nature genetics

Refining analyses of copy number variation identifies specific genes associated with developmental delay

Bradley P Coe[1], Kali Witherspoon[1], Jill A Rosenfeld[2], Bregje W M van Bon[3,4], Anneke T Vulto-van Silfhout[3], Paolo Bosco[5], Kathryn L Friend[4], Carl Baker[1], Serafino Buono[5], Lisenka E L M Vissers[3], Janneke H Schuurs-Hoeijmakers[3], Alex Hoischen[3], Rolph Pfundt[3], Nik Krumm[1], Gemma L Carvill[6], Deana Li[7], David Amaral[7], Natasha Brown[8,9], Paul J Lockhart[8,10], Ingrid E Scheffer[11], Antonino Alberti[5], Marie Shaw[4], Rosa Pettinato[5], Raymond Tervo[12], Nicole de Leeuw[3], Margot R F Reijnders[3], Beth S Torchia[2], Hilde Peeters[13,14], Brian J O'Roak[1,18], Marco Fichera[5,18], Jayne Y Hehir-Kwa[3], Jay Shendure[1], Heather C Mefford[6], Eric Haan[4,15], Jozef Gécz[4,16], Bert B A de Vries[3], Corrado Romano[5] & Evan E Eichler[1,17]

Copy number variants (CNVs) are associated with many neurocognitive disorders; however, these events are typically large, and the underlying causative genes are unclear. We created an expanded CNV morbidity map from 29,085 children with developmental delay in comparison to 19,584 healthy controls, identifying 70 significant CNVs. We resequenced 26 candidate genes in 4,716 additional cases with developmental delay or autism and 2,193 controls. An integrated analysis of CNV and single-nucleotide variant (SNV) data pinpointed 10 genes enriched for putative loss of function. Follow-up of a subset of affected individuals identified new clinical subtypes of pediatric disease and the genes responsible for disease-associated CNVs. These genetic changes include haploinsufficiency of *SETBP1* associated with intellectual disability and loss of expressive language and truncations of *ZMYND11* in individuals with autism, aggression and complex neuropsychiatric features. This combined CNV and SNV approach facilitates the rapid discovery of new syndromes and genes involved in neuropsychiatric disease despite extensive genetic heterogeneity.

CNVs collectively have an appreciable impact on human mental health, but their large size often precludes specifying the underlying genes involved in the disorder. The pathogenicity of many CNVs observed in the clinic is unknown because the typical variant is also extremely rare, requiring large surveys to achieve significance in case-control cohorts[1–4]. Large-scale analyses of clinical microarray data from children with developmental delay, intellectual disability and autism spectrum disorder (ASD) are now possible and have been used to catalog regions of human dosage imbalance. In most cases, multiple candidate genes still underlie the smallest region of overlap. In contrast, exome sequencing studies of parent-child trios provide the necessary specificity to discover *de novo* truncating mutations—that is, nonsense and frameshift indel mutations—with gene-level specificity[5–14]. Because of the extreme locus heterogeneity of such diseases, however, relatively few recurrences have been reported, as surveys of tens of thousands of exomes are still prohibitively expensive. As large-scale deletions and truncating mutations result in the same dosage imbalance for critical genes, we reasoned that systematically integrating both classes of mutation would improve power in discovering genes associated with developmental delay. Here we have constructed one of the largest CNV morbidity maps of individuals with intellectual disability, developmental delay and/or ASD, both as a clinical resource for pathogenic CNVs and also to identify genes potentially sensitive to dosage imbalance. We then integrated these data with published exome sequencing data and used next-generation sequencing methods to rapidly resequence candidate genes in individuals with unexplained developmental delay. Using this approach, we identified pathogenic mutations in new genes with both statistical significance and clinical relevance.

[1]Department of Genome Sciences, University of Washington School of Medicine, Seattle, Washington, USA. [2]Signature Genomics Laboratories, LLC, PerkinElmer, Inc., Spokane, Washington, USA. [3]Department of Human Genetics, Radboud University Medical Center, Nijmegen, the Netherlands. [4]SA Pathology, North Adelaide, South Australia, Australia. [5]IRCCS (Istituto di Ricovero e Cura a Carattere Scientifico) Associazione Oasi Maria Santissima, Troina, Italy. [6]Department of Pediatrics, University of Washington, Seattle, Washington, USA. [7]Representing the Autism Phenome Project, MIND Institute, University of California, Davis, Sacramento, California, USA. [8]Department of Paediatrics, University of Melbourne, Royal Children's Hospital, Melbourne, Victoria, Australia. [9]Barwon Child Health Unit, Barwon Health, Geelong, Victoria, Australia. [10]Murdoch Childrens Research Institute, University of Melbourne, Royal Children's Hospital, Melbourne, Victoria, Australia. [11]Florey Institute, University of Melbourne, Austin Health and Royal Children's Hospital, Melbourne, Victoria, Australia. [12]Division of Developmental and Behavioral Pediatrics, Mayo Clinic, Rochester, Minnesota, USA. [13]Center for Human Genetics, University Hospitals Leuven, KU Leuven, Leuven, Belgium. [14]Leuven Autism Research (LAuRes), Leuven, Belgium. [15]School of Paediatrics and Reproductive Health, University of Adelaide, Adelaide, South Australia, Australia. [16]Robinson Institute, University of Adelaide, Adelaide, South Australia, Australia. [17]Howard Hughes Medical Institute, Seattle, Washington, USA. [18]Present addresses: Molecular and Medical Genetics, Oregon Health and Science University (OHSU), Portland, Oregon, USA (B.J.O.) and Medical Genetics, University of Catania, Catania, Italy (M.F.). Correspondence should be addressed to E.E.E. (eee@gs.washington.edu).

Received 11 April; accepted 20 August; published online 14 September 2014; doi:10.1038/ng.3092

RESULTS

Construction of a CNV morbidity map

We constructed an expanded CNV morbidity map as previously described[1] using array comparative genomic hybridization (CGH) data from 29,085 primarily pediatric cases with intellectual disability, developmental delay and/or ASD in comparison to 19,584 adult population controls (Online Methods). The set included 13,318 previously unpublished cases and 11,255 new controls, providing enhanced power to detect large-scale, potentially pathogenic deletions and duplications (**Supplementary Table 1**). As expected, we observed a striking increase among cases for rare CNVs (frequency of <1%; $P < 1 \times 10^{-16}$, Peto and Peto), driven overwhelmingly by deletions (odds ratio (OR) for deletion of ≥500 kb = 5.09 versus OR for duplication = 1.76). An analysis of 2,086 transmissions showed that likely deleterious CNVs were transmitted preferentially from mothers (58%; $P = 0.008$, binomial test) (**Supplementary Fig. 1**)[15].

We identified 2,184 CNVs (1,348 deletions and 836 duplications) in 55 regions known to associate with autosomal genomic disorders, most of which (40/55) corresponded to genomic hotspots flanked by segmental duplication (**Supplementary Tables 2** and **3**). Among these regions were 19 loci (**Supplementary Table 2**) that had been suspected to be pathogenic and reached nominal significance in our new screen (7 deletion loci, 7 duplication loci and 5 loci significant for both). These loci included the 2q11.2 deletion[16] as well as several reciprocal duplications from known deletion syndromes such as a 15q24 microduplication (B to C region; $P = 0.027$, Fisher's exact test), reciprocal duplication of the 17q11.2 *NF1* deletion (seven cases versus zero controls; $P = 0.027$, Fisher's exact test) and a 16p13.11 microduplication ($P = 0.0112$, Fisher's exact test).

To identify new regions of genomic imbalance and potential candidate genes, we performed three analyses. First, we performed a gene-level (RefSeq) analysis to assess the excess of deletions or duplications in cases in comparison to controls. Overall, we detected 1,945 genes enriched for deletions and 2,633 genes enriched for duplications (3,800 unique genes in total) at a nominal level of significance ($P < 0.01$, one-tailed Fisher's exact test; **Supplementary Table 4**). Because many of these genes were clustered within specific genomic regions, we next computed enrichment in probands using a genomic windowing approach focused on case CNVs of >250 kb in length (**Supplementary Figs. 2** and **3**, and **Supplementary Data Set 1**) and a simulation-based empirical *P* value. The analysis identified 14 significant regions (most were either new or previously discussed in the context of case reports or single-gene studies[17–27]). This table also included some well-established risk loci such as *NRXN1*, *SATB2* and *MEF2C*, which reached genome-wide significance with additional refinement of incidence and deletion boundaries[18,21,22,25,27,28]. Unlike genomic hotspots (**Supplementary Tables 2** and **3**), most of these regions were not flanked by segmental duplications, and a smaller significant region of overlap (SRO) corresponding to a few genes could be identified on the basis of the multiple breakpoints (**Table 1** and **Supplementary Figs. 2** and **4**). In addition, we performed a reciprocal analysis for enrichment in controls and identified one duplication locus at 19q13.33, enriched for KRAB C2H2 zinc-finger transcription factor genes, which showed a moderate protective OR and nominal significance (**Supplementary Note**).

We next estimated the false discovery rate of our CNV calls by designing a customized microarray and independently validating a subset (39/40, or 97.5%) of the events corresponding to the 14 regions (Online Methods). Similarly, we assessed the transmission of 61 CNVs and found that 28 were *de novo* and 33 were inherited (21 maternal and 12 paternal, including 3 parental balanced carriers). In several cases, a single SRO was apparent, such as the 360-kb duplication region on chromosome 12p13.3 corresponding to 19 genes (*SCNN1A* to *PIANP*), where a focal 92.6-kb CNV highlighted 5 genes, including *CHD4*. In a few cases, a single gene was implicated (for example, *NRXN1*, *SATB2* or *MEF2C*) (**Table 1**). We observed significant enrichment at the *GAP43* gene[29] ($P = 0.0003$, simulated), with four deletions arising *de novo*. In other cases, such as the chromosome 1q24q25 microdeletion, we observed several peaks of significance, making it impossible to refine the CNVs to a single candidate gene (for example, *DNM3* versus *FMO1-FMO2*; **Supplementary Fig. 2**).

Integration of CNV and exome sequencing data

As a final analysis to identify high-impact candidate genes, we integrated our CNV deletion data with data for *de novo* truncating mutations identified in 1,879 probands from recently published exome sequencing studies of ASD, intellectual disability, congenital heart defects and schizophrenia[5–14]. Overall, we detected deletion enrichment at 17.4% of genes with at least one truncating mutation (43/247 with CNV deletion; $P < 0.05$, Fisher's exact test), a frequency similar

Table 1 New CNVs and smallest regions of overlap

Region	Chr.	Start (hg18, Mb)	End (hg18, Mb)	Type[a]	State	Cases[b]	Controls[b]	Inheritance[c]		Window *q* value[d]	Simulated *P* value[e]
								De novo	Inherited		
1q24 (*FMO* deletions and *DNM3*)[17]	1	167.00	172.00	MB	Deletion	12	0		2	0.0324	0.011
2q33.1 (*SATB2*)[22,25]	2	199.87	200.22	MB	Deletion	13	0	1		0.0211	0.0002
2p16.1 (*NRXN1*)[18,28]	2	50.00	51.11	MB	Deletion	30	9	4	8	Focal	0.00005
2p15-16.1 proximal (*PEX13* to *AHSA2*)	2	59.50	63.00	MB	Duplication	9	0	1		0.285	0.00001
3p25.3 (*JAGN1* to *TATDN2*)	3	9.50	11.00	MB	Duplication	10	0	1	3	0.036	0.00103
3p11.2 (*CHMP2B* to *POU1F1*)	3	87.32	87.64	MB	Deletion	9	0		3	0.0489	0.000075
3q13 (*GAP43*)[19,29]	3	116.72	117.13	MB	Deletion	9	0	4		0.0489	0.0003
3q28-29 (*FGF12*)	3	193.00	194.50	MB	Deletion	13	1		3	Focal	0.00005
4q21 (*BMP3*)	4	81.00	83.50	MB	Deletion	11	0	2		0.0324	0.00025
5q14 (*MEF2C*)[21,27]	5	88.00	88.26	MB	Deletion	10	0	2		Focal	0.00005
9p13	9	32.00	39.00	MB	Duplication	18	0	2	2[f]	0.00216	–
10q11[23]	10	49.06	52.06	HS, MB	Duplication	10	0		5	0.036	–
10q23.1 (*SFTPD* to *GLUD1*, *NRG3* inclusive)[24,55]	10	81.68	88.93	HS, MB	Deletion	11	0	5		0.0211	–
12p13 (*SCNN1A* to *PIANP*)[20]	12	6.34	6.68	MB	Duplication	23	1	3	1[f]	0.00115	–

[a]Hotspot (HS) or multiple-breakpoint (MB) locus. [b]Owing to complex CNV structure, the case-control counts are representative of the region but might vary throughout. [c]*De novo* counts also include cases from Hehir-Kwa *et al.*[56]. [d]The window *q* value is the weighted median for unique segments in the critical region. [e]Reported as the median simulation *P* value for all genes in the region (**Supplementary Table 4**). [f]Carrier of a balanced translocation.

Table 2 Intersection of CNV and exome data

		Exome data			Array CGH			
Gene	Isoform	1,879 published cases LoF	1,879 published cases *de novo* LoF (ESP average read depth >20, Dustmasked)	6,500 ESP LoF (ESP average read depth >20, Dustmasked)	Signature deletions (n = 29,085)	Control deletions (n = 19,584)	Combined LoF P value	Combined LoF q value[e]
ANK2[a]	NM_020977.3[b]	1	1	0	5	0	0.0171	0.169
ARHGAP5	NM_001030055.1	1	1	0	7	0	0.0061	0.0833
BCL11A	NM_022893.3	1	0	0	4	0	0.0286	0.244
CAPRIN1	NM_005898.4	1	1	0	4	0	0.0286	0.244
CARKD	NM_001242881.1[c]	1	1	0	12	4	0.0363	0.28
CHD2[a]	NM_001271.3	3	3	0	0	0	0.0113	0.127
CHD8[a]	NM_001170629.1	3	3	0	2	0	0.00402	0.0703
CSDE1	NM_001130523.2	1	1	0	3	0	0.0479	0.311
CUL3[a]	NM_003590.4	2	2	0	5	0	0.00383	0.0703
DLL1	NM_005618.3	1	0	0	32	1	2.17×10^{-7}	2.68×10^{-5}
DYRK1A[a]	NM_001396.3	2	2	0	11	0	1.74×10^{-4}	8.60×10^{-3}
FAM8A1	NM_016255.2	1	1	0	5	0	0.0171	0.169
FOXP1[a]	NM_001244810.1	1	1	0	4	0	0.0286	0.244
GRIN2B[a]	NM_000834.3	3	3	0	2	0	0.00402	0.0703
GTPBP4	NM_012341.2	1	1	0	3	0	0.0479	0.311
LTN1	NM_015565.2	1	1	0	6	0	0.0102	0.12
MBD5[a]	NM_018328.4	1	1	0	16	6	0.0343	0.273
MYT1L	NM_015025.2	1	1	0	8	0	0.00365	0.0703
NAA15	NM_057175.3	2	2	0	5	3	0.0296	0.244
NCKAP1	NM_205842.1	2	2	0	7	0	0.00137	0.0564
NFIA	NM_001134673.3	1	1	0	3	0	0.0479	0.311
NRXN1[a]	NM_001135659.1	1	1	0	30	9	0.00427	0.0703
NTM	NM_001144058.1	1	1	0	40	0	2.53×10^{-10}	6.25×10^{-8}
PCOLCE	NM_002593.3	1	1	0	7	0	0.0061	0.0833
PHF2	NM_005392.3	1	1	0	4	0	0.0286	0.244
RAB2A	NM_002865.2	1	1	0	3	0	0.0479	0.311
SCN1A[a]	NM_001165963.1	4	4	0	10	1	7.36×10^{-5}	4.55×10^{-3}
SCN2A[a]	NM_021007.2	6	5	0	10	0	7.34×10^{-7}	6.04×10^{-5}
SLC6A1	NM_003042.3	1	1	0	6	0	0.0102	0.12
SRM	NM_003132.2	1	1	0	9	0	0.00218	0.0703
STXBP1[a]	NM_003165.3	2	2	0	4	0	0.00641	0.0833
SUV420H1	NM_016028.4[d]	1	1	0	3	0	0.0479	0.31135
SYNGAP1[a]	NM_006772.2	4	4	0	0	1	0.00252	0.0703
TBR1	NM_006593.2	2	2	0	7	1	0.00522	0.0806
UBN2	NM_173569.3	1	1	0	5	0	0.0171	0.169
WAC	NM_016628.4	1	1	0	3	0	0.0479	0.31135
WDFY3	NM_014991.4	1	1	0	8	0	0.00365	0.0703
ZMYND11	NM_006624.5	1	1	0	8	0	0.00365	0.0703

LOF, loss of function.
[a]Disease gene in OMIM. [b]Variant 2; this is the major form of ankyrin in the adult brain. [c]Variant 2; this isoform and variants 3 and 4 are shorter than variant 1. [d]Variant 2; this isoform is shorter and has a distinct C terminus in comparison to isoform 1. [e]Please see the **Supplementary Note** for discussion of the q values shown.

to that expected with intersections by random chance (OR = 1.15, 95% confidence interval (CI) = 0.8–1.6; P = 1, Fisher's exact test). However, when we limited our analysis to the 21 genes with 2 or more truncating mutations in probands, we observed significant deletion enrichment for 33.3% of the genes (7 of 21 genes; OR = 2.72; P = 0.034, Fisher's exact test), supporting the notion that integrating CNV data and exome sequencing data increases power to detect disease-related genes. Using a statistical framework based on a hypergeometric distribution, we computed a joint probability of putative loss of function (Online Methods), combining the CNV data with the SNV data for the 6,500 individuals from the Exome Sequencing Project (ESP) (ESP6500 controls) and published *de novo* loss-of-function mutations in probands. This analysis highlighted 38 of the 247 genes with nominally significant increases in loss-of-function events in cases in comparison to controls (19 with q value ≤ 0.01), including 13 genes previously identified as disease causing (Online Mendelian Inheritance in Man, OMIM) (**Table 2**).

Targeted resequencing of candidate genes in ASD and intellectual disability

On the basis of the analyses above, we selected a set of 26 candidate genes with significant CNV enrichment, rare focal CNVs with *de novo* mutations from exome sequencing studies and top candidates from targeted resequencing in ASD and/or intellectual disability (**Table 3**). For three of these regions, we selected at least two adjacent genes mapping within the SRO; we also selected six genes (*GRIN2B*, *ARID1B*, *MBD5*, *PTEN*, *SCN1A* and *KANSL1*) known to be associated with ASD and/or intellectual disability as positive controls[30–35]. We used molecular inversion probe (MIP)-based capture[36] to sequence the 26 genes in 3,387 cases of intellectual disability and/or developmental

Table 3 Combined CNV and targeted sequencing

		CNV deletions		CNV duplications		Severe variants (nonsense, start loss, frameshift, splice site) ESP read depth >20, Dustmasked				Joint *P* value[b]		Joint *q* value[c]		
Gene	RefSeq	Cases	Controls	Cases	Controls	ID + DD (n = 3,387)	ASD (n = 1,329)	ESP6500	Simons siblings (n = 2,193)	ID + DD	ID + DD + ASD	ID + DD	ID + DD + ASD	Truncation *P* value
ACACA	NM_198839.1	28	4	34	10	1	1	7	1	0.611	0.517	0.691	0.64	0.73
ADNP	NM_015339.2	1	0	4	0	5	0	1	1	**0.0138**	**0.0376**	**0.0326**	**0.0698**	0.044
ARID1B[a]	NM_017519.2	5	1	3	2	9	0	1	0	**0.0000205**	**0.000151**	**0.000183**	**0.000654**	0.00028
CHD1L	NM_004284.3	78	7	71	8	12	1	40	0	0.419	0.628	0.545	0.71	0.94
CYFIP1	NM_014608.2	230	69	175	98	0	1	1	0	**2.75×10^{-8}**	**3.10×10^{-9}**	**7.15×10^{-7}**	**8.06×10^{-8}**	–
DIP2A	NM_015151.3	13	3	74	26	1	1	43	1	1	1	1	1	–
DNM3	NM_015569.3	11	3	2	0	2	1	0	0	**0.0095**	**0.00524**	**0.0247**	**0.0142**	–
DYRK1A	NM_001396.3	11	0	66	2	2	1	0	0	**0.00027**	**0.00015**	**0.00117**	**0.000654**	–
FOXP1	NM_032682.5	4	0	6	4	1	0	0	0	**0.0358**	**0.0449**	**0.0665**	**0.0778**	–
GRIN2B[a]	NM_000834.3	2	0	17	1	2	2	0	0	**0.0281**	**0.00546**	**0.0562**	**0.0142**	–
KANSL1[a]	NM_001193466.1	32	3	4	8	4	2	2	1	**0.00251**	**0.000418**	**0.00816**	**0.00155**	–
MAPT	NM_016835.4	32	1	4	3	1	0	6	0	0.33	0.35	0.452	0.455	–
MBD5[a]	NM_018328.4	16	6	8	5	1	0	0	0	**0.0429**	0.054	**0.0744**	**0.0878**	0.095
NRG3	NM_001165973.1	18	7	9	23	2	1	1	0	**0.0468**	**0.0307**	**0.0761**	**0.0614**	–
NRXN1	NM_004801.4	30	9	6	0	0	1	0	0	**0.019**	**0.00669**	**0.0412**	**0.0158**	–
PTEN[a]	NM_000314.4	1	1	0	5	1	0	0	0	0.235	0.295	0.339	0.404	–
SCN1A[a]	NM_006920.4	10	1	5	0	2	0	0	0	**0.00229**	**0.00361**	**0.00816**	**0.0117**	–
SCN2A	NM_021007.2	10	0	6	0	3	1	0	0	**0.000128**	**0.0000888**	**0.000666**	**0.000577**	–
TTC21B	NM_024753.3	10	0	9	0	1	0	51	0	1	1	1	1	–
SETBP1	NM_015559.2	2	0	28	1	5	0	2	0	**0.0093**	**0.0262**	**0.0247**	**0.0568**	0.011 (ID only)
SLC1A1	NM_004170.5	33	3	26	1	0	0	0	0	**0.0000221**	**0.0000221**	**0.000183**	**0.000287**	–
SOX5	NM_006940.4	15	4	17	3	0	1	2	0	0.512	0.292	0.605	0.404	–
TBL1XR1	NM_024665.4	3	0	4	5	0	0	0	0	0.2134	0.2134	0.326	0.326	–
TSPAN17	NM_012171.2	12	2	7	0	0	0	3	0	0.64	0.711	0.693	0.77	–
DIP2C	NM_014974.2	10	0	36	6	0	0	2	0	0.48	0.557	0.594	0.658	–
ZMYND11	NM_006624.5	8	0	25	15	5	0	0	0	**0.0000281**	**0.0000874**	**0.000183**	**0.000577**	–

ID, intellectual disability; DD, developmental delay.
[a]Gene known to be associated with ASD and intellectual disability[30–35]. [b]Bolded genes represent genes passing nominal significance. [c]Bolded entries represent $q < 0.1$. For details, please see the **Supplementary Note**.

Table 4 Brief phenotypic description of cases with *SETBP1* loss-of-function variants

Case	Age at examination	Sex	Alteration	Inheritance	Cognitive	Hyperactive or ADHD	Social difficulties	Other behavioral difficulties	Speech delay	Motor delay	Facial dysmorphism	Seizures or EEG abnormalities
DNA03-00335	14 years	M	p.Ile822Tyrfs*13	*De novo*	Normal IQ			+	+	+	+	
DNA-008897	73 years	M	p.Leu411Glyfs*6		Profound ID		+	+	+	+	+	
Troina 1274	19 years	M	p.Trp532*	*De novo*	Severe ID			+	+	+	+	–
Troina 1512	17 years	M	p.Ser1011*	*De novo*	Mild ID	+ (3y 8m)	+		+	+	+	–
Troina 3097	34 years	F	p.Arg143Valfs*64		Severe ID				+	+	+	+
DNA11-21308Z	36 years	F	p.Arg625*		Mild to moderate ID	+	+	+	+	+	+	
DNA11-19324Z	9 years	F	p.Arg626*		2- to 2.5-year delay at 9 years old				+	–	+	–
DNA08-08272	9 years	M	p.Gly15Argfs*47		Mild ID	+		+	+	+	+	+
Rauch *et al.*	13 years	F	p.Lys592*		Mild ID	+	+		+	–	+	
9886269	5 years	M	Deletion	*De novo*	Global delay	+			+	+	+	+
Marseglia *et al.*	15 years	M	Deletion	*De novo*	Mild ID	+	+	+	+	+	+	+
Filges *et al.* pt. 1	7 years	M	Deletion	*De novo*	Moderate ID				+	+	+	+
Filges *et al.* pt. 2	4 years	M	Deletion	*De novo*					+	+	+	

ID, intellectual disability; EEG, electroencephalogram; M, male; F, female.

delay and 1,329 cases of ASD, totaling 4,716 cases. Putative loss-of-function SNVs and indels were validated by Sanger sequencing and assessed in parental DNA, when available, to determine inheritance. Genes with significant enrichment were identified by comparison with the MIP resequencing data for 2,193 unaffected siblings from the Simons Simplex Collection[37] and ESP6500. We tested each gene for combined enrichment of loss-of-function variation across CNV and SNV data (Online Methods) and identified 16 genes (**Table 3** and **Supplementary Table 5**) with a significant enrichment of disruptive mutations in cases. Additionally, to control for the differential effects of terminal truncating events, we applied a statistical model based on predicted protein lengths for genes with truncating or splice-site events in ESP6500 (**Supplementary Fig. 5**), as this approach was complementary to the results from case-control comparison (**Table 3**).

Among the positive controls, our analysis confirmed the pathogenicity of five genes using a nominal threshold of significance on the joint *P* values: *ARID1B* (five CNVs and nine truncating SNVs, one of which was confirmed to be *de novo*; $P = 1.51 \times 10^{-4}$, $q = 6.54 \times 10^{-4}$), *GRIN2B* (two CNVs—one focal CNV disrupting the distal end of *GRIN2B*—and four new truncating variants; $P = 0.00546$, $q = 0.0142$) and *MBD5* ($P = 0.0429$, $q = 0.0744$), as well as *SCN1A* ($P = 0.0036$, $q = 0.0117$), in comparison to the adjacent gene *TTC21B* ($P = 1.00$, $q = 1.00$). Integration of SNV and CNV data confirmed *KANSL1* as the gene responsible for the 17q21.31 deletion syndrome[32] ($P = 0.000418$, $q = 0.00155$) in comparison to the adjacent gene *MAPT* ($P = 0.36$, $q = 0.455$) (**Table 3** and **Supplementary Table 5**). Clinical follow-up for one *KANSL1* case with a severe frameshift demonstrated striking phenotypic resemblance to microdeletion carriers, confirming this gene as the major contributor to the phenotype of 17q21.31 microdeletion (Koolen–de Vries) syndrome[32,38].

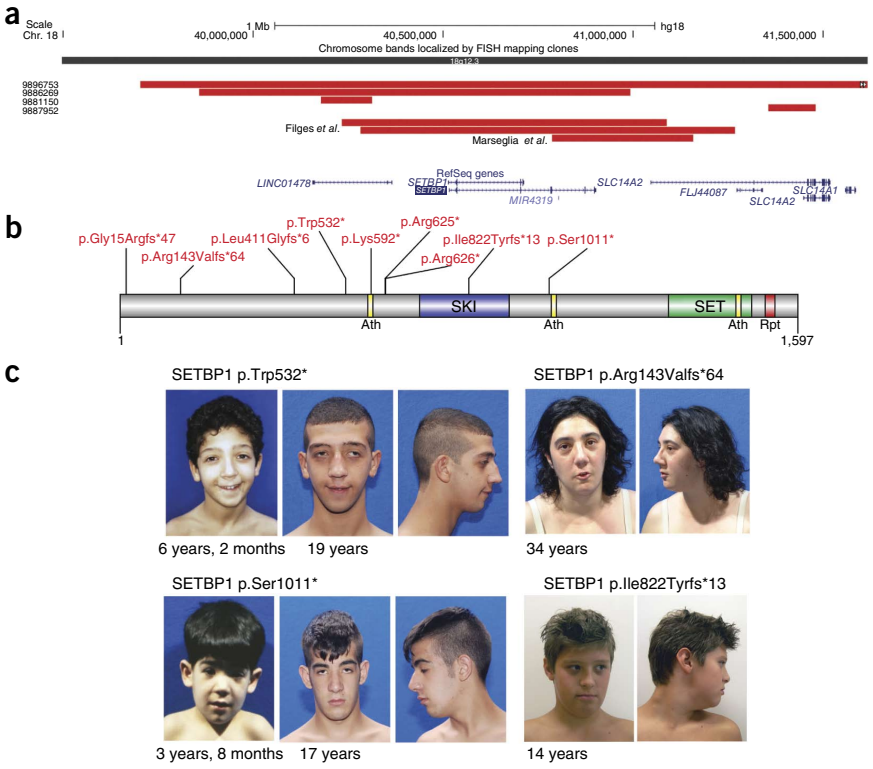

Figure 1 Truncating *SETBP1* mutations and associated phenotypes. (**a**) CNV data define a focal CNV region around *SETBP1*. Combining a focal *de novo* deletion observed in our study (9886269) with CNVs from Filges *et al.*[41] and Marseglia *et al.*[42] (red bars) highlights minimal common regions, including *SETBP1* and *LINC01478*. (**b**) Targeted resequencing identified eight truncating variants in SETBP1 and none in controls. Integration of published exome data identified one additional case and no truncating events in controls. Ath, AT hook; SKI, SKI-homologous region; SET, SET-binding domain; Rpt, repeat. (**c**) Phenotypic assessment (summarized in **Table 4**) identified a recognizable phenotype, including IQ deficits ranging from mild to severe, impaired speech and distinctive facial features. See the **Supplementary Note** for additional photographs of affected individuals and clinical descriptions. We obtained informed consent to publish the photographs.

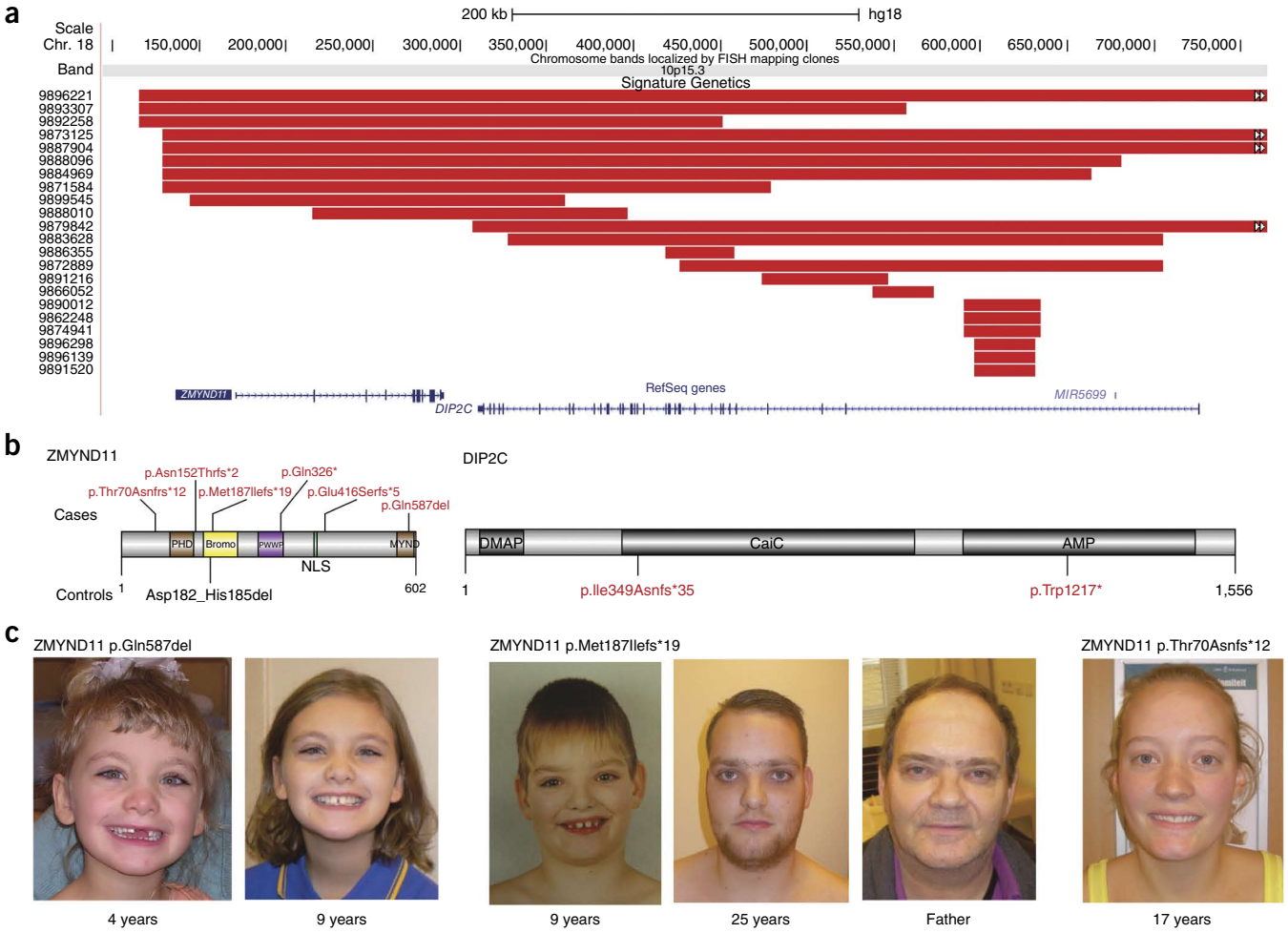

Figure 2 Truncating *ZMYND11* mutations and associated phenotypes. (**a**) CNV data refine a focal CNV deletion region (red bars) containing two genes (*ZMYND11* and *DIP2C*). (**b**) Targeted resequencing identified five truncating variants and one single-amino-acid deletion predicted to behave as loss-of-function variants by removing a critical binding residue in the MYND domain (Gln587). Analysis of control resequencing and exome data identified no additional truncating events in *ZMYND11* but highlighted two truncating mutations in *DIP2C*. PHD, plant homeodomain; Bromo, bromodomain; PWWP, conserved ProTrpTrpPro motif; NLS, nuclear localization sequence; MYND, zinc finger MYND type (myeloid, Nervy and DEAF-1); DMAP, DNA methyltransferase–associated protein; CaiC, crotonobetaine/carnitine-CoA ligase; AMP, AMP-dependent synthetase/ligase. (**c**) Phenotypic assessment (summarized in **Table 5**) showed a consistent phenotype characterized by mild intellectual disability accompanied by speech and motor delays, as well as complex neuropsychiatric behavioral and characteristic facial features. See the **Supplementary Note** for additional photographs of the affected individuals and clinical descriptions. We obtained informed consent to publish the photographs.

An enrichment of loss-of-function mutations in cases was observed for ten additional genes (*ADNP*, *DYRK1A*, *NRXN1*, *NRG3*, *SETBP1*, *ZMYND11*, *DNM3*, *CYFIP1*, *FOXP1* and *SCN2A*) (**Table 4**). In one case with a *de novo DYRK1A* splice-site mutation (see Troina1818 in **Supplementary Table 5**), the affected individual presented with severe microcephaly, consistent with published autism-related *de novo* truncating mutations and CNVs from earlier studies[36,39]. Among the genes for which there was no enrichment in cases versus controls, two were notable: *CHD1L* and *ACACA*—candidates for the 1q21 deletion and 17q12 deletion syndromes, respectively[40]. In our resequencing study of *CHD1L*, for example, we identified 14 likely truncating variants (**Table 3**) in comparison to 9 independent truncating variants in controls, which indicates that rare truncating mutations of *CHD1L* are not uncommon (**Table 3** and **Supplementary Table 5**). There was also no significant decrease in the predicted protein size in cases in comparison to controls ($P = 0.94$, log-rank test).

Phenotypic examination of cases with *SETBP1* and *ZMYND11* truncations

Among the significant genes, we focused on *SETBP1* and *ZMYND11* for further phenotypic characterization. We confirmed a focal *de novo* deletion and five cases with truncating mutations (three tested and confirmed to be *de novo*) in the *SETBP1* gene (encoding SET-binding protein 1). Disruptive mutations were absent in controls, with the exception of a splice-site alteration predicted to lead to the loss of an in-frame exon encoding 18 amino acids. Notably, all truncating mutations in cases occurred in cohorts of intellectual disability, where we observed an enrichment of mutations ($P = 0.0093$, joint loss of function) and decreased predicted protein size ($P = 0.011$, log-rank test) (**Fig. 1**, **Table 3** and **Supplementary Tables 5** and **6**). Integration of our variants from cases with 2 additional truncating variants found in a separate genetic screen for intellectual disability ($n = 847$) with the same MIPs, as well as published small deletions and *de novo* variants, highlighted a similar phenotype for the affected

Table 5 Brief phenotypic description of cases with *ZMYND11* loss-of-function variants

Case	Age at examination	Sex	Alteration	Inheritance	Cognitive	Speech delay	Social difficulties	Behavioral problems	Facial dysmorphism
Adelaide20124	4 and 9 years	F	p.Gln587del	*De novo*	Global DD	+	+		+
Adelaide3553	22 years	M	p.Asn152Thrfs*26		Global DD	+		+	
DNA-017151	17 years	F	p.Thr70Asnfs*12	*De novo*	Normal IQ	+	+	+	+
DNA04-02424	41 years	M	p.Gln326*		Mild ID	+	+	+	+
DNA05-04370		M	p.Glu416Serfs*5		Severe ID	+	+		+
DNA-013587	25 years	M	p.Met187Ilefs*19	Inherited	Global DD	+	+	+	+
Father of DNA-013587		M	p.Met187Ilefs*19	Carrier	DD			+	

DD, developmental delay; ID, intellectual disability; M, male; F, female.

individuals[12,41,42]. The majority of cases demonstrated IQ and language deficits (completely absent or substantially impaired speech in 92% (12/13) of the cases). Cases positive for mutation in *SETBP1* also frequently exhibited impairment of fine motor skills ($n = 8$), hyperactivity and/or ADHD (attention deficit and hyperactivity disorder) ($n = 7$) and autistic features and/or poor social skills ($n = 4$). We also observed a dysmorphism typified by a long face ($n = 10$), characteristic eyebrows and, less frequently, low-set ears ($n = 4$) and café-au-lait spots ($n = 4$) (**Fig. 1**, **Table 4** and **Supplementary Table 6**).

The smallest region of overlap for the 10p15.3 microdeletion predicted two possible candidate genes[43]: *ZMYND11* and *DIP2C* (**Fig. 2**). We resequenced both candidates and detected five truncating variants in *ZMYND11* (two confirmed to be *de novo* and one inherited from an affected father) and none in *DIP2C*. In contrast, concurrent examination of controls identified truncating mutations only for *DIP2C* (**Fig. 2**, **Table 3** and **Supplementary Table 5**). Integration of CNV and truncating SNV data strongly supports *ZMYND11* (developmental delay $P = 2.81 \times 10^{-5}$, joint loss of function) as opposed to *DIP2C* (developmental delay $P = 0.48$, joint loss of function) as the critical gene. Comparing the phenotypes of affected individuals with truncating SNVs in *ZMYND11* (**Fig. 2**, **Table 5** and **Supplementary Table 7**) showed a striking resemblance to the 10p15.3 microdeletion cases described previously[43] and highlighted a consistent set of behavioral features, mild intellectual disability and subtle facial features, including hypertelorism ($n = 6$), ptosis ($n = 3$) and a wide mouth ($n = 4$). The most consistent features seen in all subjects were speech and motor delays, which were observed in all affected individuals for whom information was available, including in cases with CNVs[43]. Interestingly, a psychiatric phenotype was apparent in three of five affected individuals, including aggression in three of four males. Three cases were accessible for parental DNA testing, by which we determined that two variants were *de novo* and one was paternally inherited. The paternal carrier of the variant encoding p.Met187Ilefs*19 also had developmental delay, including walking at 3–4 years of age and learning problems, in addition to aggression in childhood with mood swings. We also detected a *de novo* in-frame deletion (encoding p.Gln587del) in the MYND domain (Gln587), which represents a critical residue in corepressor binding (including NCoR)[44–46]. Examination of this individual identified similarities with published 10p15.3 microdeletion syndrome cases (**Fig. 2** and **Supplementary Table 7**), including characteristic facial dysmorphisms, global developmental delay and speech delay. Taking this evidence together, we propose that *ZMYND11* is the critical gene associated with the 10p15.3 microdeletion syndrome.

DISCUSSION

In this study, we leverage the large sample size of cases available from CNV clinical microarrays and the precision of next-generation sequencing to identify specific genes associated with neurodevelopmental disease. The expanded CNV morbidity map offers clinical usefulness as a resource to assess the pathogenic relevance of rare events, as well as a research tool to prioritize genes discovered from exome sequencing studies that are currently too underpowered to achieve statistical significance[5–14,36]. It is important to note that the large sample size (nearly 50,000 cases and controls) has begun to highlight regions that map outside of recurrent CNVs mediated by segmental duplications. The sample size is thus sufficient to survey the background level of CNVs, identifying critical regions outside of regions with elevated mutation rates (**Table 2**). In addition, the sample size has allowed the identification of various recurrent duplications (**Supplementary Tables 2** and **3**) that are neither necessary nor sufficient to cause disease but are more likely to act as genetic modifiers or risk factors similar to the 15q11.2 microdeletion[47]. It is possible that copy number polymorphisms occurring at a frequency of >1% might also contribute as weaker risk factors, but such events are typically smaller and have not been sufficiently assayed by microarrays. We identify, for example, the 16p13.11 microduplication among 68 cases in comparison to 27 controls, giving a likelihood ratio of 1.7 (95% CI = 1.13–2.56). Exploring these high-impact risk factors will be important in understanding the genetic architecture of ASD and developmental delay and its relationship to that for other neuropsychiatric features.

Under the assumption that different classes of genetic mutation (microdeletions and truncating SNVs and indels) will expose the same genic haploinsufficiency, we developed a joint probability statistic to identify 38 specific genes (**Table 3**) with a higher prior of disease involvement. Although we have not explored it here, a similar approach might be useful in assessing microduplications and hypermorphic missense mutations. Although it is clear that not all CNVs are monogenic and will be amenable to this integrated strategy, forward resequencing of 23 candidate regions (including 6 controls) identified 11 genes where there is an excess of deletions and truncating mutations in cases in comparison to controls (**Table 3**). Targeted resequencing, in particular, allows the discrimination of adjacent genes within an SRO (that is, *SCN1A* versus *TTC21B*, *KANSL1* versus *MAPT* or *ZMYND11* versus *DIP2C*). A comparison of the frequency of truncating mutations in cases and controls also reduces the likelihood that specific genes highlighted by case reports of atypical CNVs are pathogenic (for example, *ACACA* and *CHD1L*)[40].

Follow-up and phenotypic evaluation in cases provide the most compelling evidence that we have identified genes that likely underlie CNV haploinsufficiency. Studies of cases with microdeletion and translocation originally narrowed a 1-Mb deletion region on chromosome 18q12.3 to a 372-kb critical region spanning three genes (*SETBP1*, *SLC14A2* and *MIR4319*)[41,42]. We identified five truncating mutations (three of three tested and confirmed to be *de novo*) in *SETBP1* among cases with moderate to severe intellectual disability. The phenotypic similarity among microdeletion cases and cases with

truncating SNVs and indels, including intellectual disability, craniofacial dysmorphism and the almost complete absence of expressive language (92% of cases), strongly suggests that loss of function of *SETBP1* underlies this condition. Interestingly, gain-of-function mutations result in a completely different phenotype known as Schinzel-Giedion syndrome. In contrast to the likely loss-of-function mutations, gain-of-function mutations cluster within a 12-amino-acid domain and result in more severe developmental delay with multiple congenital abnormalities and death in infancy[48,49]. In addition, identical somatic mutations in this hotspot region have recently also been reported in a variety of myeloid malignancies[50,51].

Similarly, a study of 19 unrelated developmental delay cases with submicroscopic deletions in chromosome 10p15.3 (as well as data from the CNV morbidity map in this study, which has six shared samples) narrowed the critical region to two genes (*DIP2C* and *ZMYND11*)[43]. Our targeted sequencing identified truncating *ZMYND11* mutations exclusively in cases but none in *DIP2C*. *ZMYND11* (encoding zinc-finger MYND domain 11) is a tumor suppressor gene whose corresponding protein recognizes chromatin trimethylated at lysine 36 of histone H3.3 (H3.3K36me3) and regulates elongation by RNA polymerase II (ref. 52). It is associated with highly expressed genes and might be an important transcriptional corepressor early in development. Additionally, ZMYND11 has been demonstrated to have an inhibitory role in neuronal differentiation[53]. Cases with truncating mutations show borderline IQ and a mild dysmorphism similar to microdeletion cases. Interestingly, both females studied have been described as having autistic tendencies, whereas the three males in this study have been identified as having aggressive behaviors, temper tantrums and rage. The oldest male in this study (45 years of age) has, in fact, had differing psychiatric diagnoses, including borderline personality disorder, bipolar disorder, psychosis, depression, low frustration tolerance leading to aggression and ADHD. In this regard, it is noteworthy that Frommer and colleagues recently reported a *de novo* frameshift mutation of *ZMYND11* in an individual with schizophrenia[54]. We suggest that truncating mutations in *ZMYND11* are likely to be associated with other more complex neuropsychiatric disorders as children age. Early diagnoses of such carriers as children might be critical to improving their prognosis and outcome.

In conclusion, we have demonstrated that a genotype-first approach, combining copy number and mutation screening across a broad range of neurodevelopmental phenotypes, has the potential to discover new syndromes and to identify the critical genes underlying pathogenic CNVs. Given the large number of exome sequencing studies that are projected and the locus heterogeneity underlying neurocognitive disease, this CNV-SNV integrated approach in conjunction with forward resequencing in large cohorts will serve to identify additional high-impact genes and pathways important in neurodevelopment.

URLs. Exome Variant Server, http://evs.gs.washington.edu/EVS/; Wellcome Trust Case Control Consortium 2, http://www.wtccc.org.uk/ccc2/.

METHODS

Methods and any associated references are available in the online version of the paper.

Accession codes. CNV calls for the combined cases and new controls have been deposited in dbVar under accession nstd100.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

We thank F. Hormozdiari, M. Dennis and T. Brown for useful discussions and for editing the manuscript. B.P.C. is supported by a fellowship from the Canadian Institutes of Health Research. This study makes use of data generated by the Wellcome Trust Case Control Consortium. A full list of the investigators who contributed to the generation of the data is available from http://www.wtccc.org.uk/. J.A.R. and B.S.T. are employees of Signature Genomics Laboratories, LLC, a subsidiary of PerkinElmer, Inc. This work was supported by US National Institute of Mental Health grant MH101221 and Paul G. Allen Family Foundation Award 11631 to E.E.E. E.E.E. is an Allen Distinguished Investigator and an investigator of the Howard Hughes Medical Institute.

AUTHOR CONTRIBUTIONS

B.P.C. and E.E.E. designed the study. B.P.C. performed the data analysis. B.P.C., K.W. and C.B. performed array CGH, MIP sequencing and Sanger validation. J.A.R. and B.S.T. supervised array CGH experiments and coordinated clinical data collection at Signature Genomics. B.W.M.v.B., A.T.V.-v.S., P.B., K.L.F., S.B., L.E.L.M.V., J.H.S.-H., A.H., D.L., D.A., N.B., P.J.L., I.E.S., A.A., R. Pettinato, R.T., N.d.L., M.R.F.R., H.P., M.F., M.S., H.C.M., E.H., C.R., J.G. and B.B.A.d.V. provided clinical samples for resequencing, clinical reports and inheritance testing. J.Y.H.-K., R. Pfundt and N.d.L. curated the Nijmegen *de novo* CNV calls. B.P.C., K.W., C.B., B.J.O., J.S., and E.E.E. designed the MIP gene panel. G.L.C. and H.C.M. identified two *SETBP1* variants in an independent screen. N.K. curated published *de novo* mutations. B.P.C. and E.E.E. wrote the manuscript. All authors have read and approved the final version of the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare competing financial interests: details are available in the online version of the paper.

1. Cooper, G.M. *et al.* A copy number variation morbidity map of developmental delay. *Nat. Genet.* **43**, 838–846 (2011).
2. Kaminsky, E.B. *et al.* An evidence-based approach to establish the functional and clinical significance of copy number variants in intellectual and developmental disabilities. *Genet. Med.* **13**, 777–784 (2011).
3. Moreno-De-Luca, D. *et al.* Using large clinical data sets to infer pathogenicity for rare copy number variants in autism cohorts. *Mol. Psychiatry* **18**, 1090–1095 (2013).
4. Vulto-van Silfhout, A.T. *et al.* Clinical significance of *de novo* and inherited copy-number variation. *Hum. Mutat.* **34**, 1679–1687 (2013).
5. Allen, A.S. *et al. De novo* mutations in epileptic encephalopathies. *Nature* **501**, 217–221 (2013).
6. de Ligt, J. *et al.* Diagnostic exome sequencing in persons with severe intellectual disability. *N. Engl. J. Med.* **367**, 1921–1929 (2012).
7. Gulsuner, S. *et al.* Spatial and temporal mapping of *de novo* mutations in schizophrenia to a fetal prefrontal cortical network. *Cell* **154**, 518–529 (2013).
8. Iossifov, I. *et al. De novo* gene disruptions in children on the autistic spectrum. *Neuron* **74**, 285–299 (2012).
9. Jiang, Y.H. *et al.* Detection of clinically relevant genetic variants in autism spectrum disorder by whole-genome sequencing. *Am. J. Hum. Genet.* **93**, 249–263 (2013).
10. Neale, B.M. *et al.* Patterns and rates of exonic *de novo* mutations in autism spectrum disorders. *Nature* **485**, 242–245 (2012).
11. O'Roak, B.J. *et al.* Sporadic autism exomes reveal a highly interconnected protein network of *de novo* mutations. *Nature* **485**, 246–250 (2012).
12. Rauch, A. *et al.* Range of genetic mutations associated with severe non-syndromic sporadic intellectual disability: an exome sequencing study. *Lancet* **380**, 1674–1682 (2012).
13. Sanders, S.J. *et al. De novo* mutations revealed by whole-exome sequencing are strongly associated with autism. *Nature* **485**, 237–241 (2012).
14. Zaidi, S. *et al. De novo* mutations in histone-modifying genes in congenital heart disease. *Nature* **498**, 220–223 (2013).
15. Jacquemont, S. *et al.* A higher mutational burden in females supports a "female protective model" in neurodevelopmental disorders. *Am. J. Hum. Genet.* **94**, 415–425 (2014).
16. Rudd, M.K. *et al.* Segmental duplications mediate novel, clinically relevant chromosome rearrangements. *Hum. Mol. Genet.* **18**, 2957–2962 (2009).
17. Burkardt, D.D. *et al.* Distinctive phenotype in 9 patients with deletion of chromosome 1q24-q25. *Am. J. Med. Genet. A.* **155A**, 1336–1351 (2011).
18. Dabell, M.P. *et al.* Investigation of *NRXN1* deletions: clinical and molecular characterization. *Am. J. Med. Genet. A.* **161A**, 717–731 (2013).
19. Gimelli, S. *et al.* A rare 3q13.31 microdeletion including *GAP43* and *LSAMP* genes. *Mol. Cytogenet.* **6**, 52 (2013).
20. Madrigal, I., Martinez, M., Rodriguez-Revenga, L., Carrio, A. & Mila, M. 12p13 rearrangements: 6 Mb deletion responsible for ID/MCA and reciprocal duplication without clinical responsibility. *Am. J. Med. Genet. A.* **158A**, 1071–1076 (2012).

21. Paciorkowski, A.R. *et al. MEF2C* haploinsufficiency features consistent hyperkinesis, variable epilepsy, and has a role in dorsal and ventral neuronal developmental pathways. *Neurogenetics* **14**, 99–111 (2013).
22. Rosenfeld, J.A. *et al.* Small deletions of *SATB2* cause some of the clinical features of the 2q33.1 microdeletion syndrome. *PLoS ONE* **4**, e6568 (2009).
23. Stankiewicz, P. *et al.* Recurrent deletions and reciprocal duplications of 10q11.21q11.23 including *CHAT* and *SLC18A3* are likely mediated by complex low-copy repeats. *Hum. Mutat.* **33**, 165–179 (2012).
24. van Bon, B.W. *et al.* The phenotype of recurrent 10q22q23 deletions and duplications. *Eur. J. Hum. Genet.* **19**, 400–408 (2011).
25. Döcker, D. *et al.* Further delineation of the *SATB2* phenotype. *Eur. J. Hum. Genet.* **22**, 1034–1039 (2014).
26. Thorsson, T. *et al.* Chromosomal imbalances in patients with congenital cardiac defects: a meta-analysis reveals novel potential critical regions involved in heart development. *Congenit. Heart Dis.* doi:10.1111/chd.12179 (11 April 2014).
27. Le Meur, N. *et al. MEF2C* haploinsufficiency caused by either microdeletion of the 5q14.3 region or mutation is responsible for severe mental retardation with stereotypic movements, epilepsy and/or cerebral malformations. *J. Med. Genet.* **47**, 22–29 (2010).
28. Ching, M.S. *et al.* Deletions of *NRXN1* (neurexin-1) predispose to a wide spectrum of developmental disorders. *Am. J. Med. Genet. B. Neuropsychiatr. Genet.* **153B**, 937–947 (2010).
29. Shuvarikov, A. *et al.* Recurrent HERV-H–mediated 3q13.2-q13.31 deletions cause a syndrome of hypotonia and motor, language, and cognitive delays. *Hum. Mutat.* **34**, 1415–1423 (2013).
30. Endele, S. *et al.* Mutations in *GRIN2A* and *GRIN2B* encoding regulatory subunits of NMDA receptors cause variable neurodevelopmental phenotypes. *Nat. Genet.* **42**, 1021–1026 (2010).
31. Goffin, A., Hoefsloot, L.H., Bosgoed, E., Swillen, A. & Fryns, J.P. *PTEN* mutation in a family with Cowden syndrome and autism. *Am. J. Med. Genet.* **105**, 521–524 (2001).
32. Koolen, D.A. *et al.* Mutations in the chromatin modifier gene *KANSL1* cause the 17q21.31 microdeletion syndrome. *Nat. Genet.* **44**, 639–641 (2012).
33. Lossin, C. A catalog of *SCN1A* variants. *Brain Dev.* **31**, 114–130 (2009).
34. Santen, G.W. *et al.* Mutations in SWI/SNF chromatin remodeling complex gene *ARID1B* cause Coffin-Siris syndrome. *Nat. Genet.* **44**, 379–380 (2012).
35. Talkowski, M.E. *et al.* Assessment of 2q23.1 microdeletion syndrome implicates *MBD5* as a single causal locus of intellectual disability, epilepsy, and autism spectrum disorder. *Am. J. Hum. Genet.* **89**, 551–563 (2011).
36. O'Roak, B.J. *et al.* Multiplex targeted sequencing identifies recurrently mutated genes in autism spectrum disorders. *Science* **338**, 1619–1622 (2012).
37. Fischbach, G.D. & Lord, C. The Simons Simplex Collection: a resource for identification of autism genetic risk factors. *Neuron* **68**, 192–195 (2010).
38. Sharp, A.J. *et al.* Segmental duplications and copy-number variation in the human genome. *Am. J. Hum. Genet.* **77**, 78–88 (2005).
39. van Bon, B.W. *et al.* Intragenic deletion in *DYRK1A* leads to mental retardation and primary microcephaly. *Clin. Genet.* **79**, 296–299 (2011).
40. Girirajan, S. *et al.* Refinement and discovery of new hotspots of copy-number variation associated with autism spectrum disorder. *Am. J. Hum. Genet.* **92**, 221–237 (2013).
41. Filges, I. *et al.* Reduced expression by *SETBP1* haploinsufficiency causes developmental and expressive language delay indicating a phenotype distinct from Schinzel-Giedion syndrome. *J. Med. Genet.* **48**, 117–122 (2011).
42. Marseglia, G. *et al.* 372 kb microdeletion in 18q12.3 causing *SETBP1* haploinsufficiency associated with mild mental retardation and expressive speech impairment. *Eur. J. Med. Genet.* **55**, 216–221 (2012).
43. DeScipio, C. *et al.* Subtelomeric deletion of chromosome 10p15.3: clinical findings and molecular cytogenetic characterization. *Am. J. Med. Genet. A.* **158A**, 2152–2161 (2012).
44. Ansieau, S. & Leutz, A. The conserved Mynd domain of BS69 binds cellular and oncoviral proteins through a common PXLXP motif. *J. Biol. Chem.* **277**, 4906–4910 (2002).
45. Kateb, F. *et al.* Structural and functional analysis of the DEAF-1 and BS69 MYND domains. *PLoS ONE* **8**, e54715 (2013).
46. Masselink, H. & Bernards, R. The adenovirus E1A binding protein BS69 is a corepressor of transcription through recruitment of N-CoR. *Oncogene* **19**, 1538–1546 (2000).
47. Stefansson, H. *et al.* CNVs conferring risk of autism or schizophrenia affect cognition in controls. *Nature* **505**, 361–366 (2014).
48. Hoischen, A. *et al. De novo* mutations of *SETBP1* cause Schinzel-Giedion syndrome. *Nat. Genet.* **42**, 483–485 (2010).
49. Schinzel, A. & Giedion, A. A syndrome of severe midface retraction, multiple skull anomalies, clubfeet, and cardiac and renal malformations in sibs. *Am. J. Med. Genet.* **1**, 361–375 (1978).
50. Makishima, H. *et al.* Somatic *SETBP1* mutations in myeloid malignancies. *Nat. Genet.* **45**, 942–946 (2013).
51. Piazza, R. *et al.* Recurrent *SETBP1* mutations in atypical chronic myeloid leukemia. *Nat. Genet.* **45**, 18–24 (2013).
52. Wen, H. *et al.* ZMYND11 links histone H3.3K36me3 to transcription elongation and tumour suppression. *Nature* **508**, 263–268 (2014).
53. Yu, B. *et al.* BS69 undergoes SUMO modification and plays an inhibitory role in muscle and neuronal differentiation. *Exp. Cell Res.* **315**, 3543–3553 (2009).
54. Fromer, M. *et al. De novo* mutations in schizophrenia implicate synaptic networks. *Nature* **506**, 179–184 (2014).
55. Alliman, S. *et al.* Clinical and molecular characterization of individuals with recurrent genomic disorder at 10q22.3q23.2. *Clin. Genet.* **78**, 162–168 (2010).
56. Hehir-Kwa, J.Y. *et al. De novo* copy number variants associated with intellectual disability have a paternal origin and age bias. *J. Med. Genet.* **48**, 776–778 (2011).

ONLINE METHODS

Microarray platforms and samples. We combined the 15,767 cases previously published in Cooper *et al.*[1] with 13,318 new cases with intellectual disability and/or developmental delay and related phenotypes that were submitted to Signature Genomics Laboratories, LLC, for clinical microarray-based CGH. Array CGH was performed on nine different CGH platforms (**Supplementary Table 8**). All arrays were reanalyzed from the underlying raw data for CNVs (**Supplementary Note**). The majority of samples were profiled on an array with 135,000 or more probes (64%) with increased density in regions associated with known disorders[1,57]. Initial CNV calls were generated as previously described[57]. Cases were filtered by the following criteria. First, CNVs were filtered for an absolute $\log_2$ ratio of >0.3. Second, to account for excess segmentation, CNVs were manually inspected for potential merging when two CNVs of the same state were within 10% of the larger CNV's length of each other. Cases were also filtered on the basis of the following criteria: $\sigma >$ 0.29722 or excess CNVs (quartile 3 + 3 times the interquartile range (IQR) per array platform). Cases with >3 large (≥500-kb) subtelomeric events (initiating in the first 1.5 subtelomeric megabases of the p or q arm) or with more than 11 CNVs (1.5 times the IQR across all cases) were manually inspected to account for wave artifacts in low-quality samples. Finally, we inspected CNVs completely contained in regions prone to low-ratio CNVs due to wave artifacts (**Supplementary Table 9**). CNVs highlighting new regions of interest were validated on a custom 8-plex Agilent array (**Supplementary Note**). In addition, 5,531 cases previously published by Vulto-Van Silfhout *et al.*[4] were screened for *de novo* CNVs overlapping regions of interest.

We constructed a CNV atlas map by combining 8,329 controls from Cooper *et al.*[1] (dbVar study accession nsdt54) with 11,255 new controls profiled on Affymetrix SNP6 arrays from the Wellcome Trust Case Control Consortium 2 (WTCCC2) 58C cohort, as well as the Atherosclerosis Risk in Communities (ARIC) Community Surveillance Cohort (database of Genotypes and Phenotypes (dbGaP) accession phs000090.v1.p1) (**Supplementary Table 1**). All CNV calling for the ARIC and WTCCC2 58C cohorts was performed using GTC4.1 with default parameters, except for the minimum CNV size and minimum number of probes, which were set to 10 and 20 kb, respectively. One array batch with very low ratio responses (with $\log_2$ ratios at most 16.8% of those expected) was removed from the ARIC study because of poor CNV calling. Additional filtering was applied to remove cases with excessive CNV counts, and a threshold of >72 CNVs per case was established using an outlier detection method for skewed data[58]. Finally, we trimmed CNV calls that falsely extended across centromeric gaps due to small polymorphisms on both arms.

A total of 29,415 rare autosomal CNVs in cases and 741,729 (289,359 new) control CNVs were detected (**Supplementary Table 1**) and deposited into dbVar (study accession nstd100). Informed consent was obtained to publish clinical information and photographs and to further characterize the CNVs present in the individuals with detailed information presented in this paper using a protocol approved by the Signature Genomics Laboratories, LLC, Institutional Review Board–Spokane. Controls were not ascertained specifically for neurological disorders, but all controls were obtained from adult samples providing informed consent, so severe developmental phenotypes should be exceedingly rare in this group.

Statistical analysis. CNV burden was compared between cases and controls for rare CNVs (frequency of <1%) using CNV length excluding gaps and regions annotated as segmental duplications (hg18). The distribution of these CNVs is indicated in **Supplementary Figure 6**. Burden was defined using only the largest CNV to account for the large number of bases encompassed by small CNVs and the difference in array resolution between cases and controls. Statistical comparisons used the Peto and Peto modification of the Gehan-Wilcoxon test (because of non-proportional hazard ratios) to assess overall burden. For significance at specific thresholds, we used the Fisher's exact test. Significance for CNV enrichment was enumerated for all RefSeq genes (NCBI Build 36). All isoforms for each gene were combined into a single entry representing all possible coding bases. Rare CNVs from cases and all control CNVs were then enumerated for only cases where the CNV intersected with an exon. The resulting counts were compared using the one-tailed Fisher's exact test. Likelihood ratios were calculated with standard formulae, and confidence bounds were estimated using the binomial confidence interval for case and control counts calculated by the Clopper-Pearson exact tail area method as described[59]. Additionally, we calculated an empirical *P* value for genes affected by rare CNVs. To do so, we first excluded CNVs residing in regions with elevated mutation rates or unreliable CNV detection. These regions included subtelomeric CNVs initiating in the first 1.5 Mb of each chromosome, over 75% of bases intersecting with hotspots (145.1 Mb across 58 sites) and segmental duplications (130.4 Mb across 7,264 sites), initiating or terminating in a centromere gap region. All CNVs under 10 Mb in length were then randomly shuffled (chromosome selection was weighted by the number of bases not filtered) under these constraints for cases and controls, and Fisher's exact tests were calculated 20,000 times for deletions and duplications of each gene. The empirical *P* value was defined as the number of simulations more significant than observed plus one divided by the number of simulations plus one. CNV burden for regions was also enumerated using a windowed analysis of rare case CNVs over 250 kb (**Supplementary Data Set 1**). Window start and end points were defined on the basis of all unique breakpoints in the signature array. Breakpoint pairs under 50 kb in length were then filtered out, as these represent uncertainty in the edges of signature calls. Counts for *P* values were based on 40% coverage of each window by cases (over 250 kb) or controls (all CNVs). Significance was calculated using the one-tailed Fisher's exact test, and **Supplementary Figure 2** shows the negative logarithm of the *P* value. In many cases, the critical region might represent multiple subregions that individually reached significance. Here we report the larger region where smaller subregions are indicated by a number of additional CNVs over the background, preventing refinement to a single candidate gene. Because of the high prior probability of pathogenicity for large CNVs, the lack of independence between genes disrupted by CNVs and the high OR estimate for most pathogenic loci, we have chosen to report nominal significance in all cases in addition to the Benjamini-Hochberg *q* value, which represents an overestimate of the false discovery rate in our analyses[60]. Please see the **Supplementary Note** for details on our interpretation of *q* values in this study.

Joint CNV and SNV haploinsufficient mutation probabilities. We developed a model based on the hypergeometric distribution for event counts to calculate the probability of gene enrichment by integration of truncating SNV mutations and CNV deletions. For each gene, we enumerated the total number of loss-of-function events observed: cases with and without deletion CNVs (*a* and *b*); controls with and without deletion CNVs (*c* and *d*); cases with and without truncating SNV and indel mutations (*a*2 and *b*2); and controls with and without truncating SNV and indel mutations (*c*2 and *d*2). We computed the observed frequency (*Z*) of loss-of-function events (CNVs and SNVs) (Eq. 1). We assumed that mutations and CNVs were independent (as supported by the rare nature of these events); however, in cases with more frequent observations, the interaction term could be included in the calculation of *Z*. This threshold was applied to calculate probabilities with equation 2. When CNV or truncating SNV and indel mutation counts were 0 for both cases and controls, the *P* value reduces to the equivalent of the one-tailed Fisher's exact test for the assay with counts. This method also has the benefit of allowing negative observations from one assay to decrease the significance of a gene. For example, a gene with no CNVs in controls but many truncating SNV mutations would be negatively affected by those events.

$$Z = \frac{a}{a+b} + \frac{a2}{a2+b2} \tag{1}$$

$$p = \sum_{i=0}^{a+c} \sum_{j=0}^{a2+c2} \left(\frac{\binom{a+b}{i}\binom{c+d}{a+c-i}}{\binom{a+b+c+d}{a+c}} \times \frac{\binom{a2+b2}{j}\binom{c2+d2}{a2+c2-j}}{\binom{a2+b2+c2+d2}{a2+c2}} \right), \quad for \frac{i}{a+b} + \frac{j}{a2+b2} \geq Z \tag{2}$$

Truncation *P* values. For genes with truncating mutations in controls, we also compared the effect on protein length (in the context of retained

doi:10.1038/ng.3092

wild-type amino acids) in cases and controls on the basis of annotated isoforms. Although early stop-gain mutations might lead to either nonsense-mediated decay or truncated proteins, this model does not discriminate between these outcomes as both result in proteins without wild-type function. For splice-site mutations, we extracted the most likely lost exon and determined the likely protein effect (in-frame loss or introduction of a frameshift or stop codon). Predicted protein lengths for ESP6500 and cases were compared using the log-rank test.

MIP sequencing and sample cohorts. Targeted sequencing of candidate genes was accomplished using the MIP resequencing method as described[36]. In total, we successfully targeted the coding sequence and splice-donor and splice-acceptor sites of 26 genes with 1,388 MIPs. We barcoded and sequenced 192 samples for each Illumina HiSeq lane, and all analyses were performed as described[36]. We included 192 samples in each Illumina HiSeq 2000 lane with 1,388 MIP probes covering 26 genes. Details on the MIP probes used, their individual performance and concentrations in the pool are detailed in **Supplementary Table 10**.

To compare data between exome and MIP sequencing, we calculated statistics only for sites (case and control) with an average read depth of >20 in ESP6500 and no intersection with low-complexity repeat sequence (as defined by Dustmasker).

In total, we screened 8,060 unique samples, including 5,633 probands and 2,427 unaffected siblings from the Simons Simplex Collection. In addition to variant-level filtering, samples were filtered by quality control on the basis of the percentage of MIPs with at least 20 reads (our minimum for variant calling). Probands were required to have sufficient coverage for 75% of targets, whereas control samples were required to have 90% of targets covered. This resulted in the inclusion of 2,193/2,427 controls and 4,716/5,633 cases in the final analysis (**Supplementary Fig. 7**).

Cases were consented for resequencing and recontact for inheritance testing. Samples were acquired from the Autism Phenome Project (D.A.), Leuven (H.P.), Murdoch (I.E.S.), Adelaide (J.G), Nijmegen (B.B.A.d.V.), SAGE (R.B.) and Troina (C.R.) (**Supplementary Table 11**).

57. Duker, A.L. *et al.* Paternally inherited microdeletion at 15q11.2 confirms a significant role for the SNORD116 C/D box snoRNA cluster in Prader-Willi syndrome. *Eur. J. Hum. Genet.* **18**, 1196–1201 (2010).
58. Hubert, M. & Van der Veeken, S. Outlier detection for skewed data. *J. Chemometr.* **22**, 235–246 (2008).
59. Rosenfeld, J.A., Coe, B.P., Eichler, E.E., Cuckle, H. & Shaffer, L.G. Estimates of penetrance for recurrent pathogenic copy-number variations. *Genet. Med.* **15**, 478–481 (2013).
60. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J.R. Stat. Soc.* **57**, 289–300 (1995).