Prediction of Autism by Translation and Immune/Inflammation Coexpressed Genes in Toddlers From Pediatric Community Practices

Tiziano Pramparo, PhD; Karen Pierce, PhD; Michael V. Lombardo, PhD; Cynthia Carter Barnes, PhD; Steven Marinero, BS; Clelia Ahrens-Barbeau, BS; Sarah S. Murray, PhD; Linda Lopez, MA; Ronghui Xu, PhD; Eric Courchesne, PhD

Importance The identification of genomic signatures that aid early identification of individuals at risk for autism spectrum disorder (ASD) in the toddler period remains a major challenge because of the genetic and phenotypic heterogeneity of the disorder. Generally, ASD is not diagnosed before the fourth to fifth birthday.

Objective To apply a functional genomic approach to identify a biologically relevant signature with promising performance in the diagnostic classification of infants and toddlers with ASD.

Design, Setting, and Participants Proof-of-principle study of leukocyte RNA expression levels from 2 independent cohorts of children aged 1 to 4 years (142 discovery participants and 73 replication participants) using Illumina microarrays. Coexpression analysis of differentially expressed genes between Discovery ASD and control toddlers were used to define gene modules and eigengenes used in a diagnostic classification analysis. Independent validation of the classifier performance was tested on the replication cohort. Pathway enrichment and protein-protein interaction analyses were used to confirm biological relevance of the functional networks in the classifier. Participant recruitment occurred in general pediatric clinics and community settings. Male infants and toddlers (age range, 1-4 years) were enrolled in the study. Recruitment criteria followed the 1-Year Well-Baby Check-Up Approach. Diagnostic judgment followed DSM-IV-TR and Autism Diagnostic Observation Schedule criteria for autism. Participants with ASD were compared with control groups composed of typically developing toddlers as well as toddlers with global developmental or language delay.

Main Outcomes and Measures Logistic regression and receiver operating characteristic curve analysis were used in a classification test to establish the accuracy, specificity, and sensitivity of the module-based classifier.

Results Our signature of differentially coexpressed genes was enriched in translation and immune/inflammation functions and produced 83% accuracy. In an independent test with approximately half of the sample and a different microarray, the diagnostic classification of ASD vs control samples was 75% accurate. Consistent with its ASD specificity, our signature did not distinguish toddlers with global developmental or language delay from typically developing toddlers (62% accuracy).

Conclusions and Relevance This proof-of-principle study demonstrated that genomic biomarkers with very good sensitivity and specificity for boys with ASD in general pediatric settings can be identified. It also showed that a blood-based clinical test for at-risk male infants and toddlers could be refined and routinely implemented in pediatric diagnostic settings.

Published online March 4, 2015.

Copyright 2015 American Medical Association. All rights reserved.
Autism spectrum disorder (ASD) is a neurodevelopmental disorder of complex etiology with early onset and generally is not diagnosed before a median age of 3 years in the United States. Early and long-term intervention is the most effective strategy to reduce or reverse the core features of toddlers and children with autism. While current treatment strategies are rapidly progressing for a more effective intervention in autism and other developmental disorders, the hunt for biological markers or genetic signatures is ongoing. Together with the complex nature of the disorder, the identification of biomarkers or molecular signatures is limited by the inaccessibility of neural tissue and the few young postmortem samples available.

Peripheral blood of living individuals is a preferable and more accessible tissue for such screening. It is expected to carry autism-relevant signatures that can be used to detect the disorder at very young ages and might also reflect aspects of the disrupted biology underlying neural defects.

A few studies have investigated cohorts of blood-derived samples, both in vivo and in vitro, to describe sets of differentially expressed (DE) genes that distinguish individuals with ASD vs control subjects. The largest in vivo study achieved a validated classification accuracy of 68% in children with a mean age of approximately 8 years. Despite these promising efforts, additional studies are needed to support or improve genetic signatures with high specificity and sensitivity and, most important, to push the prediction power at very young ages, when intervention is most effective. In the long run, a practical clinical test will require these signatures to be effective in the general pediatric population and not just in pre-selected syndromic patients or patients with ASD from multiplex families.

Using a systems biology approach, we conducted a proof-of-principle study using leukocyte gene expression aimed to identify a genomic signature that classified with good accuracy 2 independent cohorts of infants and toddlers with ASD (mean age, approximately 2 years) recruited through community pediatric clinics and other community sources. Several genomic signatures are expected to coexist. However, until very large sample sizes of individuals of different genetic backgrounds at young ages become available, our signature represents an unprecedented study outcome that is based on a general pediatric population. With the identification of consistent dysregulated gene pathways and gene sets with predictive roles in ASD, it is expected that biomarker discovery with high specificity is possible and that a blood-based clinical test can be implemented in a routine diagnostic setting.

Methods

Participant Recruitment, Tracking, and Developmental Evaluation

Given the substantial 4:1 male to female bias in ASD and several reasons to suspect that potentially important sex differential factors underlie etiological aspects of autism (eg, in the review by Schaafsma and Pfaff on this topic), we chose to focus on boys only to reduce the potential increase in genomic heterogeneity that would accompany a mixed-sex design. Included in the study were 220 participants aged 1 to 4 years, including 147 toddlers in a discovery sample (91 ASD and 56 control) and 73 toddlers (44 ASD and 29 control) in a replication sample. Sample collection occurred from 2009 to 2011, and diagnostic evaluation occurred from 2009 to 2013. The replication sample largely overlapped with individuals in our previous pilot study of leukocyte gene expression, but the discovery toddlers represent a new and independent sample.

All toddlers were developmentally evaluated by a PhD-level psychologist (C.C.B.), and those younger than 3 years at the time of blood draw were tracked every 6 months until their third birthday, when a final diagnosis was given. Only toddlers with a provisional or confirmed ASD diagnosis were included in the present study. Toddlers were recruited via the 1-Year Well-Baby Check-Up Approach from community pediatric clinics. This approach enables a general naturalistic population screening approach for prospective study of ASD, typically developing patients, and patients with contrast developmental delay (eg, language, global developmental, or motor delay). In this approach, parents of toddlers completed a broadband developmental screen at their pediatrician’s office, and toddlers were referred, evaluated, and tracked over time. This approach provided an unbiased recruitment of toddlers representing a wide range and variety of ability and disability. Blood samples for gene expression and DNA analysis were collected from a subset of participants at the time of referral, regardless of referral reason, and before final diagnostic evaluations. No blood draws were performed if participants showed signs of influenza, a cold, or infections or if any illnesses were present or suspected 72 hours before visits. Every participant was evaluated using multiple tests, including the appropriate module of the Autism Diagnostic Observation Schedule and the Mullen Scales of Early Learning. Diagnoses were determined via these assessments and the DSM-IV-TR. Parents were interviewed using the Vineland Adaptive Behavior Scales and underwent a medical history interview. Both the discovery and replication cohorts included individuals with ASD and control participants. The control group was composed of typically developing toddlers and toddlers with contrast developmental delay (Table). Institutional review board approval from the University of California, San Diego, was obtained for the study. Written informed consent was obtained from the parents of the participants. Additional methodological information is provided in the eMethods in Supplement 1.

Blood Sample Collection, RNA Extraction, and Processing

Four to six milliliters of blood was collected into ethylenediaminetetraacetic-coated tubes from all toddlers, passed over a filter (LeukoLOCK; Ambion) to capture and stabilize leukocytes, and immediately placed in a −20°C freezer. Total RNA was extracted following standard procedures and manufacturer’s instructions (Ambion). The RNA samples in the discovery data set were tested using one platform (HT-12; Illumina), while the RNA samples in the replication data set were tested using another platform (WG-6; Illumina). Several quality cri-


teria were used to exclude low-quality arrays as previously described.\cite{14,20} Five low-quality arrays in the discovery data set were identified and excluded from statistical analyses. All arrays from the replication data set were of high quality. Both the discovery and replication data sets underwent the same filtering and normalization steps. Final samples represented 87 toddlers with ASD and 55 control participants (total, 142 participants) in the discovery cohort and 44 toddlers with ASD and 29 control participants in the replication cohort (Table). Raw and normalized data are deposited in the Gene Expression Omnibus (GSE42133). Additional methodological information is provided in the eMethods in Supplement 1.

**Statistical and Bioinformatic Analysis**

Figure 1 shows a schematic representation of the main statistical and bioinformatics analyses. Statistical analyses were performed on normalized and filtered expression data. Class comparison analysis was performed to identify DE genes using a standard univariate 2-sample t test model with 10,000 random permutations using a software package (BRB-Array Tools; Biometric Research Branch, National Cancer Institute). The significant threshold of univariate tests was .05 (Supplement 2). Differentially expressed genes were then used for enrichment pathway analysis using an available tool (Metacore; GeneGo) and coexpression analysis. A weighted gene network coexpression analysis package\cite{16,17} was used to identify coexpression modules in an unsupervised fashion from DE genes (ie, clusters of DE genes that are tightly coexpressed across all discovery sample participants) and to calculate the first principal component of each module, herein called the module eigengene (ME). The ME is a value that summarizes each module’s expression profile and can be understood as a weighted average of the gene expression profiles within a module.\cite{18}

Coexpression analysis was run by selecting the lowest power for which the scale-free topology fit index reached 0.90 (soft power threshold, 0.5) and by constructing a signed (ie, bidirectional) network with a hybrid dynamic branch-cutting method to assign individual genes to modules.\cite{19} Hypergeometric probability was used to test the significance in gene overlap vs random gene sets of equal size. A software program (CNVision; Yale University) was used to call copy number variations (CNVs) in misclassified individuals with ASD as previously described.\cite{15,20} Additional methodological information, including the differential expression analysis of cortical tissue, is provided in the eMethods in Supplement 1.

**Classification Analysis**

Twelve MEs were obtained from the weighted gene network coexpression analysis of 2765 DE genes in the discovery sample. Feature selection\cite{6} based on the MEs was achieved by running logistic regression on the ME values. We began by first identifying a pair of modules that performed best at distinguishing participants with ASD from control participants, followed by recursively adding one extra module at a time and retaining it if it increased performance. From this process, 4 of 12 modules were identified (M1 is blue, M2 is black, M3 is purple, and M4 is green yellow in eFigure 1 in Supplement 1) that displayed the best area under the curve performance. These 4 modules were further used to test the classification accuracy on the independent replication sample. To validate the accuracy of the classifier on an inde-
A dependent set of samples (replication set), gene weights were computed from the discovery cohort as the correlation between each gene in the 4 modules and their respective eigengene values. Weights were applied to the gene expression data from all available samples were processed to select for high-quality data and stronger gene expression signal. Processed data were first analyzed by differential expression and then by coexpression analysis to reduce dimensionality and leverage characteristics based on network biology. Feature selection and model building were performed on the discovery samples only (training data). The model obtained from the training data was then applied, without any further modifications, to the replication samples (validation data set) to independently test classification performance. Network-based analyses were also applied to the classification signature to identify its functional characteristics. ASD indicates autism spectrum disorder; PPI, protein-protein interaction.
Module-Based Classification Efficiently Distinguishes Individuals With ASD From Control Participants

Class comparison of discovery cohort members identified 2765 unique DE genes, with top enrichment in apoptosis, immune/inflammation response, and translation networks (Figure 2 and Supplement 2). Coexpression analysis identified 12 modules (eFigure 1 in Supplement 1), and eigengenes were calculated for each discovery cohort member and each module. Four modules’ eigengenes were used in the classification analysis together with each individual’s age as a predictor. These modules contained 762 unique genes (423 M1 genes, 191 M2 genes, 90 M3 genes, and 58 M4 genes) (Supplement 2). Logistic regression of diagnosis with age as the predictor produced an odds ratio of 1.07 (95% CI, 1.03-1.12; P < .05), and classification without age was 3% to 4% less accurate. Logistic regression and receiver operating characteristic analyses displayed a high area under the curve in both the discovery cohort (training set on Illumina HT-12) and the replication cohort (independent test set on Illumina WG-6), with 83% and 75% classification accuracy, respectively (Figure 3 and Supplement 2). While the specificity remained high across the different class comparisons, the accuracy and sensitivity decreased as the sample size was reduced (Figure 3). We questioned whether misclassified individuals carried known pathogenic CNVs for failing correct diagnostic prediction. At a 0.5 threshold, 12 of 14 misclassified individuals with ASD were genotyped for CNV analyses. A rare CNV of known ASD etiology, CNTNAP2 duplication, was found in only one individual (eTable in Supplement 1). No clinical, behavioral, or magnetic resonance imaging–based measures indicated subphenotypic differences between correctly classified and misclassified individuals.

Characterization of the Classification Signature

Metacore GeneGo analysis of the 4-module classifier displayed significant enrichment in translation and immune/inflammation genes (Figure 4 and Supplement 2). We sought to independently validate these findings by querying the DAPPLE database (http://www.broadinstitute.org/mpg/dapple), which looks for significant physical connectivity...

Results

Most discovery and replication cohort members were of white race/ethnicity. Pearson χ² test showed no significant difference in racial/ethnic characteristics between individuals with ASD and control participants (χ² = 7.98, P = .16 for the discovery cohort and χ² = 7.19, P = .21 for the replication cohort). After filtering across all discovery cohort members, 12 208 gene probes were used for downstream analyses. Multivariable regression analysis showed no variance explained by differences in racial/ethnic characteristics between individuals with ASD and control participants, and 4% of variance was explained by age (P < .05).

Figure 2. Pathway Enrichment Using Metacore GeneGo of Differentially Expressed Genes From the Discovery Sample

Pathway enrichment of the discovery sample genes

Apoptosis/Apoptotic nucleus
Apoptosis/Domain receptors and caspases
Immune response/Antigen presentation
Immune response/Phagocytosis
Immune response/TCR signaling
Translation/Translation initiation
Inflammation/Interferon signaling
Apoptosis/Anti-apoptosis via NF-κB
Cell adhesion/Leukocyte chemotaxis
Inflammation/IFN-gamma signaling

![Bar chart showing pathway enrichment](chart.png)
among proteins encoded by genes in loci associated with the disorder according to protein-protein interactions (PPIs) reported in the literature. Indeed, these gene modules revealed a statistical enrichment for PPI ($P < .001$). Using this PPI information, we created a classification network to map the genes with the highest number of PPIs. Consistent with enrichment findings of the 762-gene signature, the PPI gene list displayed translation initiation as the top process network ($P = 4e-18$). A substantial number of ribosomal and translation genes were positioned at the center of the PPI network (Figure 5), suggesting that these central genes may have important regulative roles.

To address whether the PPI network also included genes that are potentially relevant to the disrupted biology of the ASD brain, we next performed differential expression analysis (eMethods in Supplement 1) using data from a recent post-mortem tissue study and mapped the cortex-specific DE genes in the PPI network.
overlap (hypergeometric P = 1.92e-7) between brain DE genes and the PPI gene list. Indeed, 45 cortex-specific DE genes were mapped in the network, and 16 (36%) of them were located at the very core (Figure 5 and eFigure 2 in Supplement 1). Of these 45 cortex-specific DE genes within the PPI network, 62% were upregulated, while 38% were downregulated in ASD cortex. Of the 16 genes within the core, 69% were upregulated, and 31% were downregulated in ASD cortex. Pathway enrichment on the full set of DE genes upregulated in ASD cortex showed prominent overlap with processes also observed in our classifier dealing with translation, immune, and cell cycle processes, whereas there was little overlap in enrichment of DE genes downregulated in ASD cortex with enrichment observed in our classifier (eFigure 2 in Supplement 1).

Finally, comparison with recently reported ASD diagnostic classifiers5–6 displayed modest to low overlap at the gene level. In total, 12 of 55 and 18 of 43 reported genes were DE genes in the discovery cohort members, with only 6 genes and 1 gene, respectively, present in our ASD diagnostic classifier (Supplement 2). At the pathway level, translation genes have also been found in a classifier from a recent in vitro study.7

Discussion

Our research design, which used the 1-Year Well-Baby Check-Up Approach, allowed the unbiased prospective recruitment and study of individuals with ASD and control participants as they occur in community pediatric clinics, which has not previously been done by other research groups to our knowledge. Our toddlers with ASD reflect the wide clinical phenotypic range expected in community clinics, while our control toddlers reflect the mix of toddlers commonly seen in community clinics with typical development, mild language delay, transient language delay, and global developmental delay. Against this challenging control group, our signature that was derived from the application of a functional genomic approach correctly identified 83% of discovery toddlers with ASD. This candidate signature performed well in the independent replication cohort despite limitations resulting from the difference in microarray platform, experimental processing used with that cohort, and nearly half the sample size. This very good level of accuracy outperforms other behavioral and genetic screens for infants and toddlers with ASD reported in the lit-
erature, especially compared with the performance of other tests applied to the young general pediatric population (as opposed to preselected syndromic patients or patients with ASD from multiplex families). For example, the Modified Checklist for Autism in Toddlers, a commonly used parent-report screen, has very low specificity (27%)\(^2\) and positive predictive value (11%–54%) when used in general populations.\(^2\)\(^4\)\(^5\)

While important strides have been made to understand possible genetic risk factors in autism,\(^2\)\(^6\) current DNA tests detect only rare autism cases and lack specificity\(^2\)\(^7\) or confirm autism at older ages, and these tests have not been demonstrated to be effective in infants and toddlers with ASD.\(^6\)

Although the present study was designed and performed as proof of principle to stimulate research in the identification of an early biomarker of ASD from a general pediatric population, the candidate functional genomic signature reported herein has shown far greater diagnostic potential than other blood-based or behavior-based candidate classifiers in infants and toddlers with ASD. Nonetheless, the candidate genomic signature we describe represents only the first step toward a practical and accurate first-tier screen. Larger validation studies are needed, as are further studies of the specificity and sensitivity relative to other neurodevelopmental disorders.

The gene list of this new signature has little overlap with the candidate gene list in our group’s previous, small pilot study.\(^3\) This may be due to differences in the overall study design (eg, sample size, sex), feature selection strategy (fold-change vs module-based coexpression), and classification algorithm (machine learning vs standard logistic regression), as well as because of the clinical heterogeneity of the individuals investigated. While the individuals in the replication cohort in the present study largely overlapped with those in the previous pilot study, the module-based signature was derived from an independent, newly collected sample set and was tested on a more recent microarray platform. It is likely that higher concordance of findings between the 2 studies would occur if results were compared at the pathway level rather than the single-gene level. For example, both studies found signals related to genes involved in the immune response pathway. Dysregulation of immune/inflammation mechanisms has been described in a large number of autism studies.\(^2\)\(^8\) To our knowledge, the present study is the first to detect strong significant dysregulation of immune and inflammation gene networks at approximately the age at the first emergence of the clinical risk signs of ASD. Blood cell–derived gene expression studies of older children and adults with ASD also report dysregulation of immune/inflammation genes. Although evidence of immune involvement has been argued to be a later secondary abnormality in ASD, there is no experimental evidence to favor that idea over the possibility that ASD may involve both prenatal immune (eg, maternal immune activation) and genetic factors.\(^2\)\(^9\)\(^3\)\(^0\)

In addition, our signature revealed a central role of protein synthesis in the diagnostic classification of autism. The finding of translation genes at the core in the classification network is a strong reminder of the mechanism underlying the most common single-gene mutation in ASD.\(^3\)\(^1\)–\(^3\)\(^2\) Although it was not the focus of this study to address the level at which protein synthesis is altered, it seems plausible to think that global regulation of translation may be affected. A model of such global dysregulation may well explain the heterogeneity of gene networks and pathways that are involved and disrupted in autism\(^3\)\(^3\) and this model has been proposed for transcription genes in children with idiopathic autism.\(^3\)\(^4\) In contrast, we have not detected a statistically significant signal for synaptic pathways, although de novo mutations of genes with roles in synaptic function and localization have been identified\(^2\)\(^9\) and represent a point of convergence in ASD.\(^3\)\(^6\) We argue that the lack of significant signal for synaptic pathways may be due to the use of blood tissue, which is more likely to reveal changes in genes expressed at the systemic level rather than genes with high expression levels that are specific to neuronal tissue. However, we detected a significant overlap between genes in the PPI classifier network and DE genes in brain tissue from ASD cases, and a substantial proportion of these overlapping genes displayed a high number of PPIs (see the network core in Figure 5). Therefore, these results indicate that among genes in our in vivo blood classifier are those that can have effects on brain development, with functional enrichment in translation, immune, and cell cycle processes. The importance of blood as an in vivo measurement should be underscored because it is of substantial practical importance to the potential advancement of clinical uses for the expression signatures we report herein.

Conclusions

In conclusion, knowledge of these common pathways and changes in hub gene connectivity patterns will facilitate research into biological targets for biotherapeutic intervention. The findings will aid the development of accurate biomarkers for detecting risk for ASD among infants in the general pediatric population.
Drafting of the manuscript: Pramparo, Pierce, Lombardo, Xu, Courchesne.

Critical revision of the manuscript for important intellectual content: Pierce, Lombardo, Xu, Courchesne.

Statistical analysis: Xu.

Administrative, technical, or material support: Pierce, Carter Barnes, Ahrens-Barbeau, Murray, Lopez.

Study supervision: Pierce, Xu, Courchesne.

Conflict of Interest Disclosures: Drs Pramparo and Courchesne reported a patent pending that includes data from this study. No other disclosures were reported.

Funding/Support: This research was supported by grants P50-MH081775 and R01-MH036840 (Dr Courchesne), R01-MH080134 (Dr Pierce), 1U54RR025204-01 (Mr Marinerio), and 1ULIRR031980-01 (Dr Xu) from the National Institutes of Health and by grant KL2TR000899 from the University of California, San Diego Clinical and Translational Research Institute (Dr Pramparo).

Role of the Funder/Sponsor: The funding sources had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication.

Additional Contributions: Roxana Hazin, PhD, at the UC San Diego Autism Center of Excellence helped with participant recruitment. We thank all the families for making this study possible.

REFERENCES


