Genes, Brain and Behavior
Official publication of the International Behavioural and Neural Genetics Society

Genes, Brain and Behavior (2014) **13:** 276–285 doi: 10.1111/gbb.12117

Genetically meaningful phenotypic subgroups in autism spectrum disorders

O. J. Veatch[†,*], J. Veenstra-VanderWeele[‡], M. Potter[†], M. A. Pericak-Vance[§] and J. L. Haines[†]

[†]Center for Human Genetics Research, [‡]Departments of Psychiatry, Pediatrics, and Pharmacology, Vanderbilt University Medical Center, Nashville, TN, USA, and [§]Hussman Institute for Human Genomics, Miller School of Medicine, University of Miami, Miami, FL, USA
**Corresponding author: O. J. Veatch, Center for Human Genetics Research, Vanderbilt University Medical Center, 519 Light Hall Nashville, TN 37232, USA. E-mail: olivia.j.veatch@vanderbilt.edu*

Autism spectrum disorder (ASD) is a complex neurodevelopmental disorder with strong evidence for genetic susceptibility. However, the effect sizes for implicated chromosomal loci are small, hard to replicate and current evidence does not explain the majority of the estimated heritability. Phenotypic heterogeneity could be one phenomenon complicating identification of genetic factors. We used data from the Autism Diagnostic Interview-Revised, Autism Diagnostic Observation Schedule, Vineland Adaptive Behavior Scales, head circumferences, and ages at exams as classifying variables to identify more clinically similar subgroups of individuals with ASD. We identified two distinct subgroups of cases within the Autism Genetic Resource Exchange dataset, primarily defined by the overall severity of evaluated traits. In addition, there was significant familial clustering within subgroups (odds ratio, OR ≈ 1.38–1.42, $P < 0.00001$), and genotypes were more similar within subgroups compared to the unsubgrouped dataset (Fst = 0.17 ± 0.0.0009). These results suggest that the subgroups recapitulate genetic etiology. Using the same approach in an independent dataset from the Autism Genome Project, we similarly identified two distinct subgroups of cases and confirmed this severity-based dichotomy. We also observed evidence for genetic contributions to subgroups identified in the replication dataset. Our results provide more effective methods of phenotype definition that should increase power to detect genetic factors influencing risk for ASD.

Keywords: ASD, autism spectrum disorders, biomarkers, diagnosis, differential, genetics, multivariate, phenotypes, phenotypic subgroups, statistical analyses

Received 9 August 2013, revised 21 October 2013 and 16 December 2013, accepted for publication 18 December 2013

Autism spectrum disorder (ASD, OMIM 209850) is characterized by impairments in social communication and the presence of restricted and repetitive behavioral patterns (American Psychiatric Association 2012). Within this unified definition, the severity of clinical presentation is quite variable (Persico & Bourgeron 2006). Overwhelming evidence suggests strong genetic susceptibility factors underlying ASD (Bailey *et al.* 1995; Cook 2001; Hallmayer *et al.* 2011). However, it has been difficult to identify any individual genetic factors that confer even moderate risk and replicate in multiple ASD cohorts. One hypothesis is that the wide variability in clinical manifestation can be explained by underlying genetic heterogeneity (Bruining *et al.* 2010; Geschwind 2011; Hu *et al.* 2009). Identification of more phenotypically homogeneous subgroups of ASD may allow detection of more replicable genetic mechanisms conferring larger risk effects.

Various attempts have been made to reduce heterogeneity in large-scale genetic studies of ASD. One approach is to analyze genetic data from individuals who meet Diagnostic and Statistical Manual-IV (DSM-IV) criteria for strict autistic disorder separately from those meeting only some criteria (i.e. DSM-IV pervasive developmental disorder-not otherwise specified, or Asperger disorder; American Psychiatric Association 2000; Anney *et al.* 2010; Liu *et al.* 2001; Yonan *et al.* 2003). While this dichotomous categorization of ASD has advanced our knowledge of potential genetic risk factors, via detection of statistically associated and/or linked chromosomal regions, it has still not implicated genetic variants with replicable effects (Anney *et al.* 2012). In fact, the change in criteria between the DSM-IV to the new DSM-5 was motivated by the lack of empirical data supporting separate disorders within the autism spectrum, highlighting the need for practical approaches to identifying phenotypic subgroups even within more refined diagnostic categories (Frazier *et al.* 2012; Snow *et al.* 2009).

When attempting to identify subgroups within ASD, previous studies have emphasized the importance of evaluating multiple sources of behavioral data (Funalot *et al.* 2004; Georgiades *et al.* 2013; Leyfer *et al.* 2006; Veenstra-Vanderweele *et al.* 2004). These data have been used previously to identify between two and four phenotypically distinct ASD subgroups. Categories used to distinguish these subgroups were severe, moderate, or mild ASD or severe intellectual disabilities (Eaves *et al.* 1994; Fein *et al.* 1999; Prior *et al.* 1998; Sevin *et al.* 1995; Siegel *et al.* 1986; Spiker *et al.* 2002; Stevens *et al.* 2000; Wiggins *et al.* 2012). The most consistent findings across these different studies are subgroups defined as either high- or low-functioning based on the level of symptom severity and some measure of intellectual capability. When age at exam is controlled for, fewer distinct clusters are identified and functional level (as indicated by non-verbal IQ, Wing Autistic Disorder Interview Checklist, Peabody Picture Vocabulary Tests and Vineland

Adaptive Behavior Scales) stands out as a distinct identifier of subgroups (Fein *et al.* 1999; Stevens *et al.* 2000). Despite these data, most studies have not evaluated whether there are genetic contributions to these phenotypic subgroups. One notable exception is a study where subsequent genetic analyses were performed in subgroups defined by cluster analysis of the Autism Diagnostic Interview-Revised (ADI-R; Hu & Steinberg 2009; Hu *et al.* 2009, 2011). Novel genetic factors were associated with distinct subgroups, providing further support for phenotypic subgroups being genetically meaningful (Hu *et al.* 2011). It is notable that in this case cluster analysis was limited to a single source of behavioral information.

Many previous subgrouping efforts also lacked ascertainment of biomarkers or comorbidities commonly observed in individuals with ASD. As quantitative traits that are associated with ASD but not required for diagnosis, this data may improve our ability to identify more genetically similar subgroups(Gottesman & Gould 2003; Hu *et al.* 2009; Leyfer *et al.* 2006; Veenstra-Vanderweele & Cook 2004; Wiggins *et al.* 2012). For example, multiple studies have implicated the same chromosomal region, 7q35, and candidate gene, *CNTNAP2*, by focusing on expression of specific language impairment (SLI) in ASD. This parallels findings in isolated SLI (Alarcon *et al.* 2008; El-Fishawy & State 2010; Peter *et al.* 2011; Vernes *et al.* 2008; Whitehouse *et al.* 2011). Other comorbidities of potential interest to ASD studies include macrocephaly, or substantially increased head growth compared to typically developing individuals. Head circumference (HC) is of particular interest as a potential biomarker in ASD because larger HC has been associated with overall ASD symptom severity and developmental regression (Chaste *et al.* 2013). Mutations in the same gene (*PTEN*) have also been consistently identified in numerous ASD cases defined by marked comorbid macrocephaly, indicating HC is a genetically meaningful biomarker in ASD (Butler *et al.* 2005; Varga *et al.* 2009). Head circumference is also monitored during most pediatrician visits throughout the first year of life, allowing this trait to be more easily incorporated into large-scale phenotype analyses (Veenstra-Vanderweele *et al.* 2004).

We hypothesized that subgrouping individuals with ASD using multiple sources of behavioral and biomarker data would create more genetically meaningful phenotype definitions, increasing power to detect genetic effects in future studies. We used novel applications of multivariate statistics to explore these data in two large, publicly available ASD datasets.

Methods

Dataset demographics

The discovery dataset consisted of individuals in the Autism Genetic Resource Exchange (AGRE) family based study (Geschwind *et al.* 2001). Individuals not meeting DSM-IV criteria (American Psychiatric Association 2000) for an ASD diagnosis on both of the two main diagnostic instruments, the Autism Diagnostic Observation Schedule (ADOS; Gotham *et al.* 2009; Lord *et al.* 1989) and the ADI-R (Lord *et al.* 1994) were excluded. We also excluded individuals with potentially non-idiopathic autism (e.g. known neurogenetic disorders, known chromosomal abnormalities, prematurity <35 weeks). The final discovery dataset had 1261 ASD cases, who were between 2 and 21 years old at the time of the ADI-R. Genetic ancestry as determined by Structure (Pritchard *et al.* 2000) was 73% European American (EA), 17.8% Mexican American, 2.7% African American and 6.5% unknown due to missing genome-wide data. This dataset was 80% male, and 95% of the families were multiplex.

The dataset used for replication consisted of individuals in the Autism Genome Project (AGP; Hu-Lince *et al.* 2005). This dataset was comprised of 2563 ASD cases who met ASD criteria on both the ADI-R and ADOS, and were between 2 and 21 years old at the time of the ADI-R. Genetic ancestry was 64.6% European American, 3% Mexican American, 2% African American and 30.4% unknown due to missing genome-wide data. This dataset was 84% male and 54% of the cases were from multiplex families. The de-identified IDs for the final datasets are available in Table S1, Supporting Information.

Phenotype data comparisons

We included social, communication and restricted repetitive behavior (RRB) domain scores from the ADOS and the ADI-R. The ADOS is a semi-structured observational assessment of individuals suspected of having ASD (Lord *et al.* 1989). The ADI-R is an interview given by trained experts to caregivers of individuals suspected of having ASD (Lord *et al.* 1994). The ADI-R generates scores in each of four content areas that are used in diagnosis. We also included domain scores from the fourth area, 'abnormality of development evident at or before 36 months' (DevAb). The ADI-R communication measure is divided into verbal and non-verbal scores. These are measured on different, finite scales, making direct comparisons of the severity of communication deficits between non-verbal and verbal individuals inappropriate. Since every person receives a non-verbal score but not a verbal score, and these scores were strongly correlated ($\rho_{AGRE} = 0.86$, $\rho_{AGP} = 0.88$), we only incorporated non-verbal scores. We also included domain standard scores for socialization, communication, daily living skills and motor skills from the Vineland Adaptive Behavior Scales (VABS; Sparrow & Cicchetti 1985; Sparrow *et al.* 1984, 2005) for evaluation of adaptive function- an important distinguishing factor in ASD (Bolte & Poustka 2002; Matson & Shoemaker 2009). Ages at exam for all behavioral tests were also incorporated into analyses. We evaluated HCs taken at one time point, as an indicator of either developmental or persistent macrocephaly. Raw HC measures were obtained directly from publicly available data provided by the AGRE and AGP. HCs reported by AGRE are measurements taken three times, the two most similar are maintained and averaged to compute a final HC (Davis *et al.* 2013). The AGP provides body measurement information collected at various international sites. These are uploaded by each site to the AGP database. For this dataset, 31% of HCs were acquired during a visit and 69% were acquired via medical records. We excluded HC measures taken when individuals were <1 month old. We generated *z*-scores for HCs by standardizing for age and sex using a typically developing population (Roche *et al.* 1987). Unfortunately, the VABS and HC data were missing in 25–46% of the samples across these datasets; however, the statistical methods we used allow for and are robust to missing data. We chose to retain cases missing VABS and HC information as these are not considered ASD-specific diagnostic criteria. Details on availability of phenotypes of interest are provided in Table S2.

Traits included in our analyses represent different types of statistical variables including ordinal, finite variables (i.e. ADI-R, ADOS), continuous, finite variables (i.e. VABS) and continuous, infinite variables (i.e. age at exam, HC). To allow more comparable measures, we transformed variables into Hazen percentile ranks using Stata 11.2 (College Station, TX, USA; Hazen 1914; Statacorp 2009). The ADOS has available four modules used for assessment, based on the individual's age and current developmental level. The same evaluated trait often has a different domain score range depending on the module used for assessment. Therefore, ADOS domain scores were modified prior to percentile rank calculations to be comparable across the four possible modules by reducing raw values to that of the module with the smallest scale for each domain. VABS data were ranked inversely to account for the inverse relationship of these severity scores when compared to the other diagnostic methods used in analyses. We determined

the correlation structure by calculating pairwise Spearman's rank correlation coefficients (ρ; STATA 11.2; Table S3a).

Multivariate statistical analyses

Since many variables were correlated, we developed a weighting scheme for input variables to ensure that inter-correlated phenotype data did not overly influence results. If a variable was correlated with another variable at $\rho \geq 0.50$ those variables were weighted to allow for only partial variable contributions to analyses. It is notable that the strongest correlations observed for the motor skills domain standard scores from the VABS are with the communication domain standard scores at $\rho = 0.49$, which did not quite meet our threshold for non-independence. The correlations observed by the VABS developers for the motor skills domain standard scores indicated dependence on the communication domain standard score ($\rho = 0.56–0.61$; Sparrow *et al.* 2005). As such, we chose to incorporate only a partial weight for motor skills domain standard scores in our analyses. The cumulative number of variables incorporated into analyses using this weighting scheme equaled eight variables (Fig. S1).

Prior to clustering, Gower dissimilarity matrices were calculated using percentile ranked data with the 'FD' package in R (Laliberte & Legendre 2010; Laliberte & Shipley 2011). Seven different clustering methods (i.e. k means, agglomerative hierarchical, model-based, partitioning around medoids, divisive hierarchical, self-organizing tree algorithm and clustering large applications) were evaluated for internal validity(Handl *et al.* 2005) and stability(Walesiak & Dudek 2007) while partitioning the dissimilarity matrix into anywhere from 2 to 15 clusters using the 'CLVALID' package in R (Brock *et al.* 2008). The final clustering solution was performed with the most valid method, agglomerative hierarchical, using the 'CLUSTER' package in R (Maechler *et al.* 2012). Validity of the final clustering solution was further assessed by permuting phenotype data across individuals, clustering the permuted data and calculating the Adjusted Hubert-Arabie Rand index (AHARI) to compare clustering of the real data to the permuted data (Hubert & Arabie 1985). This was carried out for 1000 data permutations and the AHARIs were averaged.

Sensitivity analyses were performed to help identify input variables that were most important to definition of the final cluster solutions. We removed one variable at a time, reapplied weights to account for the missing variable and clustered, as described above using the 'CLVALID' package in R. Kruskal–Wallis tests were performed, *post-hoc*, to evaluate main cluster and subcluster-level differences in the distributions of untransformed input variables, using STATA 11.2 (Statacorp 2009).The proportion of variance accounted for by each of the input variables was computed directly from the reported Chi-square (χ^2) value using the following equation: $\text{eta}^2 = (\chi^2/(N_{obs}-1))$.

To help determine if clusters were phenotypically different on input variables not used to define cluster assignments, we also performed clustering following removal of all variables where data was missing for a portion of individuals (i.e. VABS, HCs).

We examined potential predictors of main cluster and subcluster membership by performing multinomial logistic regression in STATA 11.2. While all variables were used as input for clustering, the clustering methods we used retained cases with missing VABS and/or HC data in analyses. Logistic regression, however, handles missing data by removing individuals who are missing any of the input variables from the analysis. Therefore, main cluster and subcluster assignments were regressed on each input variable, adjusting only for other input variables that were correlated in these datasets. The Wald Chi-square statistic was then calculated to test the effect of each input variable on assignment to main clusters and subclusters (Mavandadi *et al.* 2009).

Genetic contribution to cluster assignment

We estimated genetic relationships using single nucleotide polymorphism (SNP) markers previously genotyped in these datasets. Autosomal markers were pruned based on the linkage disequilibrium structure of founders. We set an r^2 threshold of 0.16, within a 500-SNP window, sliding five SNPs at a time (Purcell *et al.* 2007). We subsequently created pedigree files with only cases included in our cluster datasets. Wright's *F*-statistic (Fst) was then calculated using PLATO (Grady *et al.* 2010). We grouped individuals into subpopulations based on cluster assignment. For each genetic marker, the correlation between individuals drawn from the subpopulation relative to the total population was determined. We then took the average Fst calculated across all informative markers.

Intra-cluster family structure was also evaluated by calculating the odds of individuals being assigned to the same cluster given a familial relationship. We calculated an odds ratio (OR) via the Chi-square statistic. 'Case' status was defined as a full sibling relationship and 'exposure' was defined as assignment to the same phenotype cluster. Each individual was manually scored for the number of full sibling relationships in the dataset. Since there are substantially more unrelated individuals than related in the datasets, we randomly sampled groups of unrelated individuals representing the same number of available familial relationships. We calculated an OR for related 'cases' and each randomly sampled unrelated 'control' group. This was performed 10 times. The reported ORs represent the range for these calculations.

Results

Discovery (AGRE) dataset

Multivariate statistical analyses

The overall best validity scores were calculated when using agglomerative hierarchical clustering to group the AGRE dataset into two clusters, one with 443 cases and one with 818 cases (Fig. 1; Table S4). The agglomerative coefficient was 0.78, suggesting that a strong clustering structure was identified. The average AHARI showed that the final clustering solution of the real data was significantly different than clustering permuted datasets ($\text{AHARI} = -6.14\text{x}10^{-5}$). Kruskal–Wallis tests indicated that all input variable distributions, except ADI-R RRB and HC, were significantly different between these two main clusters (Table 1). Examination of summary statistics for the input variables by cluster showed that individuals with scores indicating increased severity for social and communication deficits assessed on all instruments (i.e. ADI-R, ADOS and VABS), ADOS-assessed RRBs, earlier evidence of onset (i.e. ADI-R DevAb), and who were younger at the time of the exams were placed into the larger cluster (Table S5). We also note that this larger cluster had proportionally more individuals with greater severity scores on the overall calibrated severity scores from the ADOS ($z_{\text{Mann–Whitney}} = 7.90$, degrees of freedom; $\text{df} = 1,185$, $P < 0.00001$). These scores are used as measures of the breadth of ASD severity (Gotham *et al.* 2009) and were not included as input variables for clustering. Thus, we referred to the larger cluster as 'more severe' and the smaller cluster as 'less severe'. Increasing severity in social and communication scores from the ADI-R and ADOS, RRB scores from the ADOS, DevAb scores from the ADI-R, and communication and motor skills scores from the VABS were all associated with membership in the 'more severe' cluster (Table S6a). Sensitivity analyses showed that ADI-R DevAb scores had the overall largest effect on main cluster stability. The remaining input variables had similarly modest effects (Table S7). Wright's Fst indicated that genotype frequencies were more similar within the two main clusters than in the entire unclustered dataset. Also, familial relationships were significantly associated with assignment to the two main phenotype clusters (Table 2).

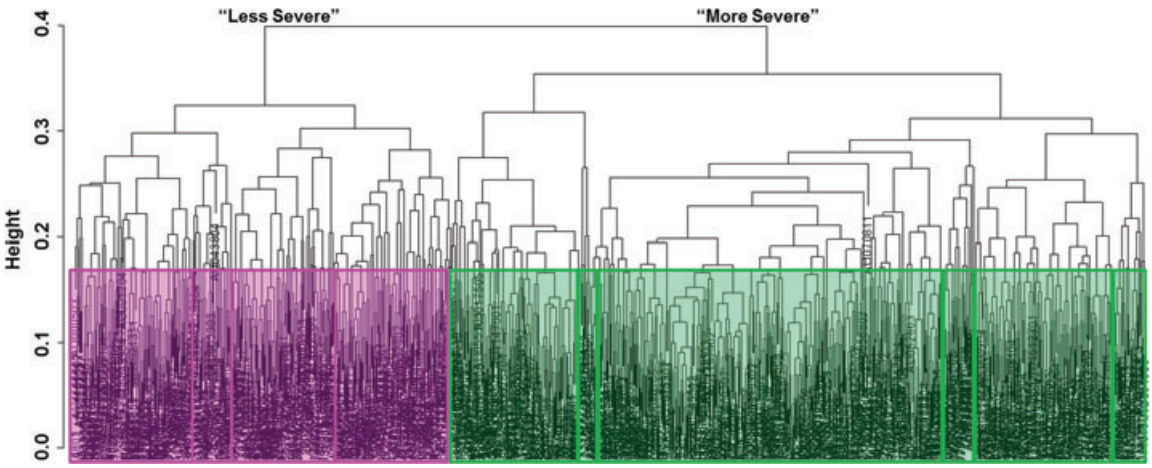

Figure 1: Agglomerative clustering AGRE dataset. Height indicates distance between merging clusters at successive stages of clustering and is related to dissimilarities among clusters. The main cluster highlighted in green represents individuals with more severe ASD phenotypes compared to individuals assigned to the main cluster highlighted pink. Subclusters are indicated with corresponding boxes.

Table 1: Cluster differences in the AGRE dataset

	Main clusters			Subclusters		
Phenotype variable	Chi^2 (df = 1)	*P*-value	eta^2	Chi^2 (df = 9)	*P*-value	eta^2
ADI-R Social	167.42	< 0.0001	$1.33E-01$	322.77	< 0.0001	$2.56E-01$
ADI-R Comm	176.61	< 0.0001	$1.40E-01$	386.11	< 0.0001	$3.06E-01$
ADI-R RRB	***0.05***	***0.8324***	***3.97E−05***	***461.90***	***<0.0001***	***3.67E−01***
ADI-R DevAb	786.18	< 0.0001	$6.24E-01$	1145.27	< 0.0001	$9.09E-01$
ADOS Social	185.94	< 0.0001	$1.48E-01$	457.91	< 0.0001	$3.63E-01$
ADOS Comm	131.45	< 0.0001	$1.04E-01$	355.32	< 0.0001	$2.82E-01$
ADOS RRB	203.29	< 0.0001	$1.61E-01$	644.55	< 0.0001	$5.12E-01$
VABS Social	102.71	< 0.0001	$1.08E-01$	242.46	< 0.0001	$2.54E-01$
VABS Comm	160.47	< 0.0001	$1.68E-01$	308.29	< 0.0001	$3.23E-01$
VABS MotorSkills	140.49	< 0.0001	$1.56E-01$	221.53	< 0.0001	$2.46E-01$
VABS DailyLiving	129.33	< 0.0001	$1.36E-01$	246.33	< 0.0001	$2.59E-01$
HC	***0.05***	***0.8258***	***7.37E−05***	***40.71***	***<0.0001***	***6.00E−02***
ADI-R Age	18.86	< 0.0001	$1.50E-02$	253.22	< 0.0001	$2.01E-01$
Ethnicity*	3.89	0.0486	$3.09E-03$	29.32	0.0006	$2.33E-02$
Sex*	0.03	0.8692	$2.38E-05$	11.38	0.2507	$9.03E-03$
ADOS CSS*	62.33	< 0.0001	$4.95E-02$	209.46	< 0.0001	$1.66E-01$

Kruskal–Wallis comparisons of variable distributions between the two main clusters and across the ten subclusters. All input variable distributions, except ADI-R RRB and HC, were significantly different between the main clusters. ADI-R and HC distributions were significantly different across subclusters and are indicated in bold italics.
ADOS CSS, calibrated severity score; df, degrees of freedom.
*Information not used as input variables.

The next best validity scores were calculated when using agglomerative hierarchical clustering to further subgroup the two main clusters into 10 subclusters; the 'more severe' main cluster into six, and the 'less severe' main cluster into four. Kruskal–Wallis tests showed that the previously non-significant HCs and ADI-R RRBs were very different across these 10 subclusters (Table 1; Table S12a). Head circumference distributions were statistically different across the four subclusters comprising the 'less severe' main cluster ($P = 0.0034$) and the six subclusters comprising the 'more severe' main cluster ($P < 0.0001$). ADI-R RRB score distributions were also statistically different between the four subclusters comprising the 'less severe' main cluster and the six subclusters comprising the 'more severe' main cluster ($P < 0.0001$). Sensitivity analyses showed that, with the exception of the DevAb scores from the ADI-R, removal of any other input variable had similar effects on stability of the subclusters (Table S7). Input variables that were significantly associated with membership in each of the 10 subclusters are detailed in Table S6b and c.

Table 2: Results Evaluating Genetics Underlying Cluster Assignments. (a) Odds ratios represent increased odds of cases being assigned to the same cluster given a familial relationship. (b) Average Wright's *F*-statistic (Fst) across informative autosomal markers comparing cluster subpopulations to total unclustered population suggest individuals with more similar genetic architecture cluster together

(a) Odds of same cluster assignment given sibling relationship

Dataset	Odds ratio range	*P*-value
AGRE	1.38–1.42	<0.00001
AGP	1.19–1.35	<0.00001

(b) *F*-Statistic comparing clusters to unclustered dataset

Dataset	Mean Fst	Standard error	95% Conf. interval
AGRE	0.1664	9.13×10^{-4}	(0.1646, 0.1682)
$AGRE_{EA}$	0.1281	7.97×10^{-4}	(0.1265, 0.1296)
AGP	0.1251	7.53×10^{-4}	(0.1236, 0.1266)
AGP_{EA}	0.1031	6.86×10^{-4}	(0.1018, 0.1045)

Fst are reported for the entire clustering dataset and the European Americans (EA) only. Frequency (f) of full sibling relationships in main clusters used for odds ratio calculations are: fAGRE = 0.91; fAGP = 0.38.

Upon removal of input variables with some level of missing data, we observed that VABS domain standard scores were still different between main clusters and across subclusters. Head circumferences were no longer different across subclusters (Table S8). AGRE also had available non-verbal IQ scores, as measured by the Raven Coloured Matrices Scale, for 44% of the evaluated dataset (Raven 1956). Non-verbal IQ scores were therefore evaluated as external correlates. Distributions between AGRE main clusters and across subclusters were significantly different, and the average non-verbal IQ was lower in the 'more severe' main cluster (Table S11).

Replication (AGP) dataset

We tested for replication in the independent, non-overlapping AGP dataset. We observed a similar correlation structure for input variables as in the AGRE dataset (Table S3b).

Agglomerative hierarchical clustering grouped the AGP dataset into two main clusters and 15 subclusters (Fig. S2). Cases with increased severity for most variables tended to group into the larger main cluster ($n = 1527$) compared to the smaller main cluster ($n = 1036$) (Table S5; Table S10a). The 'more severe' cluster had proportionally more individuals with higher ADOS calibrated severity scores ($z_{\text{Mann–Whitney}} = 11.88$, df = 2,299, $P < 0.00001$). The agglomerative coefficient was 0.79 and the average AHARI = -4.10×10^{-6}. All input variable distributions were significantly different between the two main clusters, with the exception of HC (Table S9). Increased severity for all input variables, excepting VABS motor skills and daily living skills, and HCs were associated with membership in the 'more severe' cluster (Table S10). Sensitivity analyses again showed that ADI-R DevAb scores had the overall largest effect on main cluster stability (Table S7). The remaining input variables had similarly modest effects. Siblings had increased odds of going into the same main cluster and clusters contained cases with more similar genotype frequencies than did the unclustered dataset (Table 2).

The distributions of HC were significantly different across the 15 subclusters (Table S9; Table S12b). Head circumferences were statistically different between the six subclusters comprising the 'less severe' main cluster ($P = 0.0007$), but not the nine subclusters comprising the 'more severe' main cluster ($P = 0.37$). Removal of any input variable had similar effects on subcluster stability (Table S7).

Kruskal–Wallis tests showed that upon removal of both the VABS and HCs data as variable input, VABS scores were still significantly different between the main clusters and across subclusters. Head circumferences were still different across subclusters (Table S8). Non-verbal IQ scores from the Raven Coloured Matrices Scale were only available for 1% of the evaluated AGP dataset. As such, we did not evaluate non-verbal IQ across cluster-defined subgroups in this dataset.

Discussion

The extensive phenotypic variability within ASD may hinder our ability to identify genotype–phenotype associations. To address this problem, we used a novel approach to multivariate statistical analyses to take advantage of ASD-related behavioral information from multiple sources and to include quantitative data relevant to macrocephaly. This approach allowed effective evaluation of a broad array of data, enabling potentially more accurate phenotype definitions for large ASD datasets. We demonstrated that ASD phenotypic subgroups exist and can be replicated. Further, we demonstrated that these subgroups were genetically relevant.

Clustering defined more phenotypically similar subgroups of ASD

The strongest and most obvious clustering aggregated ASD traits into two major clusters grouped on overall symptom severity. This is consistent with many other phenotype-focused studies of ASD, where the unifying theme is subgroups defined as either high- or low-functioning based on the level of symptom severity and some measure of intellectual capability (Eaves *et al.* 1994; Fein *et al.* 1999; Prior *et al.* 1998; Sevin *et al.* 1995; Siegel *et al.* 1986; Spiker *et al.* 2002; Stevens *et al.* 2000; Wiggins *et al.* 2012). There is some divergence in our results compared to one previous subgrouping study performed using item-level scores from the ADI-R available in the AGRE dataset (Hu & Steinberg 2009). On the basis of these data, this previous clustering identified four main groups of individuals with ASD. Clusters were defined by: milder symptoms across domains, intermediate symptom severity across all domains, severe language deficits and higher frequencies of savant skills. While our results similarly identified clusters defined by overall symptom severity, we notably did not identify any

clusters defined by higher frequencies of savant skills. This is most likely due to the fact that we incorporated information from multiple other sources in addition to the ADI-R. To avoid issues related to overfitting (Tetko *et al.* 1995) we chose to include domain scores instead of item-level scores. By using domain scores, we did not include information relevant to savant skills. Our results may also be different from this previous study because we excluded individuals who did not meet diagnostic criteria on both the ADI-R and ADOS. The previous study evaluated a total of 1954 individuals with ASD, presumably diagnosed using the ADI-R alone. It is very likely we excluded individuals in our analyses that were included in the previous study.

Familial clustering suggests subgroups are genetically meaningful

Odds ratios showed significantly increased odds for affected siblings to cluster together into the two main clusters when compared to unrelated cases. These calculations are indicative of underlying genetic architecture. Further supporting this assumption, Wright's Fst calculations suggest individuals with more similar genetic architecture clustered together into the two main clusters. Although Fst can be confounded by genetic ancestry, we obtained similar results using only individuals with European ancestry. It is notable that there is still evidence for significant genetic and phenotypic heterogeneity within the reported ASD subgroups, especially the main clusters defined by overall symptom severity. This is in agreement with many previous studies indicating phenotype differences within the same multiplex family and the growing body of evidence reporting the involvement of *de novo* mutations arising in the germline (Beaudet 2007; Sanders *et al.* 2011, 2012). However, the relationship of genotype to phenotype should be somewhat independent of inheritance patterns. While our results supported an underlying genetic influence on overall main cluster assignment, to decipher genetic factors contributing to cluster assignment it will be necessary to perform further genetic analyses focused on these cluster groupings.

Severity of social/communication deficits predict cluster assignments

Not surprisingly, severity of both social and communication scores greatly influence main cluster and subcluster definitions. Increasing severity for social scores from the ADI-R and ADOS were significantly associated with assignment to the 'more severe' main cluster defined in both datasets. While these data from the VABS were not associated with the 'more severe' cluster in the AGRE dataset, they were associated with the 'more severe' cluster in the larger AGP dataset. Increasing severity for communication deficits measured on all three behavioral instruments were associated with the 'more severe' cluster from both datasets, and relative risk ratio calculations indicated a stronger effect than social deficits. There are notably more non-verbal individuals assigned to the 'more severe' main clusters in both datasets ($z_{AGRE} = 13.39$, $P < 0.00001$; $z_{AGP} = 10.26$, $P < 0.00001$). This may suggest that while social and communication deficits are strongly correlated, the individual's communication level has a greater influence on ASD subgroup definition than do impairments in social interaction.

Repetitive behavior measures are distinct from social and communication deficits

Repetitive behavior scores stood out from other input variables in their contribution to cluster and subcluster definitions for both datasets. Restricted repetitive behavior scores were not strongly correlated with any other evaluated trait. ADOS RRB score distributions were significantly different between the two main clusters from both the AGRE and AGP datasets. On the other hand, ADI-R RRBs were not different between the two main AGRE dataset clusters but were between the two main AGP dataset clusters. Even with this difference, ADI-R RRB scores were more noticeably distinct across the subclusters when compared to the main clusters from both datasets.

It is interesting that RRB measures had different levels of influence on definition of the two main clusters, based on whether they were evaluated with the ADOS or the ADI-R. One explanation is that RRBs are not as extensively evaluated with the ADOS. RRBs observed on the ADOS are more likely to be simple repetitive behaviors, easily observed in a brief interaction (Hus *et al.* 2012). In contrast, the ADI-R captures a broader array of RRBs and provides information for more complex repetitive behaviors. It is notable that by including ADI-R domain scores and not item-level scores we were not fully distinguishing simple vs. complex repetitive behaviors.

Our results suggest that presence of RRBs is important to ASD phenotype definitions in these datasets and that this behavior is unique from the social and communication deficits for definition of ASD subphenotypes. This is in line with numerous other studies (Bolton *et al.* 1994; Buxbaum *et al.* 2004; Happe *et al.* 2006; Hus *et al.* 2012; Mandy & Skuse 2008; Piven *et al.* 1997; Ronald *et al.* 2005, 2006; Shao *et al.* 2003; Silverman *et al.* 2002).

Early observations of developmental abnormality from the ADI-R strongly influenced ASD subgroup definitions

The DevAb score from the ADI-R also stood out as having a strong influence on ASD subgroup definitions. We observed that these measures were not correlated with any other evaluated phenotype measure. DevAb scores had consistently different distributions between clusters and across subclusters and the largest overall effect on cluster and subcluster stability. In both the AGRE and AGP datasets, the resulting 'more severe' main clusters contained 59–80% of individuals who received the highest score possible for this measure, compared to 0–0.4% of individuals in the 'less severe' main clusters. These results suggest very severe abnormality of development observed early in life is indicative of overall severity in clinical manifestation.

HC effects smaller subclusters, not main clusters

Head circumference distributions were not significantly different between the two main clusters grouped by overall

ASD severity, from either dataset. Head circumferences had a stronger influence on subcluster assignment. Our results indicated HC was important in defining ASD subphenotypes (i.e. subclusters), but not in determining overall severity (i.e. main clusters). It is notable that for both evaluated datasets, mean normalized HC was above average compared to normally developing individuals (Table S12). It is conceivable that most individuals with ASD have larger than normal HCs (Kanner 1968) and that this is not a distinguishing trait for the level of ASD symptom severity, but rather a trait related to the broader ASD diagnostic classification.

The convergence of results using two independent ASD datasets demonstrated the utility of this approach. That we were able to show defined subgroups of phenotypic expression appearing to be genetically meaningful in the AGRE dataset and replicate these findings in an independent AGP dataset lends further support to the validity of the resulting cluster groupings and the idea that the phenotype clusters recapitulate underlying genetic mechanisms in autism spectrum disorders.

References

Alarcon, M., Abrahams, B.S., Stone, J.L., Duvall, J.A., Perederiy, J.V., Bomar, J.M., Sebat, J., Wigler, M., Martin, C.L., Ledbetter, D.H., Nelson, S.F., Cantor, R.M. & Geschwind, D.H. (2008) Linkage, association, and gene-expression analyses identify CNTNAP2 as an autism-susceptibility gene. *Am J Hum Genet* **82**, 150–159.

American Psychiatric Association (2000) *Diagnostic and Statistical Manual of Mental Disorders*, 4th edn. American Psychiatric Association, Washington, DC.

American Psychiatric Association. (2012) Report of DSM-5 proposed criteria for autism spectrum disorder designed to provide more accurate diagnosis and treatment. American Psychiatric Association, Arlington, VA.

Anney, R., Klei, L., Pinto, D. *et al.* (2010) A genome-wide scan for common alleles affecting risk for autism. *Hum Mol Genet* **19**, 4072–4082.

Anney, R., Klei, L., Pinto, D. *et al.* (2012) Individual common variants exert weak effects on the risk for autism spectrum disorderspi. *Hum Mol Genet* **21**, 4781–4792.

Bailey, A., Le, C.A., Gottesman, I., Bolton, P., Simonoff, E., Yuzda, E. & Rutter, M. (1995) Autism as a strongly genetic disorder: evidence from a British twin study. *Psychol Med* **25**, 63–77.

Beaudet, A.L. (2007) Autism: highly heritable but not inherited. *Nat Med* **13**, 534–536.

Bolte, S. & Poustka, F. (2002) The relation between general cognitive level and adaptive behavior domains in individuals with autism with and without co-morbid mental retardation. *Child Psychiatry Hum Dev* **33**, 165–172.

Bolton, P., Macdonald, H., Pickles, A., Rios, P., Goode, S., Crowson, M., Bailey, A. & Rutter, M. (1994) A case–control family history study of autism. *J Child Psychol Psychiatry* **35**, 877–900.

Brock, G., Pihur, V., Datta, S. & Datta, S. (2008) clValid: an R package for cluster validation. *J Stat Softw* **25**.

Bruining, H., de, S.L., Swaab, H., de, J.M., Kas, M., van, E.H. & Vorstman, J. (2010) Dissecting the clinical heterogeneity of autism spectrum disorders through defined genotypes. *PLoS One* **5**, e10887.

Butler, M.G., Dasouki, M.J., Zhou, X.P., Talebizadeh, Z., Brown, M., Takahashi, T.N., Miles, J.H., Wang, C.H., Stratton, R., Pilarski, R. & Eng, C. (2005) Subset of individuals with autism spectrum disorders and extreme macrocephaly associated with germline PTEN tumour suppressor gene mutations. *J Med Genet* **42**, 318–321.

Buxbaum, J.D., Silverman, J., Keddache, M., Smith, C.J., Hollander, E., Ramoz, N. & Reichert, J.G. (2004) Linkage analysis for autism in a subset families with obsessive-compulsive behaviors: evidence for an autism susceptibility gene on chromosome 1 and further support for susceptibility genes on chromosome 6 and 19. *Mol Psychiatry* **9**, 144–150.

Chaste, P., Klei, L., Sanders, S.J. *et al.* (2013) Adjusting head circumference for covariates in autism: clinical correlates of a highly heritable continuous trait. *Biol Psychiatry*.

Cook, E.H. Jr. (2001) Genetics of autism. *Child Adolesc Psychiatr Clin N Am* **10**, 333–350.

Davis, J.M., Keeney, J.G., Sikela, J.M. & Hepburn, S. (2013) Mode of genetic inheritance modifies the association of head circumference and autism-related symptoms: a cross-sectional study. *PLoS One* **8**, e74940.

Eaves, L.C., Ho, H.H. & Eaves, D.M. (1994) Subtypes of autism by cluster analysis. *J Autism Dev Disord* **24**, 3–22.

El-Fishawy, P. & State, M.W. (2010) The genetics of autism: key issues, recent findings, and clinical implications. *Psychiatr Clin North Am* **33**, 83–105.

Fein, D., Stevens, M., Dunn, M., Waterhouse, L., Allen, D., Rapin, I. & Feinstein, C. (1999) Subtypes of pervasive developmental disorder: clinical characteristics. *Child Neuropsychol* **5**, 1–23.

Frazier, T.W., Youngstrom, E.A., Speer, L., Embacher, R., Law, P., Constantino, J., Findling, R.L., Hardan, A.Y. & Eng, C. (2012) Validation of proposed DSM-5 criteria for autism spectrum disorder. *J Am Acad Child Adolesc Psychiatry* **51**, 28–40.

Funalot, B., Varenne, O. & Mas, J.L. (2004) A call for accurate phenotype definition in the study of complex disorders. *Nat Genet* **36**, 3–4.

Georgiades, S., Szatmari, P., Boyle, M., Hanna, S., Duku, E., Zwaigenbaum, L., Bryson, S., Fombonne, E., Volden, J., Mirenda, P., Smith, I., Roberts, W., Vaillancourt, T., Waddell, C., Bennett, T. & Thompson, A. (2013) Investigating phenotypic heterogeneity in children with autism spectrum disorder: a factor mixture modeling approach. *J Child Psychol Psychiatry* **54**, 206–215.

Geschwind, D.H., Sowinski, J., Lord, C., Iversen, P., Shestack, J., Jones, P., Ducat, L. & Spence, S.J. (2001) The autism genetic resource exchange: a resource for the study of autism and related neuropsychiatric conditions. *Am J Hum Genet* **69**, 463–466.

Geschwind, D.H. (2011) Genetics of autism spectrum disorders. *Trends Cogn Sci* **15**, 409–416.

Gotham, K., Pickles, A. & Lord, C. (2009) Standardizing ADOS scores for a measure of severity in autism spectrum disorders. *J Autism Dev Disord* **39**, 693–705.

Gottesman, I.I. & Gould, T.D. (2003) The endophenotype concept in psychiatry: etymology and strategic intentions. *Am J Psychiatry* **160**, 636–645.

Grady, B.J., Torstenson, E., Dudek, S.M., Giles, J., Sexton, D. & Ritchie, M.D. (2010) Finding unique filter sets in plato: a precursor to efficient interaction analysis in gwas data. *Pac Symp Biocomput*, 315–326.

Hallmayer, J., Cleveland, S., Torres, A., Phillips, J., Cohen, B., Torigoe, T., Miller, J., Fedele, A., Collins, J., Smith, K., Lotspeich, L., Croen, L.A., Ozonoff, S., Lajonchere, C., Grether, J.K. & Risch, N. (2011) Genetic heritability and shared environmental factors among twin pairs with autism. *Arch Gen Psychiatry* **68**, 1095–1102.

Handl, J., Knowles, J. & Kell, D.B. (2005) Computational cluster validation in post-genomic data analysis. *Bioinformatics* **21**, 3201–3212.

Happe, F., Ronald, A. & Plomin, R. (2006) Time to give up on a single explanation for autism. *Nat Neurosci* **9**, 1218–1220.

Hazen, A. (1914) Storage to be provided in impounding reservoirs for municipal water supply. *Trans Am Soc Civil Eng* **77**, 1539–1640.

Hu, V.W. & Steinberg, M.E. (2009) Novel clustering of items from the autism diagnostic interview-revised to define phenotypes within autism spectrum disorders. *Autism Res* **2**, 67–77.

Hu, V.W., Sarachana, T., Kim, K.S., Nguyen, A., Kulkarni, S., Steinberg, M.E., Luu, T., Lai, Y. & Lee, N.H. (2009) Gene expression profiling differentiates autism case-controls and phenotypic

variants of autism spectrum disorders: evidence for circadian rhythm dysfunction in severe autism. *Autism Res* **2**, 78–97.
Hu, V.W., Addington, A. & Hyman, A. (2011) Novel autism subtype-dependent genetic variants are revealed by quantitative trait and subphenotype association analyses of published GWAS data. *PLoS One* **6**, e19067.
Hubert, L. & Arabie, P. (1985) Comparing partitions. *J Classif* **2**, 193–218.
Hu-Lince, D., Craig, D.W., Huentelman, M.J. & Stephan, D.A. (2005) The Autism Genome Project: goals and strategies. *Am J Pharmacogenomics* **5**, 233–246.
Hus, V., Gotham, K., & Lord, C. (2012) Standardizing ADOS domain scores: separating severity of social affect and restricted and repetitive behaviors. *J Autism Dev Disord*.
Kanner, L. (1968) Autistic disturbances of affective contact. *Acta Paedopsychiatr* **35**, 100–136.
Laliberte, E. & Legendre, P. (2010) A distance-based framework for measuring functional diversity from multiple traits. *Ecology* **91**, 299–305.
Laliberte, E. & Shipley, B. (2011) Measuring functional diversity (FD) from multiple traits, and other tools for functional ecology. R package version 1.0-11. 2011.
Leyfer, O.T., Folstein, S.E., Bacalman, S., Davis, N.O., Dinh, E., Morgan, J., Tager-Flusberg, H. & Lainhart, J.E. (2006) Comorbid psychiatric disorders in children with autism: interview development and rates of disorders. *J Autism Dev Disord* **36**, 849–861.
Liu, J., Nyholt, D.R., Magnussen, P., Parano, E., Pavone, P., Geschwind, D., Lord, C., Iversen, P., Hoh, J., Ott, J. & Gilliam, T.C. (2001) A genomewide screen for autism susceptibility loci. *Am J Hum Genet* **69**, 327–340.
Lord, C., Rutter, M., Goode, S., Heemsbergen, J., Jordan, H., Mawhood, L. & Schopler, E. (1989) Autism diagnostic observation schedule: a standardized observation of communicative and social behavior. *J Autism Dev Disord* **19**, 185–212.
Lord, C., Rutter, M. & Le, C.A. (1994) Autism diagnostic interview-revised: a revised version of a diagnostic interview for caregivers of individuals with possible pervasive developmental disorders. *J Autism Dev Disord* **24**, 659–685.
Maechler, M., Rouseeuw, P., Struyf, A., Hubert, M., & Hornik, K. (2012) Cluster: cluster analysis basics and extensions. R package version 1.14.3.
Mandy, W.P. & Skuse, D.H. (2008) Research review: What is the association between the social-communication element of autism and repetitive interests, behaviours and activities? *J Child Psychol Psychiatry* **49**, 795–808.
Matson, J.L. & Shoemaker, M. (2009) Intellectual disability and its relationship to autism spectrum disorders. *Res Dev Disabil* **30**, 1107–1114.
Mavandadi, S., Nazem, S., Ten Have, T.R., Siderowf, A.D., Duda, J.E., Stern, M.B. & Weintraub, D. (2009) Use of latent variable modeling to delineate psychiatric and cognitive profiles in Parkinson disease. *Am J Geriatr Psychiatry* **17**, 986–995.
Persico, A.M. & Bourgeron, T. (2006) Searching for ways out of the autism maze: genetic, epigenetic and environmental clues. *Trends Neurosci* **29**, 349–358.
Peter, B., Raskind, W.H., Matsushita, M., Lisowski, M., Vu, T., Berninger, V.W., Wijsman, E.M. & Brkanac, Z. (2011) Replication of CNTNAP2 association with nonword repetition and support for FOXP2 association with timed reading and motor activities in a dyslexia family sample. *J Neurodev Disord* **3**, 39–49.
Piven, J., Palmer, P., Jacobi, D., Childress, D. & Arndt, S. (1997) Broader autism phenotype: evidence from a family history study of multiple-incidence autism families. *Am J Psychiatry* **154**, 185–190.
Prior, M., Eisenmajer, R., Leekam, S., Wing, L., Gould, J., Ong, B. & Dowe, D. (1998) Are there subgroups within the autistic spectrum? A cluster analysis of a group of children with autistic spectrum disorders. *J Child Psychol Psychiatry* **39**, 893–902.
Pritchard, J.K., Stephens, M. & Donnelly, P. (2000) Inference of population structure using multilocus genotype data. *Genetics* **155**, 945–959.
Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A., Bender, D., Maller, J., Sklar, P., de Bakker, P.I., Daly, M.J. & Sham, P.C. (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* **81**, 559–575.
Raven, J. (1956) *Guide to using the Coloured Progressive Matrices*. H.K. Lewis, London.
Roche, A.F., Mukherjee, D., Guo, S.M. & Moore, W.M. (1987) Head circumference reference data: birth to 18 years. *Pediatrics* **79**, 706–712.
Ronald, A., Happe, F. & Plomin, R. (2005) The genetic relationship between individual differences in social and nonsocial behaviours characteristic of autism. *Dev Sci* **8**, 444–458.
Ronald, A., Happe, F., Bolton, P., Butcher, L.M., Price, T.S., Wheelwright, S., Baron-Cohen, S. & Plomin, R. (2006) Genetic heterogeneity between the three components of the autism spectrum: a twin study. *J Am Acad Child Adolesc Psychiatry* **45**, 691–699.
Sanders, S.J., Ercan-Sencicek, A.G., Hus, V. *et al.* (2011) Multiple recurrent de novo CNVs, including duplications of the 7q11.23 Williams syndrome region, are strongly associated with autism. *Neuron* **70**, 863–885.
Sanders, S.J., Murtha, M.T., Gupta, A.R. *et al.* (2012) De novo mutations revealed by whole-exome sequencing are strongly associated with autism. *Nature* **485**, 237–241.
Sevin, J.A., Matson, J.L., Coe, D., Love, S.R., Matese, M.J. & Benavidez, D.A. (1995) Empirically derived subtypes of pervasive developmental disorders: a cluster analytic study. *J Autism Dev Disord* **25**, 561–578.
Shao, Y., Cuccaro, M.L., Hauser, E.R., Raiford, K.L., Menold, M.M., Wolpert, C.M., Ravan, S.A., Elston, L., Decena, K., Donnelly, S.L., Abramson, R.K., Wright, H.H., DeLong, G.R., Gilbert, J.R. & Pericak-Vance, M.A. (2003) Fine mapping of autistic disorder to chromosome 15q11-q13 by use of phenotypic subtypes. *Am J Hum Genet* **72**, 539–548.
Siegel, B., Anders, T.F., Ciaranello, R.D., Bienenstock, B. & Kraemer, H.C. (1986) Empirically derived subclassification of the autistic syndrome. *J Autism Dev Disord* **16**, 275–293.
Silverman, J.M., Smith, C.J., Schmeidler, J., Hollander, E., Lawlor, B.A., Fitzgerald, M., Buxbaum, J.D., Delaney, K. & Galvin, P. (2002) Symptom domains in autism and related conditions: evidence for familiality. *Am J Med Genet* **114**, 64–73.
Snow, A.V., Lecavalier, L. & Houts, C. (2009) The structure of the Autism Diagnostic Interview-Revised: diagnostic and phenotypic implications. *J Child Psychol Psychiatry* **50**, 734–742.
Sparrow, S. & Cicchetti, D. (1985) Diagnostic uses of the Vineland Adaptive Behavior Scales. *J Pediatr Psychol* **10**, 215–225.
Sparrow, S., Balla, D. & Cicchetti, D. (1984) *Vineland Adaptive Behavior Scales: Survey form Manual*. American Guidance Service, Circle Pines, MN.
Sparrow, S., Cicchetti, D. & Balla, D. (2005) *Vineland Adaptive Behavior Scales*, 2nd edn. Pearson Assessments, Minneapolis, MN.
Spiker, D., Lotspeich, L.J., Dimiceli, S., Myers, R.M. & Risch, N. (2002) Behavioral phenotypic variation in autism multiplex families: evidence for a continuous severity gradient. *Am J Med Genet* **114**, 129–136.
Statacorp (2009) *Stata Statistical Software [11]*. Statacorp LP, College Station, TX.
Stevens, M.C., Fein, D.A., Dunn, M., Allen, D., Waterhouse, L.H., Feinstein, C. & Rapin, I. (2000) Subgroups of children with autism by cluster analysis: a longitudinal examination. *J Am Acad Child Adolesc Psychiatry* **39**, 346–352.
Tetko, I.V., Livingstone, D.J. & Luik, A.I. (1995) Neural network studies. 1. Comparison of overfitting and overtraining. *J Chem Inf Comput Sci* **35**, 826–833.
Varga, E.A., Pastore, M., Prior, T., Herman, G.E. & McBride, K.L. (2009) The prevalence of PTEN mutations in a clinical pediatric

cohort with autism spectrum disorders, developmental delay, and macrocephaly. *Genet Med* **11**, 111–117.
Veenstra-Vanderweele, J. & Cook, E.H. Jr. (2004) Molecular genetics of autism spectrum disorder. *Mol Psychiatry* **9**, 819–832.
Veenstra-Vanderweele, J., Christian, S.L. & Cook, E.H. Jr. (2004) Autism as a paradigmatic complex genetic disorder. *Annu Rev Genomics Hum Genet* **5**, 379–405.
Vernes, S.C., Newbury, D.F., Abrahams, B.S., Winchester, L., Nicod, J., Groszer, M., Alarcon, M., Oliver, P.L., Davies, K.E., Geschwind, D.H., Monaco, A.P. & Fisher, S.E. (2008) A functional genetic link between distinct developmental language disorders. *N Engl J Med* **359**, 2337–2345.
Walesiak, M. & Dudek, A. (2007) Searching for optimal clustering procedure for a data set. [R package version 0.30-7].
Whitehouse, A.J., Bishop, D.V., Ang, Q.W., Pennell, C.E. & Fisher, S.E. (2011) CNTNAP2 variants affect early language development in the general population. *Genes Brain Behav* **10**, 451–456.
Wiggins, L.D., Robins, D.L., Adamson, L.B., Bakeman, R. & Henrich, C.C. (2012) Support for a dimensional view of autism spectrum disorders in toddlers. *J Autism Dev Disord* **42**, 191–200.
Yonan, A.L., Alarcon, M., Cheng, R., Magnusson, P.K., Spence, S.J., Palmer, A.A., Grunn, A., Juo, S.H., Terwilliger, J.D., Liu, J., Cantor, R.M., Geschwind, D.H. & Gilliam, T.C. (2003) A genomewide screen of 345 families for autism-susceptibility loci. *Am J Hum Genet* **73**, 886–897.

Acknowledgments

The authors have no conflicts of interest to disclose. We gratefully acknowledge the resources provided by both the Autism Genetic Resource Exchange Consortium and Autism Genome Project Consortium and the participating 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 C.M. Lajonchere (PI), and formerly by grant MH64547 to D.H. Geschwind (PI). The Autism Genome Project is funded by Autism Speaks, the Medical Research Council (UK) and the Health Research Board (Ireland).

Supporting Information

Additional supporting information may be found in the online version of this article at the publisher's web-site:

Table S1: De-identified sample IDs for phenotype analyses datasets (AGRE & AGP). Main cluster (Main_C) and subcluster (Sub_C) assignments are indicated.

Table S2: Availability of phenotypes. Reported are percentage breakdowns of trait-specific information in both datasets.

Table S3: Spearman's correlation coefficients. Spearman's rho correlations calculated using percentile ranked phenotype data in (a) AGRE discovery dataset, and (b) AGP replication dataset. MS, MotorSkills; DL, DailyLiving.

Table S4: Validity scores for clustering. The most valid number of partitions, and cluster method are indicated in bold italics. agnes, agglomerative hierarchical; diana, divisive hierarchical; pam, partitioning around medoids; sota, self-organizing tree algorithm; clara, clustering large applications. Dunn Index – smallest inter-cluster distance/largest intra-cluster distance, should be maximized (scale = 0, ∞). Connectivity – to what extent are individuals placed in the same cluster as the most similar individuals, should be minimized (scale = 0, ∞). Silhouette Width – overall average of average distance between individual and others in same cluster compared to different cluster (−1, 1); well-clustered = 1. Clustering of the datasets into 2:15 clusters were tested. The best validity scores are indicated in bold italics and support in the: (a) AGRE dataset, a two cluster solution with hierarchical clustering methods and a 10 subcluster solution with the agglomerative hierarchical method, and (b) in the AGP dataset a two cluster solution and a 15 subcluster solution with the agglomerative hierarchical method.

Table S5: Summary statistics for unclustered vs clustered datasets. Reported are medians and modes observed in the unclustered dataset compared to the two main clusters. Continuous variables are starred to indicate that the mean is reported in place of the median for this trait. Cases with scores indicating increased ASD severity preferentially cluster into the second, larger cluster in both datasets. Age is reported in years. Disc = AGRE dataset; Rep = AGP dataset.

Table S6: Membership in clusters and subclusters as a function of input variables: AGRE dataset. Relative risk ratios (R.R.R.) represent the risk for assignment to (a) the 'more severe' main cluster relative to the 'less severe' main cluster, (b) the noted subcluster relative to risk for assignment to all other subclusters within the 'less severe' main cluster and (c) the noted subcluster relative to risk for assignment to all other subclusters within the 'more severe' main cluster. Variables that were significantly associated with membership in each cluster and subcluster are indicated in bold italics. For variables from the ADI-R and ADOS, increasing values indicate greater levels of severity; for variables from the VABS decreasing values indicate greater levels of severity. Correlated variables, as defined in Fig. S1, were evaluated in the same model. Reported *P*-values are uncorrected.

Table S7: Sensitivity Analyses. Reported statistics evaluating cluster stability upon removal of each variable are: APN, average proportion of non-overlap or number of individuals not placed in same cluster when variable is removed (scale = 0,1); AD, average distance between individuals placed in same cluster when variable is removed (scale = 0, ∞); ADM, average distance between means between cluster centers for individuals placed in same cluster when variable is removed (scale = 0, ∞); FOM, figure of merit or average intra-cluster variance of the removed variable where clustering is based on remaining variables (scale = 0, ∞). For the stability measures calculated, smaller values indicate more stable cluster results. Disc, AGRE datset; Rep, AGP dataset.

Table S8: Cluster differences in datasets following removal of variables with some missing data. Kruskal–Wallis comparison of variable distributions between the two main clusters and across the subclusters upon removal of the VABS and HC data from cluster input for the (a) AGRE dataset, and (b) AGP dataset. VABS data was still significantly different across all ASD subgroupings even when not used as input for cluster analyses. ADOS CSS, calibrated severity score; df, degrees of freedom.

Table S9: Cluster differences in AGP dataset. Kruskal–Wallis comparison of variable distributions between the two main clusters and across the 15 subclusters. All input variable distributions, except HC, are significantly

different between the main clusters. Head circumference distributions are significantly different across subclusters. Asterisks indicate information not used as input variable. ADOS CSS, calibrated severity score; df, degrees of freedom.

Table S10: Membership in clusters and subclusters as a function of input variables: AGP dataset. Relative risk ratios (R.R.R.) represent the risk for assignment to (a) the 'more severe' main cluster relative to the 'less severe' main cluster, (b) the noted subcluster relative to risk for assignment to all other subclusters within the 'less severe' main cluster and (c) the noted subcluster relative to risk for assignment to all other subclusters within the 'more severe' main cluster. Variables that were significantly associated with membership in each cluster and subcluster are indicated in bold italics. For variables from the ADI-R and ADOS, increasing values indicate greater levels of severity; for variables from the VABS decreasing values indicate greater levels of severity. Correlated variables, defined similar to Fig. S1, were evaluated in the same model. Reported *P*-values are uncorrected.

Table S11: Details on non-verbal IQ in AGRE main clusters and subclusters. Reported are the frequencies of cases with non-verbal IQ information assigned to each main cluster and subcluster defined in the AGRE datasets. Mean scores and standard deviations around these means are also reported. *T*-Tests reported on top compare means between the two main clusters and below are corresponding *t*-statistics comparing subcluster means. Asterisks indicate a statistically significant difference ($P < 0.05$).

Table S12: Details on head circumference in main clusters and subclusters. Reported are the frequencies of cases with head circumference information assigned to each main cluster and subcluster defined in both the (a) AGRE and (b) AGP datasets. Mean *z*-scores and standard deviations around these means are also reported. NA indicates this information is not applicable for AGP subcluster 2.4. This subcluster only contained two individuals. Main cluster 1 = 'less severe', Main cluster 2 = 'more severe'. Reported also are t-statistics comparing HC means by subcluster. Asterisks indicate a statistically significant difference ($P < 0.05$).

Figure S1: Variable correlation structure in discovery dataset. Plot of Spearman's correlation coefficients used in variable weighting scheme for PCA and clustering. The eight variable contributions to multivariate analyses are indicated.

Figure S2: Agglomerative clustering AGP dataset. Height indicates distance between merging clusters at successive stages of clustering and is related to dissimilarities among clusters. The main cluster highlighted in green represents individuals with more severe ASD phenotypes compared to individuals assigned to the main cluster highlighted pink. Subclusters are indicated with corresponding boxes. One subcluster only contained two individuals and is not denoted with a highlighted box.