OPEN

Molecular Psychiatry (2014), 1–7

www.nature.com/mp

ORIGINAL ARTICLE

The phenotypic manifestations of rare genic CNVs in autism spectrum disorder

AK Merikangas[1], R Segurado[2], EA Heron[1], RJL Anney[1], AD Paterson[3], EH Cook[4], D Pinto[5], SW Scherer[6], P Szatmari[7], M Gill[1], AP Corvin[1] and L Gallagher[1]

Significant evidence exists for the association between copy number variants (CNVs) and Autism Spectrum Disorder (ASD); however, most of this work has focused solely on the diagnosis of ASD. There is limited understanding of the impact of CNVs on the 'sub-phenotypes' of ASD. The objective of this paper is to evaluate associations between CNVs in differentially brain expressed (DBE) genes or genes previously implicated in ASD/intellectual disability (ASD/ID) and specific sub-phenotypes of ASD. The sample consisted of 1590 cases of European ancestry from the Autism Genome Project (AGP) with a diagnosis of an ASD and at least one rare CNV impacting any gene and a core set of phenotypic measures, including symptom severity, language impairments, seizures, gait disturbances, intelligence quotient (IQ) and adaptive function, as well as paternal and maternal age. Classification analyses using a non-parametric recursive partitioning method (random forests) were employed to define sets of phenotypic characteristics that best classify the CNV-defined groups. There was substantial variation in the classification accuracy of the two sets of genes. The best variables for classification were verbal IQ for the ASD/ID genes, paternal age at birth for the DBE genes and adaptive function for *de novo* CNVs. CNVs in the ASD/ID list were primarily associated with communication and language domains, whereas CNVs in DBE genes were related to broader manifestations of adaptive function. To our knowledge, this is the first study to examine the associations between sub-phenotypes and CNVs genome-wide in ASD. This work highlights the importance of examining the diverse sub-phenotypic manifestations of CNVs in ASD, including the specific features, comorbid conditions and clinical correlates of ASD that comprise underlying characteristics of the disorder.

Molecular Psychiatry advance online publication, 25 November 2014; doi:10.1038/mp.2014.150

INTRODUCTION

Advances in our understanding of the genetics underlying Autism Spectrum Disorder (ASD) from linkage,[1–3] genome-wide association studies,[4–7] and next-generation sequencing studies[8–10] have identified several potential genetic pathways to ASD including chromatin remodeling, metabolism, mRNA translation and synaptic function, and others.[11] However, the bulk of the genetic research has focused on categorical classification of the presence or absence of an ASD diagnosis, which is characterized by a broad range of core and associated features and heterogeneity in its manifestations.[12,13] Investigation of the role of susceptibility genes for specific sub-phenotypes of ASD is a critical next step in integrating this diverse research. The term 'sub-phenotypes' as used here describes the specific features, comorbid conditions or clinical correlates of ASD that comprise underlying phenotypic characteristics of the disorder.

Copy number variation (CNV), the most prevalent type of structural variation in the human genome,[14,15] has been shown to contribute to genetic heterogeneity[16] and may also provide an important tool to identify sources of phenotypic heterogeneity.[17] Identification of specific components of the ASD phenotype or its correlates that are associated with CNVs could provide etiologic clues to the development of the ASD phenotype. There is growing evidence that certain recurrent CNVs, while predisposing toward ASD, also influence broader phenotypic manifestations, such as intellectual disability (ID), physical characteristics and seizures.[18] This suggests that the impact of CNVs might extend beyond the current clinical diagnostic entities in neuropsychiatry, as demonstrated by the wide range of manifestations of well-characterized genetic disorders, such as the 22q deletion syndrome.[19] Some CNVs might also act as modifiers or vulnerability factors for these phenotypes rather than influencing specific disease outcomes.[20,21] Since multiple genes can be disrupted by a single CNV, one might expect greater evidence of syndromal presentations that transcend standard clinical features of the specific disorder with multi-systemic effects. To date, most of the CNV research in ASD has focused on disorder-specific CNV identification, and few studies[22,23] have systematically examined the impact of CNVs genome-wide on specific features, comorbid conditions or clinical correlates of ASD. With few exceptions (for example, refs. 1,5,24–28), there

[1]Department of Psychiatry, Neuropsychiatric Genetics Research Group, Institute of Molecular Medicine, Trinity College Dublin, Dublin 8, Ireland; [2]Centre for Support and Training in Analysis and Research, University College Dublin, Dublin 4, Ireland; [3]Program in Genetics and Genome Biology, Hospital for Sick Children, Division of Biostatistics, Dalla Lana School of Public Health, University of Toronto, Toronto, ON, Canada; [4]Institute for Juvenile Research, Department of Psychiatry, University of Illinois at Chicago, Chicago, IL, USA; [5]Departments of Psychiatry, and Genetics and Genomic Sciences, Seaver Autism Center, The Mindich Child Health & Development Institute, Icahn School of Medicine at Mount Sinai, New York, NY, USA; [6]Department of Molecular Genetics, The Centre for Applied Genomics and Program in Genetics and Genomic Biology, The Hospital for Sick Children, University of Toronto, Toronto, ON, Canada and [7]The Division of Child and Adolescent Psychiatry, Centre for Addiction and Mental Health, The Hospital for Sick Children, University of Toronto, Toronto, ON, Canada. Correspondence: Dr AK Merikangas, Department of Psychiatry, Neuropsychiatric Genetics Research Group, Trinity College Dublin, Institute of Molecular Medicine, St. James's Hospital, James's Street, Dublin, Dublin 8, Ireland.
E-mail: merikana@tcd.ie
Received 7 April 2014; revised 10 August 2014; accepted 19 September 2014

has been limited investigation of sub-phenotypic features of ASD in linkage studies, genome-wide association studies or next-generation sequencing studies.

The aim of this work was to examine the phenotypic outcome of variation in genic CNVs in ASD, using a case-only design. The purpose of investigating genic CNVs was to narrow the analytic focus and target functional regions of the genome. We employed a data-driven approach (random forests) to determine whether there were more homogeneous phenotypic sub-groups within the sample that would be more amenable to genetic association analyses. Unlike parametric statistical approaches that are hypothesis driven, and assume a standard distribution of the data, the random forest approach does not assume a Normal distribution of the predictor data.[29] We hypothesized that individuals carrying rare CNVs that impact genes implicated in either ASD or ID, or in differentially brain expressed (DBE) genes, would be more likely to present with general developmental anomalies, including non-verbal status, seizures, gait disturbances, lower IQ and adaptive function than those cases with CNVs impacting other genes.

MATERIALS AND METHODS

Subjects

This study is a secondary analysis of data from the Autism Genome Project (AGP), a collaborative data collection and genetics research initiative.[30] The original AGP CNV study sample comprises 2705 parent-affected child trios (2384 of European ancestry) derived from a pooled sample, and a subset of those are included in Pinto *et al.*[31,32] The sample included in the analyses presented here consisted of 1590 cases of European ancestry with a diagnosis of an ASD with at least one rare CNV impacting any gene. Diagnostic inclusion was defined by DSM-IV[33] criteria at all sites, assessed by the Autism Diagnostic Observation Schedule (ADOS)[34] and the Autism Diagnostic Interview-Revised.[35] Overall, the sample was 85.5% male, and the mean age at assessment was 8.8 years (105 months, s.d. 60.9 months, range 22–565 months). Ethical approval was granted at each participating AGP site. Both parents provided written informed consent, and consent was also obtained from subjects who were considered competent to provide it, based on the parents' or caregivers' assessment of competency. Ethnically matched convenience controls used to estimate CNV frequencies were obtained from unrelated control subjects with no obvious psychiatric history from three additional studies (that is, Study of Addiction Genetics and Environment;[36] Ontario Colorectal Cancer case–control study;[37,38] and Health, Aging and Body Composition[39]). See Supplementary Materials Section 1.1 for additional information.

Clinical measures

These analyses were based on phenotypes of possible neurodevelopmental origin derived from the Autism Diagnostic Interview that were available across AGP sites including verbal status, age at first words and phrases (also collapsed into a composite language delay), gait disturbance and non-febrile seizures (did not require a diagnosis of epilepsy). Additionally, the ADOS severity score,[40] Vineland Adaptive Behavior Scales (VABS),[41] a selected composite intelligence quotient derived by an AGP sub-committee (IQ; verbal, performance, and full scale; see Supplementary Materials for additional information), maternal and paternal age at birth and family type were analyzed when they were available. Family type classification ascertained included simplex (no known affected individual among the first- to third-degree relatives), multiplex (at least one first- to third-degree relative of the proband with a validated, clinical ASD diagnosis) and unknown (where family history was not explicitly assessed). More detail on the clinical measures is included in the Supplementary Materials (Section 1.2), and phenotype missingness is reported in Supplementary Table S1.

Genotyping and CNV Calling

Genotyping procedures and CNV calling were thoroughly described in Pinto *et al.*[32,42] Briefly, genotyping was completed on the Illumina Infinium 1 M or Human1M-Duo SNP microarray. CNV discovery included computational calling *via* QuantiSNP,[43] iPattern[44] and PennCNV[45] algorithms. After initial CNV calling, the sample was then limited to those individuals of European ancestry and to CNVs that occurred in $<1\%$ of cases and controls. There were a total of 1369 complete trios in whom the inheritance of their CNVs could be established (for example, where one or both the parents were not missing data due to QC filtering or the absence of parental DNA entirely, or at specific markers. N.B. this was not person-specific, but CNV-specific). The case and control CNV frequencies are reported in Pinto *et al.*[42]

Gene lists

Because individual CNV events were rare, we examined two candidate gene lists that were used to aggregate these rare CNV events for statistical analysis. The lists were (1) An ASD- and ID-implicated gene list 'expert-curated' *via* literature searches and database reviews through December 2009, and used in the original AGP CNV analyses (see Pinto *et al*[32] for review) and (2) A DBE list (see Raychaudhuri *et al.*[46] for review). Briefly, the ASD-implicated gene list contains 36 genes strongly implicated in ASD, and the ID-implicated gene list contains 110 genes known to be implicated in ID but not yet in ASD. The ASD and ID-implicated lists were pooled for the analyses presented here, because the ASD list was too small for meaningful analysis. The 3268 DBE genes were defined by Raychaudhuri *et al.*[46] using data from the Gene Expression Omnibus,[47] to have specific, differential expression in the brain and central nervous system as compared with other healthy body tissues. A total of 42 genes overlap in the ASD/ID-implicated candidate list and DBE list. More details are provided in Supplementary Materials Section 1.3.

Statistical analyses

In this case-only study, three analytic techniques were employed to explore these data: (1) Latent profile analysis (LPA) to examine the structure of the phenotype data and to determine whether there were more homogeneous subgroups among the highly correlated phenotypes; (2) Recursive partitioning *via* random forests to determine whether phenotypic subgroups could be predicted by CNV carrier status; and (3) Univariable and multivariable (that is, to control for potential confounders) association analyses *via* linear and logistic regression to validate the variables highlighted in the two data-driven analytic approaches.

The purpose of the LPA was to identify more homogeneous subgroups of patients based on the highly correlated phenotype data. The LPA was completed in the Mplus software package,[48] using the mixture model option, and model fit was evaluated for one through twelve class solutions. On the basis of the optimal Bayesian Information Criterion solution, there was no obvious solution for the sample as a whole, leading us to conclude that there were no obvious phenotypic subgroups within this sample. The results of this analysis are reported in the Supplementary Materials (Supplementary Figure S1).

Random forests, a type of classification and regression tree recursive partitioning, were employed to identify more homogeneous subsets of the heterogeneous clinical groups based on their CNV carrier status,[29] and to address potential multicollinarity of the phenotypic variables. They also serve as a data reduction strategy, where the 'important' variables can be selected for additional analyses. Random forests can be used when there are missing data, with both continuous and categorical data, with non-Normal data, in the presence of interactions, and do not require reduction of the predictor space before classification.[29] Briefly, a classification tree is constructed by choosing an initial splitting variable (here, CNV carrier status) to determine the optimal splits of the data based on the randomly selected predictor variables (here, clinical phenotype variables). A collection of each of these trees is termed a forest.[49–51] Here, a random forest analysis using the DOS command line version of the Willows software package[52] was used to investigate the phenotypic differences between cases with and without CNVs impacting ASD/ID or DBE genes. Forests were created from 10 000 trees with a minimum terminal node size (smallest group of cases when the tree construction terminates) of 10 (0 terminal node size in the *de novo* analyses due to the rarity of *de novo* events). The number of variables chosen to split a node was selected by the program automatically (here it was four). Results presented here are the *variable importance* measure, an assessment of which variable is the best classifier of the data, and an *out-of-bag (OOB)* error measure. When constructing a tree, one-third of the cases are held out to test the tree; these are the OOB samples. Once the tree is constructed, the OOB sample is dropped through the tree, and each case is classified. The proportion of cases correctly classified is the accuracy, and 1 minus the accuracy is the error. The variable importance score is calculated by comparing

classification accuracy when the case status in the original and OOB samples is randomly shuffled and then run through the tree. If the error rates do not increase when the cases status is permuted, then the variable under examination is not a useful predictor of case status. This accuracy is averaged over all of the trees in the forest, and the overall classification accuracy of the forest is reflected by the reported OOB error. For reference, no variable importance score has a particular meaning; instead it is a relative ranking of the examined predictors. Therefore, the variable importance scores must be interpreted in reference to one another, and should not be compared across studies.[51] Limitations of the method include a potential bias toward continuous variable types[53] and highly correlated variables[54] in the analyses.

The association analyses examined the relationship between rare CNVs impacting genes implicated in either the ASD/ID or DBE gene list and the phenotypes of interest, which serves to validate the findings resulting from the random forest analyses. Typically one would follow-up the results of only the 'important' variables discovered in the random forest analysis; however, we completed association analyses on all variables since we were testing the random forest method. In the models presented here, CNV carrier status predicted the outcome phenotype, except in the case of parental age, where that predicted CNV status. Either linear or logistic regression was employed, depending upon the outcome variable. Preliminary analyses were performed univariable, unadjusted, and further analyses were multivariable, adjusted for relevant covariates (age at test, genotyping stage and age of the other parent for parental age). Full details of the statistical test and covariates used for each phenotype are described in Supplementary Table 2. Male-only analyses were completed, but without a change in the results, so only the combined sex analyses are shown here. The results presented here are not corrected for multiple comparisons, and all analyses were completed in SAS v9.2.[55]

RESULTS

Frequency of CNVs

Table 1 shows the frequency of CNVs impacting genes in the ASD/ID gene list and the DBE gene list by CNV type among participants who carried at least one rare genic CNV. Deletions and duplications are not mutually exclusive; therefore, one individual can have both a deletion and a duplication in different portions of their genome. This results in overlapping, non-independent analyses for deletions, duplications and any CNVs. It is important to note that the results here only reflect a subset of those included in Pinto *et al.*[42] where the rates of CNVs in cases and controls are presented.

The frequency of cases with CNVs impacting the ASD/ID list genes was low, with 6.6% of the sample having a CNV that impacted a gene in this list (3.21% deletions, 3.58% duplications). The frequency of cases with CNVs impacting DBE list genes was substantially higher than that of the ASD/ID list, with 51.32% of the sample having a CNV that impacted a gene (or genes) in this list. Overall, 26.73% of the sample had deletions that impacted a gene in the DBE list and 30.75% had duplications that impacted a gene in the list.

Table 1. Number and percentage of cases who are carriers of specific CNV types among AGP rare, genic CNV carriers

Gene list	*CNV type*	*Number (%)*
Any gene	Deletion	988 (62.14)
	Duplication	1147 (72.14)
	De Novo	106 (7.74)
	Inherited	1339 (97.81)
ASD/ID gene	Any CNV	105 (6.6)
	Deletion	51 (3.21)
	Duplication	57 (3.58)
	De Novo	21 (1.53)
	Inherited	73 (5.33)
DBE gene	Any CNV	816 (51.32)
	Deletion	425 (26.73)
	Duplication	489 (30.75)
	De Novo	65 (4.75)
	Inherited	669 (48.87)

Abbreviations: ASD/ID, Autism Spectrum Disorder or Intellectual Disability; AGP, Autism Genome Project; CNV, copy number variation; DBE, differentially brain expressed. Deletions and duplications are not mutually exclusive categories.

CNV inheritance information was only available for a subset of the 1590 rare, genic CNV carriers; 1369 cases had full trio data available for analysis (86% of the sample). Overall, 7.7% of these cases had a *de novo* CNV. Only 1.5% of the sample had a *de novo* CNV that impacted a gene in the ASD/ID gene list. The frequency of *de novo* CNVs impacting genes in the DBE list was higher, with 4.75% having a *de novo* CNV that impacted a gene in this list.

Recursive partitioning results

The results of the random forest exploratory analyses are presented in Table 2. Overall, there was great variability of the classification accuracy for the selected measures in this study. The ASD/ID list was a substantially better classifier than was the DBE list, with OOB error rates nearly tenfold lower for the ASD/ID list for all CNV types (3–7% in the ASD/ID list vs 27–52% in the DBE list over 10 000 permutations). The scale of the variable importance statistics was similar between both lists, but there were a greater number of variables that hindered classification (negative importance) in the DBE list than in the ASD/ID list. Among the ASD/ID forests, deletions showed better classification accuracy than did all CNVs and duplications. Verbal IQ was the best classifier for deletions and all CNVs, while VABS daily living skills were the best classifier for duplications. Conversely, sex and seizures were universally poor classifiers.

With respect to the DBE genes, based on the OOB error, deletions tended to be better classifiers than either duplications or all CNVs. Paternal age, performance IQ, VABS composite and daily living skills were the only phenotypes that were universally positive classifiers across CNV types, whereas others were quite variable. Notably, age at first words and phrases, full-scale IQ, maternal age and VABS socialization classification importance varied between positive and negative values depending on the CNV type. Seizures, family type, gait disturbance, language delay and the ADOS severity score hindered classification.

In the *de novo* random forest analyses, the best phenotypes for classifying *de novo* CNV carriers were the VABS scores, specifically the composite and daily living skills. ASD/ID *de novo* CNVs provided more robust classification values than the any *de novo* and DBE *de novo* CNVs.

Association analyses

Table 3 presents summaries of the associations between the presence or absence of CNVs impacting ASD/ID genes, DBE genes and the phenotypes described above. Table 4 shows the summary of associations between parental age and CNVs in the gene lists. In the tables, statistical significance was designated at the $\alpha = 0.05$ level and 95% confidence limits are reported; there were nine statistical tests per phenotype.

Statistically significant associations between deletions impacting ASD/ID genes predicting the phenotypes under investigation were found for language delay (all ASD/ID CNVs and deletions, unadjusted and adjusted), verbal IQ (all ASD/ID CNVs unadjusted, duplications unadjusted and adjusted), VABS communication (duplications, unadjusted and adjusted), socialization (deletions, unadjusted) and composite scores (deletions unadjusted, and duplications unadjusted and adjusted).

Statistically significant associations between CNVs impacting DBE genes were found for VABS communication (duplications, unadjusted and adjusted), socialization (duplications, unadjusted and adjusted), daily living skills (all DBE CNV and duplications,

Table 2. Random forest variable importance scores and 'Out of Bag' (OOB) error rates

Variable	*ASD/ID* (N = *1590*)			*DBE* (N = *1590*)			*De Novo* (N = *1369*)		
	All	*Del*	*Dup*	*All*	*Del*	*Dup*	*All*	*ASD/ID*	*DBE*
ADOS severity score	0.05	0.06	0.08	−0.16	−0.25	−0.27	0.00	0.13	0.15
Family type	0.04	0.03	0.02	−0.37	−0.11	0.00	0.23	0.15	0.46
First phrases	**1.15**	0.32	0.32	−0.42	0.61	−0.31	0.67	0.51	0.55
First words	0.20	0.45	−0.04	−0.71	0.57	−0.74	0.00	0.42	0.67
Full-scale IQ	0.90	0.37	0.19	−0.93	1.22	0.32	1.38	**1.10**	1.57
Gait disturbance	0.06	0.05	0.00	−0.41	−0.07	−0.32	−0.08	−0.01	0.04
Language delay	0.94	**0.54**	0.05	−0.09	−0.13	-0.06	0.22	0.29	0.39
Maternal age	0.83	0.32	**0.48**	−0.47	**1.29**	0.39	0.91	0.43	0.33
Paternal age	**0.99**	0.42	0.43	**0.34**	**2.69**	**0.58**	1.27	0.31	0.41
Performance IQ	0.75	0.30	0.29	**0.65**	0.37	0.29	1.01	0.08	0.71
Seizures	−0.12	−0.05	−0.01	−0.29	−0.04	−0.18	0.07	−0.23	−0.05
Sex	−0.03	−0.01	0.00	0.02	−0.06	−0.11	−0.01	−0.01	0.08
VABS communication	0.80	0.23	**0.48**	−0.21	1.14	0.99	**3.72**	1.03	**3.15**
VABS composite	0.59	0.62	**0.53**	**0.29**	1.14	**0.52**	**6.29**	**1.79**	**6.34**
VABS daily living skills	0.95	**0.61**	**0.56**	0.04	0.44	0.35	**4.30**	**1.51**	**3.22**
VABS socialization	0.82	0.34	0.35	−0.04	−0.51	**1.54**	3.37	0.84	2.70
Verbal status	0.77	0.22	0.11	−0.21	0.21	0.36	0.28	0.14	0.10
Verbal IQ	**2.01**	**0.87**	0.36	−0.23	**1.89**	−0.20	1.46	0.40	1.40
OOB error	0.07	0.03	0.04	0.52	0.27	0.32	0.08	0.02	0.05

Abbreviations: ADOS, Autism Diagnostic Observation Schedule; ASD/ID, Autism Spectrum Disorder or Intellectual Disability; DBE, differentially brain expressed; Del, Deletion; Dup, Duplication; IQ, intelligence quotient; VABS, Vineland Adaptive Behavior Scales. The top three classifiers in each CNV type are in bold typeface. The OOB error rate reflects classification accuracy over 10 000 permutations. The variable importance score reflects the difference between the OOB error rates when the CNV status is and is not permuted over 10 000 iterations. If the error rates do not increase when the CNV status is permuted, then that variable is a poor classifier. Negative variable importance indicates that the variable hinders classification. Ultimately, variable importance demonstrates which variables have a major role in discriminating participants with and without CNVs.

Table 3. Summary of findings on significant associations of CNVs impacting Autism Spectrum Disorder or Intellectual Disability (ASD/ID) or differentially brain expressed (DBE) genes with clinical phenotypes

Independent variable (gene list)	*Dependent variable (phenotype)*	N	*OR*	*LCL*	*UCL*	P
ASD/ID All	Language delay	1559	0.53	0.34	0.80	0.003
Deletions in ASD/ID	Language delay	1559	0.43	0.24	0.77	0.005
Duplications in ASD/ID	VABS composite	1247	0.48	0.25	0.90	0.022
	VABS communication	1261	0.45	0.24	0.83	0.010
All CNVs in DBE	VABS daily living skills	1254	0.73	0.57	0.93	0.012
	VABS composite	1247	0.63	0.49	0.82	0.001
Deletions in DBE	VABS composite	1247	0.63	0.49	0.82	<0.001
Duplications in DBE	VABS socialization	1273	0.66	0.51	0.86	0.002
	VABS daily living skills	1254	0.65	0.50	0.84	0.001
	VABS omposite	1247	0.60	0.46	0.79	<0.001
	VABS communication	1261	0.73	0.57	0.93	0.010
All *De novo*	Seizures	1237	2.02	1.17	3.50	0.012
ASD/ID *De novo*	Seizures	1237	3.47	1.26	9.59	0.017
DBE *De novo*	Family type (Simplex vs Multiplex)	1168	2.50	1.22	5.11	0.012

Abbreviations: CNV, copy number variation; VABS, Vineland Adaptive Behavior Scales. Only statistically significant findings ($\alpha = 0.05$) in adjusted models are shown, along with their associated odds ratio (OR), lower (LCL) and upper (UCL) 95% confidence limits and *P*-value (*P*).

Table 4. Summary of findings on significant associations of parental age with CNVs impacting Autism Spectrum Disorder or Intellectual Disability (ASD/ID) or differentially brain expressed (DBE) genes

Independent variable (parental age)	*Dependent variable (gene list)*	N	*OR*	*LCL*	*UCL*	P
Maternal age	All CNVs in DBE	1292	0.969	0.939	0.999	0.046
Paternal age	Deletions in DBE	1292	1.034	1.004	1.064	0.025

Abbreviation: CNV, copy number variation. Only statistically significant findings ($\alpha = 0.05$) in adjusted models are shown, along with their associated odds ratio (OR), lower (LCL) and upper (UCL) 95% confidence limits and *P*-value (*P*).

unadjusted and adjusted) and composite scores (all CNV and duplications, unadjusted and adjusted) with the CNV effect being for a less severe presentation with better adaptive function than those without these CNVs. Associations were also found for maternal (any CNV, adjusted) and paternal age (deletions, adjusted).

Among *de novo* CNVs, significant associations emerged for verbal status, seizures and family type. Interestingly, these measures were not associated in the ASD/ID and DBE univariable analyses. Non-febrile seizures were associated with all of the investigated *de novo* CNV types, with the exception of the DBE CNVs in the adjusted model. A statistically significant association between family type and DBE *de novo* CNVs was found in both the unadjusted and adjusted models. However, a statistically significant association between verbal status and *de novo* ASD/ID CNVs was found only in the adjusted model.

Full details of the means or proportions for each phenotype by CNV group are detailed in Supplementary Tables 3–5.

DISCUSSION

CNVs in the ASD/ID list were primarily associated with communication and language domains, whereas CNVs in DBE genes were related to broader manifestations of adaptive function. There was substantial variation in the classification accuracy of the two sets of genes in the random forest analysis demonstrating specific versus broad impact. The novel use of random forests (validated by additional standard association analyses) highlights the importance of defining homogenous subgroups for genotype–phenotype correlation research, and the promise of their utility as a data reduction strategy in future research.

The greater classification accuracy of the ASD/ID gene list than the DBE list in the random forests suggests that CNVs impacting genes on the ASD/ID list are more closely associated with the sub-phenotypes of ASD. Further, the contribution of the language-related variables in these analyses supports the role of genes in the ASD/ID list in language and communication. By contrast, the statistically significant univariable associations between the clinical sub-phenotypes of interest and CNVs in the DBE gene list occurred consistently for duplications that influenced adaptive function. Therefore, the CNVs in the DBE gene list appear to have an impact on broader manifestations of neurodevelopmental disorders, rather than on specific features of ASD, thereby suggesting a more generalized role in function. Duplications tended to be associated with a less severe presentation than deletions, confirming earlier work that demonstrates duplications are less deleterious than deletions.[56] A possible explanation for this finding is that the absence of a protein product cannot be easily corrected downstream, while the over production of a protein product can be regulated (for example, 17p11.2, where the duplication at this locus causes milder Potocki–Lupski syndrome, while the deletion causes the more severe Smith–Magenis syndrome[57]).

There was some specificity of the sub-phenotypes associated with the ASD/ID, DBE candidate gene lists and *de novo* CNVs. In the association analyses, the CNVs in ASD/ID genes were primarily associated with communication and language domains (that is, verbal IQ, language delay and VABS communication) that are usually considered as core features of the ASD phenotype. The lack of uniformity of the associations between CNV types and the sub-phenotypes in the association analyses (deletions with language delay and duplications for the other measures) indicates the importance of investigating deletions and duplications separately.

The lack of consistent findings between the total and *de novo* CNVs could be attributable to lower statistical power of the reduced sample size with *de novo* CNVs. However, the analysis of the *de novo* subset produced findings that appeared relevant to ASD. For example, despite the low rate of seizures in this sample, *de novo* CNVs in the ASD/ID genes were associated with seizures and a lower overall level of language, suggesting that *de novo* events could be associated with greater severity of symptoms. The relationship between CNVs, epilepsy and ASD has been highlighted in other investigations such as recent studies that distinguish clinical characteristics of ASD cases with and without epilepsy.[58,59] Seizures were also associated with the presence of any *de novo* CNV, while verbal status was limited to *de novo* CNVs in the ASD/ID list, indicating a specific effect of genes in the ASD/ID list.

The association of the simplex family type with *de novo* mutations is expected in light of the assumption that those with inherited mutations are more likely to have a family history of ASD or other mental illness. However, the link between CNVs and the simplex family type emerged solely for DBE *de novo* CNVs rather than with ASD/ID CNVs. This finding suggests that family type may be an important source of heterogeneity in ASD. Systematic incorporation of family history is necessary to provide a comprehensive depiction of the role of both genetic risk factors and CNVs associated with ASD. Confirmation of these findings will require follow-up in an independent sample with a larger number of *de novo* mutations to afford sufficient statistical power for these comparisons.

These findings add to previous reports by specifically implicating deletions, as opposed to duplications, associated with advanced paternal age. Although chromosomal aberrations are associated with advanced paternal age,[60] a well-documented risk factor for ASD,[61–64] as well as neurodevelopmental disorders in general,[65] we did not demonstrate an association between paternal age and *de novo* mutations.[66] Somewhat surprisingly, advanced maternal age was negatively associated with DBE CNVs. However, there is a complex relationship between joint influence of maternal and paternal age on neurodevelopmental disorders.[67] Further support for the univariable associations was illustrated by the classification analyses where paternal age attained the highest classification accuracy for DBE deletions. Whereas maternal age did not have as high classification accuracy as paternal age and some other variables, it still retained relatively high levels of accuracy in a number of the models.

This study has several limitations including: (1) Lack of consistent collection of sub-phenotypic measures across sites that resulted in a large amount of missing phenotypic data (see Supplementary Materials for details of data completeness on core measures, Supplementary Table S1); (2) Potential bias in the candidate gene lists, due to publication bias or biased candidate gene selection in the samples contributed to the Gene Expression Omnibus, or omissions of important associated variants discovered after 2009; and (3) Potential bias toward continuous variable types,[53] and highly correlated variables[54] in the random forest analyses. In addition, we did not apply correction for multiple comparisons, so the association analysis results should be interpreted with caution. Traditional methods such as Bonferroni[68] or False Discovery Rate[69] are not valid in the context of non-independent CNV types, and a complex correlation structure within the phenotypic variables, which is why we attempted to complete data reduction *via* LPA and the random forest analyses. To date, no appropriate methods of correction for multiple comparisons for such a complex data structure have been established, and the best method would be replication of the findings in an independent sample. Such efforts are underway. Furthermore, the benefit of the random forest method is its use as a data reduction strategy, thereby reducing the need for multiple comparisons correction. Future research would not require the completion of association tests for all phenotypes, rather only for the ones with a high variable importance relative to the others.

The complexity of the role of CNVs (e.g., deletions versus duplications; *de novo* versus inherited) and their associations with different sub-phenotypes, as well as their environmental

moderators, highlights the importance of careful dissection of phenotypes and genotypes in elucidating the etiology of ASD.[70] This work documents the importance of investigation of the core features and correlates of ASD phenotypes in elucidating the sources of heterogeneity in the genetic architecture of ASD and other neurodevelopmental disorders. It also documents the need for inclusion of standardized measures of the specific aspects of neurodevelopment in future genetic studies. In spite of the difficulty in establishing direct clinical consequences of CNVs, deeper studies of CNVs to investigate their function could inform our understanding of mechanisms underlying disorders, syndromes, sub-phenotypes or clinical correlates (for example, McLysaght *et al.*[71]).

Although this work focuses solely on CNVs, it illustrates the complexity of the genetic architecture of ASD. The influence of environmental risk factors may have differential influence on those who harbor either rare or common susceptibility variants for ASD, and CNVs may also combine with other genetic risk factors to influence the expression of ASD. Our findings suggest that CNVs may have a role both in specific components of ASD and in their severity. These findings highlight the need for future research that integrates different genetic pathways to ASD, as well as environmental exposures that may either have additive or interactive influences on neurodevelopmental disorders.

CONFLICT OF INTEREST

The authors declare no conflicts of intrest.

ACKNOWLEDGMENTS

The authors wish to acknowledge the support and input of the AGP and their funders: Autism Speaks (USA), the Medical Research Council (UK), the Health Research Board (Ireland), the National Institutes of Health (USA), Genome Canada, the Canadian Institutes of Health Research and the Hilibrand Foundation. We thank all the children and their families who participated in this study, and the staff who facilitated this participation. Clinical data were coordinated by to a centralized Data Coordinating Center at the Research Institute at Nationwide Children's Hospital, Columbus, Ohio. We acknowledge the assistance of Ms. Ann Thompson and Ms. La Vonne Mangin in the completion of the phenotype data set required to complete this work. We appreciate the work of Dr Judith Miller in coordinating the AGP IQ committee and creating a single selected, composite IQ data set for the AGP. We also acknowledge technical support from the Trinity Centre for High Performance Computing. AKM was funded by the Irish Research Council for Science, Engineering and Technology, and the Meath Foundation (Ireland). The database of Genotypes and Phenotypes (dbGaP, http://www.ncbi.nlm.nih.gov/gap) accession number for the raw data from the ASD-affected families is phs0000267.v4.

REFERENCES

1 Schellenberg GD, Dawson G, Sung YJ, Estes A, Munson J, Rosenthal E *et al.* Evidence for multiple loci from a genome scan of autism kindreds. *Mol Psychiatry* 2006; **1111**: 1979.
2 Trikalinos TA, Karvouni A, Zintzaras E, Ylisaukko-oja T, Peltonen L, Jarvela I *et al.* A heterogeneity-based genome search meta-analysis for autism-spectrum disorders. *Mol Psychiatry* 2006; **11**: 29–36.
3 Szatmari P, Paterson AD, Zwaigenbaum L, Roberts W, Brian J, Liu XQ *et al.* Mapping autism risk loci using genetic linkage and chromosomal rearrangements. *Nat Genet* 2007; **39**: 319–328.
4 Weiss LA, Arking DE, Gene Discovery Project of Johns Hopkins and the Autism Consortium, Daly MJ, Chakravarti A. A genome-wide linkage and association scan reveals novel loci for autism. *Nature* 2009; **461**: 802–808.
5 Anney R, Klei L, Pinto D, Almeida J, Bacchelli E, Baird G *et al.* Individual common variants exert weak effects on the risk for autism spectrum disorders. *Hum Mol Genet* 2012; **21**: 4781–4792.
6 Anney R, Klei L, Pinto D, Regan R, Conroy J, Magalhaes TR *et al.* A genome-wide scan for common alleles affecting risk for autism. *Hum Mol Genet* 2010; **19**: 4072–4082.
7 Wang K, Zhang H, Ma D, Bucan M, Glessner JT, Abrahams BS *et al.* Common genetic variants on 5p14.1 associate with autism spectrum disorders. *Nature* 2009; **459**: 528–533.
8 Sanders SJ, Murtha MT, Gupta AR, Murdoch JD, Raubeson MJ, Willsey AJ *et al.* *De novo* mutations revealed by whole-exome sequencing are strongly associated with autism. *Nature* 2012; **485**: 237–241.
9 Neale BM, Kou Y, Liu L, Ma'ayan A, Samocha KE, Sabo A *et al.* Patterns and rates of exonic *de novo* mutations in autism spectrum disorders. *Nature* 2012; **485**: 242–245.
10 O'Roak BJ, Vives L, Girirajan S, Karakoc E, Krumm N, Coe BP *et al.* Sporadic autism exomes reveal a highly interconnected protein network of *de novo* mutations. *Nature* 2012; **485**: 246–250.
11 Huguet G, Ey E, Bourgeron T. The genetic landscapes of autism spectrum disorders. *Annu Rev Genomics Hum Genet* 2013; **14**: 191–213.
12 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**: 188–196.
13 Szatmari P, Merette C, Emond C, Zwaigenbaum L, Jones MB, Maziade M *et al.* Decomposing the autism phenotype into familial dimensions. *Am J Med Genet B Neuropsychiatr Genet* 2008; **147B**: 3–9.
14 Redon R, Ishikawa S, Fitch KR, Feuk L, Perry GH, Andrews TD *et al.* Global variation in copy number in the human genome. *Nature* 2006; **444**: 444–454.
15 Sebat J, Lakshmi B, Troge J, Alexander J, Young J, Lundin P *et al.* Large-scale copy number polymorphism in the human genome. *Science* 2004; **305**: 525–528.
16 Pinto D, Marshall C, Feuk L, Scherer SW. Copy-number variation in control population cohorts. *Hum Mol Genet* 2007 16 Spec No **2**: R168–R173.
17 Girirajan S, Rosenfeld JA, Coe BP, Parikh S, Friedman N, Goldstein A *et al.* Phenotypic heterogeneity of genomic disorders and rare copy-number variants. *N Engl J Med* 2012; **367**: 1321–1331.
18 O'Donovan MC, Kirov G, Owen MJ. Phenotypic variations on the theme of CNVs. *Nat Genet* 2008; **40**: 1392–1393.
19 Reiersen AM. Psychopathology in 22q11 deletion syndrome. *J Am Acad Child Adolesc Psychiatry* 2007; **46**942; author reply 942–944.
20 Guilmatre A, Dubourg C, Mosca AL, Legallic S, Goldenberg A, Drouin-Garraud V *et al.* Recurrent rearrangements in synaptic and neurodevelopmental genes and shared biologic pathways in schizophrenia, autism, and mental retardation. *Arch Gen Psychiatry* 2009; **66**: 947–956.
21 Moreno-De-Luca D, Mulle JG, Kaminsky EB, Sanders SJ, Myers SM, Adam MP *et al.* Deletion 17q12 is a recurrent copy number variant that confers high risk of autism and schizophrenia. *Am J Hum Genet* 2010; **87**: 618–630.
22 Sanders SJ, Ercan-Sencicek AG, Hus V, Luo R, Murtha MT, Moreno-De-Luca D *et al.* Multiple recurrent *de novo* CNVs, including duplications of the 7q11.23 Williams syndrome region, are strongly associated with autism. *Neuron* 2011; **70**: 863–885.
23 Girirajan S, Dennis MY, Baker C, Malig M, Coe BP, Campbell CD *et al.* Refinement and discovery of new hotspots of copy-number variation associated with autism spectrum disorder. *Am J Hum Genet* 2013; **92**: 221–237.
24 Buxbaum JD, Silverman JM, Smith CJ, Kilifarski M, Reichert J, Hollander E *et al.* Evidence for a susceptibility gene for autism on chromosome 2 and for genetic heterogeneity. *Am J Hum Genet* 2001; **68**: 1514–1520.
25 Qiao Y, Riendeau N, Koochek M, Liu X, Harvard C, Hildebrand MJ *et al.* Phenomic determinants of genomic variation in autism spectrum disorders. *J Med Genet* 2009; **46**: 680–688.
26 van Daalen E, Kemner C, Verbeek NE, van der Zwaag B, Dijkhuizen T, Rump P *et al.* Social Responsiveness Scale-aided analysis of the clinical impact of copy number variations in autism. *Neurogenetics* 2011; **12**: 315–323.
27 Connolly JJ, Glessner JT, Hakonarson H. A genome-wide association study of autism incorporating autism diagnostic interview-revised, autism diagnostic observation schedule, and social responsiveness scale. *Child Dev* 2013; **84**: 17–33.
28 Qiao Y, Tyson C, Hrynchak M, Lopez-Rangel E, Hildebrand J, Martell S *et al.* Clinical application of 2.7 M Cytogenetics array for CNV detection in subjects with idiopathic autism and/or intellectual disability. *Clin Genet* 2013; **83**: 145–154.
29 Vittinghoff E, McCulloch CE, Glidden DV, Shiboski SC. 5 Linear and non-linear regression methods in epidemiology and biostatistics. In: Rao CR, Rao DC, Miller JP (eds) *Handbook of Statistics: Epidemiology and Medical Statistics*. North Holland: Amsterdam, The Netherlands, 2007.
30 Hu-Lince D, Craig DW, Huentelman MJ, Stephan DA. The Autism Genome Project: goals and strategies. *Am J Pharmacogenomics* 2005; **5**: 233–246.
31 Pinto D, Delaby E, Merico D, Barbosa M, Merikangas A, Klei L *et al.* Convergence of genes and cellular pathways dysregulated in autism spectrum disorders. *Am J Hum Genet* 2014; **94**: 677–694.
32 Pinto D, Pagnamenta AT, Klei L, Anney R, Merico D, Regan R *et al.* Functional impact of global rare copy number variation in autism spectrum disorders. *Nature* 2010; **466**: 368–372.
33 American Psychiatric Association. *Diagnostic and Statistical Manual of Mental Disorders (DSM-IV)* 4th edn. American Psychiatric Association: Washington, DC, 1994.

34 Lord C, Rutter M, Goode S, Heemsbergen J, Jordan H, Mawhood L *et al.* Autism diagnostic observation schedule: a standardized observation of communicative and social behavior. *J Autism Dev Disord* 1989; **19**: 185–212.
35 Lord C, Rutter M, Le Couteur A. Autism Diagnostic Interview-Revised: a revised version of a diagnostic interview for caregivers of individuals with possible pervasive developmental disorders. *J Autism Dev Disord* 1994; **24**: 659–685.
36 Bierut LJ, Agrawal A, Bucholz KK, Doheny KF, Laurie C, Pugh E *et al.* A genome-wide association study of alcohol dependence. *Proc Natl Acad Sci USA* 2010; **107**: 5082–5087.
37 Figueiredo JC, Lewinger JP, Song C, Campbell PT, Conti DV, Edlund CK *et al.* Genotype-environment interactions in microsatellite stable/microsatellite instability-low colorectal cancer: results from a genome-wide association study. *Cancer Epidemiol Biomarkers Prev* 2011; **20**: 758–766.
38 Newcomb PA, Baron J, Cotterchio M, Gallinger S, Grove J, Haile R *et al.* Colon Cancer Family Registry: an international resource for studies of the genetic epidemiology of colon cancer. *Cancer Epidemiol Biomarkers Prev* 2007; **16**: 2331–2343.
39 Fox CS, Liu Y, White CC, Feitosa M, Smith AV, Heard-Costa N *et al.* Genome-wide association for abdominal subcutaneous and visceral adipose reveals a novel locus for visceral fat in women. *PLoS Genet* 2012; **8**: e1002695.
40 Gotham K, Pickles A, Lord C. Standardizing ADOS scores for a measure of severity in autism spectrum disorders. *J Autism Dev Disord* 2009; **39**: 693–705.
41 Sparrow SS, Cicchetti DV, Balla DA. *Vineland and Adaptive Behavior Scales (Vineland-II)* 2nd edn. Pearson: San Antonio, TX, 2005.
42 American Psychiatric Association. *Diagnostic and Statistical Manual of Mental Disorders (DSM-II)* 2nd edn. American Psychiatric Association: Washington, DC, 1968.
43 Colella S, Yau C, Taylor JM, Mirza G, Butler H, Clouston P *et al.* QuantiSNP: an Objective Bayes Hidden-Markov Model to detect and accurately map copy number variation using SNP genotyping data. *Nucleic Acids Res* 2007; **35**: 2013–2025.
44 Pinto D, Zhang J, Thiruv B, Wang Z, Feuk L, Hu P *et al.* A robust copy number variation discovery algorithm for multiple array platforms. *The American Society of Human Genetics Annual Meeting* 2008; **11**: 74.
45 Wang K, Li M, Hadley D, Liu R, Glessner J, Grant SF *et al.* PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome Res* 2007; **17**: 1665–1674.
46 Raychaudhuri S, Korn JM, McCarroll SA, Altshuler D, Sklar P, Purcell S *et al.* Accurately assessing the risk of schizophrenia conferred by rare copy-number variation affecting genes with brain function. *PLoS Genet* 2010; **6**: 9.
47 Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res* 2002; **30**: 207–210.
48 Muthén LK, Muthén BO. *Mplus User's Guide*. 6th edn Muthén & Muthén: Los Angeles, CA, 1998-2011).
49 Strobl C, Boulesteix AL, Kneib T, Augustin T, Zeileis A. Conditional variable importance for random forests. *BMC Bioinformatics* 2008; **9**: 307.
50 Strobl C, Boulesteix AL, Zeileis A, Hothorn T. Bias in random forest variable importance measures: illustrations, sources and a solution. *BMC Bioinformatics* 2007; **8**: 25.
51 Strobl C, Malley J, Tutz G. An introduction to recursive partitioning: rationale, application, and characteristics of classification and regression trees, bagging, and random forests. *Psychol Methods* 2009; **14**: 323–348.
52 Zhang H, Wang M, Chen X. Willows: a memory efficient tree and forest construction package. *BMC Bioinformatics* 2009; **10**: 130.
53 Altmann A, Tolosi L, Sander O, Lengauer T. Permutation importance: a corrected feature importance measure. *Bioinformatics* 2010; **26**: 1340–1347.
54 Archer KJ, Kimes RV. Empirical characterization of random forest variable importance measures. *Comput Stat Data Anal* 2008; **52**: 2249–2260.
55 SAS® 9.2SAS Institute Inc.: Cary, North Carolina, 2010.
56 Freeman JL, Perry GH, Feuk L, Redon R, McCarroll SA, Altshuler DM *et al.* Copy number variation: new insights in genome diversity. *Genome Res* 2006; **16**: 949–961.
57 Girirajan S, Campbell CD, Eichler EE. Human copy number variation and complex genetic disease. *Annu Rev Genet* 2011; **45**: 203–226.
58 Viscidi EW, Triche EW, Pescosolido MF, McLean RL, Joseph RM, Spence SJ *et al.* Clinical characteristics of children with autism spectrum disorder and co-occurring epilepsy. *PLoS ONE* 2013; **8**: e67797.
59 Mulligan CK, Trauner DA. Incidence and behavioral correlates of epileptiform abnormalities in autism spectrum disorders. *J Autism Dev Disord* 2013; **44**: 452–458.
60 Goriely A, McVean GA, Rojmyr M, Ingemarsson B, Wilkie AO. Evidence for selective advantage of pathogenic FGFR2 mutations in the male germ line. *Science* 2003; **301**: 643–646.
61 Durkin MS, Maenner MJ, Newschaffer CJ, Lee LC, Cunniff CM, Daniels JL *et al.* Advanced parental age and the risk of autism spectrum disorder. *Am J Epidemiol* 2008; **168**: 1268–1276.
62 Reichenberg A, Gross R, Weiser M, Bresnahan M, Silverman J, Harlap S *et al.* Advancing paternal age and autism. *Arch Gen Psychiatry* 2006; **63**: 1026–1032.
63 Hultman CM, Sandin S, Levine SZ, Lichtenstein P, Reichenberg A. Advancing paternal age and risk of autism: new evidence from a population-based study and a meta-analysis of epidemiological studies. *Mol Psychiatry* 2011; **16**: 1203–1212.
64 Grether JK, Anderson MC, Croen LA, Smith D, Windham GC. Risk of autism and increasing maternal and paternal age in a large north American population. *Am J Epidemiol* 2009; **170**: 1118–1126.
65 Saha S, Barnett AG, Foldi C, Burne TH, Eyles DW, Buka SL *et al.* Advanced paternal age is associated with impaired neurocognitive outcomes during infancy and childhood. *PLoS Med* 2009; **6**: e40.
66 Kong A, Frigge ML, Masson G, Besenbacher S, Sulem P, Magnusson G *et al.* Rate of *de novo* mutations and the importance of father's age to disease risk. *Nature* 2012; **488**: 471–475.
67 Croen LA, Najjar DV, Fireman B, Grether JK. Maternal and paternal age and risk of autism spectrum disorders. *Arch Pediatr Adolesc Med* 2007; **161**: 334–340.
68 Perneger TV. What's wrong with Bonferroni adjustments. *BMJ* 1998; **316**: 1236–1238.
69 Benjamini Y, Hochberg Y. Controlling the false discovery rate - a practical and powerful approach to multiple testing. *J Roy Stat Soc B Met* 1995; **57**: 289–300.
70 Lee C, Scherer SW. The clinical context of copy number variation in the human genome. *Exp Rev Mol Med* 2010; **12**: e8.
71 McLysaght A, Makino T, Grayton HM, Tropeano M, Mitchell KJ, Vassos E *et al.* Ohnologs are overrepresented in pathogenic copy number mutations. *Proc Natl Acad Sci USA* 2014; **111**: 361–366.

Supplementary Information accompanies the paper on the Molecular Psychiatry website (http://www.nature.com/mp)