromatin organization (yellow) responses were highly discriminatory between true positives and true negatives, while all of the asthmatic patients in the validation set showed substantial activation of innate immunity (mauve). Thus, while innate immunity–associated pathways are likely involved in common defense mechanisms to HRV and symptomatic vs asymptomatic infection of healthy subjects, it may be that asthmatic disease state alone predisposed all patients in our validation set to high innate activation. This is perhaps a limitation of our use of healthy patients in the training set, though it does highlight that our classifier was fully specified on the training set alone before being applied to the asthmatic patient data.

Nonetheless, despite limited additional discriminatory power from innate immunity–associated pathways, chromatin organization and acquired immunity clearly show a high propensity to be significantly upregulated in children who experienced RE relative to those who did not. As such, the star plot provides a visual tool for physicians to observe responsive mechanisms common and distinct between phenotypes, as well as individualized profiles.

DISCUSSION

In this study, we demonstrated the effectiveness of classifiers built using biological process pathways that are differentially responsive to clinically relevant stimulation (G × E classifier). Specifically, our case study uses HRV-stimulation response data from healthy patients to classify and predict the risk of exacerbation in asthmatic patients whose PBMCs were subjected to in vitro exposure of the same virus. Although the classifier was trained on healthy volunteers, we were able to accurately classify 70% of asthmatic patients as low risk (no exacerbation) or high risk (recurrent exacerbation) by using the same pathway gene sets, despite small numbers of patients in both the training and validation cohorts. A classifier trained on patient-specific dynamic responses at the pathway level

Downloaded from https://academic.oup.com/jamia/article-abstract/24/6/1116/4004728/A-genome-by-environment-interaction-classifier-for
by University of Liverpool user
on 19 October 2017

(N-of-1-pathways: 74% accuracy, 71% AUC, 2-sided $P = 0.039$; Table 2 and Figure 3A) significantly outperformed a cohort-derived classifier trained on transcript-level variation between the sets of asymptomatic vs symptomatic patients at 1 static point (52% accuracy, 41% AUC; Results section). This increase in predictive power was achieved through 4 major mechanisms: (1) control of confounded environmental variables within each subject that impact global transcriptional activity (eg, preexisting habits, medications, and additional chronic diseases); (2) control of preexisting baseline differences in gene or pathway expression across patients that may introduce bias or lead to spurious conclusions; (3) reduced statistical dimension through aggregation of responsive transcripts into enriched gene sets with common biological properties prior to cross-group comparisons; and (4) a focus on analysis of clinically relevant elicited responses within each patient rather than total variability. Within a given clinical class, between-patient heterogeneity in specific transcript expression resulting from unique genetic architecture can also be subsumed into pathway scores if the summative effect on a functional pathway remains in common.

Unsurprisingly, these results suggest that dynamic changes in the personal transcriptome are more predictive than static measurements. This agrees with what is observed in clinical practice, where provocation tests are routinely shown as more correlated to the patient's response to therapy than unprovoked phenotypes (eg, bronchial histamine challenge test with spirometry is superior to peak flow measurements). The stimulation assay we outline here corresponds to such a provocation assay, revealing a dynamic phenotype at the transcriptome level.

Along with the improved classifier performance, the N-of-1-pathways strategy provides an additional benefit over conventional classifiers, as pathways in each patient are individually scored. This permits researchers or clinicians to meaningfully examine each patient's individual responsivity independently (Figure 3C). As our knowledge of the mechanistic involvement of each pathway grows, these star plots could be used in aggregate not only to stratify patients according to their classified risk, but also to suggest a more effective precision medicine approach to disease management for any given individual.

It is of great interest that several of the GO biological processes identified were highly relevant for the viral challenge under study. Acquired and innate immunity, for example, are critical elements in acute responses to viruses. Moreover, type I interferons (ie, α and β interferons) play a major role in activating antiviral responses.[40] Activation of innate immunity in particular is also a hallmark of hypersensitivity in asthmatic patients,[41,42] so it is both expected and reassuring that our classifier included multiple GO terms related to this process. However, that same hallmark of hypersensitivity in asthma reduced the discriminatory power of innate immunity processes in classifying already asthmatic patients as exacerbating vs non-exacerbating (Figure 3). Not many classifiers of asthmatic exacerbation have been published, but when we compare the biological processes revealed by our classifier to one identified by Bjornsdottir et al. (2011),[43] we see both similarities and differences. The classifier in their study used a larger dataset of 118 adult asthmatic patients, with RNA samples obtained during asthma-quiescent vs mid-exacerbation in the same cohort of patients, using the natural course of each patient's disease progression rather than a provoked stimulus response assay. Their asthmatic exacerbation classifier, which was run through Ingenuity Pathway Analysis, revealed entirely immune-mediated responses, including Toll-like receptor signaling, T cell receptor signaling, B cell receptor signaling, interferon signaling, interferon responsive factors, and IL15 signaling. While we could not do a direct comparison of GO pathways or underlying transcript participation in classification due to the proprietary nature of Ingenuity Pathway Analysis, these appear to show substantial convergence with the GO pathways of acquired and innate immunity that we observed (Supplementary Table S5). Notably, however, while innate immunity–associated genes were significantly represented in their study, our study found that activation of the innate immune response was a characteristic of all asthmatic patients exposed to HRV, with little predictive discriminatory power to classify exacerbation from non-exacerbation. In addition, the classifier built by Bjornsdottir did not substantially prioritize transcripts or pathways associated with morphogenesis, chromatin organization, or generalized response to stimulus, which were highly discriminatory in our analysis. We found that these may provide insights into additional exacerbation pathology–associated processes that occur in the subset of predisposed patients who are exposed to triggering stimuli.

Another notable result of this study was that it underscores the specific clinical relevance of HRV infection to asthmatic exacerbation, as had been repeatedly described in clinical and epidemiological studies.[12–17] While observational associations could be biased by the ubiquity of HRV (the common cold) relative to other respiratory infections, we found that the transcriptional response of PBMCs to HRV exposure was predictive of the outcome in our asthma cohort, while a classifier based on influenza A–provoked responses was not. This also highlights the importance of choosing a highly clinically relevant stimulus during design of the bioassay component in order to optimize the chance of success.

Though the training set data were the same, the classifier developed here using single-subject metrics of responsive pathways (N-of-1-pathways scores) diverges substantially from the classifier produced by Zaas et al.[20] Their classifier was developed by first selecting a common gene signature indicative of acute respiratory infection across 3 distinct viral challenges (HRV, respiratory syncytial virus, and influenza A). This gene signature was then used to differentiate influenza infection from bacterial infection or asymptomatic status in an independent clinical dataset. On the other hand, our classifier was designed to predict a different phenotype – exacerbated asthma in patients using responsiveness of their PBMCs to HRV in an in vitro assay (virogram). It remains to be seen if the common gene signature of viral infection derived by Zaas et al.[20] is predictive of asthmatic exacerbation, and this could be determined in future studies.

Although previous studies attempting to create asthmatic exacerbation risk classifiers used differential expression of pathway gene sets to classify clinical phenotypes,[44,45] these did not use a matched sample design (eg, stimulated vs unstimulated) and thus were not able to assess individual dynamic changes as features and/or true pathway-level signals with underlying transcript heterogeneity. We showed that developing an unbiased $G \times E$ classifier based on the genomic profiles of healthy individuals experiencing rhinovirus stimulation enabled risk assessment of exacerbation in children with asthma. Classification of asthma exacerbation was successfully demonstrated using the dynamic change of expression in pathways of single subjects, whereas conventional methods of gene-level profiling failed.

LIMITATIONS AND FUTURE STUDIES

One limitation of the current study is the small cohort size in both our training (healthy) and validation (asthmatic) sets. Designing

Downloaded from https://academic.oup.com/jamia/article-abstract/24/6/1116/4004728/A-genome-by-environment-interaction-classifier-for
by University of Liverpool user
on 19 October 2017

Table 4. Gene ontology (GO) gene sets of the G × E classifier obtained in the training set

Class of GO-BP gene sets	GO ID	GO term
I. Acquired Immunity	GO:0002768*	immune response–regulating cell surface receptor
	↪GO:0002429	immune response–activating cell surface receptor
	↪GO:0050851	antigen receptor–mediated signaling pathway
	↪GO:0050852	T cell receptor signaling pathway
II. Innate Immune Response	GO:0045087*	innate immune response
	↪GO:0060337	type I interferon-mediated signaling pathway
	(p)↪GO:0034340	response to type I interferon
	(i)↪GO:0071357	cellular response to type I interferon
III. Morphogenesis	GO:0050807	regulation of synapse organization
	GO:0001658	branching involved in ureteric bud morphogenesis
	GO:0060688	regulation of morphogenesis of a branching structure
	GO:2000027	regulation of organ morphogenesis
IV. Response to Stimulus	GO:0043279	response to alkaloid
	GO:0050795	regulation of behavior
	GO:0009581	detection of external stimulus
V. Chromatin Organization	GO:0006325*	chromatin organization
	↪GO:0016568	chromatin modification
	↪GO:0016569	covalent chromatin modification
	↪GO:0016570	histone modification
	GO:0006913	nucleocytoplasmic transport

Twenty GO biological processes (GO-BPs) responsive to HRV stimulation were selected by the best classifier after evaluation in the validation set (Materials and Methods section). GO-BPs were organized into 5 categories by an unbiased information theoretic similarity score (Materials and Methods section), and these classes were manually assigned representative names. Of note, some GO-BPs were ordered according to their GO hierarchy when available, with the parent term annotated with an asterisk and located above the child term. For example, GO:0002768 is the parent of GO:0002429. Legend: ↪ = "is a" (parent-child relationship); (p)↪ = "part of" (parent-child relationship); (i)↪inferred as "is a" (parent-child relationship).

more accurate classifiers on a larger and/or asthmatic cohort may improve patient stratification. Accurate classifiers using N-of-1-pathways analytics are powered to identify dynamic changes in gene expression patterns arising from 2 samples of a single subject and offer additional insights on individual responsive pathways.[7,9,10] These improved genomic signals at the subject level can be compiled for a cross-patient statistical analysis to uncover common pathways and genes as well as allow patient risk stratification. In contrast, many conventional transcriptome analyses operating on single and static measurement profiles utilize statistics that generally require larger cohorts to combat cross-patient variability and random noise in the raw expression data, unrelated to the measure of interest.

As this study was conducted as a proof-of-concept, the number of subjects was limited and did not allow for designing a classifier from groups of children with asthma. The current accuracy of the classifier is sufficient for a proof-of-concept, but was generated from symptomatic vs asymptomatic non-asthmatic subjects and then applied a classifier to assess propensity for exacerbations in children with asthma. Further, the PBMC samples were obtained after the asthma exacerbations occurred, so it is not possible in this study to determine whether these responsive pathways lead to or result from recurrent asthma exacerbations. In future prospective studies, we plan to generate both training and validation cohorts from groups of children with asthma, which should increase the accuracy of the classifier and its clinical utility. We are also improving the N-of-1-pathways framework to identify better responsive transcripts in each patient, which may provide greater insight into the molecular mechanism(s) underpinning each pathway of the classifier. While expression-level classifiers implicitly harbor co-transcribed and co-regulated transcripts, the GO terms of our G × E classifier comprise more obvious overlap of mRNAs (Table 4). We will investigate methods to identify more independent classifier features in future studies. Per standard operating procedures, PBMC RNA is extracted from PBMCs as 1 homogenized sample without counts of cell-type subpopulations, thus the RNA expression changes between 2 samples confound changes in expressed transcripts with changes to cellular composition of the sample. In future studies, deconvoluting the signal according to cell type could be addressed for cell type–specific transcription, using expression change and proportion of cell type change between paired samples.

CONCLUSION

Given the importance of monitoring the dynamic gene expression driven by genome-by-environment interplay that influences the direction of biological processes and the course of a disease, improvements are needed in the design of advanced experimental and analytics assays intended for individual patients' point of care and disease assessment in order to tailor treatment. Our proposed framework addresses this challenge by differing from traditional classifiers that employ mRNA expression as the features. In particular, we introduce the use of the N-of-1-pathways framework to first identify the responsive pathways induced gene-by-environment interaction via a paired stimulated vs unstimulated assay, where the stimulation is relevant to the clinical phenotype of interest. These responsive pathways are then utilized as features in classification algorithms to predict the phenotype of interest. In the context of asthma, at least 48% of pediatric admissions for exacerbated asthma are due to HRV, and, as such, our in vitro HRV-stimulated assay was designed ab initio for a single-subject analytic framework and is seen to identify children prone to asthma exacerbations. The type of stimulated PBMC assay proposed could scale to predict responses to therapy of a large number of clinical conditions that are immune-mediated, or for which PBMC is a justified target tissue. For example, chemical or biological ligands rather than intact viruses could be used to

Downloaded from https://academic.oup.com/jamia/article-abstract/24/6/1116/4004728/A-genome-by-environment-interaction-classifier-for
by University of Liverpool user
on 19 October 2017

stimulate specific pathways to study drug-targeting receptors. N-of-1-pathways analytics provides an informative, quantitative, and visual characterization of subject-specific response to stimulation, influenced by the interaction of genomic variability and stimulus exposure, and is, in principle, applicable not only to the transcriptome but also to other omics measures (eg, proteome). When applied to easily accessible tissue in a minimally invasive way, these types of in vitro stimulation assays and associated analytics could potentially supplant more invasive in vivo prognosis procedures that also assay responses to the environment in the emerging field of precision medicine.

FUNDING

FDM and YAL are supported in part by the Arizona Health Sciences Center and the BIO5 Institute. This study was funded in part by the following grants: K22 LM008308-04, 5U10HL064307, the University of Arizona Cancer Center (P30CA023074), and the Center for Biomedical Informatics and Biostatistics of the University of Arizona Health Sciences.

COMPETING INTERESTS

None.

CONTRIBUTORS

MJH, DJJ, and FDM conceived the bioassay experiments; VG, YAL, IA, AB, JL, HL, and DS conceived and conducted the computational biology and high-throughput experiments; VG, AB, IA, YAL, GS, JB, FDM, DJJ, MJH, JL, HL, DS, and CK contributed to Materials and Methods; VG, IA, YAL, AB, AB, JB, FDM, CK, QL, and YS contributed to figures, tables, and additional files; VG, IA, YAL, JL, JB, DS, MJH, FDM, QL, and YS analyzed the results; GS, JB, YAL, IA, CK, VG, JL, DS, DJJ, and FDM wrote and revised the manuscript.

DATA-SHARING

All software/programs are provided at http://Lussierlab.org/publications/N-of-1-pathways.

SUPPLEMENTARY MATERIAL

Supplementary material is available at *Journal of the American Medical Informatics Association* online.

ACKNOWLEDGMENTS

We thank the Genomics Shared Service of the Arizona Cancer Center for arraying the paired samples.

REFERENCES

1. Chaussabel D. Assessment of immune status using blood transcriptomics and potential implications for global health. *Semin Immunol*. 2015;27(1):58–66.
2. Duffy D, Rouilly V, Libri V, *et al*. Functional analysis via standardized whole-blood stimulation systems defines the boundaries of a healthy immune response to complex stimuli. *Immunity*. 2014;40(3):436–50.
3. Collins FS, Varmus H. A new initiative on precision medicine. *New Engl J Med*. 2015;372(9):793–95.
4. Khera AV, Emdin CA, Drake I, *et al*. Genetic risk, adherence to a healthy lifestyle, and coronary disease. *N Engl J Med*. 2016;375(24):2349–58.
5. Frost HR, Shen L, Saykin AJ, Williams SM, Moore JH, Alzheimer's Disease Neuroimaging Initiative. Identifying significant gene-environment interactions using a combination of screening testing and hierarchical false discovery rate control. *Genet Epidemiol*. 2016;40(7):544–57.
6. Khatri P, Sirota M, Butte AJ. Ten years of pathway analysis: current approaches and outstanding challenges. *PLoS Comput Biol*. 2012;8(2):e1002375.
7. Gardeux V, Achour I, Li J, *et al*. "N-of-1-pathways" unveils personal deregulated mechanisms from a single pair of RNA-Seq samples: towards precision medicine. *J Am Med Inform Assoc*. 2014;21(6):1015–25.
8. Gardeux V, Arslan A, Achour I, Ho T-T, Beck W, Lussier Y. Concordance of deregulated mechanisms unveiled in underpowered experiments: PTBP1 knockdown case study. *BMC Med Genomics*. 2014;7(Suppl 1):S1.
9. Gardeux V, Bosco A, Li J, *et al*. Towards a PBMC "virogram assay" for precision medicine: Concordance between ex vivo and in vivo viral infection transcriptomes. *J Biomed Inform*. 2015;55:94–103.
10. Schissler AG, Gardeux V, Li Q, *et al*. Dynamic changes of RNA-sequencing expression for precision medicine: N-of-1-pathways Mahalanobis distance within pathways of single subjects predicts breast cancer survival. *Bioinformatics*. 2015;31(12):i293–302.
11. Schissler AG, Li Q, Chen JL, *et al*. Analysis of aggregated cell-cell statistical distances within pathways unveils therapeutic-resistance mechanisms in circulating tumor cells. *Bioinformatics*. 2016;32(12):i80–89.
12. Busse WW, Lemanske RF Jr, Gern JE. Role of viral respiratory infections in asthma and asthma exacerbations. *Lancet*. 2010;376(9743):826–34.
13. Corne JM, Marshall C, Smith S, *et al*. Frequency, severity, and duration of rhinovirus infections in asthmatic and non-asthmatic individuals: a longitudinal cohort study. *Lancet*. 2002;359(9309):831–34.
14. Johnston SL, Pattemore PK, Sanderson G, *et al*. The relationship between upper respiratory infections and hospital admissions for asthma: a time-trend analysis. *Am J Respir Critical Care Med*. 1996;154(3 Pt 1):654–60.
15. Lemanske RF Jr, Busse WW. Asthma: clinical expression and molecular mechanisms. *J Allergy Clin Immunol*. 2010;125(2 Suppl 2):S95–102.
16. Sears MR, Johnston NW. Understanding the September asthma epidemic. *J Allergy Clin Immunol*. 2007;120(3):526–29.
17. Miller EK, Lu X, Erdman DD, *et al*. Rhinovirus-associated hospitalizations in young children. *J Infect Dis*. 2007;195(6):773–81.
18. Wu P, Hartert TV. Evidence for a causal relationship between respiratory syncytial virus infection and asthma. *Expert Rev Anti Infect Ther*. 2011;9(9):731–45.
19. Forno E, Celedon JC. Predicting asthma exacerbations in children. *Curr Opin Pulmonary Med*. 2012;18(1):63–69.
20. Zaas AK, Chen M, Varkey J, *et al*. Gene expression signatures diagnose influenza and other symptomatic respiratory viral infections in humans. *Cell Host Microbe*. 2009;6(3):207–17.
21. Ashburner M, Ball CA, Blake JA, *et al*. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genetics*. 2000;25(1):25–29.
22. Chen J, Sam L, Huang Y, *et al*. Protein interaction network underpins concordant prognosis among heterogeneous breast cancer signatures. *J Biomed Inform*. 2010;43(3):385–96.
23. Chen JL, Hsu A, Yang X, *et al*. Curation-free biomodules mechanisms in prostate cancer predict recurrent disease. *BMC Med Genomics*. 2013;6 (Suppl 2):S4.
24. Martinez FD, Chinchilli VM, Morgan WJ, *et al*. Use of beclomethasone dipropionate as rescue treatment for children with mild persistent asthma (TREXA): a randomised, double-blind, placebo-controlled trial. *Lancet*. 2011;377(9766):650–57.
25. Lemanske RF Jr, Mauger DT, Sorkness CA, *et al*. Step-up therapy for children with uncontrolled asthma receiving inhaled corticosteroids. *New Engl J Med*. 2010;362(11):975–85.
26. Irizarry RA, Hobbs B, Collin F, *et al*. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*. 2003;4(2):249–64.
27. Groza T, Kohler S, Doelken S, *et al*. Automatic concept recognition using the human phenotype ontology reference and test suite corpora. *Database (Oxford)*. 2015;2015:ii.

Downloaded from https://academic.oup.com/jamia/article-abstract/24/6/1116/4004728/A-genome-by-environment-interaction-classifier-for
by University of Liverpool user
on 19 October 2017

28. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics*. 2007;8(1):118–27.
29. Gene Ontology C. The Gene Ontology in 2010: extensions and refinements. *Nucleic Acids Res*. 2010;38(Database issue):D331–35.
30. Carlson M. org.Hs.eg.db: Genome wide annotation for Human. R package version 3.4.1. via Bioconductor; 2017.
31. Gentleman RC, Carey VJ, Bates DM, *et al*. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol*. 2004;5(10):R80.
32. R: Development core team. *R: A Language and Environment for Statistical Computing: R Foundation for Statistical Computing*. Vienna, Austria; 2004.
33. Yang X, Regan K, Huang Y, *et al*. Single sample expression-anchored mechanisms predict survival in head and neck cancer. *PLoS Comput Biol*. 2012;8(1):e1002350.
34. Lê S, Josse J, Husson F. FactoMineR: An R package for multivariate analysis. *J Stats Software*. 2008;25(1):1–18.
35. Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH. The WEKA data mining software: an update. *ACM SIGKDD Explorations Newsletter*. 2009;11(1):10–18.
36. Hsu W-R, Murphy AH. The attributes diagram: A geometrical framework for assessing the quality of probability forecasts. *Int J Forecast*. 1986;2(3):285–93.
37. Dupuy A, Simon RM. Critical review of published microarray studies for cancer outcome and guidelines on statistical analysis and reporting. *J Natl Cancer Institute*. 2007;99(2):147–57.
38. Simon R. Roadmap for developing and validating therapeutically relevant genomic classifiers. *J Clin Oncol*. 2005;23(29):7332–41.
39. Jiang J, Conrath D. Multi-word complex concept retrieval via lexical semantic similarity. Paper presented at the International Conference on Information Intelligence and Systems; 1999. doi:10.1109/ICIIS.1999.810309.
40. Samuel CE. Antiviral actions of interferons. *Clin Microbiol Rev*. 2001;14(4):778–809.
41. Chatila TA. Innate immunity in asthma. *N Engl J Med*. 2016;375(5):477–79.
42. Sly PD, Holt PG. Role of innate immunity in the development of allergy and asthma. *Curr Opin Allergy Clin Immunol*. 2011;11(2):127–31.
43. Bjornsdottir US, Holgate ST, Reddy PS, *et al*. Pathways activated during human asthma exacerbation as revealed by gene expression patterns in blood. *PloS One*. 2011;6(7):e21902.
44. Beale J, Jayaraman A, Jackson DJ, *et al*. Rhinovirus-induced IL-25 in asthma exacerbation drives type 2 immunity and allergic pulmonary inflammation. *Sci Transl Med*. 2014;6(256):256ra134.
45. Pritchard AL, White OJ, Burel JG, Carroll ML, Phipps S, Upham JW. Asthma is associated with multiple alterations in anti-viral innate signalling pathways. *PLoS One*. 2014;9(9):e106501.

Downloaded from https://academic.oup.com/jamia/article-abstract/24/6/1116/4004728/A-genome-by-environment-interaction-classifier-for
by University of Liverpool user
on 19 October 2017

Supplementary Material for

A genome by environment interaction assay for precision medicine: personal transcriptome response to rhinovirus identifies children prone to asthma exacerbations

V. Gardeux, J. Berghout, I. Achour, A.G. Schissler, Q. Li, J. Berghout, C. Kenost, Y. Shang, A. Bosco, J. Li, D. Saner, M. J. Halonen, D. J. Jackson, H. Li, F. D. Martinez, Y. A. Lussier

Supplement Tables S1-S4

Supplement Figure S1-S2

Supplementary Table S1. Primary characteristics of asthma participants who experienced no exacerbations or recurrent exacerbation during the TREXA and BADGER clinical trials. Note that the p-value (rightmost column) indicates either the result of a nonparametric significance test for a difference in central tendencies between the two phenotypes (for continuous data) or the result of a Fisher's Exact Test for difference in proportions. No p-values were smaller than an unadjusted significance threshold of 5%. Therefore, it can be inferred that there were no significant confounders measured between the two groups of interest.

At trial randomization	**No Exacerbation (NE; N_{NE} = 11)**			**Recurrent Exacerbation (RE; N_{RE} = 12)**			
		%	**Out of**	**n**	**%**	**Out of**	**p**
Sex (male)	6	54.5	11	8	66.7	12	0.7
White	8	72.7	11	10	83.3	12	0.6
Father w/ diagnosis of asthma (4 values missing)	3	27.3	11	0	0.0	8	0.087
Mother with diagnosis of asthma	6	54.5	11	2	16.7	12	0.089
Eczema	5	45.5	11	5	41.7	12	0.9
	n	**mean**	**SD**	**n**	**mean**	**SD**	**p**
Height (cm)	11	149.5	20.5	12	151.3	22.7	0.8
Weight (kg)	11	46.7	24.0	12	59.1	30.8	0.3
Age at Randomization (years)	11	12.7	3.5	12	11.3	3.6	0.6
Age at onset of asthma symptoms (years)	11	3.0	2.2	12	4.1	2.9	0.3
Age at asthma diagnosis (years)	10	6.1	4.6	12	5.0	3.5	0.5
#Positive aeroallergen skin tests	11	3.5	2.0	12	3.9	2.5	0.6
Prebronchodilator FEV_1 (% predicted)	10	102	14.5	11	92.7	10.3	0.10
Prebronchodilator FEV_1/FVC (%)	10	83.5	4.1	11	77.7	8.1	0.057
Asthma Control Test (ACT) score*	11	21.1	5.7	12	23.9	1.7	0.14
At trial extension study	**NE**			**RE**			
	n	**mean**	**SD**	**n**	**mean**	**SD**	**p**
Age (years)	11	15.0	3.7	12	13.8	3.5	0.4
Height (cm)	11	160.7	16.9	12	163.3	17.6	0.7
Weight (kg)	11	60.2	24.7	12	77.0	34.3	0.2
Body-mass index (kg/m^2)	11	22.5	6.3	12	27.7	7.5	0.087
Asthma Control Test (ACT) score	11	23.6	1.6	12	22.2	2.4	0.093
Prebronchodilator FEV_1 (% predicted)	11	94.9	13.4	12	91.6	9.8	0.5
Prebronchodilator FEV_1/FVC (%)	11	81.4	5.5	12	76.5	8.1	0.11

No Exacerbation (NE) defined as no use of oral steroids during clinical trial.

Recurrent Exacerbation (RE) defined as use of oral steroids two times or more during the year of the clinical trial.

FEV_1 = forced expiratory volume in the first second and FVC = forced vital capacity.

Supplementary Table S2. Leave-one-out cross validation performance of five different pathway-response classifiers built on the training set of nineteen healthy individuals.

Classifier	#Features (GO-BP)	Training performance in the training set								
		Accuracy	Sensitivity	Specificity	Precision	AUC	TP	FP	FN	TN
Random	20	**84.21**	80	88.89	88.89	86.67	8	1	2	8
Naïve	20	**84.21**	90	77.78	81.82	85.56	9	2	1	7
Near.	20	**78.95**	70	88.89	87.5	79.45	7	1	3	8
SVM	20	**78.95**	90	66.67	75	78.33	9	3	1	6
Decision	20	**78.95**	70	88.89	87.5	66.11	7	1	3	8

TP: True positive; FP: False positive; FN: False negative; TN: True negative

Supplementary Table S3. Pathway response classifiers of Influenza and RSV fail: leave-one-out cross validation performance of classifiers built on the Zaas et al. training set of symptomatic versus asymptomatic healthy individuals infected with Influenza A (n=20) or RSV (n=17), validated in our *in vitro* HRV-exposed asthmatic set. Using a chi-square autowrapper feature in R and the Random Forest algorithm, the number of features (distinct GO-BP terms) with the largest overall accuracy in our validation set is reported. Top 10, 20 (shown), 50 and 100 gene sets (GO-BP terms) were tested with all results similar to those reported – in essence, predicting nearly all subjects as negatives/non-exacerbating. Accuracy values were computed with the "verification"(Hsu and Murphy 1986) package in R.

Classifier algorithm	Virus	#Features (GO-BP terms)	Training performance in the training set								
			Accuracy	Sens	Spec	Prec	AUC	TP	FP	FN	TN
Random Forest	Influenza	20	52.2	8.3	100	100	50.8	1	0	11	11
Random Forest	RSV	20	47.8	0	100	0	85.2	0	0	12	11

TP: True positive; FP: False positive; FN: False negative; TN: True negative

Supplementary Table S4. mRNA-level classifiers fail: leave-one-out cross validation performance of six different transcript-level (mRNA-level) classifiers built on the Zaas et al. training set of nineteen healthy individuals (see Table 1). Using a chi-square autowrapper feature in R, the number of features (distinct mRNAs) offering the largest overall accuracy is reported. With the exception of SVM that performed worse at n=20 mRNAs, all other methods report the same results at n=20.

It is common knowledge that conventional transcript level classifiers derived from human transcriptomes require hundreds of subjects for accuracy. Thus, unsurprisingly they fail in this small dataset. They may be limited by the fact that the biological signal may not be residing in one biological state (whether the transcriptomics expression at baseline or the transcriptomics response upon infection), but in the dynamic response between these two states. Moreover, the integration and combined analysis of different microarray studies remains a challenge [1]. Indeed, levels of expressions are difficult to normalize across experiments. Therefore, technical variations, biological variations and batch effect often biase the predictor's accuracy. More subtle normalization techniques could in theory improve the prediction accuracy, but we did not find a suitable one in this case. For example transforming the gene expression values into ranks/quantiles was also tested but did not improve the predictor's performances.

The Area under the curve (AUC) and significance tests were computed with the "verification"(Hsu and Murphy 1986) package in R. The fully-specified classifier derived from the training set was then applied once on the validation set, and the accuracy of predictions were measured (Simon 2005, Dupuy and Simon 2007).

Classifier	#Features (mRNAs)	Training performance in the training set								
		Accuracy	Sensitivity	Specificity	Precision	AUC	TP	FP	FN	TN
Random Forest	26	52.2	100	0	52.2	40.9	12	11	0	0
Naïve Bayes	26	52.2	100	0	52.2	50	12	11	0	0
Near. Neighbor	26	52.2	100	0	52.2	50	12	11	0	0
SVM	26	52.2	100	0	52.2	50	12	11	0	0
Decision Tree	26	52.2	100	0	52.2	50	12	11	0	0

TP: True positive; FP: False positive; FN: False negative; TN: True negative

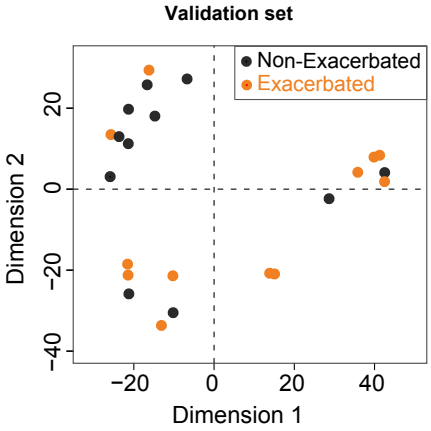
1. Buness, A., et al., *Classification across gene expression microarray studies.* BMC Bioinformatics, 2009. **10**: p. 453.

Table S5. 47 of the genes reproted by the Bjornsdottir exacerbation classifier were geneset members of the pathways identified as features in our classifier, predominantly in the acquired and innate immune categories

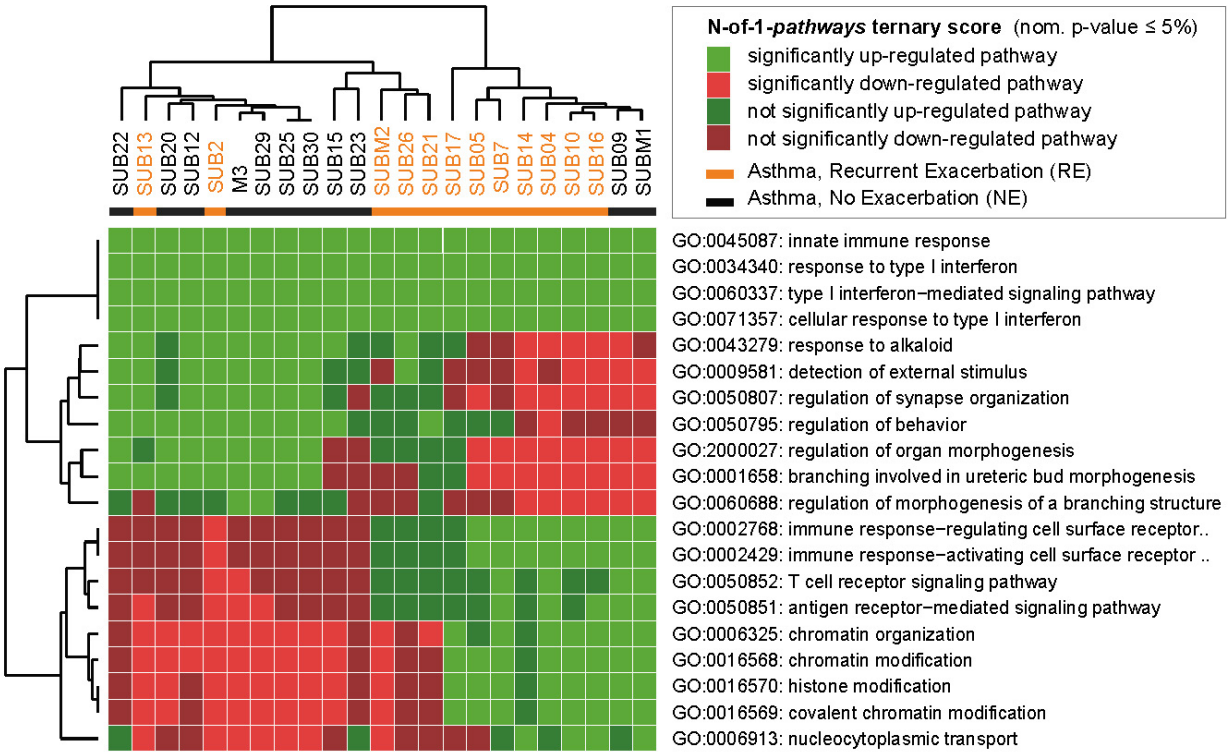
		Acquired Immunity				Innate Immune response				Morpho-genesis				Response to stimulus			Chromatin organization				
	Pathways (our classifier) / **DE genes (Bjornsdottir)***	GO:0002768	GO:0002429	GO:0050851	GO:0050852	GO:0045087	GO:0060337	GO:0034340	GO:0071357	GO:0050807	GO:0001658	GO:0060688	GO:2000027	GO:0043279	GO:0050795	GO:0009581	GO:0006325	GO:0016568	GO:0016569	GO:0016570	GO:0006913
TLR Signaling	TLR1															X					X
	TLR2	X	X													X					X
	TLR4															X					
T Cell Receptor Signaling	LCK	X	X	X	X																
	LAT	X	X	X	X																
	IKBKB	X	X	X	X																
	PIK3CA	X	X	X	X																
B Cell Receptor Signaling	SYK	X	X	X																	
	PIK3CA	X	X	X	X																
	IKBKB	X	X	X	X															X	X
Interferon Signaling	JAK2																	X	X		
	STAT1						X	X	X				X								
	STAT2						X	X	X												
	IRF9						X	X	X												
	IRF1						X	X	X												
	IRF7						X	X	X												
	IFITM1						X	X	X												
	IFI35						X	X	X												
	IFIT1						X	X	X												
	OAS3						X	X	X												
	OASL						X	X	X												
	IFIT2						X	X	X												
	IFI6						X	X	X												
	MX1						X	X	X												
	IFIT3						X	X	X												
	IFITM2						X	X	X												
	MX2						X	X	X												
	OAS2						X	X	X												
	IFITM3						X	X	X												
	ISG20						X	X	X												

	IFI27						X	X	X												
	ISG15						X	X	X												
	GBP2						X	X	X												
	OAS1						X	X	X											X	
	IRF4						X	X	X									X	X		
Interferon Response Factors	OAS1						X	X	X												
	OAS2						X	X	X												
	ISG15						X	X	X												
	STAT2						X	X	X												
	BCL2	X	X	X							X		X	X							
	IFIT3						X	X	X												
	IRF1						X	X	X												
	STAT1						X	X	X				X								
	IRF9						X	X	X											X	X
	JAK2																	X	X		
	HLA-G	X					X	X	X												
	CCL5													X	X						
	PLSCR1	X	X																		
	IRF7						X	X	X											X	
	IRF4						X	X	X									X	X		
	GATA3	X	X	X	X								X								
	IFITM1						X	X	X												
	IFI27						X	X	X												
IL15 Signaling	SP100						X	X	X												
	MX1						X	X	X												
	IFITM1						X	X	X												
	CCR2														X						
	IFI35						X	X	X												
	LYN	X	X	X											X						
	CD40	X																			
	CCL2														X						X
	TLR2	X	X													X					
	IFIT1						X	X	X												X
	SRSF7																				
	CXCR4														X						
	BCL2	X	X	X							X		X	X						X	
	LEF1																	X	X		
	GATA3	X	X	X	X								X								

68 of the differentially expressed genes identified by Bjornsdottier et al did not have an annotation to one of our 20 selected GO-BP terms, and 11 genes did not appear in our version of the master filtered (15-500 genes per geneset) GO annoation file



Supplementary Figure S1. Enriched non-exacerbated asthmatic in the top left quadrant of an PCA analysis of the validation set. The PCA was computed over the N-of-1-pathways scores of the virogram dataset derived from comparing in each pathway of each subject the dynamic change of gene expression between non-induced PBMCs and HRV-induced PBMCs. Each point represents a distinct subject. Dimensions 1 and 2 of the PCA plot suggests a clustering of non-exacerbated asthmatics in the top left quadrant (nominal two-sided hypergeometric p-value = 0.027).



Supplementary Figure S2. Validation set cohort clusters into clinical phenotypes using the top 20 pathways of the classifier selected from the training set cohort. The heatmap, constructed with default unsupervised clustering parameters of the *hclust* function of R, displays the N-of-1-*pathways* scores for the 23 validation patients (columns) corresponding to the 20 pathways (rows) identified as responsive in the *training set* study. Here the ternary states of up, down, or nonl significantly responsive (dysregulated) are represented using four values so that the directionality may also be seen for the nonl significant gene sets. As such, there are four possible values according to the ternary matrix states: 1) the pathway is significantly upl regulated in the virusl exposed sample (light green), 2) the pathway is significantly downl regulated in the virusl exposed sample (light red), 3) the pathway is not found significantly dysregulated (pl value > 5%) but the score indicate a trend of upl regulation (dark green), and 4) the pathway is not found significantly dysregulated (pl value > 5%) but the score indicate a trend of downl regulation (dark red). Of note, the threshold chosen for considering a pathway as significantly responsive is a nominal pl value ≤ 5%.