nature genetics

A copy number variation morbidity map of developmental delay

Gregory M Cooper[1,16,17], Bradley P Coe[1,17], Santhosh Girirajan[1,17], Jill A Rosenfeld[2], Tiffany H Vu[1], Carl Baker[1], Charles Williams[3], Heather Stalker[3], Rizwan Hamid[4], Vickie Hannig[4], Hoda Abdel-Hamid[5], Patricia Bader[6], Elizabeth McCracken[7], Dmitriy Niyazov[8], Kathleen Leppig[9], Heidi Thiese[9], Marybeth Hummel[10], Nora Alexander[10], Jerome Gorski[11], Jennifer Kussmann[11], Vandana Shashi[12], Krys Johnson[13], Catherine Rehder[14], Blake C Ballif[2], Lisa G Shaffer[2] & Evan E Eichler[1,15]

To understand the genetic heterogeneity underlying developmental delay, we compared copy number variants (CNVs) in 15,767 children with intellectual disability and various congenital defects (cases) to CNVs in 8,329 unaffected adult controls. We estimate that ~14.2% of disease in these children is caused by CNVs >400 kb. We observed a greater enrichment of CNVs in individuals with craniofacial anomalies and cardiovascular defects compared to those with epilepsy or autism. We identified 59 pathogenic CNVs, including 14 new or previously weakly supported candidates, refined the critical interval for several genomic disorders, such as the 17q21.31 microdeletion syndrome, and identified 940 candidate dosage-sensitive genes. We also developed methods to opportunistically discover small, disruptive CNVs within the large and growing diagnostic array datasets. This evolving CNV morbidity map, combined with exome and genome sequencing, will be critical for deciphering the genetic basis of developmental delay, intellectual disability and autism spectrum disorders.

Large CNVs are enriched in the aggregate among severe pediatric disease, including neurological and congenital birth defects[1,2] as well as neuropsychiatric diseases[3–5]. Clinical interpretation of individual loci has been problematic for several reasons. First, except for CNV 'hotspots' flanked by duplications prone to unequal crossing over and elevated *de novo* mutation rates[6,7], disease associations for many individual CNVs remain unclear because of their rarity and the need to screen extraordinarily large sample sizes. Second, even for CNVs with clear pathogenicity, the dosage-sensitive genes that underlie the phenotypes observed have generally not been identified because the CNVs are large and encompass many genes. Finally, considerable variation in expressivity is often observed, with the same lesion contributing to different disease outcomes[8–12]. Thus, although their disease risk in general is well established, the phenotypic consequences for most large CNVs are not well characterized nor have these effects been fine mapped. Here, we leverage a collection of data from 15,767 children with various developmental and intellectual disabilities and compare them to a CNV map we generated from 8,329 adult controls to produce a detailed genome-wide morbidity map of developmental delay and congenital birth defects. We report striking differences in the CNV landscape between cases and controls, highlight potentially pathogenic genes, refine known disease-causing mutations and develop methods to opportunistically discover smaller disruptive CNVs from clinical datasets.

RESULTS

Study overview

We analyzed 15,767 DNA samples from children referred to Signature Genomic Laboratories, LLC, with a general diagnosis of intellectual disability and/or developmental delay, although we note that this cohort also includes a constellation of phenotypes including, but not restricted to, congenital malformation, hypotonia and feeding difficulties, speech and motor deficits, growth retardation, cardiovascular and renal defects, epilepsy, hearing impairment, craniofacial and skeletal features and behavioral issues. Overall, 73% of the cases suffer from intellectual disability, developmental delay and/or autism spectrum disorder, and 12% of the cases were not annotated. The remaining cases were classified with various congenital abnormalities. Detailed phenotypic information was limited to the 48.4% of the cases for which specific subclassifications could be made, including

[1]Department of Genome Sciences, University of Washington, Seattle, Washington, USA. [2]Signature Genomic Laboratories, LLC, Spokane, Washington, USA. [3]Department of Pediatrics, Division of Genetics, University of Florida, Gainesville, Florida, USA. [4]Vanderbilt University Medical Center, Nashville, Tennessee, USA. [5]Department of Pediatrics, Division of Child Neurology, University of Pittsburgh, Pittsburgh, Pennsylvania, USA. [6]Northeast Indiana Genetic Counseling Center, Ft. Wayne, Indiana, USA. [7]Children's Hospital Pittsburgh, Pittsburgh, Pennsylvania, USA. [8]Ochsner Clinic, New Orleans, Louisiana, USA. [9]Group Health Cooperative, Seattle, Washington, USA. [10]West Virginia University, Morgantown, West Virginia, USA. [11]University of Missouri, Columbia, Missouri, USA. [12]Department of Pediatrics, Duke University Medical Center, Durham, North Carolina, USA. [13]Department of Pathology, Duke University Medical Center, Durham, North Carolina, USA. [14]Clinical Molecular Diagnostic Laboratory, Duke University Health System, Durham, North Carolina, USA. [15]Howard Hughes Medical Institute, University of Washington, Seattle, Washington, USA. [16]Present address: HudsonAlpha Institute for Biotechnology, Huntsville, Alabama, USA. [17]These authors contributed equally to this work. Correspondence should be addressed to E.E.E. (eee@gs.washington.edu).

Received 11 April; accepted 22 July; published online 14 August 2011; doi:10.1038/ng.909

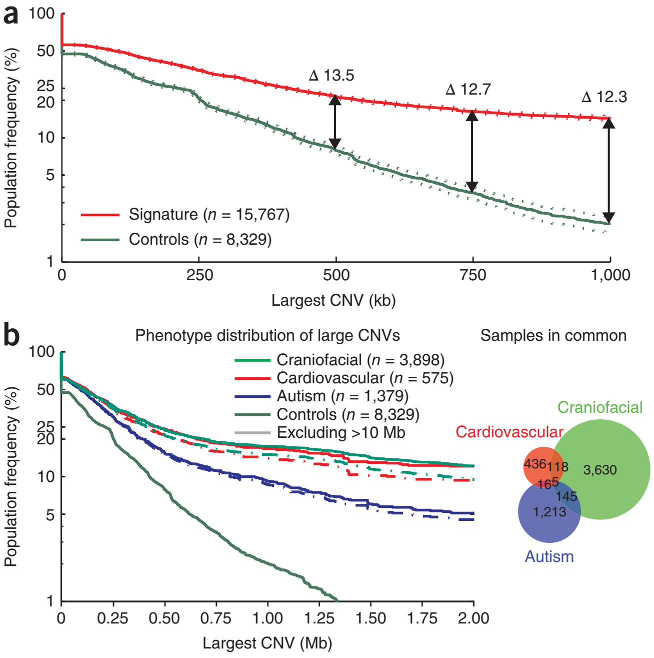

Figure 1 CNV size distributions in affected and unaffected individuals. The population frequency of the largest CNV in a sample is displayed as a survivor function with the proportion of samples carrying a CNV of a given size displayed as a curve and the 95% confidence intervals indicated by dotted lines. (**a**) The distribution of large CNVs in the Signature set (filtered to only contain events detectable by the Illumina 550K array) compared to our control population (downsampled to only events detectable by the Signature 97K array) is indicated for the overall population. After corrections for different array densities, we observed a >13.5% increase in CNV burden beyond 500 kb in cases, with a proportion of the burden representing potentially new loci. (**b**) We also performed a similar analysis on subphenotypes; in this analysis, we included all Signature CNVs in conjunction with downsampled control CNVs, as we are highlighting interphenotype differences rather than case versus control frequencies. The plot depicts autism, cardiovascular and craniofacial phenotypes, which represent fairly distinct sample sets, and shows an increased burden for the cardiovascular and craniofacial phenotypes, even after excluding karyotypically visible (>10 Mb) events.

575 cases with cardiovascular defects, 1,776 with epilepsy and/or seizure disorder, 1,379 with autism spectrum disorder and 3,898 with craniofacial defects (**Supplementary Tables 1** and **2**).

We analyzed DNA samples obtained from whole blood using customized array comparative genomic hybridization (CGH) at an average probe density of ~97,000 oligonucleotides, which is sufficient for reliable genome-wide detection of CNVs >300 kb and for targeted detection of events >40 kb for approximately one-fourth of the genome[13]. After filtering, a total of 16,526 rare (<1% population frequency) autosomal CNV calls were made with an average of 1.05 CNV events per individual (with a median CNV size of 213 kb). Using a customized higher density microarray and fluorescent *in situ* hybridization, we validated 402 of 425 CNVs (with a precision of 0.945) greater than 150 kb (**Supplementary Note** and **Supplementary Table 3**). Similarly, manual inspection of calls with low log ratios or *z*-scores (with absolute values of log ratios < 0.25 and *z*-scores < 1.5) suggests a false discovery rate of 0.0138. For comparison, we identified CNVs from a control set of 8,329 adult samples assayed using multiple Illumina genome-wide SNP microarrays. These samples were studied as part of genome-wide association studies (dbGaP) for phenotypes unrelated to neurological disease (for example, lipid concentration levels, blood pressure, asthma and so on) (**Supplementary Table 4**). We called CNVs using a Hidden Markov Model (HMM)-based discovery method[14] with an overall precision of 0.892 in identifying large CNVs (>100 kb) (with 6/6 and 19/22 validated calls)[15,16]. From this dataset, we identified 446,736 CNVs with an average of 53.6 events (rare and common) per individual (and a median size of 1.9 kb). Because of the increased probe density (most controls assayed using arrays with >550,000 probes), our control dataset provides increased CNV detection power and resolution when compared to the disease dataset, reducing the potential for spurious CNV enrichments within cases (see Online Methods).

CNV burden

We compared CNV content between the cases and controls excluding common CNVs (>1% population frequency). Consistent with previous studies of pediatric neurological disease[3–5,17,18], we found a significant excess of large CNVs among cases relative to controls. This excess is evident at 250-kb CNVs and becomes more pronounced with increasing CNV size (**Fig. 1a**). For example, at a threshold of 400 kb, ~25.7% (4,047 cases) of the children we studied with intellectual disability and/or developmental delay harbor an event of at least this size compared to 11.5% of the controls, suggesting that an estimated 14.2% of intellectual disability and/or developmental delay in this cohort is caused by the presence of CNVs >400 kb in length (odds ratio (OR) = 2.7, $P = 5.86 \times 10^{-158}$). At a threshold of 1.5 Mb, we identified 1,782 (11.3%) affected individuals compared to only 52 (0.6%) controls (OR = 20.3, $P = 6.87 \times 10^{-266}$), and at a threshold of 3.0 Mb, this odds ratio jumps to 47.7 ($P = 1.68 \times 10^{-197}$). There is a remarkably strong correlation ($r^2 = 0.97$) with the *de novo* rate as a function of increasing CNV size, with 50% of events at 1 Mb reported as inherited (**Supplementary Fig. 1**).

We detected 1,492 CNVs in 1,400 individuals within 45 known genomic disorder regions (**Table 1** and **Supplementary Table 5**). Among these individuals, deletions are twice as common ($n = 954$ deletions compared to $n = 538$ duplications) and show greater average penetrance (96.3%) when compared to duplications (94.3% penetrance). We note that 'classic' phenotypically well defined syndromes known to result from CNVs (for example, Smith-Magenis syndrome and Williams syndrome) are underrepresented here relative to other cohorts of individuals with similar phenotypes (**Supplementary Table 6**), suggesting that our estimate of CNV burden in intellectual disability and/or developmental delay is not upwardly biased by ascertainment for known CNV carriers.

Examining the size distribution of CNVs in the context of major subphenotypes shows that the large CNV burden is increased in more severe developmental phenotypes associated with multiple congenital abnormalities. We find, for example, that children also diagnosed with craniofacial and cardiovascular defects showed a significantly increased burden of large CNVs when compared to children with autism spectrum disorder ($P = 4.99 \times 10^{-10}$ and $P = 6.45 \times 10^{-5}$, respectively, at >400 kb) (**Fig. 1b**). Children with an additional diagnosis of epilepsy and/or severe seizure disorder tended to have a more intermediate CNV burden when compared to individuals with autism or more severe intellectual disability (**Supplementary Fig. 2**). These distinctions remained significant even after excluding CNVs larger than 10 Mb (which would have been detectable by karyotype analysis) and when the CNV burden among the subset of controls screened for psychiatric disease was used as the baseline, showing a role for large CNVs in more severe phenotypic variation.

Table 1 Frequencies of known genomic disorders in cases and controls

			Deletions (<10 Mb)					Duplications (<10 Mb)				
Chr.	Start (Mb)	End (Mb)	Deletion	Cases[a]	Controls[b]	*P*	Penetrance	Duplication	Cases[a]	Controls[b]	*P*	Penetrance
1	0.00	10.00	1p36 deletion syndrome (*GABRD*)[c]	79	0	2.6×10^{-15}	1.00	1p36 duplication (*GABRD*)[c]	16	1	0.0074	0.94
1	144.00	144.34	TAR deletion (*HFE2*)	13	2	0.0659	0.87	1q21.1 duplication (*HFE2*)	25	6	0.0511	0.81
1	145.04	145.86	1q21.1 deletion (*GJA5*)	47	2	3.3×10^{-7}	0.96	1q21.1 duplication (*GJA5*)	26	1	0.0002	0.96
2	96.09	97.04	2q11.2 deletion (*LMAN2L, ARID5A*)	2	0	0.4282	1.00	2q11.2 duplication (*LMAN2L, ARID5A*)	1	0	0.6543	1.00
2	100.06	107.81	2q11.2q13 deletion (*NCK2, FHL2*)	0	0	1.0000	NA	2q11.2q13 duplication (*NCK2, FHL2*)	2	0	0.4282	1.00
2	110.18	110.34	2q13 deletion (*NPHP1*)	78	30	0.0813	0.72	2q13 duplication (*NPHP1*)	118	32	0.0003	0.79
2	239.37	242.12	2q37 deletion (*HDAC4*)[c]	22	0	0.0001	1.00	2q37 duplication (*HDAC4*)[c]	0	0	1.0000	NA
3	197.23	198.84	3q29 deletion (*DLG1*)	6	0	0.0785	1.00	3q29 duplication (*DLG1*)	4	0	0.1833	1.00
4	1.84	1.98	Wolf-Hirschhorn deletion (*WHSC1, WHSC2*)[c]	21	0	0.0001	1.00	Wolf-Hirschhorn region duplication	7	0	0.0513	1.00
5	175.65	176.99	Sotos syndrome deletion (*NSD1*)	8	0	0.0336	1.00	5q35 duplication (*NSD1*)	0	0	1.0000	NA
6	100.92	101.05	6q16 deletion (*SIM1*)[c]	1	0	0.6543	1.00	6q16 duplication (*SIM1*)[c]	1	0	0.6543	1.00
7	72.38	73.78	Williams syndrome deletion (*ELN, GTF2I*)	42	0	1.8×10^{-8}	1.00	Williams syndrome duplication (*ELN, GTF2I*)	16	0	0.0011	1.00
7	74.80	76.50	WBS-distal deletion (*RHBDD2, HIP1*)	2	0	0.4282	1.00	WBS-distal duplication (*RHBDD2, HIP1*)	0	0	1.0000	NA
8	8.13	11.93	8p23.1 deletion (*SOX7, CLDN23*)	7	0	0.0513	1.00	8p23.1 duplication (*SOX7, CLDN23*)	7	0	0.0513	1.00
9	136.95	140.20	9q34 deletion (*EHMT1*)[c]	60	0	8.5×10^{-12}	1.00	9q34 duplication (*EHMT1*)[c]	4	0	0.1833	1.00
10	81.95	88.79	10q23 deletion (*NRG3, GRID1*)	8	0	0.0336	1.00	10q23 duplication (*NRG3, GRID1*)	1	0	0.6543	1.00
11	43.94	46.02	Potocki-Shaffer syndrome (*EXT2*)[c]	5	0	0.1199	1.00	11p11.2 duplication (*EXT2*)[c]	0	0	1.0000	NA
11	67.51	70.96	*SHANK2 FGFs* deletion	1	0	0.6543	1.00	*SHANK2 FGFs* duplication	0	0	1.0000	NA
12	63.36	66.93	12q14 deletion syndrome (*GRIP1, HMGA2*)[c]	2	0	0.4282	1.00	12q14 duplication (*GRIP1, HMGA2*)[c]	0	0	1.0000	NA
13	19.71	19.91	13q12 deletion (*CRYL1*)[c]	14	12	0.9240	0.54	13q12 duplication (*CRYL1*)[c]	4	0	0.1833	1.00
15	20.35	20.64	15q11.2 deletion (*NIPA1*)	94	19	2.1×10^{-5}	0.83	15q11.2 duplication (*NIPA1*)	64	36	0.6614	0.64
15	22.37	26.10	Prader-Willi/Angelman syndrome	16	0	0.0011	1.00	Prader-Willi/Angelman region duplication	27	0	1.1×10^{-5}	1.00
15	28.92	30.27	15q13.3 deletion (*CHRNA7*)	42	0	1.8×10^{-8}	1.00	15q13.3 duplication (*CHRNA7*)	20	3	0.0200	0.87
15	70.70	72.20	15q24 BP0-BP1 deletion (*BBS4, NPTN, NEO1*)	4	0	0.1833	1.00	15q24 BP0-BP1 duplication (*BBS4, NPTN, NEO1*)	1	0	0.6543	1.00
15	70.70	73.58	15q24 BP0-BP1 deletion (*PML*)	4	0	0.1833	1.00	15q24 BP0-BP1 duplication (*PML*)	4	0	0.1833	1.00
15	73.76	75.99	15q24 BP2-BP3 deletion (*FBXO22, TPSAN3*)	1	0	0.6543	1.00	15q24 BP2-BP3 duplication (*FBXO22, TPSAN3*)	0	0	1.0000	NA
15	80.98	82.53	15q25.2 deletion (*HOMER2, BNC1*)	1	0	0.6543	1.00	15q25.2 duplication (*HOMER2, BNC1*)	0	0	1.0000	NA
15	97.18	100.34	None	10	1	0.0641	0.91	None	1	0	0.6543	1.00
16	3.72	3.80	Rubinstein-Taybi syndrome[c]	7	0	0.0513	1.00	Rubinstein-Taybi region duplication	6	0	0.0785	1.00
16	15.41	16.20	16p13.11 deletion (*MYH11*)	18	3	0.0361	0.86	16p13.11 duplication (*MYH11*)	24	10	0.3315	0.71
16	21.26	29.35	16p11.2p12.1 deletion	2	0	0.4282	1.00	16p11.2p12.1 duplication	2	0	0.4282	1.00
16	21.85	22.37	16p12.1 deletion (*EEF2K, CDR2*)	37	3	0.0001	0.93	16p12.1 duplication (*EEF2K, CDR2*)	4	1	0.4368	0.80
16	28.68	29.02	16p11.2 distal deletion (*SH2B1*)	15	1	0.0107	0.94	16p11.2 distal duplication (*SH2B1*)	14	2	0.0484	0.88
16	29.56	30.11	16p11.2 deletion (*TBX6*)	64	3	3.4×10^{-9}	0.96	16p11.2 duplication (*TBX6*)	28	2	0.0004	0.93
17	0.05	2.54	17p13.3 deletion (both *YWHAE* and *PAFAH1B1*)[c]	7	0	0.0513	1.00	17p13.3 duplication (both *YWHAE* and *PAFAH1B1*)[c]	2	0	0.4282	1.00
17	0.50	1.30	17p13.3 deletion (including *PAFAH1B1*)[c]	8	0	0.0336	1.00	17p13.3 duplication (including *PAFAH1B1*)[c]	6	0	0.0785	1.00
17	2.31	2.87	17p13.3 deletion (including *YWHAE*)[c]	7	0	0.0513	1.00	17p13.3 duplication (including *YWHAE*)[c]	4	0	0.1833	1.00
17	14.01	15.44	HNPP (*PMP22*)	3	0	0.2801	1.00	CMT1A (*PMP22*)	9	2	0.2086	0.82
17	16.65	20.42	Smith-Magenis syndrome deletion	16	0	0.0011	1.00	Potocki-Lupski syndrome	9	0	0.0220	1.00
17	26.19	27.24	*NF1* deletion syndrome	5	0	0.1199	1.00	*NF1* duplication	2	0	0.4282	1.00
17	31.89	33.28	RCAD (renal cysts and diabetes) (*TCF2*)	14	2	0.0484	0.88	17q12 duplication	18	3	0.0361	0.86
17	41.06	41.54	17q21.31 deletion (*MAPT*)	23	0	0.0001	1.00	17q21.31 duplication (*MAPT*)	2	0	0.4282	1.00
22	17.40	18.67	DiGeorge/VCFS deletion	96	0	0.0000	1.00	22q11.2 duplication	50	5	1.3×10^{-5}	0.91
22	20.24	21.98	22q11.2 distal deletion (*BCR, MAPK1*)	13	0	0.0040	1.00	22q11.2 distal duplication (*BCR, MAPK1*)	7	0	0.0513	1.00
22	49.46	49.52	Phelan-McDermid syndrome deletion (*SHANK3*)[c]	45	0	0.0000	1.00	22q13 duplication (*SHANK3*)[c]	7	0	0.0513	1.00

All coordinates are according to build36. The genes in parentheses are potential candidate genes and identifiers of the genomic locations. Chr., chromosome; VCFS, velocardiofacial syndrome; WBS, Williams-Beuren syndrome; TAR, thrombocytopenia-absent radius; HNPP, hereditary neuropathy with liability to pressure palsies; CMT1A, Charcot-Marie-Tooth disease type 1A; NA, not applicable. We identified no CNVs in 2p15p16.1 (*VRK2*), 15q24 (BP1-BP2) (*CLK3*), 15q24 (*SIN3A*), 17q23 (*TUBD1*) and 17q23.1-q23.2 (*TBX2* and *TBX4*). Note that a single CNV may encompass more than one genomic disorder.

[a] Total cases, *n* = 15,767. [b] Total controls, *n* = 8,329. [c] Rearrangements not mediated by segmental duplications.

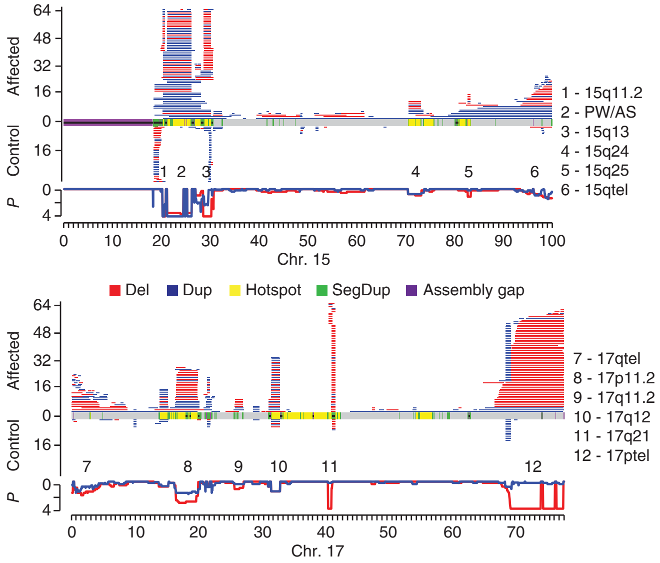

Figure 2 Maps of CNV locations for chromosomes 15 and 17. CNVs (>400 kb) in affected individuals are shown in the upper portion for each chromosome with control CNVs shown in the lower portion. Disease enrichment *P* values are plotted just below the control CNV maps, computed in 200-kb windows along each chromosome (with a step size of 50 kb). Deletions and duplications are shown in red and blue, respectively, with the *P* value wiggle plots colored accordingly and plotted on a negative log scale. In the middle of each plot, chromosomal features are colored as depicted. Significantly enriched regions are numbered and named on the right-hand side.

Locus-specific enrichments

A comparison of the CNV landscape between cases and controls shows striking differences and some general genomic architectural features (**Fig. 2**). To compensate for the effects of breakpoint imprecision and multi-platform comparisons, we contrasted the number of deletions (or duplications) present in cases versus controls in 200-kb windows along the human genome using a Fisher's exact test (**Supplementary Table 7** and **Supplementary Fig. 3**). This analysis identified 80 genomic regions that were at least weakly enriched for CNVs (counting deletions and duplications separately) among cases (with at least five windows having $P < 0.1$), 27 of which showed strong evidence for enrichment ($P < 0.001$). Notably, 27.5% (22/80) of the enriched CNV loci reside at genomic hotspots flanked by large (>10 kb) blocks of highly similar (>90%) segmental duplication and include most known genomic disorders (**Supplementary Table 7**). An additional 46 enrichments represent large CNVs near telomeres (**Supplementary Fig. 4**). Although we observed enrichments at one or both ends of all chromosomes, 12 chromosome ends showed particularly strong ($P < 0.001$) enrichment. Of the 80 CNV loci, 15 are new or are supported by isolated case reports (**Table 2**). Additional phenotypic details for CNV carriers, including ethnicity and inheritance status, at each of these 15 CNV loci are provided in **Supplementary Table 8**, in some cases with comparison to similar CNVs observed in case reports from DECIPHER (Database of Chromosomal Imbalance and Phenotype in Humans using Ensembl Resources)[19]. We note that one of these 15 CNVs (duplications at 10p15.3) appears to be enriched among cases as a consequence of allelic stratification between the African and European populations and was thus eliminated from further consideration (Online Methods and **Supplementary Note**).

Among the 14 newly discovered CNV loci for intellectual disability and/or developmental delay, we identified a 660-kb deletion mapping to chromosome 15q25.2 flanked by segmental duplications (69.8 kb, 98.6% identity) (**Fig. 3a**). The deletion is absent from the controls analyzed here and from the Database of Genomic Variants (see URLs) but is present in five affected individuals (including two siblings) among the intellectual disability and/or developmental delay sample set. Clinical aspects of the probands were variable and consisted of neurologic features and developmental delay (**Supplementary Table 9**); one female had only mild motor delay associated with a congenital myopathy but was otherwise cognitively normal. The two brothers with the deletion both had autism spectrum disorders, but additional family members were not tested (**Supplementary Note**). A previous meta-analysis found this deletion in 4 of 6,860 cases[16] with schizophrenia and autism compared to 0 of 5,674 controls (combined with this study, $P = 0.037$ after excluding one sibling). Thus, although the statistical significance remains modest and population stratification cannot be definitively ruled out (**Supplementary Note**), these data suggest a potentially new genomic disorder that will be observed at a frequency of 1/3,000 referred cases.

Table 2 New potentially pathogenic loci identified by sliding window analysis

Chr.	Start (Mb)	End (Mb)	Size (Mb)	CNV	*P* (adjusted)	Cases (adjusted)[a]	Controls (adjusted)[a]	Description	Ethnicity[b]
2[c,d]	111.05	112.95	1.9	Del	0.006 (0.032)	12 (12)	0 (1)	2q13	10C,1A
10[c]	81.6	88.9	7.3	Del	0.014 (0.064)	10 (10)	0 (1)	10q23.1	6C,1O
2	45.2	45.9	0.7	Dup	0.022 (0.022)	9 (9)	0 (0)	2p21	8C
2[b,c]	111.05	112.85	1.8	Dup	0.034 (0.022)	8 (9)	0 (0)	2q13	5C,2O
4	9.45	10.45	1.0	Dup	0.034 (0.051)	8 (7)	0 (0)	4p16.1	6C,1A,1O
4	81.95	83.35	1.4	Del	0.034 (0.034)	8 (8)	0 (0)	4q21.21–q21.22	6C,1A
2	3.25	3.45	0.2	Dup	0.051 (0.051)	7 (7)	0 (0)	2p25.3	3C,1O
2	165.4	166.1	0.7	Del	0.051 (0.051)	7 (7)	0 (0)	2q24.3	5C,1O
21	19.95	20.25	0.3	Del	0.051 (0.079)	7 (6)	0 (0)	21q21.1	1C,1A,2O
8	53.45	54.05	0.6	Dup	0.051 (0.051)	7 (7)	0 (0)	8q11.23	6C,1O
1	170	170.6	0.6	Del	0.079 (0.079)	6 (6)	0 (0)	1q24.3	5C
12	8.05	8.25	0.2	Dup	0.079 (0.051)	6 (7)	0 (0)	12p13.31	6C
15[c,d]	82.9	83.6	0.7	Del	0.079 (0.120)	6 (5)	0 (0)	15q25	1C,2A,2O
6	20.85	21.25	0.4	Del	0.079 (0.079)	6 (6)	0 (0)	6p22.3	1E,1A,1O

Chr., chromosome; CNV, copy number variants.

[a]The counts and *P* values are based on the single most significant 200-kb window, whereas the adjusted counts include all samples with a CNV overlapping the region but exclude all related samples (**Supplementary Table 7**). [b]C, Caucasian (primarily European descent); A, African-American; O, other. [c]Previously described loci[16,50] with uncertain pathogenicity. [d]Hotspot regions.

One of the most common genomic hotspots in this study is at 15q11.2 (*NIPA1*), a 292-kb deletion whose pathogenicity has been considered uncertain[4,20]. In terms of frequency, the 15q11.2 deletion is second only to the velocardiofacial/DiGeorge syndrome (VCF/DGS) deletion, and our data indicate that it is significantly enriched (OR = 2.36, $P = 2.5 \times 10^{-5}$), albeit at lower penetrance (0.83) than those in most other genomic disorders. In addition, we find support for the pathogenicity of duplications of obesity-associated 16p11.2 (*SH2B1*)[21,22] and epilepsy-associated 15q13.3 (*CHRNA7*)[23]. We also analyzed 111 regions of the human genome predicted to be prone to recurrent microdeletions and microduplications based on the presence of homologous segmental duplications at their flanks in the reference assembly[6]. Of these potential hotspots, 62 harbored CNVs likely mediated by nonallelic homologous recombination between the flanking segmental duplications ('active hotspots'), whereas the remaining 49 did not. The presence of segmental duplications in

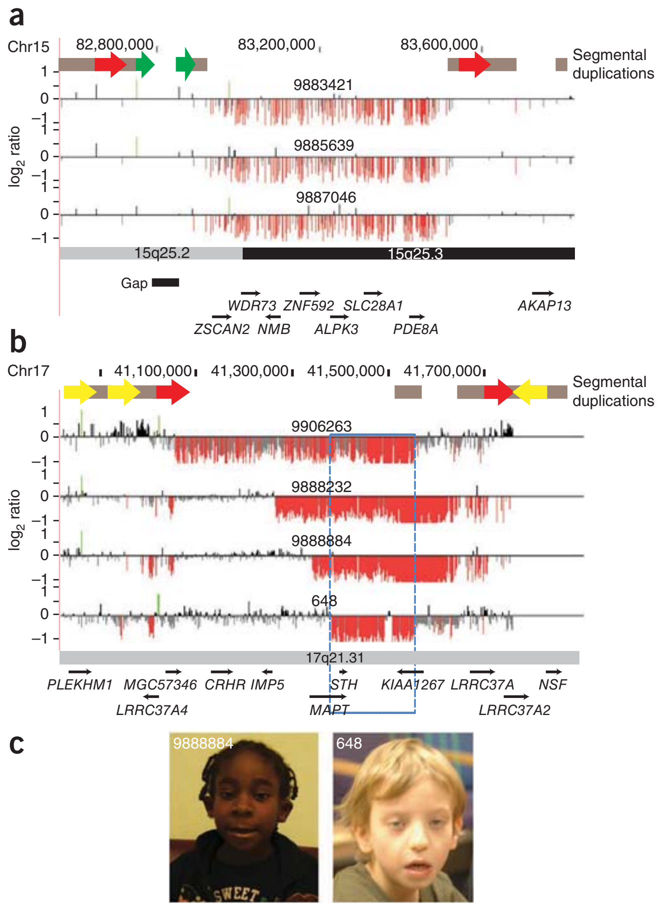

Figure 3 Discovery of new microdeletions associated with genomic disorders. (**a**) A newly discovered microdeletion on chromosome 15q25.2q25.3. Array CGH analysis for three individuals with a 660-kb (chr15:82,889,423–83,552,890) deletion is shown. This microdeletion maps within a genomic hotspot flanked by high-identity segmental duplication blocks. Intrachromosomal segmental duplications of high similarity relevant to this hotspot region are depicted as red (69.8 kb, 98.6% identity) and green (17.6 kb, 98.6% identity) block arrows. Note that the directly orientated segmental duplications (red block arrows) likely mediate the underlying 15q25 rearrangements by non-allelic homologous recombination. This region also contains a 60-kb (chr15:82,775,465–82,835,495) gap in the current builds (build 36 and build 37) of the reference genome assembly. (**b**) Atypical 17q21.31 microdeletions refine critical interval genes. High-density array CGH for the 17q21.31 microdeletion region is shown for three individuals. Probes with log2 ratios below a threshold of 1.5 standard deviation from the normalized mean log2 ratio denote deletions (red). We identified the typical deletions (top panel) in 23 individuals, whereas we identified atypical deletions in three individuals. Note that the smallest deletion (blue dashed box) refines the phenotype-associated critical region (chr17:41,356,798–41,631,306) to encompass only five RefSeq genes. (**c**) Photographs of two individuals (9888884 and 648) with atypical deletions. Subject 9888884 is a 5-year-old female child with clinical features typical of 17q21.31 microdeletion syndrome, including distinctive dysmorphic features with a bulbous nasal tip, upslanting and almond-shaped palpebral fissures, long face, strabismus, epicanthal folds and prominent ears; developmental delay with limited speech; hypotonia in infancy; and a friendly disposition. Additional features are low birth weight, short stature, microcephaly, long fingers and heart defects. This subject also presented with postaxial polysyndactyly, neonatal cholestasis, resolved leucopenia, dry skin with some hyperpigmented lesions and an anteriorly split tongue. Subject 648 is 9-year-old male child with a clinical history of generalized hypotonia, seizures, autism, intellectual disability, motor developmental delay and dysmorphic features consistent with the 17q21.31 microdeletion syndrome (epicanthal folds; ptosis; long, pear-shaped nose; and long, tapering fingers). We obtained informed consent to publish the photographs.

direct, as opposed to inverted, orientation is a key distinction between active and inactive hotspots (46/54 direct compared to 16/57 inverted in active hotspots; OR = 3.04). We also found that segmental duplications flanking active hotspots are larger and show higher sequence identity compared to inactive hotspots (Kolmogorov-Smirnov test, *P* = 0.0022) (**Supplementary Fig. 5**). Notably, we identified eight regions that showed no evidence of copy number variation in cases or controls despite the presence of large, highly similar and directly oriented segmental duplications at their flanks (**Supplementary Table 10**). These may be regions that are mutationally active but in which dosage imbalance is lethal (for example, 7p14.3, flanked by 19.9-kb duplications and containing *BBS9* and *BMPER*).

In addition to identifying new potentially pathogenic loci, the large number of cases analyzed provided the opportunity to identify atypical deletions (deletions characterized by noncanonical breakpoints and likely not generated by non-allelic homologous recombination mutational mechanism) and refine the critical region of known genomic disorders. For example, we identified three individuals with smaller, atypical deletions within the 17q21.31 microdeletion syndrome region[6,24,25] (**Fig. 3b**). The breakpoints in these cases contrast with those of 23 cases carrying the canonical 480-kb deletion mediated by unequal crossover between directly orientated segmental duplications—a genomic architecture largely restricted to individuals of European descent[26]. Detailed clinical information on two individuals with the atypical deletion (**Fig. 3c**) showed strong phenotypic similarity with the known syndrome, including a pronounced philtrum, epicanthic folds, cupped ears and skeletal defects of the hand (**Supplementary Note** and **Supplementary Table 11**). The strong phenotypic similarity refines the dosage-sensitive region to only three genes (**Fig. 3b**), including *MAPT*, which is disrupted by one of these atypical deletions.

Gene content analysis

Encouraged by the additional refinement provided by atypical deletion events, we performed a gene-based analysis on the complete intellectual disability and/or developmental delay dataset as well as on case subsets partitioned by additional phenotypic data. We identified 615 genes as significantly deleted in any phenotype (Benjamini-Hochberg corrected *P* < 0.05; **Supplementary Table 12**), the vast majority of which associated with known pathogenic loci or subtelomeric alterations. An Ingenuity Pathways Analysis (IPA) (see URLs) showed significant enrichment in expected functional categories (for example, cardiovascular disease and developmental, endocrine system and developmental disorders).

We then expanded our analysis to include candidate associations with nominal significance, as the above analysis is likely to be overly conservative because of the high level of dependence between neighboring genes. An IPA of genes with a nominal *P* < 0.02 identified the same functional categories as above, suggesting that a large proportion of the nominally significant genes are likely relevant to morbidity. In addition to identifying genes within known genomic disorders, this analysis identified genes outside of these intervals. For example, we observed an excess of smaller deletions of *SCN1A* specifically in cases with epilepsy (*P* = 0.019), consistent with the literature[27]. *CD44* deletions on 11p13 were significantly enriched in craniofacial cases (*P* = 0.010) and have previously been linked to cleft lip and palate in SNP and expression

Table 3 Validation of smaller deletions

Chr.	Start position (bp)	Stop position (bp)	Gene	Confirmation	Identical breakpoints
Tier 1					
12	113,316,929	113,317,081	*TBX5*	3 of 4	Ambiguous
1	40,001,351	40,013,297	*BMP8*	6 of 6	Ambiguous
1	233,932,670	233,932,900	*LYST*	6 of 6	Yes
12	12,868,741	12,873,755	*DDX47*	6 of 6	Yes
11[a]	43,729,037	43,732,247	*HSD17B12*	6 of 6	Yes
20	45,205,105	45,205,194	*EAB1*	6 of 6	Yes
13	21,173,329	21,173,574	*FGF9*	4 of 6	Yes
6	162,314,324	162,314,439	*PARK2*	6 of 6	No
9[a,b]	93,525,765	93,527,210	*NTRKR2*	6 of 6	No
1	166,548,570	166,548,864	*TBX19*	6 of 6	Yes
				55 of 58	
Tier 2					
18	148,699	148,714	*USP14*	3 of 4	Yes
2	166,518,441	166,518,461	*TTC21B*	0 of 5	NA
10	26,889,040	26,896,423	*APBB1IP*	2 of 3	No
4	110,114,972	110,115,164	*COL25A1*	4 of 5	Yes
4[a,c]	77,301,890	77,308,653	*SCARB2*	2 of 4	Yes
9	883,912	884,195	*DMRT1*	5 of 5	Yes
12	31,835,960	31,836,367	*H3F3C*	4 of 4	Yes
13	97,907,423	97,907,559	*MST3*	0 of 4	NA
9	86,546,627	86,546,662	*NTRK2*	5 of 5	Yes
				25 of 40	

Chr., chromosome

[a]Exon-altering variants. [b]Five samples harbor a non-exonic copy number polymorphism; one sample has a unique, exon-altering deletion. [c]Overlaps the neighboring gene (*FAM47D*). Note that annotations are based on the UCSC gene model and not RefSeq genes. NA, not applicable.

microarray studies[28,29]. A region on 9p24 containing five genes was significantly enriched in craniofacial cases, with the peak significance focused at *SLC1A1* (peak $P = 0.00172$), which encodes a high affinity glutamate transporter previously implicated in multiple neurological conditions[30]. This peak, specific to *SLC1A1*, was also significantly enriched in neurological, craniofacial and epilepsy cases. A 2q37 deletion immediately proximal to the 2q37 deletion region (**Table 1**) containing 15 genes was enriched primarily in the neurological (modal $P = 0.00479$) and epilepsy (modal $P = 0.00542$) phenotypes and contains genes associated with neurodevelopmental and sleep phase disturbances (*GBX2* and *PER2*)[31,32]. Finally, the deletion of *PARD3* was significantly enriched in autism cases ($P = 0.01023$). *PARD3* has been previously associated with bipolar disease[33] and is involved in both tight junction formation and axonal fate determination[34].

We also identified 325 duplicated genes (**Supplementary Table 12**) significantly enriched among the cases (Benjamini-Hochberg corrected $P < 0.05$). Similar to deletions, nearly all genes enriched among duplications at this stringent threshold were within known pathogenic duplications and were overrepresented (according to IPA) in categories that fit well with the expected phenotypic abnormalities (for example, cardiovascular disease, developmental, endocrine system and developmental disorders). Expanding our analysis to enrichments with nominal significance identified IPA functions identical to the conservative approach as well as several promising candidate gene regions. We observed duplications containing three genes (*SH3YL1*, *ACP1* and *FAM150B*) on chromosome 2p in cases with craniofacial disorders ($P = 0.01032$). Notably, large 2p distal duplications have been associated with facial dysmorphism in multiple case reports[35,36]. Similarly, we observed duplication of two genes (*RSPO4* and *PSMF1*) on distal chromosome 20p in cases with cardiac defects ($P = 0.01195$), and larger duplications of 20p have been associated with cardiac defects[37]. The results suggest a potential role for these small subtelomeric regions in disease. Finally, we observed duplication of proximal 8p extending to include two genes in cases with neurological disorders ($P = 0.00479$), one of which (*FNTA*) has been shown to be more highly expressed in schizophrenia[38].

Discovery of smaller gene-disrupting CNVs

Although the data suggest that as much as 14.2% of developmental delay may be explained by large CNVs, many causal mutations remain to be identified. We sought to determine if previously unreported smaller CNVs could be identified among these cases, assuming that

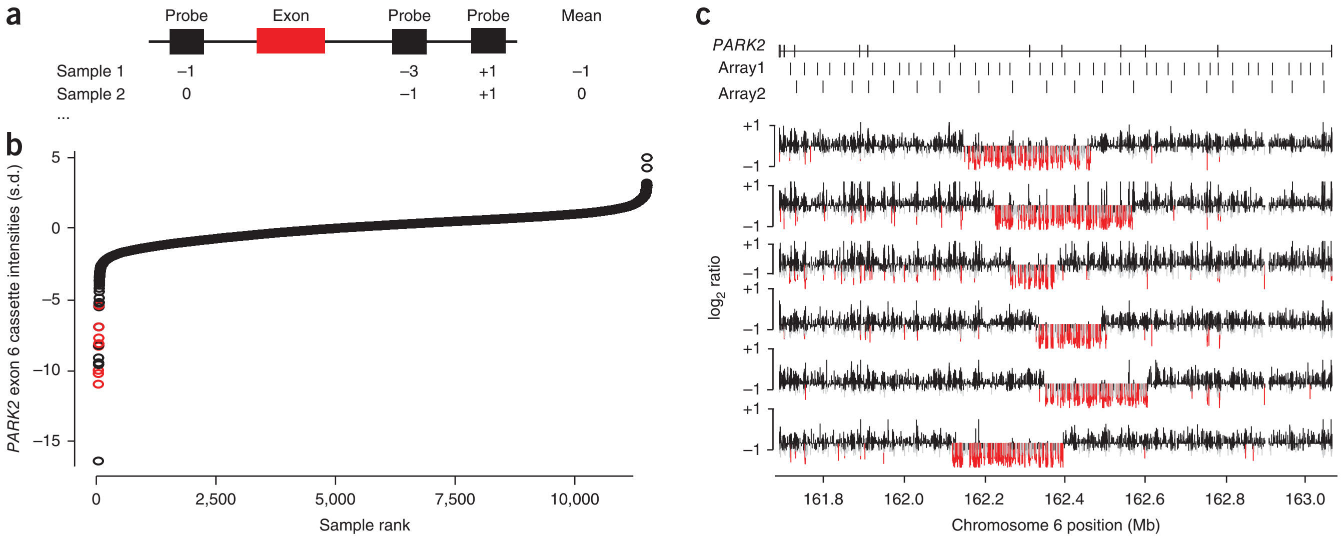

Figure 4 Discovery of new, exon-altering CNVs using the Signature CGH data. (**a**) For each coding exon (red bar), we used the three probes (black rectangles) nearest the exon for any given individual to define a cassette score. (**b**) Distribution of cassette intensities for exon 6 of *PARK2* are sorted from lowest to highest (measured in standard deviation; *y* axis) across all samples (*x* axis). Red open circles correspond to known large deletion events that span the exon. (**c**) Validation results for the most strongly negative samples from **b** not previously known to carry deletions. Log2 ratio values (*y* axis for each individual row) for *PARK2* (coordinates on the *x* axis) in each of six tested samples are shown. Probes with very low intensities (<–0.5) are colored red and those with moderately low values (<–0.3) are shown in gray. The locations of *PARK2* exons and probes on two of the most commonly used original oligonucleotide arrays are shown at the top.

breakpoints would not necessarily be recurrent and that individually relevant events would be rare (<0.1%); such variants may, in principle, identify new candidate genes, refine the molecular basis for the phenotypic consequences of larger CNVs and broaden the predictive power of a given microarray experiment. Therefore, we conducted a directed search for small, exon-affecting CNVs, reasoning that such variants are more likely to have disease relevance and would be amenable to follow up. For each consensus coding sequence (CCDS) exon[39], we determined the average intensity for the three closest probes (termed a 'cassette') in each sample and, in turn, identified cassettes with outlier intensities that may be indicative of deletions (Online Methods and **Supplementary Fig. 6**). Because this strategy is exon centric, it is partially platform and breakpoint independent. We analyzed 186,014 autosomal coding exons using 65,704 cassettes (multiple exons are often targeted by the same cassette) excluding exons within known common CNVs[16,40,41]. After a series of data normalization and quality-control steps, we identified 829 cassettes in which a small (10–100) set of samples had probe intensities that clustered well below the population-wide mean. Each of these cassettes was manually reviewed to eliminate artifacts and select for genes with greater potential for disease involvement; we selected 19 of these genes for follow up and organized them into two subjectively defined tiers of quality (**Table 3**).

Among the first tier of predicted deletions, we found that 55 of 58 individual (sample-level) predictions validated, with at least one validated event for all ten examined genes, and for the second tier, we found that 25 of 40 predictions validated across seven of the nine examined genes. A total of 44 of the validated deletions spanned only a single probe on the originally used array (**Supplementary Fig. 7**). We determined deletion events at three genes to be polymorphisms[42–44]. Notably, we found *PARK2* to contain at least six distinct exon-affecting deletions ranging in size from 118–315 kb (**Fig. 4**, **Supplementary Fig. 8** and **Supplementary Note**). However, there is no evidence for CNV enrichment at this locus among cases, as this phenomenon also holds true for control samples (**Supplementary Fig. 9**), suggesting that *PARK2* is a fragile gene prone to recurrent deletion events. We also identified small deletions in *TBX5*, a gene known to cause Holt-Oram syndrome[45] (a disorder characterized by upper limb abnormalities and congenital heart defects; MIM#142900). We found that 7 of 15 samples predicted to harbor a *TBX5* event were fetal samples, a rate significantly greater than the background proportion of fetal samples (13.4%, $P = 0.0019$), consistent with the observations that *TBX5* mutations can result in prenatal abnormalities detectable by ultrasound[46].

DISCUSSION

We present one of the largest studies investigating the role of rare CNVs in intellectual disability and developmental delay, analyzing data from 15,767 affected individuals and 8,329 controls. These data quantify the massive contribution of large CNVs to pediatric disease, with 25.7% of affected individuals harboring CNVs >400 kb compared with only 11.5% of controls. Disease risk increases steadily in relation to CNV size, with an OR > 20 for carriers of CNVs larger than 1.5 Mb and an OR of nearly 50 at a threshold of 3 Mb. We find that the CNV burden differs significantly depending on the nature of the primary clinical referral, with craniofacial abnormalities and structural defects of the heart being especially enriched for large CNVs relative to epilepsy and autism spectrum disorder (**Fig. 1** and **Supplementary Fig. 2**). As has been observed in model organisms and predicted based on theory[47,48], haploinsufficiency appears more common and penetrant than triplosensitivity for severe developmental phenotypes. Although this cohort does not represent a random sampling of individuals with intellectual disability and/or developmental delay and includes some individuals without these phenotypes, our estimates are likely applicable to intellectual disability and/or developmental delay in general. For example, in a literature survey[49], the average CNV burden across 15 genome-wide studies of intellectual disability and/or developmental delay (with a combined sample size of 1,021) was estimated to be ~13.7%, which is similar to our estimate of 14.2% (note that this estimate was derived by averaging the diagnostic yields for all studies with a genome-wide resolution of 1 Mb or better as indicated in **Table 2** of Miller *et al.*[49]). Furthermore, the observed enrichment for many loci known to contribute to intellectual disability and/or developmental delay risk (**Table 1**) and individual genes previously identified to be disrupted among affected individuals (**Supplementary Table 12**) clearly supports the applicability of the inferences generated here for both intellectual disability and/or developmental delay specifically and for neurological disease (for example, schizophrenia and autism) in general.

Practically, these data serve as a clinical resource that will be useful in diagnostics (**Tables 1** and **2**). The large number of controls and cases used here provides estimates of penetrance for 59 pathogenic CNVs (accounting for ~10% of cases) and sheds light on either ambiguous or previously unknown pathogenic variants, including 14 new or previously marginally supported CNV loci that collectively represent ~0.7% (112/15,767; **Table 2** and **Supplementary Note**) of the individuals studied here. We note that although one CNV locus (10p15.3 duplications) appeared to be enriched among cases as a result of ancestry differences between the cases and controls, the aggregate ethnic composition of the 14 loci in **Table 2** closely matched our control dataset (**Supplementary Note** and **Supplementary Figs. 10** and **11**), suggesting that population stratification for rare variants is unlikely to explain the enrichment at these loci. The size distribution (median of 940 kb), inheritance rate (15 of 34 tested CNVs are *de novo*, with at least one *de novo* variant observed in 6 of the 14 loci) and overlap with DECIPHER entries further support a role for these CNV loci in disease risk.

Among these new potentially pathogenic CNVs, we provide additional support for a genomic disorder mapping to 15q25.2, which we found in five affected individuals (including two affected siblings) and zero controls (**Supplementary Fig. 12**). Our results, combined with earlier studies of schizophrenia and autism (four cases compared to zero controls)[16], implicate this CNV as a high-risk allele for pediatric neurological disease with variable outcomes (**Supplementary Note** and **Supplementary Table 9**) as well as neuropsychiatric disease ($P = 0.037$). In addition, our data support the pathogenicity of CNVs at 2q13 whose significance was uncertain because they were observed in a small number of control samples[50]. In our study, we observed 12 deletions ($P = 0.032$) and 9 duplications ($P = 0.022$) on chromosome 2q13 in cases but only 1 deletion in controls. We furthermore find an enrichment of the deletion in cardiovascular cases (peak $P = 0.012$) and the duplication in cases with craniofacial features (peak $P = 0.010$). These results are consistent with two previously reported deletion cases with multiple heart defects and two duplication cases with various facial and skeletal features[50]. Additionally, our data support the pathogenicity of duplications at 16p11.2 (*SH2B1*), duplications at 15q13.3 segmental duplication breakpoints BP3–BP5 (*CHRNA7*), and deletions at 15q11.2 involving segmental duplication breakpoints BP1–BP2 (*NIPA1*). The latter are present in ~1 in 167 affected individuals studied here and, although incompletely penetrant (0.83), are likely strong risk factors for developmental delay in addition to schizophrenia[4,51].

Finally, the discovery of atypical and smaller deletions among cases with virtually identical phenotypes helps to refine the smallest region of overlap for known syndromes. The atypical deletions of 17q21.31 exclude deletions of *CRHR1* as playing a role in this syndrome (although deletions of long-range regulatory elements that change *CRHR1* expression cannot be ruled out) and narrow the likely candidates to three genes, including *MAPT*, which was disrupted by proximal breakpoints in two cases (**Fig. 3b**). Overall, we identified 615 deleted genes and 325 duplicated genes significantly enriched in cases when compared to controls. The dosage imbalance of these genes should not be considered as proven but, rather, these genes should be considered as candidates with higher prior probability of dosage sensitivity for future studies. It is encouraging that this set includes a number of previously hypothesized and new associations between genes and particular traits (**Supplementary Table 12**). In addition, our data show that even older, low-resolution microarray data afford discovery opportunities for CNVs that have not previously been detectable. Indeed, we successfully identified and confirmed dozens of small deletion events, several of which have plausible disease roles (for example, *TBX5* deletions in Holt-Oram syndrome), including many detected by only a single probe in the original microarray experiment. As the underlying raw data from diagnostic laboratories is released, prospectively, there will be great potential for finding additional exon-altering deletions. Further validation of these and other candidates will yield new insights into the specific phenotypes affected by the loss or gain of individual genes. Although most arrays cannot robustly capture the small deletions we identified, such as those adjacent to exons of *FGF9* and *LYST* (associated with Chediak-Higashi syndrome), control screening using PCR or other targeted high-throughput assays may be used to follow up individually interesting candidates (**Supplementary Note**).

We predict that this map of CNVs and potentially dosage-sensitive genes will be invaluable for both clinical and research purposes in the future. For example, researchers in a previous study[52] used an exon-targeted microarray to identify a number of individual gene disruptions in individuals with intellectual disability and/or developmental delay that were of plausible but uncertain pathogenicity given their rarity. We find support for a number of these genes, including two (*CREBBP* and *SLC1A1*) that are significantly enriched among individuals here with similar phenotypes to those previously described (**Supplementary Note**). As genomic discovery efforts (especially exome sequencing) expand, the results described here should prove increasingly important to clinicians and researchers faced with the challenges of linking rare disruptive mutations to pediatric diseases.

URLs. Database of Genomic Variants, http://projects.tcag.ca/variation/; Ingenuity Pathway Analysis, http://www.ingenuity.com/; InCHIANTI, http://www.inchiantistudy.net/; UCSC LiftOver tool, http://genome.ucsc.edu/.

METHODS

Methods and any associated references are available in the online version of the paper at http://www.nature.com/naturegenetics/.

Accession codes. All CNV calls have been submitted to dbVar under accession nstd54.

Note: Supplementary information is available on the Nature Genetics website.

ACKNOWLEDGMENTS

We thank N. Krumm, M. Malig, L. Vives and J. Luu for assistance in validation experiments. We also thank M. Dennis, C. Alkan, E. Karakoc 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/. Funding for the project was provided by the Wellcome Trust under awards 076113 and 085475. We also thank A. Aragaki, C. Kooperberg and R. Jackson for access to SNP data (Fred Hutchinson Cancer Research Center (FHCRC) control dataset) generated as part of the ongoing genome-wide association study to identify genetic components of hip fracture in the Women's Health Initiative. This work was supported by US National Institutes of Health HD065285 to E.E.E. E.E.E. is an investigator of the Howard Hughes Medical Institute.

AUTHOR CONTRIBUTIONS

G.M.C., B.P.C., S.G., E.E.E., J.A.R., B.C.B. and L.G.S. designed the study. L.G.S. supervised array-CGH experiments at Signature Genomics. J.A.R. and B.C.B. coordinated clinical data collection. G.M.C. and B.P.C. performed data analysis and curated control CNV data. S.G. curated genomic disorders data. S.G., T.H.V. and C.B. performed array CGH and PCR validations. C.W., H.S., R.H., V.H., H.A.-H., P.B., E.M., D.N., K.L., H.T., M.H., N.A., J.G., J.K., V.S., K.J. and C.R. provided clinical information. G.M.C., B.P.C., S.G. 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 accompany the full-text HTML version of the paper at http://www.nature.com/naturegenetics/.

Published online at http://www.nature.com/naturegenetics/.
Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. Greenway, S.C. *et al. De novo* copy number variants identify new genes and loci in isolated sporadic tetralogy of Fallot. *Nat. Genet.* **41**, 931–935 (2009).
2. Mefford, H.C. *et al.* Recurrent reciprocal genomic rearrangements of 17q12 are associated with renal disease, diabetes, and epilepsy. *Am. J. Hum. Genet.* **81**, 1057–1069 (2007).
3. Sebat, J. *et al.* Strong association of *de novo* copy number mutations with autism. *Science* **316**, 445–449 (2007).
4. Stefansson, H. *et al.* Large recurrent microdeletions associated with schizophrenia. *Nature* **455**, 232–236 (2008).
5. Walsh, T. *et al.* Rare structural variants disrupt multiple genes in neurodevelopmental pathways in schizophrenia. *Science* **320**, 539–543 (2008).
6. Sharp, A.J. *et al.* Discovery of previously unidentified genomic disorders from the duplication architecture of the human genome. *Nat. Genet.* **38**, 1038–1042 (2006).
7. Gu, W., Zhang, F. & Lupski, J.R. Mechanisms for human genomic rearrangements. *Pathogenetics* **1**, 4 (2008).
8. Girirajan, S. *et al.* A recurrent 16p12.1 microdeletion supports a two-hit model for severe developmental delay. *Nat. Genet.* **42**, 203–209 (2010).
9. Mefford, H.C. *et al.* Recurrent rearrangements of chromosome 1q21.1 and variable pediatric phenotypes. *N. Engl. J. Med.* **359**, 1685–1699 (2008).
10. van Bon, B.W. *et al.* Further delineation of the 15q13 microdeletion and duplication syndromes: a clinical spectrum varying from non-pathogenic to a severe outcome. *J. Med. Genet.* **46**, 511–523 (2009).
11. Shprintzen, R.J. Velocardiofacial syndrome and DiGeorge sequence. *J. Med. Genet.* **31**, 423–424 (1994).
12. Karayiorgou, M. *et al.* Schizophrenia susceptibility associated with interstitial deletions of chromosome 22q11. *Proc. Natl. Acad. Sci. USA* **92**, 7612–7616 (1995).
13. Coe, B.P. *et al.* Resolving the resolution of array CGH. *Genomics* **89**, 647–653 (2007).
14. Cooper, G.M., Zerr, T., Kidd, J.M., Eichler, E.E. & Nickerson, D.A. Systematic assessment of copy number variant detection via genome-wide SNP genotyping. *Nat. Genet.* **40**, 1199–1203 (2008).
15. Itsara, A. *et al. De novo* rates and selection of large copy number variation. *Genome Res.* **20**, 1469–1481 (2010).
16. Itsara, A. *et al.* Population analysis of large copy number variants and hotspots of human genetic disease. *Am. J. Hum. Genet.* **84**, 148–161 (2009).
17. de Vries, B.B. *et al.* Diagnostic genome profiling in mental retardation. *Am. J. Hum. Genet.* **77**, 606–616 (2005).
18. Sharp, A.J., Cheng, Z. & Eichler, E.E. Structural variation of the human genome. *Annu. Rev. Genomics Hum. Genet.* **7**, 407–442 (2006).
19. Firth, H.V. *et al.* DECIPHER: Database of Chromosomal Imbalance and Phenotype in Humans Using Ensembl Resources. *Am. J. Hum. Genet.* **84**, 524–533 (2009).
20. Mefford, H.C. *et al.* A method for rapid, targeted CNV genotyping identifies rare variants associated with neurocognitive disease. *Genome Res.* **19**, 1579–1585 (2009).
21. Walters, R.G. *et al.* A new highly penetrant form of obesity due to deletions on chromosome 16p11.2. *Nature* **463**, 671–675 (2010).

22. Bochukova, E.G. *et al.* Large, rare chromosomal deletions associated with severe early-onset obesity. *Nature* **463**, 666–670 (2010).
23. Helbig, I. *et al.* 15q13.3 microdeletions increase risk of idiopathic generalized epilepsy. *Nat. Genet.* **41**, 160–162 (2009).
24. Koolen, D.A. *et al.* A new chromosome 17q21.31 microdeletion syndrome associated with a common inversion polymorphism. *Nat. Genet.* **38**, 999–1001 (2006).
25. Shaw-Smith, C. *et al.* Microdeletion encompassing *MAPT* at chromosome 17q21.3 is associated with developmental delay and learning disability. *Nat. Genet.* **38**, 1032–1037 (2006).
26. Zody, M.C. *et al.* Evolutionary toggling of the *MAPT* 17q21.31 inversion region. *Nat. Genet.* **40**, 1076–1083 (2008).
27. Suls, A. *et al.* Microdeletions involving the *SCN1A* gene may be common in *SCN1A*-mutation–negative SMEI patients. *Hum. Mutat.* **27**, 914–920 (2006).
28. Baroni, T. *et al.* Human cleft lip and palate fibroblasts and normal nicotine-treated fibroblasts show altered *in vitro* expressions of genes related to molecular signaling pathways and extracellular matrix metabolism. *J. Cell. Physiol.* **222**, 748–756 (2010).
29. Park, J.W. *et al.* High throughput SNP and expression analyses of candidate genes for non-syndromic oral clefts. *J. Med. Genet.* **43**, 598–608 (2006).
30. McCullumsmith, R.E. & Meador-Woodruff, J.H. Striatal excitatory amino acid transporter transcript expression in schizophrenia, bipolar disorder, and major depressive disorder. *Neuropsychopharmacology* **26**, 368–375 (2002).
31. Chen, L., Chatterjee, M. & Li, J.Y. The mouse homeobox gene *Gbx2* is required for the development of cholinergic interneurons in the striatum. *J. Neurosci.* **30**, 14824–14834 (2010).
32. Toh, K.L. *et al.* An hPer2 phosphorylation site mutation in familial advanced sleep phase syndrome. *Science* **291**, 1040–1043 (2001).
33. Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**, 661–678 (2007).
34. Brajenovic, M., Joberty, G., Kuster, B., Bouwmeester, T. & Drewes, G. Comprehensive proteomic analysis of human Par protein complexes reveals an interconnected protein network. *J. Biol. Chem.* **279**, 12804–12811 (2004).
35. Stalker, D.J., Vigneswaren, S., Sharples, P.M. & Lunt, P.W. Distal trisomy 2p and arachnodactyly. *J. Med. Genet.* **37**, 974–976 (2000).
36. Li, F., Batista, D.A., Maumenee, I. & Wang, T. An unbalanced translocation between chromosomes 2p and 6p associated with Axenfeld-Rieger anomaly type 3, hearing loss, developmental delay, and distinct facial dysmorphism. *Am. J. Med. Genet. A.* **152A**, 1318–1321 (2010).
37. Chaabouni, M. *et al. De novo* trisomy 20p of paternal origin. *Am. J. Med. Genet. A.* **143A**, 1100–1103 (2007).
38. Bowden, N.A., Scott, R.J. & Tooney, P.A. Altered gene expression in the superior temporal gyrus in schizophrenia. *BMC Genomics* **9**, 199 (2008).
39. Pruitt, K.D. *et al.* The consensus coding sequence (CCDS) project: identifying a common protein-coding gene set for the human and mouse genomes. *Genome Res.* **19**, 1316–1323 (2009).
40. Conrad, D.F., Andrews, T.D., Carter, N.P., Hurles, M.E. & Pritchard, J.K. A high-resolution survey of deletion polymorphism in the human genome. *Nat. Genet.* **38**, 75–81 (2006).
41. McCarroll, S.A. *et al.* Common deletion polymorphisms in the human genome. *Nat. Genet.* **38**, 86–92 (2006).
42. McCarroll, S.A. *et al.* Integrated detection and population-genetic analysis of SNPs and copy number variation. *Nat. Genet.* **40**, 1166–1174 (2008).
43. Conrad, D.F. *et al.* Origins and functional impact of copy number variation in the human genome. *Nature* **464**, 704–712 (2010).
44. Kidd, J.M. *et al.* Mapping and sequencing of structural variation from eight human genomes. *Nature* **453**, 56–64 (2008).
45. Basson, C.T. *et al.* Mutations in human *TBX5* cause limb and cardiac malformation in Holt-Oram syndrome. *Nat. Genet.* **15**, 30–35 (1997).
46. Brons, J.T. *et al.* Prenatal ultrasound diagnosis of the Holt-Oram syndrome. *Prenat. Diagn.* **8**, 175–181 (1988).
47. Turner, D.J. *et al.* Germline rates of *de novo* meiotic deletions and duplications causing several genomic disorders. *Nat. Genet.* **40**, 90–95 (2008).
48. Fisher, E. & Scambler, P. Human haploinsufficiency—one for sorrow, two for joy. *Nat. Genet.* **7**, 5–7 (1994).
49. Miller, D.T. *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.* **86**, 749–764 (2010).
50. Rudd, M.K. *et al.* Segmental duplications mediate novel, clinically relevant chromosome rearrangements. *Hum. Mol. Genet.* **18**, 2957–2962 (2009).
51. International Schizophrenia Consortium. Rare chromosomal deletions and duplications increase risk of schizophrenia. *Nature* **455**, 237–241 (2008).
52. Boone, P.M. *et al.* Detection of clinically relevant exonic copy-number changes by array CGH. *Hum. Mutat.* **31**, 1326–1342 (2010).

ONLINE METHODS

Cases. Samples from individuals with intellectual disability and/or developmental delay and related phenotypes were submitted to Signature Genomic Laboratories, LLC, mostly from the United States and Canada, for clinical microarray-based CGH. A total of 15,767 samples were analyzed, and 16,526 rare autosomal CNV calls were detected (**Supplementary Table 1**) and deposited into dbVar (dbVar study accession nstd54)[53]. 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 Genomic Laboratories, LLC, institutional review board. Although not a random set of children with intellectual disability and/or developmental delay, the presentations in our set are representative of those observed in a clinical diagnostic setting. The majority of the individuals in our set have intellectual disability and/or developmental delay; however, clinical features such as craniofacial and skeletal features, growth retardation, cardiovascular and renal defects, hypotonia, speech and motor deficits, hearing impairment, epilepsy and behavioral problems were also documented. We identified 575 cases with cardiovascular defects, 1,776 cases with epilepsy and/or seizure disorder, 1,379 cases with autism spectrum disorder, 3,898 cases with craniofacial defects and 8,772 cases with general neurological defects; many individuals had multiple subclassifications (**Supplementary Table 2**). Self-reported ethnicity was available for 144 individuals, with 75% (108/144; 95% CI 67.3–81.4%) reporting Caucasian (primarily European descent), 6.9% (10/144; 95% CI 3.8–12.3%) reporting African American and 18.1% (26/144; 95% CI 12.6–25.1%) reporting other ethnicity. These samples were analyzed across nine custom array-CGH platforms, with most being tested on an Agilent array with ~97,000 probes (**Supplementary Fig. 13**).

Controls. Controls were not ascertained specifically for neurological disorders, but all controls were obtained from adult samples providing informed consent, so developmental disorders should be exceedingly rare in this group. Of individuals with known ethnicity, 81.2% are Caucasian (primarily European descent), 2% are African or African American and 16.5% are other or mixed ancestry. Because of the slight enrichment of African-American cases compared to our control samples, we modeled the potential impact of large CNV stratification and found no evidence for an overall enrichment of unique large CNVs in the African-ancestry cohort (**Supplementary Fig. 10**). DNA was obtained from cell lines and blood-derived samples generated for association studies of various phenotypes. The data sets are detailed in **Supplementary Table 4**. Data were obtained from the following sources: HGDP[16,54]; NINDS (dbGaP accession number phs000089)[16,55], PARC/PARC2[56,57]; London (parents of asthmatic children)[15]; FHCRC (pre-release data provided courtesy of A. Aragaki, C. Kooperberg and R. Jackson as part of an ongoing genome-wide association study to identify genetic components of hip fracture in the Women's Health Initiative); InCHIANTI (data provided by InCHIANTI study of aging; see URLs)[15,58]; and Wellcome Trust Case Control Consortium 2, National Blood Services Cohort (WTCCC2 NBS)[59]. Control CNV arrays were analyzed as described previously[16]. Briefly, a Hidden Markov Model (HMM) based on both allele frequencies and total intensity values (logR) was used to identify putative alterations, followed by manual inspection of large CNVs (>100 probes and >1 Mb) in conjunction with user guided merging of nearby (<1 Mb between for arrays with <1 million probes and <200 kb for arrays with >1 million probes) calls, which represent a single region broken up by the HMM, or gaps. All samples on arrays with densities <1 million probes were filtered by a maximal genome-wide logR standard deviation (s.d.) of 0.25, whereas the high-density 1.2 million probe WTCCC2 data was filtered using an increased s.d. cutoff of 0.37. Large alterations with non-canonical allele frequencies indicative of mosaics were excluded because of the high likelihood of these resulting from cell culture immortalization. For the two datasets where the Illumina array mapping corresponded to build35 (National Human Genome Research Institute), we used the autosomal calls generated previously[16] and mapped the coordinates to build36 using the UCSC LiftOver tool (see URLs).

Multi-platform CNV comparison. Microarray platform heterogeneity may yield false CNV enrichments signals as a function of differential detection power related to probe density, data quality, analysis methods, or other factors. We made a number of efforts to control for such potential effects, and we believe our study design is robust to this source of error for a number of reasons. First, the control data for this study were generated on higher resolution platforms (317,000–1,200,000-probe Illumina SNP arrays, with 88% of controls being profiled on 550,000-probe or higher density platforms) compared to the case data (where the median array is ~97,000 probes and the highest density is ~130,000 probes). As a result, our CNV detection power is substantially higher for cases than controls; notably, such differences will tend to manifest as false positive enrichments for CNVs in controls whereas we are focused exclusively on enrichments within cases. Second, we rigorously eliminated potential sources of errors in the case CNV data with a combination of both manual and automated filters, including calls with low probe counts, high degrees of overlap with segmental duplications in the reference assembly and likely reference-sample CNVs. Third, for the sliding-window enrichment tests, we eliminated all CNVs in cases that spanned fewer than ten probes on the lowest resolution (HH317K) control SNP array. Fourth, we validated 402 of 425 CNVs and determined the precision in cases to be high in general (0.945) and higher in cases relative to controls (0.892). Fifth, we specifically analyzed the 14 potentially pathogenic CNVs (**Table 2**) for control SNP microarray performance. Eleven of 14 loci harbored small CNV calls within the region of interest from multiple control studies; as CNV calling algorithms tend to show increased sensitivity to larger alterations, we consider this to indicate sufficient control sensitivity within these loci to detect large CNVs. The remaining three loci are split between the minimal common region on 1q24.3, which shows a single 72-kb CNV in controls (again suggesting detectability of larger events), and two loci that harbor very small CNVs detectable only on the highest resolution 1.2 million probe arrays. These two regions have high probe coverage on the 550,000 control array (46 probes within the smallest 6p22.3 Signature call and 40 probes in the minimal common region of 2q24.3). Further, all of these regions have *de novo* CNVs in our samples, supporting the hypothesis that these are pathogenic loci and not simply common CNVs that we failed to detect with SNP platforms.

Control CNV burden. Control CNVs were merged into copy number variant regions (CNVRs) by comparing each CNV to all of its overlapping partners and merging those with 50% reciprocal overlap. These CNVRs were then analyzed in the context of sliding 300-kb genomic windows to identify regions of high variability (**Supplementary Fig. 9** and **Supplementary Table 13**). Regions of high SNP diversity were obtained from a previous study[44] and used to identify regions where the breakpoint variability is likely to result from general sequence variation (such as the *HLA* locus on 6p). To perform a gene-based search for highly variable loci, we first generated a merged RefSeq list that combined overlapping splice variants into a single, large gene definition. We then analyzed these loci in the context of overlapping gain and loss CNVs that contained the entire gene, overlapped the transcript (gene-breaking or exon hits) or were contained within an intron. Finally, we analyzed each gene in the context of the number of unique CNVRs that overlapped the gene space (exonic or intronic).

Discovery of new exon-altering CNVs. For a subset of 11,529 samples, we identified for each coding exon[39] the three closest probes, requiring at least one probe on both sides within 100 kb of the exon. We required that all probes map within 200 kb, yielding 65,704 unique cassettes targeting 186,014 autosomal coding exons. We then determined the average cassette intensity for each sample and normalized it by array type. Subsequently, we considered filtered cassettes by the following criteria: 10–100 samples with scores at least 5 s.d. below average; the subset of samples at less than 5 s.d. below average compose at least 10% of samples with scores less than 3 s.d. below average (a measure of cluster separation); and no overlap of the target exon (note that individual probes were not filtered given the heterogeneity of platforms and the potential for atypical CNVs) with common copy number polymorphisms or deletions seen in multiple control individuals[16,42,43,60]. This yielded 829 candidates for follow up, each of which was manually reviewed to eliminate cassettes in which all candidate deletions clustered within a single array type suggestive of a batch artifact and noisy cassettes resulting from probes embedded within segmental duplications (for examples, see **Supplementary Fig. 6**). Subsequently, 19 cassettes were chosen for validation, manually divided into two qualitative tiers based on the totality of the evidence (for example, follow-up potential of the affected gene, visual analysis of probe intensity distributions, and so on).

doi:10.1038/ng.909

We designed a custom NimbleGen oligonucleotide array spanning each of the 19 genes and their flanks at very high density (**Supplementary Note**) and performed CGH on 98 samples chosen by cassette score and availability and predicted to carry a deletion at 1 of the 19 genes.

53. Sayers, E.W. *et al.* Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* **39**, D38–D51 (2011).
54. Li, J.Z. *et al.* Worldwide human relationships inferred from genome-wide patterns of variation. *Science* **319**, 1100–1104 (2008).
55. Simon-Sanchez, J. *et al.* Genome-wide SNP assay reveals structural genomic variation, extended homozygosity and cell-line induced alterations in normal individuals. *Hum. Mol. Genet.* **16**, 1–14 (2007).
56. Albert, M.A., Danielson, E., Rifai, N. & Ridker, P.M. Effect of statin therapy on C-reactive protein levels: the pravastatin inflammation/CRP evaluation (PRINCE): a randomized trial and cohort study. *J. Am. Med. Assoc.* **286**, 64–70 (2001).
57. Simon, J.A. *et al.* Phenotypic predictors of response to simvastatin therapy among African-Americans and Caucasians: the Cholesterol and Pharmacogenetics (CAP) Study. *Am. J. Cardiol.* **97**, 843–850 (2006).
58. Melzer, D. *et al.* A genome-wide association study identifies protein quantitative trait loci (pQTLs). *PLoS Genet.* **4**, e1000072 (2008).
59. Wellcome Trust Case Control Consortium. Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls. *Nature* **464**, 713–720 (2010).
60. Redon, R. *et al.* Global variation in copy number in the human genome. *Nature* **444**, 444–454 (2006).

doi:10.1038/ng.909