Published in final edited form as:

Autism Res. 2009 April ; 2(2): 67–77. doi:10.1002/aur.72.

Novel clustering of items from the Autism Diagnostic Interview-Revised to define phenotypes within autism spectrum disorders

Valerie W. Hu¹ and Mara E. Steinberg^{1,2,*}

¹ Dept. of Biochemistry and Molecular Biology, The George Washington University Medical Center, 2300 Eye St., N.W., Washington, DC 20037

² Dept. of Hearing and Speech Sciences, 0100 Lefrak Hall, Univ. of Maryland; College Park, MD 20742

Abstract

Heterogeneity in phenotypic presentation of ASD has been cited as one explanation for the difficulty in pinpointing specific genes involved in autism. Recent studies have attempted to reduce the “noise” in genetic and other biological data by reducing the phenotypic heterogeneity of the sample population. The current study employs multiple clustering algorithms on 123 item scores from the Autism Diagnostic Interview-Revised (ADI-R) diagnostic instrument of nearly 2000 autistic individuals to identify subgroups of autistic probands with clinically relevant behavioral phenotypes in order to isolate more homogeneous groups of subjects for gene expression analyses. Our combined cluster analyses suggest optimal division of the autistic probands into 4 phenotypic clusters based on similarity of symptom severity across the 123 selected item scores. One cluster is characterized by severe language deficits, while another exhibits milder symptoms across the domains. A third group possesses a higher frequency of savant skills while the fourth group exhibited intermediate severity across all domains. Grouping autistic individuals by multivariate cluster analysis of ADI-R scores reveals meaningful phenotypes of subgroups within the autistic spectrum which we show, in a related (accompanying) study, to be associated with distinct gene expression profiles.

Keywords

ADI-R; multivariate cluster analyses; ASD phenotypes

Introduction

Autism spectrum disorders (ASD) are developmental disabilities resulting from dysfunction in the central nervous system and are characterized by impairments in three behavioral areas: communication (notably spoken language), social interactions, and repetitive behaviors or restricted interests (Volkmar et al., 1994). ASD usually manifest before three years of age and the severity can vary greatly. Idiopathic ASD include autism, which is considered to be the most severe form, pervasive developmental disorders not otherwise specified (PDD-NOS), and Asperger’s syndrome, a milder form of autism in which persons can have relatively normal intelligence and communication skills but still experience great difficulty with social interactions. ASD with defined genetic etiologies or chromosomal

* Corresponding author: Valerie Hu, Ph.D., Dept. of Biochemistry and Molecular Biology, The George Washington University Medical Center, 2300 Eye St., N.W., Washington, DC 20037. Email: bcmvwh@gwumc.edu.

Contributions to this manuscript VWH conceived of and designed the approach to ASD phenotype analyses and wrote the manuscript. MES assisted in the analyses and contributed to manuscript preparation.

aberrations include Rett's syndrome, tuberous sclerosis, Fragile X syndrome, and chromosome 15 duplication (reviewed in (Muhle, Trentacoste, & Rapin, 2004)). Familial studies provide evidence that individuals closely related to an autistic individual (i.e. mother, father, and siblings) may have "autistic tendencies" but do not meet criterion for ASD, suggesting that a broad autism phenotype (BAP) may also exist (Piven, Palmer, Jacobi, Childress, & Arndt, 1997).

Previous studies establish a strong genetic component for the etiology of autism, and many loci have been proposed as autism susceptibility regions, including loci on chromosomes 1, 2, 7, 11, 13, 15, 16, 17 (reviewed in (Gupta & State, 2007; Polleux & Lauder, 2004; Santangelo & Tsatsanis, 2005; Yonan et al., 2003)). However, the specific genes involved within each locus have not been determined to date. Available data further suggests that multiple gene interactions, epigenetic factors, and environmental risk factors may also be at the core of autism etiology (del Gaudio et al., 2006; Geschwind, 2008; Herbert et al., 2006; Jiang et al., 2004; Lathe, 2006; Varki, Geschwind, & Eichler, 2008).

Heterogeneity in phenotypic presentation of ASD has been offered as one explanation for the difficulty in pinpointing chromosomal loci and genes involved in autism. Thus, recent studies have attempted to reduce the "noise" in genetic data by reducing the phenotypic heterogeneity of the sample population using a variety of approaches. Some of the earlier studies stratified samples for genetic analyses primarily on language deficits of the proband (eg., age at first word, phrase speech delay), while other studies focused on other attributes of autistic disorder, such as compulsions, or Restricted and Repetitive Stereotyped Behaviors (RRSB) to restrict phenotypic heterogeneity (Alarcon, Cantor, Liu, Gilliam, & Geschwind, 2002; Bradford et al., 2001; Hollander et al., 2000; Silverman et al., 2001). Another strategy for increasing the probability of observing genetic linkage was based upon the use of "endophenotypes" for specific autism-associated behaviors which were present in nonaffected family members (Spence et al., 2006). Using this approach, Alarcon et al. and Chen et al. reported quantitative trait loci (QTL) for language and nonverbal communication deficits, respectively (Alarcon, Yonan, Gilliam, Cantor, & Geschwind, 2005; Chen, Kono, Geschwind, & Cantor, 2006).

The Autism Diagnostic Interview-Revised (ADI-R) is a comprehensive assessment instrument for ASD which is a clinician-administered interview that probes for language, social, behavioral, and functional abnormalities that are inconsistent with a specific child's stage of development (Lord, Rutter, & Couteur, 1994). Principal components analysis (PCA) of 98 items from the Autism Diagnostic Interview-Revised (ADI-R) has also been used as a means to isolate genetically relevant phenotypes (Tadevosyan-Leyfer et al., 2003). This study identified 6 "factors" which accounted for 41% of the variation in the autistic population studied. Reexamination of genetic data from individuals defined by presence or absence "savant skills" (one of the factors) and from their respective family members showed an increase in LOD score (0.4 → 2.6) in the chromosome 15q11-q13 region relative to the combined unsegregated sample population (Nurmi et al., 2003). However, this finding could not be replicated by another group (Ma et al., 2005). Recent analyses of the use of the ADI-R to increase phenotypic homogeneity summarize the major studies which have attempted to stratify autism samples and further caution that such stratification based upon just a few defined attributes can also lead to unintended associations with other variables, such as age, gender, race, etc. (Hus, Pickles, Cook Jr., Risi, & Lord, 2007; Lecavalier et al., 2006).

In this paper, we demonstrate the use of multiple clustering methods applied to a broad range of ADI-R items from a large population (1954 individuals) to identify subgroups of autistic individuals with clinically relevant behavioral phenotypes. We further select

individual male samples based on these cluster methods for gene expression analyses, demonstrating that the selected samples are indeed representative of the clusters identified within the broader autistic population, and cover a broad range in terms of age and symptom severity of ASD. In the accompanying manuscript, we show that the selected lymphoblastoid cell lines derived from individuals who fall within 3 of the phenotypic subgroups show distinct differences in gene expression profiles that in part relate to the severity of the phenotype. Functional and pathway analyses of the gene expression data also suggest distinct differences in the biological phenotypes that associate with these subgroups.

Methods

Analysis of data from ADI-R questionnaires to identify phenotypic subgroups

ADI-R score sheets were downloaded for 1954 individuals with autism from the Autism Genetic Research Exchange (AGRE) phenotype database. The gender and age profile of the individuals whose score sheets were used were as follows: 1526 males [age range: 1.85 – 47.68 yrs; mean age: 8.3 yrs; median age: 7.2 yrs]; 428 females [age range: 2.04 – 44.63 yrs; mean age: 8.15 yrs; median age: 7.12 yrs]. A total of 123 items that were identical or comparable on both 1995 and 2003 versions of the ADI-R were included. Following the example of Tadevosyan-Leyfer et al. (2003), “current” and “ever” scores were used for most of these items to provide some redundancy in the data and increase the robustness of the symptomatic profile of each individual. Only items scored numerically (0 = normal; 3 = most severe) were incorporated into our analyses. A score of 8 for items in the spoken language subgroup indicated that the items were not applicable because of insufficient language and was replaced with a rating of 3. Scores of 8 or 9 for other items (excluding those from the spoken language subgroup), which indicated the item was not asked or not applicable, were replaced with blanks to reflect that no information was available for that item. A score of 1 or 2 on item 19 (LEVELL) indicated an overall language deficit and, as a result, scores for items 20-28 were assigned a score of 3 to reflect impaired language skills, as previously done by others (Tadevosyan-Leyfer et al., 2003). Items with a score of 4 for the savant skills, which meant that the individual possessed an isolated though meaningful skill/knowledge above that of his general functional level or the population norm, were replaced with 3 to maintain consistency of the 0-3 scale across all items. Scores of 7 for some items were changed to a score between 0 and 3 depending on the nature of the question and how it reflected severity with respect to that specific item. A score of -1 indicated missing data (according to AGRE) and was replaced with a blank. It should be noted that the missing scores were random among clusters and did not appear to be an obvious factor in the cluster analyses. Supplementary Table 1 summarizes the score modifications for each item used in our cluster analyses of autistic individuals.

Data from ADI-R score sheets for 1954 individuals were loaded into MeV (Saeed et al., 2003), a software program created by John Quackenbush and colleagues to analyze microarray gene expression data. Each individual is represented by a horizontal row in the data matrix while ADI-R items are represented by vertical columns. Multiple clustering analyses were employed to subgroup individuals on the basis of similarity of ADI-R item scores, and included principal components analysis (PCA), hierarchical clustering (HCL), and k-means clustering (KMC), which is a “supervised” clustering method, for which the number of clusters (K) is specified. A fitness of merit (FOM) analysis (Yeung, Haynor, & Ruzzo, 2001) was also conducted to estimate the optimal number of clusters, while correspondence analysis (COA) was used to visualize the association of specific items with the different clusters of individuals. A description of each of these analytical methods is summarized by Saeed et al. (Saeed et al., 2003)

Selection of samples for large-scale gene expression analyses

Lymphoblastoid cell lines (LCL) for DNA microarray analyses were selected on the basis of phenotypic clustering of autistic individuals using the methods described above. As described in the results, the application of multiple clustering algorithms to the selected ADI-R items from scoresheets of 1954 individuals resulted in 4 reasonably distinct phenotypic subgroups. Samples were selected from 3 of the 4 groups for gene expression analyses. These groups included those with severe language impairment, those with milder symptoms across all domains, and those defined by presence of notable savant skills. The intermediate group was not included because we first wished to test the concept that the extreme phenotypes of ASD (severe and mild) could be distinguished by gene expression profiling. The savant phenotype was included for gene expression analyses not only because savant skills are of general interest, but also because they were a dominant feature of the third principal component in the PCA analysis of probands (data not shown). Because we wanted to reduce the heterogeneity of subjects for our gene expression studies on *idiopathic* autism, we chose to exclude all probands whose autism could be attributed to a known genetic cause (Fragile X, chromosome 15 duplication, Rett's Syndrome) and to avoid confounding factors due to diagnosed comorbid conditions (OCD, bipolar, etc.) or prematurity. In a previous and separate study on autistic-nonautistic male siblings (manuscript submitted), we observed differential expression of genes involved in steroid hormone biosynthesis (particularly androgens) and, wishing to avoid the complication of hormonal (gender) effects on gene expression, excluded females from the gene expression studies. Clearly, females with ASD need to be studied as well. Interestingly, separate cluster analysis of the ADI-R scores of male and female subjects were very similar, suggesting that they exhibit much of the same behavioral/functional phenotypes as males. In addition, a score < 80 on the Peabody Picture Vocabulary Test (PPVT) was used to confirm language deficits for those in the group identified by cluster analysis as having severe language impairment. For the accompanying gene expression study, 26-31 cell lines were obtained for each study group, along with 29 cell lines from "control" individuals who were nonautistic siblings of individuals with autism, matched roughly in age to the autistic probands, the majority of which were unrelated to the controls. In this study, we also applied cluster analyses to the ADI-R scores of the ASD individuals whose LCL were selected for gene expression analysis and demonstrate that applying our exclusion criteria did not change the cluster assignment for these samples. Supplementary Table 2 provides a demographic profile of the subjects selected for gene expression analyses which includes pedigree, age, race, ethnicity as well as standard PPVT and Raven's scores.

Results

To reduce the phenotypic heterogeneity of autism for gene expression analyses, we applied several different clustering methods to the scores from ADI-R questionnaires (from the AGRE database) describing 1954 autistic individuals. For these analyses, we selected 123 item scores that covered a broad spectrum of behaviors and functions in order to identify phenotypic subgroups of individuals with idiopathic ASD who were characterized by combined symptoms across multiple domains. These domains included language, nonverbal communication, social interactions, play skills, interests and behaviors, physical sensitivities and mannerisms, aggression, and savant skills. The specific items and score adjustments are shown in Supplementary Table 1.

Principal components analysis of the subjects based on their ADI-R scores shows separation of the autistic individuals into 2 main clusters, but did not clarify the phenotypic nature of each group of subjects (Fig. 1A). Hierarchical clustering (HCL) was therefore performed to obtain a broader sense of the structure of the ASD population as revealed by their respective scores on the selected ADI-R items. This analysis clearly shows separation of the

individuals into more than 2 clusters, based upon symptomatic profile across the different items (Fig. 1B). A Figure of Merit (FOM) analysis which was employed to estimate the optimal number of clusters for supervised clustering analysis (Fig. 1C) suggested 3-5 clusters. We then performed K-means clustering of the subjects using each of these K-values (3-5), and concluded that 4 clusters gave optimal separation of recognizable phenotypes (Fig. 2A). For example, one group is characterized by severe language deficits (samples within this group were assigned the color Red for ease of individual identification), while another group (Blue) exhibits milder symptoms across the domains, as indicated by more black in the matrix, reflecting ADI-R severity scores of 0 (normal). A third group (Yellow) possesses noticeable savant skills, which are represented by the last 12 columns on the right of the score matrices, while the fourth group (Green) exhibited intermediate severity across the domains, but with relatively lower frequency of savant skills. When the subgroup color coding from the KMC analyses was applied to the graph obtained by principal components analysis (Fig. 1A), a clear, though not perfect, separation among the groups is observed (Fig. 2B). It is worth noting that the first 3 components of the PCA capture 38% of the variation among the samples (with 42% represented within the first 4 components). These results indicate that there is a large amount of variability in the ADI-R data as is evident from the many small branches in the hierarchical cluster which collectively contributes to the residual variance beyond that accounted for by the first 3-4 principal components. Yet, the separation of the ASD population into severely language impaired, intermediate, mild, and highly savant phenotypes is quite clear. A correspondence analysis (COA) of the data further suggests that specific clusters of items (e.g., savant skills, aggression, or ritualistic behaviors/resistance to change) are more strongly associated on the basis of higher severity scores with individuals in certain subgroups than in others (Fig. 3) and Table 1. For example, savant skills (turquoise squares) associate more with the “savant” (yellow) and mild (blue) individuals, while severe deficits in spoken language, nonverbal communication, and social skills (pink squares) are concentrated in the group with severe language impairment (red). Not surprisingly, we also see association of circumscribed interests and unusual preoccupations (lavender squares) with the mild (blue) and “savant” (yellow) groups, but these behavioral traits are also clustered with compulsions and ritualistic behaviors (Table 1). Interestingly, aggression and physiological symptoms (lime-colored squares) are associated with individuals exhibiting the mild (blue) phenotype of ASD.

Based upon these combined clustering methods, we selected LCL from individuals represented in 3 of the 4 phenotypic groups for gene expression analyses. These groups included those with severe language impairment, those with a milder phenotype (~40% of whom had clinical diagnoses of Asperger’s Syndrome or PDD-NOS), and those with notable savant skills. Because of the relatively low number of individuals in the “savant” category once other exclusion criteria were applied, we selected a few samples from the group with severe language impairment who also exhibited high scores on a majority of savant skills. It should be pointed out that those with savant skills were a minor fraction of the group with severe language impairment. Principal components and K-means cluster analyses of the ADI-R item scores for the individuals selected for the microarray studies confirm the separation of the selected samples into 4 phenotypic groups as described in the figure legend (Figs. 4A and 4B), with the fourth phenotypic group representing individuals with severe language deficits and savant skills (depicted by orange color in Fig. 4A).

Figure 5 shows the sum of ADI-R scores across all of the items used in this study for the selected individuals, as well as the sum of item scores specific for different functional domains. As shown by the inset within several of the graphs, the group selected for gene expression analysis typically mirrors that of the 1954 individuals from the repository (inset), suggesting that the selected individuals were phenotypically representative of the general autistic population. The profiles for other functional domains (e.g., nonverbal

communication, play skills, restricted interests and behaviors) are similar to that representing the sum of all items, for all the individuals in the repository as well as the ones selected for microarray analyses. The average of item scores for each group across the items in each domain as well as the group averages of combined ADI-R scores across all items also confirms the phenotypic distinction among the groups (Figs. 6 and 7). Although there is no significant difference between the average of the sums of the ADI-R scores for the mild (blue) and savant (yellow) groups, the ADI-R score profiles in Figure 4B as well as in Fig. 7 show that there are indeed quantitative differences in severity among the phenotypic groups across multiple functional/behavioral domains, with the savant group showing lower severity scores than the mild group for almost all items except for savant skills. It is also interesting to note that while individuals in the mild AS group (blue) exhibit lower severity scores in the language domain, most of their scores in the social, nonverbal, and play categories are nearly as severe as those for individuals with severe language impairment (red), suggesting that higher language abilities do not necessarily correlate well with improved social skills (Fig. 4B and Fig. 7).

Discussion

The primary goal of this study was to develop a method of directly clustering autistic probands according to similarity of severity scores across a broad range of behavioral and functional symptoms probed by the ADI-R in order to reduce the heterogeneity of samples for biological (specifically, gene expression) analyses. In this respect, our study differs from many other studies which have attempted to analyze the factor structure of the ADI-R (Constantino et al, 2004; Georgiades et al., 2007; Tadevosyan et al, 2003; Van Lang et al., 2006). These studies are comprehensively discussed in a recent study by Snow et al. who performed factor analyses on a majority of ADI-R items from scoresheets of both verbal and non-verbal autistic probands and concluded that autistic symptomatology can be best described by a two-domain model (Snow, Lecavalier, & Houts, 2008). The items comprising the larger of the two domains (Factor I) correspond to the items associated with the “pink” cluster in Fig. 3 (this study) which include all of the spoken language and nonverbal/social communication items (Table 1). The items comprising the second domain (Factor II) correspond roughly to the items identified by the “lavender” cluster in Fig. 3 and include restricted/repetitive behaviors, circumscribed interests, unusual preoccupations, and stereotypes (Table 1). Interestingly, the correspondence analysis (COA), which associates clusters of items with groups of individuals on the basis of severity scores, shows an association between the latter set of items (Factor II) with the mild and savant groups of ASD. In addition to identifying item clusters that correlate with Factors I and II in the Snow study, our COA also identified 2 additional clusters of items that appear to separate distinctly from those of the other 2 clusters (Fig. 3). One of these item clusters involves the savant skills items (turquoise colored squares in Fig. 3) which were not included by Snow et al. (Snow, Lecavalier, & Houts, 2008), but which were identified as a factor in the ADI-R by Tadevosyan-Leyfer et al. (2003). A subsequent genetic linkage analysis by Nurmi et al. (2003) based upon this factor demonstrated the value of subdividing the autistic probands and their families according to the savant phenotype. Not surprisingly, the items in this cluster associate most strongly with the savant and mild ASD individuals in our study. The fourth cluster of items in Fig. 3 (lime colored squares), which has not been extensively explored in the context of autism, includes items related to aggression and self-injury, and associates predominantly with a minority of individuals exhibiting the mild ASD phenotype. It will be of further interest to explore the co-expression and significance of these behavioral traits in specific phenotypes of ASD in future studies.

Another recent study by Rapin et al. argues that there are 2 major subtypes of language disorder in autistic children, differentiated mainly by impaired expressive phonology, with

each subtype subdivided by comprehension ability (Rapin, Dunn, Allen, Stevens, & Fein, 2009). In this respect, the group with low phonology may be comparable to the group that we identify as “severely language impaired”, although there is no direct comparison of items analyzed. Indeed, we show in the accompanying article on gene expression analyses of several subtypes of ASD defined by this study, that the subgroup with severe language impairment exhibits the most differentially expressed genes relative to nonautistic controls and is the only subtype with significant dysregulation of circadian rhythm genes.

In contrast to the many studies which have sought to identify discrete phenotypes of autism which can be used to reduce heterogeneity for biological studies, Ring et al. have recently proposed a continuous gradient model in which the differences between autistic individuals is more quantitative than qualitative (Ring, Woodbury-Smith, Watson, Wheelwright, & Baron-Cohen, 2008). The study of Constantino et al. which shows a continuum in terms of a range of deficits based on scores on the Social Responsiveness Scale and ADI-R (Constantino et al., 2004) may also be used in support of this gradient concept. Our gene expression analyses of several of the ASD phenotypes identified here through cluster analyses of autistic probands based on their ADI-R scores, essentially offers support for both the discrete phenotype and the gradient models by identifying sets of genes that are differentially expressed either quantitatively or qualitatively among ASD subgroups relative to controls. These results respectively provide evidence for genes (common to 2 or more groups) responsible for core deficits across the spectrum that differ mainly in symptom severity as well as for genes (unique to a given subgroup) which implicate the involvement of different metabolic and/or signaling pathways among the phenotypes.

With respect to diagnosis of autism, the ADI-R is one of the most widely used and comprehensive diagnostic instruments for autism (Lord et al., 1994) and, to many, represents the “gold standard” for identifying individuals with ASD. However, it is only administered after a child presents with abnormal development (e.g., delayed speech) or aberrant behaviors, which typically is noticed between the ages of 2 and 3. Although many studies are currently attempting to identify even earlier signs of abnormal social development (e.g., lack of eye contact, pointing, or shared attention in toddlers (Landa, Holman, & Garrett-Mayer, 2007)), there is still a need to identify definitive molecular markers of ASD that may be used to screen for autism even earlier (pre- or postnatally) as well as to provide targets for therapeutic intervention. We have therefore embarked upon a series of studies to identify expressed biomarkers of ASD through the use of large-scale gene expression analyses. Because ADI-R scores are the most widely available phenotypic data for the majority of autistic children (particularly within the AGRE repository), we sought to use the information in this test instrument as a starting point to subdivide diagnosed individuals for genomics analyses. We demonstrate in the accompanying manuscript that subgrouping of autistic individuals by multivariate cluster analysis of ADI-R scores which captures the breadth of the disorder within each individual reveals meaningful subgroups or phenotypes of idiopathic autism that can be separated from controls as well as distinguished from each other by gene expression profiling. Detailed bioinformatics analyses of the differentially expressed genes from the resulting subgroups reveal similarities as well as differences in pathways and functions associated with the different ASD phenotypes. Based on these combined and complementary analyses, we suggest that multivariate analysis of the ADI-R data using a broad spectrum of the ADI-R items and a combination of clustering methods that are typically employed in DNA microarray analyses may be an effective means of reducing the phenotypic heterogeneity of the sample population without restricting the phenotype to only one or a few items which, as pointed out by others (Hus et al., 2007; Lecavalier et al., 2006), may associate coincidentally with other variables. Such an approach towards stratification of individuals, which utilizes the full spectrum of autism-associated behaviors and can be easily tailored to include additional relevant scored items, is

expected to aid in the association of genetic and other biological factors with specific forms of idiopathic autism. Finally, we suggest that similar cluster analyses of scored behavioral and functional evaluations may also be useful in reducing the heterogeneity of other complex, heterogeneous psychiatric disorders for genetic and other biological analyses.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Grant sponsors: National Institute of Mental Health, NIH, Grant # R21 MH073393 (VWH); Autism Speaks, Grant # 2381 (VWH)

We gratefully acknowledge the resources provided by the Autism Genetic Resource Exchange (AGRE) Consortium^{*} and the participating AGRE families. The Autism Genetic Resource Exchange is a program of Autism Speaks and is supported, in part, by grant 1U24MH081810 from the National Institute of Mental Health to Clara M. Lajonchere (PI).

We especially thank Dr. Vlad Kustanovich of AGRE for providing us with additional information about the samples which were not easily retrievable in the database.

References

- Alarcon M, Cantor RM, Liu J, Gilliam TC, Geschwind DH. Evidence for a language quantitative trait locus on chromosome 7q in multiplex autism families. *Am J Hum Genet.* 2002; 70(1):60–71. [PubMed: 11741194]
- Alarcon M, Yonan AL, Gilliam TC, Cantor RM, Geschwind DH. Quantitative genome scan and ordered-subsets analysis of autism endophenotypes support language QTLs. *Mol Psychiatry.* 2005; 10(8):747–57. [PubMed: 15824743]
- Bradford Y, Haines J, Huthcheson H, Gardiner M, Braun T, Sheffield V, et al. Incorporating language phenotypes strengthens evidence of linkage to autism. *American Journal of Medical Genetics - Neuropsychiatric Genetics.* 2001; 105(6):539–547. [PubMed: 11496372]
- Chen GK, Kono N, Geschwind DH, Cantor RM. Quantitative trait locus analysis of nonverbal communication in autism spectrum disorder. *Mol Psychiatry.* 2006; 11(2):214–20. [PubMed: 16189504]
- Constantino JN, Gruber CP, Davis S, Hayes S, Passanante N, Przybeck T. The factor structure of autistic traits. *Journal of Child Psychology and Psychiatry and Allied Disciplines.* 2004; 45(4):719–726.

*The AGRE Consortium:

Dan Geschwind, M.D., Ph.D., UCLA, Los Angeles, CA;
Maja Bucan, Ph.D., University of Pennsylvania, Philadelphia, PA; W.Ted Brown, M.D., Ph.D., F.A.C.M.G., N.Y.S. Institute for Basic Research in Developmental Disabilities, Long Island, NY;
Rita M. Cantor, Ph.D., UCLA School of Medicine, Los Angeles, CA;
John N. Constantino, M.D., Washington University School of Medicine, St. Louis, MO; T.Conrad Gilliam, Ph.D., University of Chicago, Chicago, IL;
Martha Herbert, M.D., Ph.D., Harvard Medical School, Boston, MA;
Clara Lajonchere, Ph.D., Cure Autism Now, Los Angeles, CA;
David H. Ledbetter, Ph.D., Emory University, Atlanta, GA;
Christa Lese-Martin, Ph.D., Emory University, Atlanta, GA;
Janet Miller, J.D., Ph.D., Cure Autism Now, Los Angeles, CA;
Stanley F. Nelson, M.D., UCLA School of Medicine, Los Angeles, CA;
Gerard D. Schellenberg, Ph.D., University of Washington, Seattle, WA;
Carol A. Samango-Sprouse, Ed.D., George Washington University, Washington, D.C.;
Sarah Spence, M.D., Ph.D., UCLA, Los Angeles, CA;
Matthew State, M.D., Ph.D., Yale University, New Haven, CT.
Rudolph E. Tanzi, Ph.D., Massachusetts General Hospital, Boston, MA.

- del Gaudio D, Fang P, Scaglia F, Ward PA, Craigen WJ, Glaze DG, et al. Increased MECP2 gene copy number as the result of genomic duplication in neurodevelopmentally delayed males. *Genet Med*. 2006; 8(12):784–92. [PubMed: 17172942]
- Geschwind DH. Autism: Many genes, common pathways? *Cell*. 2008; 135(3):391–395. [PubMed: 18984147]
- Georgiades S, Szatmari P, Zwaigenbaum L, Duku E, Bryson S, Roberts W, et al. Structure of the autism symptom phenotype: A proposed multidimensional model. *J Am Acad Child Adolesc Psychiatry*. 2007; 46(2):188–96. [PubMed: 17242622]
- Gupta AR, State MW. Recent advances in the genetics of autism. *Biol Psychiatry*. 2007; 61(4):429–37. [PubMed: 16996486]
- Herbert MR, Russo JP, Yang S, Roohi J, Blaxill M, Kahler SG, et al. Autism and environmental genomics. *Neurotoxicology*. 2006; 27(5):671–84. [PubMed: 16644012]
- Hollander E, Novotny S, Allen A, Aronowitz B, Cartwright C, DeCaria C. The relationship between repetitive behaviors and growth hormone response to sumatriptan challenge in adult autistic disorder. *Neuropsychopharmacology*. 2000; 22(2):163–167. [PubMed: 10649829]
- Hus V, Pickles A, Cook EH Jr, Risi S, Lord C. Using the autism diagnostic interview-revised to increase phenotypic homogeneity in genetic studies of autism. *Biological Psychiatry*. 2007; 61(4):438–448. [PubMed: 17276746]
- Jiang Y, Sahoo T, Michaelis RC, Bercovich D, Bressler J, Kashork CD, et al. A mixed epigenetic/genetic model for oligogenic inheritance of autism with a limited role for UBE3A. *American Journal of Medical Genetics*. 2004; 131 A(1):1–10. [PubMed: 15389703]
- Landa RJ, Holman KC, Garrett-Mayer E. Social and communication development in toddlers with early and later diagnosis of autism spectrum disorders. *Archives of General Psychiatry*. 2007; 64(7):853–864. [PubMed: 17606819]
- Lathe, R. Autism, brain, and environment. Jessica Kingsley Publishers; London: 2006.
- Lecavalier L, Aman MG, Scahill L, McDougle CJ, McCracken JT, Vitiello B, et al. Validity of the autism diagnostic interview-revised. *American Journal on Mental Retardation*. 2006; 111(3):199–215+228. [PubMed: 16597187]
- Lord C, Rutter M, Couteur AL. Autism diagnostic interview-revised: A revised version of a diagnostic interview for caregivers of individuals with possible pervasive developmental disorders. *Journal of Autism and Developmental Disorders*. 1994; 24(5):659–685. [PubMed: 7814313]
- Ma DQ, Jaworski J, Menold MM, Donnelly S, Abramson RK, Wright HH, et al. Ordered-subset analysis of savant skills in autism for 15q11-q13. *American Journal of Medical Genetics - Neuropsychiatric Genetics*. 2005; 135 B(1):38–41.
- Muhle R, Trentacoste SV, Rapin I. The genetics of autism. *Pediatrics*. 2004; 113(5):e472–86. [PubMed: 15121991]
- Nurmi EL, Dowd M, Tadevosyan-Leyfer O, Haines JL, Folstein SE, Sutcliffe JS. Exploratory subsetting of autism families based on savant skills improves evidence of genetic linkage to 15q11-q13. *Journal of the American Academy of Child and Adolescent Psychiatry*. 2003; 42(7):856–863. [PubMed: 12819446]
- Piven J, Palmer P, Jacobi D, Childress D, Arndt S. Broader autism phenotype: Evidence from a family history study of multiple-incidence autism families. *American Journal of Psychiatry*. 1997; 154(2):185–190. [PubMed: 9016266]
- Polleux F, Lauder JM. Toward a developmental neurobiology of autism. *Ment Retard Dev Disabil Res Rev*. 2004; 10(4):303–17. [PubMed: 15666334]
- Rapin I, Dunn MA, Allen DA, Stevens MC, Fein D. Subtypes of language disorders in school-age children with autism. *Dev Neuropsychol*. 2009; 34(1):66–84. [PubMed: 19142767]
- Ring H, Woodbury-Smith M, Watson P, Wheelwright S, Baron-Cohen S. Clinical heterogeneity among people with high functioning autism spectrum conditions: Evidence favouring a continuous severity gradient. *Behavioral and Brain Functions*. 2008; 4:11. [PubMed: 18289376]
- Saeed AI, Sharov V, White J, Li J, Liang W, Bhagabati N, et al. TM4: A free, open-source system for microarray data management and analysis. *BioTechniques*. 2003; 34(2):374–378. [PubMed: 12613259]

- Santangelo SL, Tsatsanis K. What is known about autism: Genes, brain, and behavior. *Am J Pharmacogenomics*. 2005; 5(2):71–92. [PubMed: 15813671]
- Silverman JM, Smith CJ, Schmeidler JM, Buxbaum JD, Lawlor BA, Fitzgerald M. Symptom domains in autism and related conditions: Evidence for familiarity. *American Journal of Medical Genetics - Neuropsychiatric Genetics*. 2001; 105(7):593.
- Snow A,V, Lecavalier L, Houts C. The structure of the Autism Diagnostic Interview-Revised: diagnostic and phenotypic implications. *J. of Child Psychology and Psychiatry*. 2008; (Dec. 9) [*E-pub ahead of print*].
- Spence SJ, Cantor RM, Chung L, Kim S, Geschwind DH, Alarcón M. Stratification based on language-related endophenotypes in autism: Attempt to replicate reported linkage. *American Journal of Medical Genetics, Part B: Neuropsychiatric Genetics*. 2006; 141(6):591–598.
- Tadevosyan-Leyfer O, Dowd M, Mankoski R, Winklosky B, Putnam S, McGrath L, et al. A principal components analysis of the autism diagnostic interview-revised. *Journal of the American Academy of Child and Adolescent Psychiatry*. 2003; 42(7):864–872. [PubMed: 12819447]
- van Lang NDJ, Boomsma A, Sytema S, de Bildt AA, Kraijer DW, Ketelaars C, et al. Structural equation analysis of a hypothesised symptom model in the autism spectrum. *Journal of Child Psychology and Psychiatry and Allied Disciplines*. 2006; 47(1):37–44.
- Varki A, Geschwind DH, Eichler EE. Human uniqueness: Genome interactions with environment, behaviour and culture. *Nature Reviews Genetics*. 2008; 9(10):749–763.
- Volkmar FR, Klin A, Siegel B, Szatmari P, Lord C, Campbell M, et al. Field trial for autistic disorder in DSM-IV. *Am J Psychiatry*. 1994; 151(9):1361–7. [PubMed: 8067493]
- Yeung KY, Haynor DR, Ruzzo WL. Validating clustering for gene expression data. *Bioinformatics*. 2001; 17(4):309–18. [PubMed: 11301299]
- Yonan AL, Palmer AA, Smith KC, Feldman I, Lee HK, Yonan JM, et al. Bioinformatic analysis of autism positional candidate genes using biological databases and computational gene network prediction. *Genes Brain Behav*. 2003; 2(5):303–20. [PubMed: 14606695]

Fig. 1.

A) Graphical representation of the results of a principal components analysis (PCA) of 1954 autistic probands based on their severity scores on 123 items from their respective ADI-R scoresheets which were obtained from the AGRE repository. The X, Y, and Z axes represent the first, second, and third principal components, respectively. Each point on the graph represents an individual whose location on the graph is determined by his respective severity profile on the selected ADI-R items. B) Hierarchical clustering (HCL) analysis of the 1954 individuals based on similarity of profile of 123 ADI-R item scores. Each row represents an individual while each column represents an item. The intensity of the red color correlates with severity on an item, with the most intense red representing a severity score of 3. Black denotes a score of “0”, which is considered “normal”. Gray denotes the absence of data for a particular item. Missing data was generally sporadic and was not an obvious factor in the cluster analyses. C) A Figure of Merit (FOM) analysis, which was used to estimate the optimal number of clusters or subgroups of individuals as determined by similar severity profiles across the 123 items, suggested approximately 3-5 clusters. We then performed K-means cluster analyses on the data using $K = 3, 4$, or 5 , and decided, after viewing the results, that the 4 cluster solution gave the best separation of recognizable phenotypes based on the selected ADI-R items (see Fig. 2A).

Fig. 2.

A) K-means cluster (KMC) analysis of the dataset described in Fig. 1 using $K = 4$. Visual inspection of the ADI-R data matrix in the 4 clusters reveals a subgroup (designated “Red”) which has severe deficits in spoken language (represented by the broad, bright red band of columns), a subgroup that has a milder phenotype across most items (blue) as indicated by a greater number of black items (severity score = 0), a group with a notable higher frequency of savant skills (yellow), and a group which we considered intermediate in severity (green). B) Superposition of the colors for each individual onto the respective points on the PCA graph (from Fig. 1A) shows that the clearly separated group of individuals are those with severe language deficits while the other cluster can be divided into the mild, “savant”, and intermediate phenotypes.

Fig. 3.

A) A 3-d graphical representation of a correspondence analysis (COA) which associates different clusters of items (shown as colored squares) with different individuals (shown as colored points corresponding to colors assigned to the KMC clusters), on the basis of severity scores of the items within the phenotypic clusters. B) A 2-d representation of the same data as in Fig. 3A. Supplemental Table 3 lists the ADI-R items which separate into 4 distinct clusters (turquoise — savant skills; lime — aggression and physiological symptoms; lavender — compulsions and rituals; pink — verbal and nonverbal communication and social skills).

Fig. 4.

A) Principal components analysis (PCA) of ADI-R item scores from 87 autistic individuals whose lymphoblastoid cell lines were selected for gene expression profiling (see accompanying manuscript). The red points on the graph represent individuals with severe language deficits; blue points represent those with mild symptoms; yellow points represent those with notable savant skills; and orange points represent those with savant skills and severe language deficits. B) Centroid graphs from a K-means cluster (KMC) analysis of the item scores from the 87 individuals mentioned above in which the graphs for each group show the average scores for each of the 123 items used in this analysis. In this figure, the items corresponding to language items are shaded red, those corresponding to social skills are shaded green, and those identifying savant skills are shaded yellow. It is clear from both PCA and KMC analyses that the 4 groups are phenotypically different on the basis of the severity of ADI-R scores for the selected items.

Fig. 5.

Graphical displays of the sum of ADI-R item scores for each of the individuals whose lymphoblastoid cell line was selected for gene expression analyses. One panel shows the sum of the scores for each individual across all 123 items, while the other 3 panels show the sum of the scores for the items pertaining only to language skills, savant skills, and aggression. The inset in each graph shows the same data for 1954 autistic individuals whose ADI-R scoresheets were available from the AGRE repository at the time this study was conducted. This data is shown to demonstrate that the overall pattern of scores across all or selected items for the individuals whose gene expression profiles were evaluated paralleled that of the repository population, with the exception of the savant skills, since we deliberately selected a larger proportion of “savant” samples for our analyses to achieve statistically significant microarray data.

Fig. 6.

This graph shows the average ADI-R scores across the 123 items for each phenotypic group analyzed as well as the average cumulative score for the entire group (white bar). Red: severely language impaired; Orange: savants with severe language impairment; Blue: mild ASD; Yellow: savants. * $p \leq 0.05$; ** $p \leq 0.001$.

Fig. 7. Average ADI-R scores for specific items within functional categories for the 4 different subgroups of individuals whose LCL were selected for gene expression profiling. A) Average item scores for language skills, social development, interests and behaviors and savant skills for the severely language impaired (red), mild ASD (blue), savant (yellow), and language impaired + savant (orange) groups. B) Average item scores for nonverbal communication, play skills, physical sensitivities and mannerisms, and aggression for the 4 phenotypic groups.

Table 1

Clusters of associated items identified by correspondence analysis (COA) of the ADI-R data for 1954 individuals from the AGRE repository

Clusters of ADIR items from COA				
(turquoise)	(lime)	(lavender)	(pink)	
CVISSPZ	GAIT5	CSTEREO	LEVELL	CSHOW
EVISSPZ	CAGGFAM	ESTEREO	CCOMPSL	SHOW5
CMEMZ	EAGGFAM	CUNPROC	COMPSL5	COSH ARE
EMEMZ	CAGGOTH	EUNPROC	CUSEBOD	OSHA5
CMUSICZ	EAGGOTH	CCIRINT	EUSEBOD	CSHARE
EMUSICZ	CSLFINJ	ECIRINT	CARTIC	SHARE5
CDRAWZ	ESLFINJ	CCRIT	ARTICF5	COCOMF
EDRAWZ	CHVENT	ECRIT	CCHAT	OCOMF5
CREADZ	EHVENT	CNOISE	CHAT5	CQUALOV
EREADZ	CFAINT	ENOISE	CCONVER	QUALOV5
CCOMPUZ	EFAINT	CABINR	CONVER5	CRFACEX
ECOMPUZ		EABINR	CINAPPQ	RFACEX5
		CCHANGE	EINAPPQ	CINAPFE
		ECHANGE	CPRON	EINAPFE
		CRESIS	EPRON	CQRESP
		ERESIS	CNEOID	QRESP5
		CUATT	ENEOID	CINITIA
		EUATT	CVERRIT	INITIA5
		CMLHAND	EVERRIT	CSOPLA Y
		EMLHAND	CINR	SOPLAY5
			EINR	CINTCH
			CSPEECH	INTCH5
			SPEECH5	CRESPCH
			CPOINT	RESPCH5
			POINT5	CGRPLAY
			CNOD	GRPLAY5
			NOD5	CFRIEND
			CHSHAKE	FREND15
			HSHAKE5	CSOCDIS
			CINSGES	SOCDIS5
			INSGES5	CUSEOBJ
			AVOICE5	EUSEOBJ
			CIMIT	CUNSENS
			IMIT5	EUNSENS
			CPLAY	CHFMAN
			PLAY5	EHFMAN
			CPEERPL	COTHMAN

Clusters of ADIR items from COA				
(turquoise)	(lime)	(lavender)	(pink)	
			PEERPL5	EOTHMAN
			GAZE5	CGAIT
			CSSMILE	
			SSMILE5	