An evidence-based approach to establish the functional and clinical significance of copy number variants in intellectual and developmental disabilities

Erin B. Kaminsky, PhD[1], Vineith Kaul, MS[1], Justin Paschall, PhD[2], Deanna M. Church, PhD[2], Brian Bunke, BS[1], Dawn Kunig, BS[1], Daniel Moreno-De-Luca, MD, MSc[1], Andres Moreno-De-Luca, MD[1], Jennifer G. Mulle, MHS, PhD[1], Stephen T. Warren, PhD[1,3], Gabriele Richard, MD[4], John G. Compton, PhD[4], Amy E. Fuller, MS[4], Troy J. Gliem, BS[5], Shuwen Huang, PhD[6,7], Morag N. Collinson, BS[6], Sarah J. Beal, BS[6], Todd Ackley, BS[8], Diane L. Pickering, MS[9], Denae M. Golden, BS[9], Emily Aston, BS[10], Heidi Whitby, BS[10], Shashirekha Shetty, PhD[10], Michael R. Rossi, PhD[1], M. Katharine Rudd, PhD[1], Sarah T. South, PhD[10], Arthur R. Brothman, PhD[10], Warren G. Sanger, PhD[9], Ramaswamy K. Iyer, PhD[8], John A. Crolla, PhD[6,7], Erik C. Thorland, PhD[5], Swaroop Aradhya, PhD[4], David H. Ledbetter, PhD[1], and Christa L. Martin, PhD[1]

Purpose: Copy number variants have emerged as a major cause of human disease such as autism and intellectual disabilities. Because copy number variants are common in normal individuals, determining the functional and clinical significance of rare copy number variants in patients remains challenging. The adoption of whole-genome chromosomal microarray analysis as a first-tier diagnostic test for individuals with unexplained developmental disabilities provides a unique opportunity to obtain large copy number variant datasets generated through routine patient care. **Methods:** A consortium of diagnostic laboratories was established (the International Standards for Cytogenomic Arrays consortium) to share copy number variant and phenotypic data in a central, public database. We present the largest copy number variant case-control study to date comprising 15,749 International Standards for Cytogenomic Arrays cases and 10,118 published controls, focusing our initial analysis on recurrent deletions and duplications involving 14 copy number variant regions. **Results:** Compared with controls, 14 deletions and seven duplications were significantly overrepresented in cases, providing a clinical diagnosis as pathogenic. **Conclusion:** Given the rapid expansion of clinical chromosomal microarray analysis testing, very large datasets will be available to determine the functional significance of increasingly rare copy number variants. This data will provide an evidence-based guide to clinicians across many disciplines involved in the diagnosis, management, and care of these patients and their families.
***Genet Med* 2011:13(9):777–784.**

Key Words: *CNVs, evidence-based approach, clinical significance, ID/DD, consortium*

Copy number variation, defined as the gain or loss of genomic material >1 kb in size,[1] has been the subject of intense research in both normal and disease populations over the last several years. These investigations were made possible by the completion of the Human Genome Project, which provided a detailed physical map and high-quality reference assembly of the human genome[2] and enabled the development of whole-genome array technologies capable of accurate determination of copy number at very high resolution.

Copy number variants (CNVs) are common in normal individuals and have been identified in approximately 35% of the human genome.[1] When present as hemizygous events in normal individuals, these imbalances are considered "benign" (i.e., no major phenotypic effect on human development); however, their role as susceptibility loci in common and complex genetic diseases and traits is now being actively explored. Data from control populations are being collected in databases of normal variation, including the Database of Genomic Variants[1] and the Database of Genomic Structural Variation (dbVar) (http://www.ncbi.nlm.nih.gov/dbvar).[3] These large datasets will contribute to a human gene dosage map through exclusion by defining those regions for which single copy loss or gain is tolerated and do not produce an overtly abnormal phenotype.

CNVs have also been identified as one of the most common causes of human disease. In fact, one of the earliest and most significant clinical benefits of the Human Genome Project has been the application of whole-genome CNV analysis to evaluate individuals with developmental disabilities, including developmental delay (DD), intellectual disability (ID), autism, epilepsy,

From the [1]Department of Human Genetics, Emory University School of Medicine, Atlanta, Georgia; [2]National Center for Biotechnology Information, Bethesda, Maryland; [3]Departments of Pediatrics and Biochemistry, Emory University School of Medicine, Atlanta, Georgia; [4]GeneDx, Gaithersburg, Maryland; [5]Department of Laboratory Medicine and Pathology, Mayo Clinic College of Medicine, Rochester, Minnesota; [6]Wessex Regional Genetics Laboratory, Salisbury District Hospital, Salisbury, Wiltshire, United Kingdom; [7]National Genetics Reference Laboratory (Wessex), Wessex Regional Genetics Laboratory, Salisbury District Hospital, Salisbury, Wiltshire, United Kingdom; [8]Michigan Medical Genetics Laboratories, Ann Arbor, Michigan; [9]Human Genetics Laboratory, University of Nebraska Medical Center, Omaha, Nebraska; and [10]University of Utah School of Medicine and ARUP Laboratories, Salt Lake City, Utah.

Christa L. Martin, PhD, Department of Human Genetics, Emory University School of Medicine, 615 Michael St., Suite 301, Atlanta, GA 30322. E-mail: christa.martin@emory.edu and David H. Ledbetter, PhD, Geisinger Health System, 100 North Academy Drive, MC 22-01, Danville, PA 17822. E-mail: dhledbetter@geisinger.edu.

Shashirekha Shetty is currently at the Cleveland Clinic, Cleveland, Ohio. Ramaswamy K. Iyer is currently at Inova Translational Medicine Institute, Falls Church, Virginia. David H. Ledbetter is currently at Geisinger Health System, Danville, Pennsylvania.

Disclosure: See Acknowledgments for author disclosures.

Supplemental digital content is available for this article. Direct URL citations appear in the printed text and are provided in the HTML and PDF versions of this article on the journal's Web site (www.geneticsinmedicine.org).

Published online ahead of print August 12, 2011.

DOI: 10.1097/GIM.0b013e31822c79f9

and/or birth defects, a group of disorders representing up to 14% of the population.[4] Commonly referred to as cytogenetic or chromosomal microarrays (CMA), these technologies have quickly replaced the standard G-banded karyotype as the first-tier genetic test for the evaluation of this patient population.[5,6] There are many technology platforms available for whole-genome copy number analysis at resolutions of 100–500 kb (compared with ~5–10 Mb for karyotype), with even higher resolution at "clinical targets," such as individual genes in which haploinsufficiency leads to dominant Mendelian disorders. From numerous published studies, the yield of clinically significant or pathogenic CNVs (pCNVs) by CMA is 15–20%, compared with a yield of approximately 3–5% by standard cytogenetic analysis in the same patient population.[5]

In an important subset of CMA cases, the potential functional significance of a particular CNV may be unknown and is referred to as a variant of uncertain clinical significance (VOUS). Parental and family studies can be helpful in the clinical interpretation of these cases, as a de novo occurrence of the CNV strengthens the evidence that it is pathogenic. However, the significance of many CNVs still remains uncertain even after familial studies due to variable expressivity or incomplete penetrance. Therefore, it would be extremely beneficial to improve our knowledge of the functional significance of CNVs throughout the genome by performing comparative analyses of large datasets from case cohorts and control populations to definitively associate specific genomic regions with human disease.

Herein, we describe genome-wide CNV results from the first dataset from the International Standards for Cytogenomic Arrays (ISCA) consortium[5] (https://www.iscaconsortium.org/) that includes analysis of 15,749 cases and 10,118 controls. This study was designed to assess the frequency of CNVs in this population and initiate an evidence-based process to determine the functional significance of structural variation across the genome. Compared with individually rare CNVs, recurrent CNVs lend themselves to large case-control studies due to their relatively higher frequency. Therefore, we have focused our initial analysis on 14 recurrent CNV regions to statistically assess the correlation between rare CNVs and developmental disorders. Furthermore, ongoing analysis of the ISCA CNV dataset compared with normal structural variation will delineate genomic regions and individual genes that are subject to dosage effects resulting in intellectual and other developmental disabilities. Such efforts will result in a human gene dosage map for developmental disorders.

MATERIALS AND METHODS

Cases

This study adhered to guidelines set by the institutional review boards at the participating laboratories. CMA was performed in a subset of clinical ISCA laboratories on cases referred for diagnostic testing with various indications including unexplained DD, ID, dysmorphic features, multiple congenital anomalies, autism spectrum disorders (ASDs), or clinical features suggestive of a chromosomal syndrome. Anonymized data from 15,749 cases were included.

CNV detection

CMA was carried out following standard procedures. We used a consensus microarray design, focusing on unique genomic regions and avoiding repetitive sequences.[7] The arrays were either 44K or 105K custom-designed 60-mer oligonucleotide arrays (Agilent Technologies, Santa Clara, CA) with a whole-genome backbone plus targeted, higher density coverage of known disease-causing regions.[7] The backbone coverage included probes spaced every approximately 35–75 kb, allowing for CNVs of approximately 250 kb and greater to be detected. All clinically relevant CNVs $\geq$ 500 kb in the backbone are reported in this study. The 500 kb threshold in the backbone regions was used as this size limit was consistently used as the reporting criteria by the ISCA laboratories. For the targeted regions, we could identify imbalances of approximately 20–50 kb.

Arrays were scanned using a GenePix Autoloader 4200AL, GenePix 4000B (Molecular Devices, Sunnyvale, CA) or Agilent scanner (Agilent Technologies, Santa Clara, CA). Results were analyzed using Feature Extraction and DNA Analytics software packages (Agilent Technologies, Santa Clara, CA). Data include only those imbalances that contained at least four consecutive probes with abnormal $\log_2$ ratios. Data are presented as minimum coordinates (sequence positions of the first and last probes within the CNV) in the NCBI36 genome assembly.

CNVs were categorized by clinical laboratories as pathogenic, VOUS, or benign based on known clinically relevant regions, gene content, and inheritance pattern as described previously.[5,8] For both deletions and duplications, the genes located within the CNVs were assessed, as well as neighboring genes. Imbalances that involved large genomic segments from the chromosomal backbone coverage were considered to be likely pathogenic if they contained multiple known genes and did not overlap a confirmed benign CNV region. CNVs were classified as pathogenic if the CNV included an autosomal dominant gene known to cause a disease phenotype. The genomic regions associated with known pathogenic and benign CNVs are listed in Tables, Supplemental Digital Content 1, http://links.lww.com/GIM/A196 and were also deposited into dbVar (nstd45). Because the clinical laboratories that contributed data used different standards for reporting benign CNVs, an accurate assessment of the frequency of these benign CNVs was impossible for this dataset; therefore, benign CNVs identified in cases with otherwise normal array results were not included in this study.

Confirmation of abnormal array findings were carried out by fluorescence in situ hybridization (FISH), quantitative polymerase chain reaction, standard G-banded chromosome analysis, multiplex ligation-dependent probe amplification, or a second array analysis, depending on the size of the observed CNV. As the great majority of pathogenic changes were confirmed by an independent method, the genotypic data quality is extremely high, providing a large dataset with high fidelity. Parental studies by FISH, quantitative polymerase chain reaction, multiplex ligation-dependent probe amplification, or array analysis were conducted to determine the inheritance in a subset of cases where parental samples were referred for follow-up testing. To the best of our knowledge, results from testing of parental and siblings' samples were excluded from the final dataset if they showed the same genomic imbalance as the proband.

We developed an automated program to scan the data for inconsistencies in clinical interpretation for two or more reported genomic imbalances that overlapped in length by more than 50% but that were classified differently (as pathogenic, VOUS, or benign). This program flagged the genomic regions in which there was inconsistent annotation of CNVs, and these CNVs were subsequently reviewed and, where appropriate, assigned a single classification. For cases with complex rearrangements involving several CNVs, the interpretation was based on each individual CNV. The reported CNVs from this

study are included in Table, Supplemental Digital Content 2, http://links.lww.com/GIM/A197 and were submitted to dbVar (nstd37). The number of genes was assessed by counting partial and whole genes included in the region based on the UCSC known gene track.

Statistical analysis

Our initial approach focuses on recurrent events as they are more common and lend themselves to case-control analysis; future studies will focus on nonrecurrent CNVs as large enough case numbers become available. Recurrent rearrangements mediated by segmental duplications were identified by comparison with previously described hotspot regions.[9] Imbalances were considered recurrent if they included the critical region of the deletion/duplication event and, based on probe coverage, were likely mediated by paired, flanking segmental duplications. We carried out statistical analysis of 14 selected regions including (Table 1 for chromosome coordinates) 1q21 thrombocytopenia-absent radius region,[10,11] distal 1q21.1,[12,13] 3q29,[14,15] 5q35,[16,17] 7q11.23,[18,19] 8p23.1,[20,21] 15q11.2-q13,[22–24] 15q13,[25,26] 16p13.11,[27,28] 16p11.2,[29–31] 17p11.2,[32,33] 17q12,[34–36] 17q21.31,[37–39] and 22q11.2.[40,41] For the 1q21 regions, if the imbalance included both 1q21 thrombocytopenia-absent radius[10] and the distal 1q21.1 region,[12] the imbalance was included in the distal 1q21.1[12] frequency. In the 15q11q13 region, imbalances that spanned BP2–BP5[42] were counted in the BP2–BP3 frequency and not the BP4–BP5 frequency. Both the smaller and larger rearrangements (~1.5 and ~3.0 Mb) for 16p13.11[28] and 22q11[43] were included in their respective CNV categories. For this study, we excluded recurrent CNVs involving 17p12 (HNPP/CMT1A) as these CNVs are either not associated with cognitive defects or are late-onset in nature (and, therefore, not expected to be enriched in our mostly pediatric patient population) and 15q11 (BP1–2) which were not consistently reported by the contributing laboratories. CNV data from 10,118 individuals from control populations were obtained from several recent reports.[44–47] Processed CNV data were used directly from three of the previous control studies.[44–46] For the data from the article by Shi et al.,[47] we performed CNV analysis of the raw data for regions of interest using the Affymetrix Power Tools software (Affymetrix, Santa Clara, CA). Log_2 ratio data were extracted and analyzed using the BEAST algorithm (Satten et al., submitted). All *P* values and odds ratios for case-control analyses were calculated using Fisher's exact test.

RESULTS

CNV characterization

We analyzed data from 15,749 whole-genome oligonucleotide arrays on individuals who presented for diagnostic array testing with abnormal clinical phenotypes including DD/ID, ASD, and/or multiple congenital anomalies. We detected 4628 imbalances consistent with our reporting criteria (defined in "Materials and Methods") and classified 2691 (17.1%) as pathogenic (pCNVs), in line with prior reports of the yield from CMA in diagnostic testing.[5] As a single individual may have had multiple pCNVs (e.g., unbalanced translocations), the diagnostic yield for this dataset was 14.7% (2321 cases with pCNV/15,749 total cases). Excluding 106 whole-chromosome aneuploidies, there were 2585 pCNVs with a mean size of approximately 6.5 Mb (median of ~2.8 Mb) and a mean of approximately 69 genes per CNV (median of 44 genes). Deletions were more commonly interpreted as pathogenic than duplications, accounting for 67.9% of the imbalances.

In 9.3% of cases, an observed genomic imbalance was classified as a VOUS, as there was insufficient evidence to conclude the CNV was either pathogenic or benign. There were ultimately 1468 CNVs classified as VOUS, with a mean size of 765 kb (median of 569 kb) and a mean of approximately 10 genes per CNV (median of five genes). Duplications were more common than deletions, accounting for 68.8% of the imbalances.

The inheritance of a CNV was determined in a subset of cases to aid in the clinical interpretation and where both parental specimens were available. Of the 1412 CNVs with known inheritance, 566 (~40%) were found to be de novo. The majority of the de novo events (513 CNVs, ~91%) were classified as pathogenic, whereas 51 CNVs (~9%) were classified as uncertain. Two de novo CNVs, interpreted to be benign, were incidentally identified in the course of parental studies to determine the inheritance of other CNVs classified as VOUS. The de novo benign CNVs included a duplication of the beta-defensin cluster on chromosome 8p and a duplication of the *CHRNA7* (OMIM# 118511) gene on chromosome 15q; both of these CNVs have been observed as common polymorphisms in control populations.

Frequency of recurrent events

A subset of the imbalances identified by CMA includes recurrent imbalances that result from rearrangements between low-copy repeats, also known as segmental duplications. These rearrangements cause genomic disorders that have been recently reviewed.[48] Sharp et al.[9] described 130 rearrangement hotspots in the human genome by defining these regions as large genomic segments (50 kb-10 Mb) that are flanked by segmental duplications ≥ 10 kb in size and ≥95% identical. Of all CNVs detected in this case cohort, approximately 24% result from rearrangements between segmental duplications.

Tables 1 and 2 list the frequencies in the ISCA dataset for 14 CNV regions associated with recurrent deletions and duplications, respectively. It is important to note that many of the recognizable recurrent syndromes may still be tested for by targeted FISH studies, rather than CMA. As cases ascertained from FISH testing were not included in this study, the frequencies of such syndromes are likely underestimated.

For the 14 recurrent regions, the number of deletions and duplications were often unequal, which can be explained by ascertainment (recurrent duplications may result in milder phenotypes and, therefore, not be ascertained in our cohort of affected individuals) and mechanism (deletions generated by non-allelic homologous recombination occur more frequently than duplications).[49] Not surprisingly, the most common deletion in this cohort, with 93 cases (1 in 169 abnormal cases), was the 22q11.2 deletion (OMIM# 188400),[40] whereas the reciprocal duplication (OMIM# 608363) with a milder phenotype[41] was detected in only 32 cases. The most common recurrent duplication in our dataset was in 16p13.11, seen in 45 cases, whereas the reciprocal deletion associated with neurodevelopmental defects was detected in only 22 cases. For both deletions and duplications, the second most commonly affected region was the recurrent 16p11.2 CNV (OMIM# 611913). Both deletions and duplications of this region have been reported in individuals with an abnormal neurologic phenotype.[30] The frequency of the 16p11.2 deletion in this abnormal cohort is approximately 1 in 235. Therefore, this CNV was detected nearly as often as the 22q11.2 deletions, indicating that this CNV is also a frequent cause of intellectual and developmental disabilities.

Table 1 Frequencies of recurrent deletions

Deleted region	Syndrome/phenotype	Approximate minimum coordinates (NCBI36)	No. cases	Frequency in 15,749 cases
22q11.2	22q11.2 deletion syndrome[40] (1.5 and 3 Mb)	chr22:17,400,436–18,676,130	93	1 in 169
16p11.2	Autism[30]	chr16:29,557,497–30,107,356	67	1 in 235
1q21.1	ID, microcephaly, cardiac, and cataracts[12,13]	chr1:145,044,110–145,861,130	55	1 in 286
15q13.2-q13.3 BP4-BP5	ID and epilepsy[25]	chr15:28,924,396–30,232,700	46	1 in 342
15q11.2-q13 BP2-BP3	Prader-Willi/Angelman syndrome[22] (BP1/2–3)	chr15:21,309,483–26,230,781	41	1 in 384
7q11.23	Williams syndrome[18]	chr7:72,382,390–73,780,449	34	1 in 463
16p13.11	Autism, ID, and schizophrenia[27,28] (1.5 and 3 Mb)	chr16:15,411,955–16,199,769	22	1 in 716
17q21.31	17q21 deletion syndrome[37,38]	chr17:41,060,948–41,650,183	22	1 in 716
17q12	Renal cysts, diabetes, autism, and schizophrenia[34–36]	chr17:31,930,169–33,323,031	18	1 in 875
1q21	Thrombocytopenia-absent radius (TAR) syndrome[10]	chr1:144,097,430–144,463,097	17	1 in 926
17p11.2	Smith-Magenis syndrome[32]	chr17:16,723,271–20,234,630	16	1 in 984
8p23.1	8p23.1 deletion syndrome[20]	chr8:8,156,705–11,803,128	10	1 in 1575
3q29	3q29 deletion syndrome[14,15]	chr3:197,240,451–198,829,062	9	1 in 1750
5q35	Sotos syndrome[16]	chr5:175,661,584–176,946,567	8	1 in 1969

The frequencies may be underestimated, as clinically recognizable recurrent deletion syndromes could be tested using FISH studies, rather than aCGH.

Table 2 Frequencies of recurrent duplications

Duplicated region	Syndrome/ phenotype	No. cases	Frequency in 15,749 cases
16p13.11	Variable phenotype[27,28] (1.5 and 3 Mb)	45	1 in 350
16p11.2	Autism[30]	39	1 in 404
15q11.2-q13 BP2-BP3	Autism[23,24] (BP1/2–3)	35	1 in 450
22q11.2	Variable phenotype[41] (1.5 and 3 Mb)	32	1 in 492
1q21.1	ID and autism[12,13]	28	1 in 562
17q12	Epilepsy[34]	21	1 in 750
7q11.23	Autism[19]	16	1 in 984
17p11.2	Potocki-Lupski syndrome[33]	15	1 in 1,050
15q13.2-q13.3 BP4-BP5	Psychiatric disease[26]	14	1 in 1,125
1q21	Reciprocal duplication of TAR region[11]	9	1 in 1,750
3q29	Variable phenotype[15]	8	1 in 1,969
8p23.1	Variable phenotype[21]	6	1 in 2,625
5q35	Short stature, microcephaly, and speech delay[17]	2	1 in 7,875
17q21.31	Behavioral problems[39]	0	Unknown

Frequency of nonrecurrent events

Of all CNVs detected in this case cohort, most (~76%) were individually rare and not mediated by segmental duplications. This large group of CNVs provides a resource to examine regions of the genome that contain multiple CNVs with overlapping segments of deleted or duplicated material to define genotype-phenotype correlations. As an example, we highlight three recently described regions (2p15 deletion,[50] 16q24.3 deletion,[51] and 17p13 duplication[52]) where overlapping de novo CNVs were characterized to define the associated phenotype and identify candidate genes. In the ISCA case cohort, we found four de novo deletions in 2p15 with a smallest region of overlap (SRO) of approximately 2.4 Mb, five de novo deletions in 16q24 with a SRO of approximately 450 kb, and four de novo duplications in 17p13 with a SRO of approximately 312 kb. As the ISCA database grows, cases such as these will prove invaluable for identifying disease-causing genes.

Case-control analysis to define functional significance

The CNVs identified in this study of individuals with neurodevelopmental disorders are rare and highly heterogeneous, with no single CNV being identified in more than 1% of the cases. Therefore, methods are needed to begin to statistically assess the relationship between such rare variation and human disease. For this study, we first focused on deletions and duplications of 14 recurrent genomic regions as their relative frequency is higher than CNVs involving nonrecurrent regions. We selected 14 of the most common and clinically relevant recurrent CNVs (listed in "Materials and Methods") for a formal case-control study to initiate an evidence-based process for defining the clinical significance of structural variation across the genome. Many of these 14 regions have inconclusive or contradictory data in the literature regarding their phenotypic implications, so a targeted analysis of these regions is needed to inform their functional significance.

Tables 3 and 4 list the results of these analyses for recurrent deletions and duplications, respectively. We compared the ISCA case cohort of 15,749 cases to 10,118 combined controls from several recent publications.[44–47] These reports used microarrays with levels of resolution equivalent to or higher than

Table 3 Case-control analysis of recurrent deletions

Deleted region	Initial call	Final call	Cases	Controls	OR	Lower 95% CI	Upper 95% CI	*P*	Study by Itsara et al.[45]
22q11.2	pCNV	pCNV	93	0	∞	15.96	∞	9.15×10^{-21}	7.93×10^{-9}
16p11.2	pCNV	pCNV	67	5	8.64	3.52	27.49	6.34×10^{-10}	0.186
1q21.1	pCNV	pCNV	55	3	11.82	3.84	59.07	5.38×10^{-9}	1.67×10^{-4}
15q13.2-q13.3 BP4-BP5	pCNV	pCNV	46	0	∞	7.71	∞	1.44×10^{-10}	1.08×10^{-5}
15q11.2-q13 BP2-BP3	pCNV	pCNV	41	0	∞	6.84	∞	2.77×10^{-9}	
7q11.23	pCNV	pCNV	34	0	∞	5.62	∞	8.49×10^{-8}	
16p13.11	pCNV	pCNV	22	3	4.72	1.42	24.62	0.0063	
17q21.31	pCNV	pCNV	22	0	∞	3.52	∞	2.49×10^{-5}	
17q12	pCNV	pCNV	18	0	∞	2.83	∞	0.00015	
1q21	pCNV	pCNV	17	1	10.93	1.71	456.06	0.0026	[b]
17p11.2	pCNV	pCNV	16	0	∞	2.48	∞	0.00045	
8p23.1	pCNV	pCNV	10	0	∞	1.44	∞	0.0084	
3q29	pCNV	pCNV	9	0	∞	1.27	∞	0.0147	0.164
5q35	pCNV	pCNV	8	0	∞	1.10	∞	0.026	

[a]Itsara et al.[45] performed a meta-analysis of segmental duplication-mediated regions on 6860 cases and 5674 controls. For regions in common, the *P* value assessing the difference in CNV frequency between the cases and controls was included.
[b]The 1q21 and 1q21.1 regions were combined in the analysis of Itsara et al.[45]

Table 4 Case-control analysis of recurrent duplications

Duplicated region	Initial call	Final call	Cases	Controls	OR	Lower 95% CI	Upper 95% CI	*P*	Study by Itsara et al.[45]
16p13.11	VOUS	VOUS	45	20	1.45	0.84	2.59	0.203	
16p11.2	VOUS	pCNV	39	4	6.28	2.26	24.19	2.50×10^{-5}	0.100
15q11.2-q13 BP2-BP3	pCNV	pCNV	35	0	∞	5.79	∞	4.57×10^{-8}	2.69×10^{-4}
22q11.2	pCNV	pCNV	32	5	4.12	1.59	13.54	0.0011	0.330
1q21.1	pCNV	pCNV	28	3	6.00	1.85	30.88	0.0004	0.041
17q12	pCNV	pCNV	21	4	3.38	1.14	13.53	0.022	
7q11.23	pCNV	pCNV	16	1	10.29	1.60	430.72	0.0046	
17p11.2	pCNV	pCNV	15	0	∞	2.31	∞	0.0008	
15q13.2-q13.3 BP4-BP5	VOUS	VOUS	14	3	3.00	0.84	16.28	0.083	[b]
1q21	VOUS	VOUS	9	12	0.48	0.179	1.25	0.116	[c]
3q29	pCNV	VOUS	8	1	5.14	0.69	227.96	0.100	1
8p23.1	pCNV	VOUS	6	0	∞	0.76	∞	0.088	
5q35	pCNV	VOUS	2	0	∞	0.12	∞	0.52	
17q21.31	N/A	N/A	0	0	nd	nd	nd	nd	

[a]Itsara et al.[45] performed a meta-analysis of segmental duplication-mediated regions on 6860 cases and 5674 controls. For regions in common, the *P* value assessing the difference in CNV frequency between the cases and controls was included.
[b]The 15q11.2-q13 and 15q13.2-q13.3 duplication regions were combined in the analysis of Itsara et al.[45]
[c]The 1q21 and 1q21.1 regions were combined in the analysis of Itsara et al.[45]
N/A, not applicable; nd, not determined.

the ISCA array design; thus, there should be no significant difference in sensitivity in the calls between the case and control datasets given that the 14 regions analyzed in this study were approximately 600 kb or greater. Although not all the controls used in these studies were formally assessed for neurocognitive abnormalities, these datasets have been used before as control populations in other studies. Itsara and colleagues[45] previously performed a meta-analysis of segmental duplication mediated

regions on 6860 abnormal individuals and 5674 control individuals.[45] For regions in common with our study, the CNV *P* values from the previous study are included in Tables 3 and 4 for comparison.

All 14 recurrent deletions were significantly overrepresented in cases compared with controls (Table 3), demonstrating each is a pCNV. The 22q11.2 deletion was not seen in controls, confirming the pathogenic nature of this known disease-causing CNV ($P = 9.15^{-21}$). The 16p11.2 deletion was observed in 67 cases in the ISCA cohort, but only five 16p11.2 deletions were found among the control population, providing strong evidence for the pathogenic nature of this CNV (OR = 8.64; $P = 6.34^{-10}$).

Other recurrent deletions detected with a high frequency in the abnormal cohort include those in 1q21.1 (OMIM# 612474; OR = 11.82; $P = 5.38^{-09}$), 15q13 (OMIM# 612001; OR = ∞; $P = 1.44^{-10}$), and 15q11-q13 (breakpoint [BP] 1/2–3 of the Prader-Willi [OMIM# 176270]/Angelman [OMIM# 105830] syndromes region; OR = ∞; $P = 2.77^{-09}$). We also identified 18 deletions involving the 17q12 region (OMIM# 137920); these deletions were initially reported to have no neurocognitive phenotype.[34] More recent studies, however, have shown an association between 17q12 deletions and DDs[35] and autism/schizophrenia.[36] The absence of the 17q12 deletion in 10,118 controls is strong evidence for classifying this deletion as pathogenic ($P = 0.00015$).

We also analyzed the reciprocal duplications of the 14 recurrent deletion CNVs (Table 4). Determining the functional significance for duplications can be more challenging due to the more subtle and milder phenotypes associated with an increase in gene dosage compared with the more severe phenotypic effects of haploinsufficiency. The initial classifications for these CNVs ranged from VOUS to pathogenic events.

For six duplications initially classified as pathogenic (in 1q21.1 [OMIM# 612475], 7q11.23 [OMIM# 609757], 15q11.2-q13 [OMIM# 608636], 17p11.2 [OMIM# 610883], 17q12, and 22q11.2), the case-control analysis corroborated this classification (Table 4). The 16p11.2 duplication was initially classified as a VOUS; however, our case-control analysis demonstrates that this duplication is most likely pathogenic (OR = 6.28; $P = 2.5^{-05}$).

Several recurrent CNV regions have had equivocal reports in the literature. For example, duplications of 16p13.11 have been previously suggested to be linked with autism,[27] whereas another study proposed that the duplications may be a benign CNV.[28] Because of the uncertainty in the literature, duplications in three regions (16p13.11, 15q13 BP4–5, and proximal 1q21) were initially classified as VOUS. As these duplications were not significantly enriched in the ISCA case cohort or in controls, the classification of these CNVs remains uncertain at this time using the formal case-control assessment.

Duplications of 3q29,[15] 8p23.1,[21] and 5q35[17] have been previously reported in individuals with abnormal phenotypes. In this case-control analysis, these events were identified more often in cases than in controls. However, because of the low frequency of these duplications in the clinically affected population, the differences were not statistically significant. Therefore, as a conservative approach, we would classify these three CNVs as uncertain until larger sample sizes are available. More detailed phenotypic investigations of individuals carrying duplications of 3q29, 8p23.1, and 5q35 in the ISCA cohort and other patient cohorts will help to clarify whether the observed phenotypes are consistent with the previously reported syndromes associated with these duplications.

DISCUSSION

There are now many published reports of the significant role of rare, de novo CNVs with major phenotypic effects in various human disease populations, including intellectual disabilities, ASDs, epilepsy, and schizophrenia, among others. Many of these studies are based on well-phenotyped research cohorts that were originally collected and characterized to optimize the ability to detect small effects in genome-wide association studies. Although positive associations have been identified for a few common diseases through these efforts, a surprising and remarkable finding has been the identification of rare, de novo CNVs with major phenotypic effects, particularly in neurocognitive and behavioral disorders. Because these events are rare, obtaining adequate evidence for their functional role in disease causation requires very large sample sizes and large control populations.

An alternative model for assessing the contribution of CNVs to disease, which has been used particularly in the study of children with unexplained developmental disabilities and congenital anomalies, has been the reporting of case series from clinical laboratory testing. Most of these published studies have represented CNV data from single laboratories and were based on previous generation targeted array analysis using bacterial artificial chromosome genomic clones.[5] Compared with analysis of research cohorts of well-phenotyped patients, the amount and quality of phenotypic data associated with clinical laboratory referrals is often quite limited.

For this study, we have combined these two approaches by exploiting a large CNV dataset derived from a consortium of clinical laboratories to explore the frequency and functional significance of rare CNVs. Our analysis of the first 15,749 ISCA cases, one of the largest CNV studies to date, has confirmed the power of this approach. We have defined the frequency (17.1%) of pCNVs in a cohort of individuals with intellectual and developmental disabilities and performed formal case-control studies of selected recurrent genomic regions whose frequency was sufficient for statistical analysis.

The determination of whether a CNV contributes to an abnormal phenotype depends on many factors, including gene content, previous evidence of pCNVs in the region, type of CNV (deletion or duplication), inheritance pattern, and frequency in unaffected populations. As such, larger CNVs may be more likely to be classified as pathogenic as they have a higher chance of including a dosage-sensitive gene and/or they include a larger number of genes that cumulatively result in an abnormal phenotype. Our experience, as well as that of other groups,[53] has shown that the classification of a previously unreported CNV not associated with known disease genes can vary. To address such discrepancies, we used case-control statistical evidence for 14 selected recurrent CNV regions to objectively determine their significance.

We analyzed deletions and duplications of each region separately, resulting in 28 total recurrent CNV regions. Using this approach, we demonstrated and confirmed the pathogenic nature of 20 recurrent regions. For the 16p11.2 duplications that had previously been reported as uncertain in the literature, we were able to reclassify this CNV region as pathogenic. Overall, we conclude that 21 of the 28 recurrent CNVs examined should be considered pathogenic and provide a clinical diagnosis for any individual harboring a CNV of these regions.

The statistical approach we used to classify recurrent CNVs and the results we obtained are useful tools for researchers and the clinical community in interpreting whether a CNV has pathologic effects. However, although such statistical analysis

is possible for recurrent CNVs, where the frequency is high, this strategy is more difficult for the remaining approximately 75% of CNVs, which are not mediated by segmental duplications and are individually very rare. Therefore, other approaches need to be explored to address this class of CNVs. One possibility for these highly heterogenous CNVs is to analyze all genomic intervals of a defined size (e.g., 500 kb or 1 Mb) or to use a "sliding-window" analysis to examine overlapping genomic intervals along the length of each chromosome. By comparing structural variation observed in cases to controls, disease-causing regions can be differentiated from those associated with normal variation by using the control data to define regions of the genome where dosage changes can be tolerated without overt phenotypic effects. As nonrecurrent CNVs are very rare events, the collection of data from hundreds of thousands of cases will be needed for this type of analysis to be successful. Continued efforts of the ISCA consortium, as well as other databases such as DECIPHER (https://decipher-.sanger.ac.uk/), will be essential to this process to obtain enough overlapping CNVs to provide the power needed for statistical analyses.

The ISCA consortium is continuing to grow and now includes more than 150 clinical laboratories from across the world. Given the rapid increase in utilization of this testing on a routine clinical basis, and the ability to recruit an expanding number of collaborating labs contributing data to a central database, the size of this cohort will continue to rapidly grow, providing a highly cost-effective way to obtain very large CNV datasets. In addition, as this data will be publicly available through two NCBI resources, database of Genotypes and Phenotypes and dbVar, this resource can be readily accessed by researchers and the clinical community. Having large datasets from individuals with abnormal phenotypes will foster more objective formal scientific analyses to predict which CNVs will impact human development. Such efforts will make it possible to develop a whole-genome dosage map in humans to determine which genes and regions are subject to haploinsufficiency or triplosensitivity compared with those that are tolerant of dosage changes.

ACKNOWLEDGMENTS

This work was supported, in part, by NIH Grants HD064525 (D.H.L. and C.L.M.), MH074090 (D.H.L. and C.L.M.), MH080129 (S.T.W.), and MH083722 (S.T.W.), and the Intramural Research Program of the NIH, National Library of Medicine.

Disclosure: SA is a medical geneticist, GR is the medical director, JGC is the scientific director, and AEF is a genetic counselor at GeneDx, a subsidiary of Bioreference Laboratories. DHL is a consultant and member of the Scientific Advisory Board for Roche Nimblegen and GeneDX/BioReference Laboratories. The other authors declare no conflicts of interest.

The authors thank all members of the clinical laboratories for performing the microarray experiments, including Daniel Saul, Stephanie Warren, and Nancy Flores for technical assistance, and Angela DeLorenzo, Ken Chatterten, and Kristi De-Haai for data entry. They thank John C. Barber for helpful discussions, Eli Williams for critical reading of the manuscript, and Cheryl T. Strauss for editorial assistance.

REFERENCES

1. Iafrate AJ, Feuk L, Rivera MN, et al. Detection of large-scale variation in the human genome. *Nat Genet* 2004;36:949–951.
2. IHGSC. Finishing the euchromatic sequence of the human genome. *Nature* 2004;431:931–945.
3. Church DM, Lappalainen I, Sneddon TP, et al. Public data archives for genomic structural variation. *Nat Genet* 2010;42:813–814.
4. Boyle CA, Boulet S, Schieve LA, et al. Trends in the prevalence of developmental disabilities in US children, 1997–2008. *Pediatrics* 2011;127:1034–1042.
5. Miller DT, Adam MP, Aradhya S, et al. Consensus statement: chromosomal microarray is a first-tier clinical diagnostic test for individuals with developmental disabilities or congenital anomalies. *Am J Hum Genet* 2010;86:749–764.
6. Manning M, Hudgins L. Array-based technology and recommendations for utilization in medical genetics practice for detection of chromosomal abnormalities. *Genet Med* 2010;12:742–745.
7. Baldwin EL, Lee JY, Blake DM, et al. Enhanced detection of clinically relevant genomic imbalances using a targeted plus whole genome oligonucleotide microarray *Genet Med* 2008;10:415–429.
8. Lee C, Iafrate AJ, Brothman AR. Copy number variations and clinical cytogenetic diagnosis of constitutional disorders. *Nat Genet* 2007;39:S48–S54.
9. Sharp AJ, Hansen S, Selzer RR, et al. Discovery of previously unidentified genomic disorders from the duplication architecture of the human genome. *Nat Genet* 2006;38:1038–1042.
10. Klopocki E, Schulze H, Strauss G, et al. Complex inheritance pattern resembling autosomal recessive inheritance involving a microdeletion in thrombocytopenia-absent radius syndrome. *Am J Hum Genet* 2007;80:232–240.
11. Brunet A, Armengol L, Heine D, et al. BAC array CGH in patients with Velocardiofacial syndrome-like features reveals genomic aberrations on chromosome region 1q21.1. *BMC Med Genet* 2009;10:144.
12. Mefford HC, Sharp AJ, Baker C, et al. Recurrent rearrangements of chromosome 1q21.1 and variable pediatric phenotypes. *N Engl J Med* 2008;359:1685–1699.
13. Brunetti-Pierri N, Berg JS, Scaglia F, et al. Recurrent reciprocal 1q21.1 deletions and duplications associated with microcephaly or macrocephaly and developmental and behavioral abnormalities. *Nat Genet* 2008;40:1466–1471.
14. Willatt L, Cox J, Barber J, et al. 3q29 microdeletion syndrome: clinical and molecular characterization of a new syndrome. *Am J Hum Genet* 2005;77:154–160.
15. Ballif BC, Theisen A, Coppinger J, et al. Expanding the clinical phenotype of the 3q29 microdeletion syndrome and characterization of the reciprocal microduplication. *Mol Cytogenet* 2008;1:8.
16. Kurotaki N, Harada N, Shimokawa O, et al. Fifty microdeletions among 112 cases of Sotos syndrome: low copy repeats possibly mediate the common deletion. *Hum Mutat* 2003;22:378–387.
17. Franco LM, de Ravel T, Graham BH, et al. A syndrome of short stature, microcephaly and speech delay is associated with duplications reciprocal to the common Sotos syndrome deletion. *Eur J Hum Genet* 2010;18:258–261.
18. Bayes M, Magano LF, Rivera N, Flores R, Perez Jurado LA. Mutational mechanisms of Williams-Beuren syndrome deletions. *Am J Hum Genet* 2003;73:131–151.
19. Somerville MJ, Mervis CB, Young EJ, et al. Severe expressive-language delay related to duplication of the Williams-Beuren locus. *N Engl J Med* 2005;353:1694–1701.
20. Devriendt K, Matthijs G, Van Dael R, et al. Delineation of the critical deletion region for congenital heart defects, on chromosome 8p23.1. *Am J Hum Genet* 1999;64:1119–1126.
21. Barber JC, Bunyan D, Curtis M, et al. 8p23.1 duplication syndrome differentiated from copy number variation of the defensin cluster at prenatal diagnosis in four new families. *Mol Cytogenet* 2010;3:3.
22. Kuwano A, Mutirangura A, Dittrich B, et al. Molecular dissection of the Prader-Willi/Angelman syndrome region (15q11–13) by YAC cloning and FISH analysis. *Hum Mol Genet* 1992;1:417–425.
23. Cook EH Jr, Lindgren V, Leventhal BL, et al. Autism or atypical autism in maternally but not paternally derived proximal 15q duplication. *Am J Hum Genet* 1997;60:928–934.
24. Bolton PF, Dennis NR, Browne CE, et al. The phenotypic manifestations of interstitial duplications of proximal 15q with special reference to the autistic spectrum disorders. *Am J Med Genet* 2001;105:675–685.
25. Sharp AJ, Mefford HC, Li K, et al. A recurrent 15q13.3 microdeletion syndrome associated with mental retardation and seizures. *Nat Genet* 2008;40:322–328.
26. Miller DT, Shen Y, Weiss LA, et al. Microdeletion/duplication at 15q13.2q13.3 among individuals with features of autism and other neuropsychiatric disorders. *J Med Genet* 2009;46:242–248.
27. Ullmann R, Turner G, Kirchhoff M, et al. Array CGH identifies reciprocal 16p13.1 duplications and deletions that predispose to autism and/or mental retardation. *Hum Mutat* 2007;28:674–682.
28. Hannes FD, Sharp AJ, Mefford HC, et al. Recurrent reciprocal deletions and duplications of 16p13.11: the deletion is a risk factor for MR/MCA while the duplication may be a rare benign variant. *J Med Genet* 2009;46:223–232.

29. Kumar RA, KaraMohamed S, Sudi J, et al. Recurrent 16p11.2 microdeletions in autism. *Hum Mol Genet* 2008;17:628–638.
30. Weiss LA, Shen Y, Korn JM, et al. Association between microdeletion and microduplication at 16p11.2 and autism. *N Engl J Med* 2008;358:667–675.
31. Marshall CR, Noor A, Vincent JB, et al. Structural variation of chromosomes in autism spectrum disorder. *Am J Hum Genet* 2008;82:477–488.
32. Smith AC, McGavran L, Robinson J, et al. Interstitial deletion of (17)(p11.2p11.2) in nine patients. *Am J Med Genet* 1986;24:393–414.
33. Potocki L, Chen KS, Park SS, et al. Molecular mechanism for duplication 17p11.2- the homologous recombination reciprocal of the Smith-Magenis microdeletion. *Nat Genet* 2000;24:84–87.
34. Mefford HC, Clauin S, Sharp AJ, et al. Recurrent reciprocal genomic rearrangements of 17q12 are associated with renal disease, diabetes, and epilepsy. *Am J Hum Genet* 2007;81:1057–1069.
35. Nagamani SC, Erez A, Shen J, et al. Clinical spectrum associated with recurrent genomic rearrangements in chromosome 17q12. *Eur J Hum Genet* 2010;18:278–284.
36. Moreno-De-Luca D, Mulle JG, Kaminsky EB, 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.
37. Koolen DA, Vissers LE, Pfundt R, et al. A new chromosome 17q21.31 microdeletion syndrome associated with a common inversion polymorphism. *Nat Genet* 2006;38:999–1001.
38. Shaw-Smith C, Pittman AM, Willatt L, et al. Microdeletion encompassing MAPT at chromosome 17q21.3 is associated with developmental delay and learning disability. *Nat Genet* 2006;38:1032–1037.
39. Grisart B, Willatt L, Destree A, et al. 17q21.31 microduplication patients are characterised by behavioural problems and poor social interaction. *J Med Genet* 2009;46:524–530.
40. Desmaze C, Scambler P, Prieur M, et al. Routine diagnosis of DiGeorge syndrome by fluorescent in situ hybridization. *Hum Genet* 1993;90:663–665.
41. Ou Z, Berg JS, Yonath H, et al. Microduplications of 22q11.2 are frequently inherited and are associated with variable phenotypes. *Genet Med* 2008;10:267–277.
42. Pujana MA, Nadal M, Guitart M, Armengol L, Gratacos M, Estivill X. Human chromosome 15q11–q14 regions of rearrangements contain clusters of LCR15 duplicons. *Eur J Hum Genet* 2002;10:26–35.
43. Emanuel BS. Molecular mechanisms and diagnosis of chromosome 22q11.2 rearrangements. *Dev Disabil Res Rev* 2008;14:11–18.
44. Consortium IS. Rare chromosomal deletions and duplications increase risk of schizophrenia. *Nature* 2008;455:237–241.
45. Itsara A, Cooper GM, Baker C, et al. Population analysis of large copy number variants and hotspots of human genetic disease. *Am J Hum Genet* 2009;84:148–161.
46. Shaikh TH, Gai X, Perin JC, et al. High-resolution mapping and analysis of copy number variations in the human genome: A data resource for clinical and research applications. *Genome Res* 2009;19:1682–1690.
47. Shi J, Levinson DF, Duan J, et al. Common variants on chromosome 6p22.1 are associated with schizophrenia. *Nature* 2009;460:753–757.
48. Mefford HC, Eichler EE. Duplication hotspots, rare genomic disorders, and common disease. *Curr Opin Genet Dev* 2009;19:196–204.
49. Turner DJ, Miretti M, Rajan D, et al. Germline rates of de novo meiotic deletions and duplications causing several genomic disorders. *Nat Genet* 2008;40:90–95.
50. Rajcan-Separovic E, Harvard C, Liu X, et al. Clinical and molecular cytogenetic characterisation of a newly recognised microdeletion syndrome involving 2p15–16.1. *J Med Genet* 2007;44:269–276.
51. Willemsen MH, Fernandez BA, Bacino CA, et al. Identification of ANKRD11 and ZNF778 as candidate genes for autism and variable cognitive impairment in the novel 16q24.3 microdeletion syndrome. *Eur J Hum Genet* 2010;18:429–435.
52. Bruno DL, Anderlid BM, Lindstrand A, et al. Further molecular and clinical delineation of co-locating 17p13.3 microdeletions and microduplications that show distinctive phenotypes. *J Med Genet* 2010;47:299–311.
53. Tsuchiya KD, Shaffer LG, Aradhya S, et al. Variability in interpreting and reporting copy number changes detected by array-based technology in clinical laboratories. *Genet Med* 2009;11:866–873.

