Prioritization of Copy Number Variation Loci Associated with Autism from AutDB–An Integrative Multi-Study Genetic Database

Idan Menashe[1,2]*, Eric C. Larsen[1], Sharmila Banerjee-Basu[1]*

[1] MindSpec, McLean, Virginia, United States of America, [2] Department of Public Health, Faculty of Health Sciences, Ben Gurion University of the Negev, Beer-Sheva, Israel

Abstract

Copy number variants (CNVs) are thought to play an important role in the predisposition to autism spectrum disorder (ASD). However, their relatively low frequency and widespread genomic distribution complicates their accurate characterization and utilization for clinical genetics purposes. Here we present a comprehensive analysis of multi-study, genome-wide CNV data from AutDB (http://mindspec.org/autdb.html), a genetic database that accommodates detailed annotations of published scientific reports of CNVs identified in ASD individuals. Overall, we evaluated 4,926 CNVs in 2,373 ASD subjects from 48 scientific reports, encompassing ~2.12×10^9 bp of genomic data. Remarkable variation was seen in CNV size, with duplications being significantly larger than deletions, ($P = 3\times10^{-105}$; Wilcoxon rank sum test). Examination of the CNV burden across the genome revealed 11 loci with a significant excess of CNVs among ASD subjects ($P<7\times10^{-7}$). Altogether, these loci covered 15,610 kb of the genome and contained 166 genes. Remarkable variation was seen both in locus size (20 - 4950 kb), and gene content, with seven multigenic (≥3 genes) and four monogenic loci. CNV data from control populations was used to further refine the boundaries of these ASD susceptibility loci. Interestingly, our analysis indicates that 15q11.2-13.3, a genomic region prone to chromosomal rearrangements of various sizes, contains three distinct ASD susceptibility CNV loci that vary in their genomic boundaries, CNV types, inheritance patterns, and overlap with CNVs from control populations. In summary, our analysis of AutDB CNV data provides valuable insights into the genomic characteristics of ASD susceptibility CNV loci and could therefore be utilized in various clinical settings and facilitate future genetic research of this disorder.

Citation: Menashe I, Larsen EC, Banerjee-Basu S (2013) Prioritization of Copy Number Variation Loci Associated with Autism from AutDB–An Integrative Multi-Study Genetic Database. PLoS ONE 8(6): e66707. doi:10.1371/journal.pone.0066707

Editor: Yi-Hsiang Hsu, Harvard Medical School, United States of America

Received December 29, 2012; **Accepted** May 13, 2013; **Published** June 18, 2013

Funding: AutDB is funded by the Simons Foundation, which licenses it as SFARI Gene (https://gene.sfari.org/autdb/Welcome.do). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: menashe.idan@gmail.com (IM); sharmila@mindspec.org (SB)

Introduction

Copy number variations (CNVs) are structural chromosomal aberrations, which are giving rise to gains or losses of certain genomic loci across the human genome [1,2]. While most CNVs have no apparent phenotypic consequences, there is increasing evidence that a number of chromosomal micro deletions or duplications at specific locations are involved in the predisposition of various human diseases [3,4]. Recent advances in high-resolution, high-throughput genomics technologies have facilitated the detection of CNVs in large-scale genetic studies. Moreover, the continuous drop in labor and cost associated with these technologies promote their inclusion in genetic screening for pre- and post- pregnancy tests.

Autism spectrum disorder (ASD) constitutes a collection of clinically heterogeneous disorders that are characterized by impairments in social interactions, deficits in language and communication, and increased repetitive or stereotypic movements [5,6]. ASD is highly heritable with estimates ranging between 40-90% heritability [7,8]. However, given the genetically heterogeneous nature of ASD, the underlying genetic mechanisms of these disorders remain vague. Recent genetic studies have indicated that rare CNVs may play an important role in ASD susceptibility [9–14], and today they are considered one of the common genetic contributors of ASD [15]. Given the evidence that CNVs are a significant genetic risk factor not only for ASD, but for other developmental deficits and congenital anomalies, it has recently been proposed that chromosomal microarray (CMA) screening replace conventional cytogenetic techniques as a first-tier clinical diagnostic test for individuals with these disorders [16,17]. However, the relatively low frequency and widespread genomic distribution of these variants in ASD cases complicates the clinical utilization of CMA screening as a potential diagnostic tool.

Given the important role of CNVs in ASD genetics and the increasing usage of CMA screening for the genetic evaluation of ASD individuals, there is a tremendous need to consolidate the large amounts of CNV data that were generated from ASD subjects and subsequently prioritize the most consistent genomic loci associated with ASD susceptibility. To this end, we explored the CNV data available at the online autism genetic database AutDB (http://www.mindspec.org/autdb.html) which, to the best of our knowledge, is the most comprehensive online resource of curated genetic data of ASD from published scientific reports. Using a range of statistical and bioinformatics analyses we performed a comprehensive and rigorous assessment of CNVs that were observed in ASD cohorts across multiple published reports consisting of both large-scale whole-genome studies and smaller-scale case studies. We subsequently determined the genomic boundaries and genetic characteristics of 11 loci demonstrating significant CNV burden among ASD subjects.

Materials and Methods

CNV data

For this study, we used CNV data available at the CNV module of the AutDB database (data freeze of October 2011)[18]. As with the other modules of AutDB, content of the CNV module originates entirely from published, peer-reviewed scientific literature and is rigorously annotated by scientists. Preliminary screening of reports for inclusion in the database resulted from a search of the scientific literature using PubMed (http://www.ncbi.nlm.nih.gov/pubmed/) with the following keywords: "autism/autistic/ASD" and "copy number/CNV/deletion/duplication/chromosome/structural variant". Furthermore, CNV reports listed in ASD review articles that were not identified in the initial PubMed search were included for consideration. Next, the initial candidate CNV reports were filtered to remove those reports that did not contain at least one ASD individual in which one or more CNVS were identified. This restriction has since been relaxed to include reports describing patients with other neurodevelopmental or neuropsychiatric disorders, such as mental retardation/intellectual disability and developmental delay; in some cases, these patients also display ASD traits, but no formal diagnosis of ASD. Detailed information on ASD and control subjects from the studied cohorts was extracted from each selected CNV report for inclusion in the CNV module database.

For the purpose of CNV prioritization we aimed at analyzing a homogeneous subset of the CNV data set from AutDB by using several filtering criteria (number of CNVs removed from the data are in square brackets): only ASD cases were used [12,416 CNVs]; patients with a disease other than ASD, such as schizophrenia, developmental delay, etc. were excluded [74 CNVs]; a genome-wide CNV discovery method (array-CGH, SNP array, or solid phase hybridization) was required for inclusion; CNVs discovered by a targeted discovery method such as FISH or qPCR were excluded [215 CNVs]; we manually screened individuals to ensure minimal overlap of patients; in some cases a given CNV report may include individuals that had previously been described in another report [104 CNVs]. In such cases, we used patient ID information included in the scientific report to identify duplicate patient entries in the CNV module dataset and subsequently use the CNV data for a given individual that contained the largest number of CNV loci [104 CNVs]. Furthermore, we restricted our analysis to CNVs with defined start and end points [218 CNVs]. Finally, to maintain a uniform CNV/individual ratio in our data, we used CNVs identified in a control cohort of unaffected matched siblings as a filter to generate a set of "case-specific" CNVs in the accompanying case cohort of the Sanders *et. al* study [14] [13,036 CNVs]. This filtering process resulted in a homogeneous data set of 4,926 CNVs in 2,373 ASD subjects gathered from 48 scientific reports.

CNV burden

To search for genomic loci demonstrating excess of CNVs among ASD subjects, we divided the genome into consecutively distributed regions of 10 kb, and evaluated the CNV burden in each region as follows:

$$CNV\ burden\ (\theta) = \sum_{i=1}^{N} C_{i,j} \times W_j \qquad (1)$$

where $C_{i,j} = 1$ if a CNV (i) from a particular study (j) of all CNVs (N) in the data, overlaps with the 10 kb genomic region, and W_j is the weight associated with the study (j) calculated as the Loci/CNVs ratio in the study.

To assess the statistical significance of θ, we first randomly distributed the CNVs in our dataset in the human genome and then calculated their corresponding 10 kb θs. We repeated this procedure 10,000 times to generate a null genome-wide distribution of θs that fitted a Poisson distribution with $\lambda = 0.8363$. We then used this Poisson distribution to calculate the statistical significance of θ associated with each 10 kb region among individual with ASD in our data. Given the median CNV length in our data was ~ 43 kb, the maximal number of non-overlapping CNV loci in the human genome is $\sim 7\times10^4$. Hence, we used this number to set a Bonferroni corrected cutoff for genome-wide significance of $P<7\times10^{-7}$ ($0.05/7\times10^4$). All analyses were performed using a commercial software package (MATLAB R2011b, The MathWorks Inc., Natick, MA, 2000).

CNV loci characterization

We used the RefSeq genes track in the UCSC Genome Browser on Human Mar. 2006 (NCBI36/hg18) assembly, (http://genome.ucsc.edu) to locate genes overlapping with ASD susceptibility CNV loci. Further, we used the December 2012 release of the Human Gene module of AutDB [18] to identify genes that have been reported as containing potential susceptibility variants for ASD.

Analysis of control CNV data

The majority of CNV loci curated in the CNV module of AutDB and used in our analysis were curated from published scientific reports in which one or more filtering steps were used to remove variants previously identified in unaffected individuals within the general population, resulting in a population of case-specific or case-enriched CNVs that were subsequently annotated. However, we concluded that an independent analysis of control CNV datasets using our prioritization strategy could be useful both in more accurately defining the boundaries of the eleven susceptibility loci identified in our initial analysis, as well as in allowing us to differentiate between potential false-positives or polymorphic CNV regions that would likely confer decreased risk of ASD susceptibility than loci with little or no control CNV overlap. Therefore, we examined the overlap of these 11 ASD susceptibility loci with CNVs among 4400 individuals with no diagnosis of ASD (controls) using data from three genome-wide CNV analyses varying in their sample sizes, studied cohorts, and CNV detection method/platform [19–21]. This control CNV data were collected from the Database of Genomic Variants (DGV) [22], and were analyzed using the same 10 kb regions described above.

Results

The CNV module of AutDB

The CNV module of AutDB as of October 2011 consisted of CNV data from 72 annotated publications, encompassing 30,989 CNVs from 4,359 individuals (3,099 ASD cases, 66 cases of other neurodevelopmental or neuropsychiatric disorders, and 1,194 control individuals). These have been summarized into 2,429 unique CNV reports classified as 'major' (i.e. independently validated) or 'minor' (1,047 and 1,382 CNV reports respectively), and distributed across 1035 CNV loci genome-wide (Figure S1). The median number of reports per CNV locus was two (one 'major' and one 'minor'), with the highest number of reports per locus reaching 16 (14 and 2 'major' and 'minor' respectively) for the 16p11.2 locus (Figure S1).

CNV dataset

For the purpose of CNV prioritization, we applied a set of stringent filtering criteria (see methods) to the AutDB data to generate a uniform subset of CNVs. This resulted in a dataset containing 4,926 CNVs in 2,373 ASD subjects (Mean = 2.08, STD = 1.79 CNVs per individual) collected from 48 scientific reports (Table S1) that encompassed ~2.13×10^{9} bp of genomic data. The maximal number of CNVs per individuals seen in our data was 18. Of the 2,373 ASD subjects in our data, 1,532 were males and 330 females, which is consistent with the 4:1 reported male-to-female ratio of ASD prevalence in the general population [23]. Notably, the gender of 511 subjects in our data was not reported. Of the 4,926 CNVs in our data, 1,923 (39%) were duplications and 3,003 (61%) were deletions (Table 1). In addition, 3,377 (68.6%) of the CNVs were inherited, 239 (4.9%) were *de novo*, and 1,310 (26.6%) had no indicated inheritance (Table 1). We observed remarkable variation in CNV size and inheritance patterns, with duplications being significantly larger than deletions ($P = 3\times10^{-105}$; Wilcoxon rank sum test; Figure S2), and *de novo* CNVs tended to be more prevalent among females than males ($X^2 = 18.6$; $P<0.0001$).

CNV characterization

To identify genomic loci with an excess of CNVs among ASD subjects, we divided the genome into consecutively distributed regions of 10 kb and assessed the burden of CNVs within them (See Materials and Methods). The genomic distribution of CNV burden among individuals with ASD is depicted in Figure 1. Overall, there were eleven genomic loci displaying significant burden score ($P<7\times10^{-7}$) distributing along eight chromosomes (Figure 2; Table 2), and containing 166 RefSeq genes [24]. Of these, four loci contained only one gene (*BCL9, NLGN1, DOCK8,* and *KPNA3* in 1q21.1, 3q26.31, 9p24.3, and 13q14.3, respectively), one locus contained three genes (*TRIML1, TRIML2,* and *LOC401164*, on 4q35.2) and six loci contained ≥ seven genes (Table 2). Seven of these loci included a relatively equal number of duplications and deletions, whereas four loci contained a majority (≥75%) of duplications.

The highest CNV burden was seen in a locus on human chromosome 16p11.2 ($\theta = 30.96$; $P < 1\times10^{-20}$), with 27 duplications and 36 deletions identified in ASD subjects from 10 different studies (Figure 3). No CNVs within this region were observed in 4400 controls. This locus spanned 750 kb and contained 31 RefSeq genes, three of which (*SEZ6L2*, MAPK3, and *KCTD13*) have been reported as containing susceptibility genetic variants in ASD individuals [25–27]. Interestingly, while duplications in the 16p11.2 locus tended to be inherited, deletions were overwhelmingly *de novo* in origin (Table 2).

The next top-ranked CNV loci were located in three regions on human chromosome 15q11.2-13.3 (Figure 4). The largest of the three loci (15q11.2-q13.1) spanned 4.95 Mb between chromosomal breakpoints BP2 and BP3 and consisted predominantly of duplications among ASD subjects (31 duplications vs. 2 deletions). Two additional susceptibility loci on chromosome 15 (15q11.2 and 15q13.2-q13.3, within BP1-BP2 and BP4-BP5, respectively) were approximately half the size (~2.4 Mb) and demonstrated equivalent prevalence of duplications and deletions in ASD individuals. A total of 44 RefSeq genes reside within these three regions, 10 of which have already been associated with ASD: *CYFIP1, NIPA1, NIPA2,* and *TUBGCP5* within the 15q11.2 locus; *UBE3A, ATP10A, GABRB3, SNRPN,* and *HERC2* within the 15q11.2-q13.1 locus; and *CHRNA7* within the 15q13.2-q13.3 locus. A closer inspection of the inheritance patterns of CNVs in these three genomic regions revealed that duplications tended to be *de novo* in origin, whereas deletions tended to be inherited (Table 2).

Two multigenic ASD susceptibility CNV loci were identified on human chromosome 22 (Figure 5). The 22q11.21 and 22q13.32-q13.33 loci spanned 2.5 Mb and 1.5 Mb, and overlapped with 51 and 33 genes, respectively. The CNV locus on 22q11.21 was predominantly enriched in duplications in ASD cases (21 duplications vs. 1 deletion), with a sharp increase in the number of duplications within a 20 kb region (chr22:19,345,000-19,365,000) that was implicated as a CNV enriched region in ASD cases (Glessner et al) (Figure 5A). Of the 51 genes within the 22q11.21 locus, two (*TBX1* and *GNBL1*) have already been associated with ASD. The locus on 22q13.32-q13.33 was primarily

Table 1. CNV characteristics.

	Male			Female		
	Duplications	**Deletions**	**sum**	**Duplications**	**Deletions**	**sum**
De novo	62 (1.7%)	109 (3.1%)	171 (4.8%)	27 (3.5%)	40 (5.1%)	67 (8.9%)
Inherited	1046 (29.5%)	1663 (46.9%)	2709 (76.4%)	200 (26.5%)	343 (45.7%)	543 (71.0%)
NR*	188 (5.3%)	479 (13.5%)	667 (18.8%)	32 (4.0%)	122 (15.2%)	154 (20.1%)
Sum	1296 (36.5%)	2251 (63.5)	3547 (100.0%)	259 (34.0%)	505 (66.0%)	764(100.0%)
	NR*			**All**		
	Duplications	**Deletions**	**sum**	**Duplications**	**Deletions**	**sum**
De novo	1 (0.2%)	0 (0.0%)	1 (0.2%)	90 (1.8%)	149 (3.0%)	239 (4.9%)
Inherited	50 (8.1%)	75 (12.2%)	125 (20.3%)	1296 (26.3%)	2081 (42.3%)	3377 (68.6%)
NR*	317 (51.5%)	172 (28.0%)	489 (79.5%)	537 (10.9%)	773 (15.7%)	1310 (26.6%)
Sum	368 (59.8%)	247 (40.2%)	615 (100.0%)	1923 (39.0%)	3003 (61.0%)	4926 (100.0%)

*NR = Not reported.

doi:10.1371/journal.pone.0066707.t001

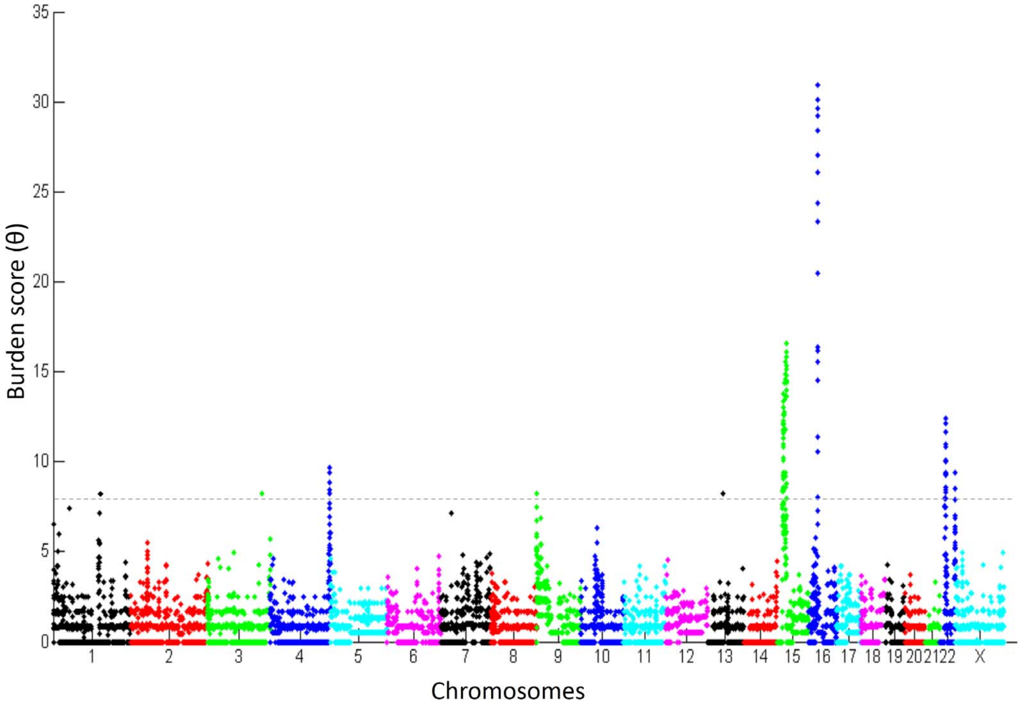

Figure 1. Whole-genome distribution of CNV burden. A Manhattan plot showing CNV burden among ASD subjects in 10 kb regions continuously distributed across the human genome. A dashed horizontal line indicate the burden score of 6.5 (0.995 quantile) that was used a threshold to determine the top ranked ASD susceptibility CNV loci.
doi:10.1371/journal.pone.0066707.g001

enriched in deletions among individuals with ASD (nine deletions vs. three duplications), and did not include CNVs in controls (Figure 5B). No previously characterized ASD-associated genes resided within the boundaries of this genomic locus; however, the ASD-associated gene SHANK3 is directly adjacent to the telomeric end of this region.

Finally, we examined the overlap of these eleven ASD susceptibility loci with CNV data from 4400 control individuals.

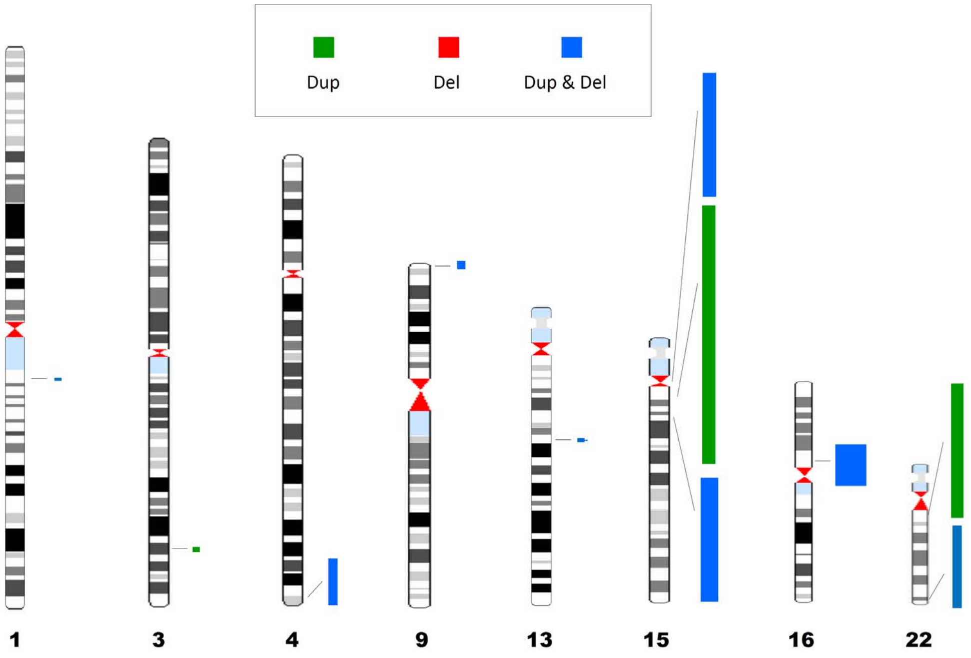

Figure 2. Physical locations of the top ranked 11 ASD susceptibility CNV loci on Human G-banded ideogram. CNV loci length and width are proportional to their genomic size and burden score respectively. Green, red, and blue are for CNV loci containing primarily copy number gains (Duplications), copy number losses (Deletions), or both Duplications and Deletions, respectively.
doi:10.1371/journal.pone.0066707.g002

Table 2. Top ASD susceptibility CNV loci.

Genomic locus					CNVs										
Locus	Start[a] (kb) (hg18)	End[a] (kb) (hg18)	Size (kb)	RefSeq Genes[b]	Duplications[c]			Deletions[c]			Total	# Studies	Θ[d]	P-value[e]	
					Inh	DN	NR	Inh	DN	NR					# CNVs among controls[f]
1q21.1	145555	145575	20	BCL9	3	3	3	-	-	4	13	4	8.15	2.61×10^{-7}	0
3q26.31	174745	174775	30	NLGN1	1	-	114	-	-	-	115	2	8.23	2.61×10^{-7}	31
4q35.2	189295	190195	900	TRIML1 TRIML2 LOC401164	1	-	8	1	1	5	15	5	9.38	2.16×10^{-8}	4
9p24.3	235	375	140	DOCK8	4	1	4	-	2	1	12	5	8.23	2.61×10^{-7}	0
13q14.3	49265	49275	10	KPNA3	4	-	-	6	-	-	10	1	8.22	2.61×10^{-7}	0
15q11.2	18525	20845	2320	9	5	6	1	6	1	-	19	7	13.38	4.31×10^{-13}	384
15q11.2-q13.1	21185	26135	4950	28	3	7	20	1	-	1	32	8	13.01	4.31×10^{-13}	159
15q13.2-q13.3	27985	30425	2440	7	4	8	2	8	1	3	26	9	16.60	1.11×10^{-16}	272
16p11.2	29495	30245	750	31	13	5	9	2	21	13	62	10	30.96	$< 1\times10^{-20}$	0
22q11.21	17265	19805	2540	51	8	3	10	-	1	-	22	7	12.15	7.24×10^{-12}	46
22q13.32-q13.33	47925	49435	1510	33	3	-	-	4	3	2	12	6	9.41	2.16×10^{-8}	0
Sum			15610	167	59	32	173	45	31	34	374				
					264			110							

[a] The boundaries of the ASD susceptibility CNV loci were determined as the midpoint of the 10kb region with a $\theta>6.5$.

[b] RefSeq genes including both protein-coding and non-coding RNA genes, and excluding Pseudogenes. For loci with >3 genes, only the number of RefSeq genes is given.

[c] CNV type. Inh = Inherited, DN = De-Novo, NR = Not Reported.

[d] CNV burden score.

[e] P-value calculated based on a Poisson distribution of CNV burden scores.

[f] 4400 individuals with no ASD diagnosis from three large genome-wide studies.

doi:10.1371/journal.pone.0066707.t002

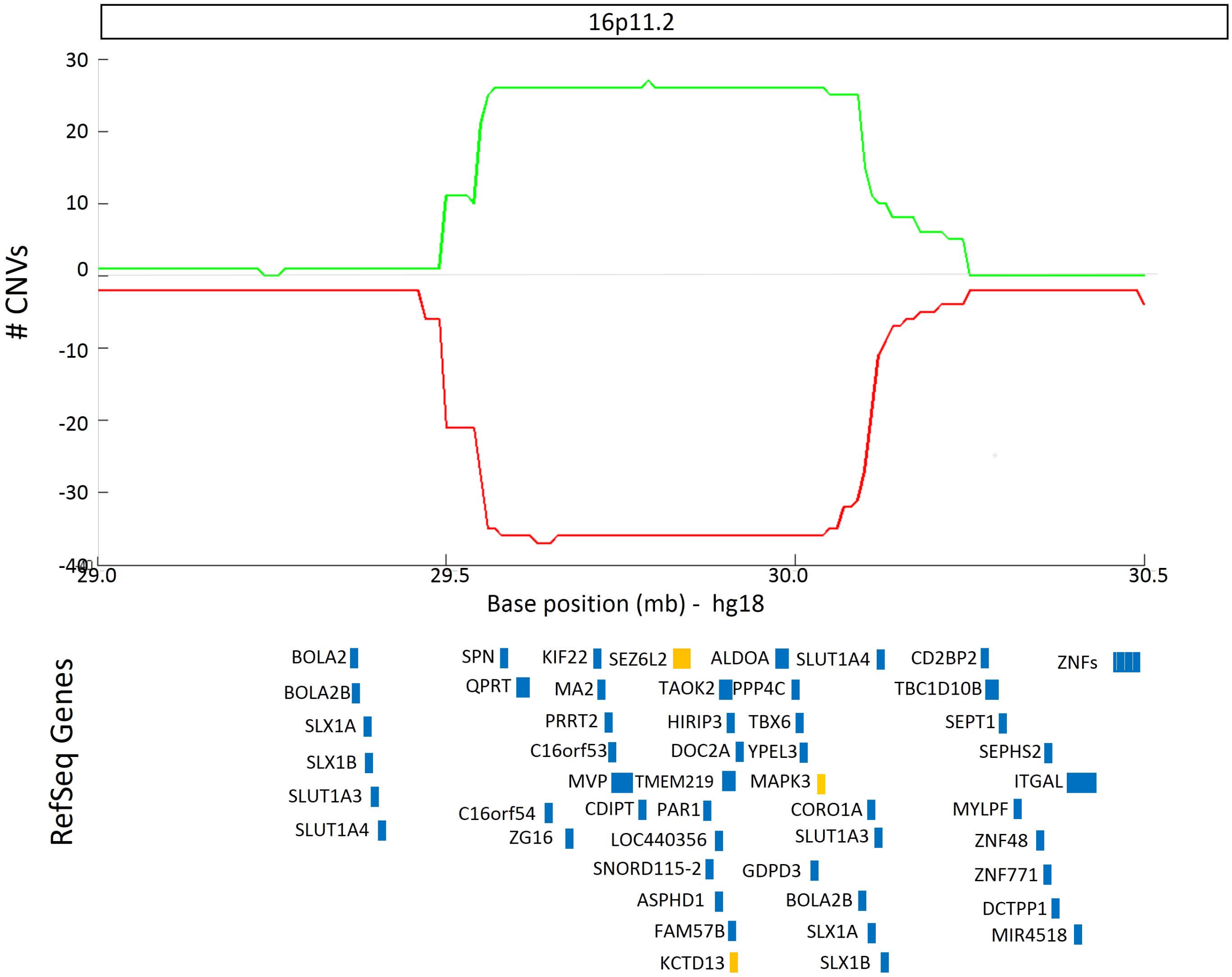

Figure 3. ASD susceptibility CNV locus on human chromosome 16. The number of individuals with duplications (green) and deletions (red) are plotted along human chromosome 16p11.2. RefSeq Genes overlapping with this region are depicted in blue rectangles. Genes that have been associated with ASD according to AutDB [18] are colored in orange.
doi:10.1371/journal.pone.0066707.g003

Consequently, three of these loci (3q26.31, 4q35.2; Figure S3, and 15q11.2; Figure 4) showed a complete overlap with CNVs among controls, suggesting potential false positives, or polymorphic loci that confer lesser risk than initially estimated. In addition, three other ASD susceptibility loci (15q11.2-q13.1, 15q13.2-q13.3, and 22q11.21; Figures 4 and 5) demonstrated partial overlap with CNVs among controls, allowing us to refine the boundaries of this susceptibility loci.

Discussion

The major goals of this study were to: (A) prioritize ASD susceptibility loci based on their CNV burden among ASD individuals, and (B) determine the genomic and genetic characteristics of these loci. The CNV dataset used for our analysis was curated from published scientific reports and primarily included CNVs that have been exclusively seen among individual with ASD in each of these studies. Yet, we compared these data with a control CNV dataset derived from three genome-wide CNV reports [19–21] to further refine the boundaries of our susceptibility loci and identify potential false-positives or lower-risk susceptibility loci. In addition, we contrasted our results with those of two other similar studies [28,29], and a comprehensive review of the scientific literature describing ASD-associated CNVs [30]. We found that five (45.45%) of the susceptibility loci reported in this study were also identified in all of these other reports, and two other loci (18.2%) overlapped with one large population study of CNVs in human genetic disease [28] (Table S2). These overlapping findings support the validity of our analysis, especially given the relatively low frequency and widespread genomic distribution of ASD susceptibility CNV loci, as well as the differences in population makeup in the other studies.

The highest CNV burden in our study was seen in the 16p11.2 locus. This genomic locus have long been known as, a genetic risk factor for ASD [31], as well as other disorders including schizophrenia [32], developmental delay and cognitive impairment [33], major depressive disorder [34], and obesity [35]. Three genes within this locus (*SEZ6L2, MAPK3,* and *KCTD13)* have been independently identified as ASD genetic risk factors [25–27], but the genetic mechanisms by which deletions or duplications within the 16p11.2 locus contribute to ASD susceptibility remain unknown. Initial hints for potential functional mechanism might be found in a recent study in zebrafish [27] demonstrated that over-expression of the ASD-associated gene, *KCTD13,* led to

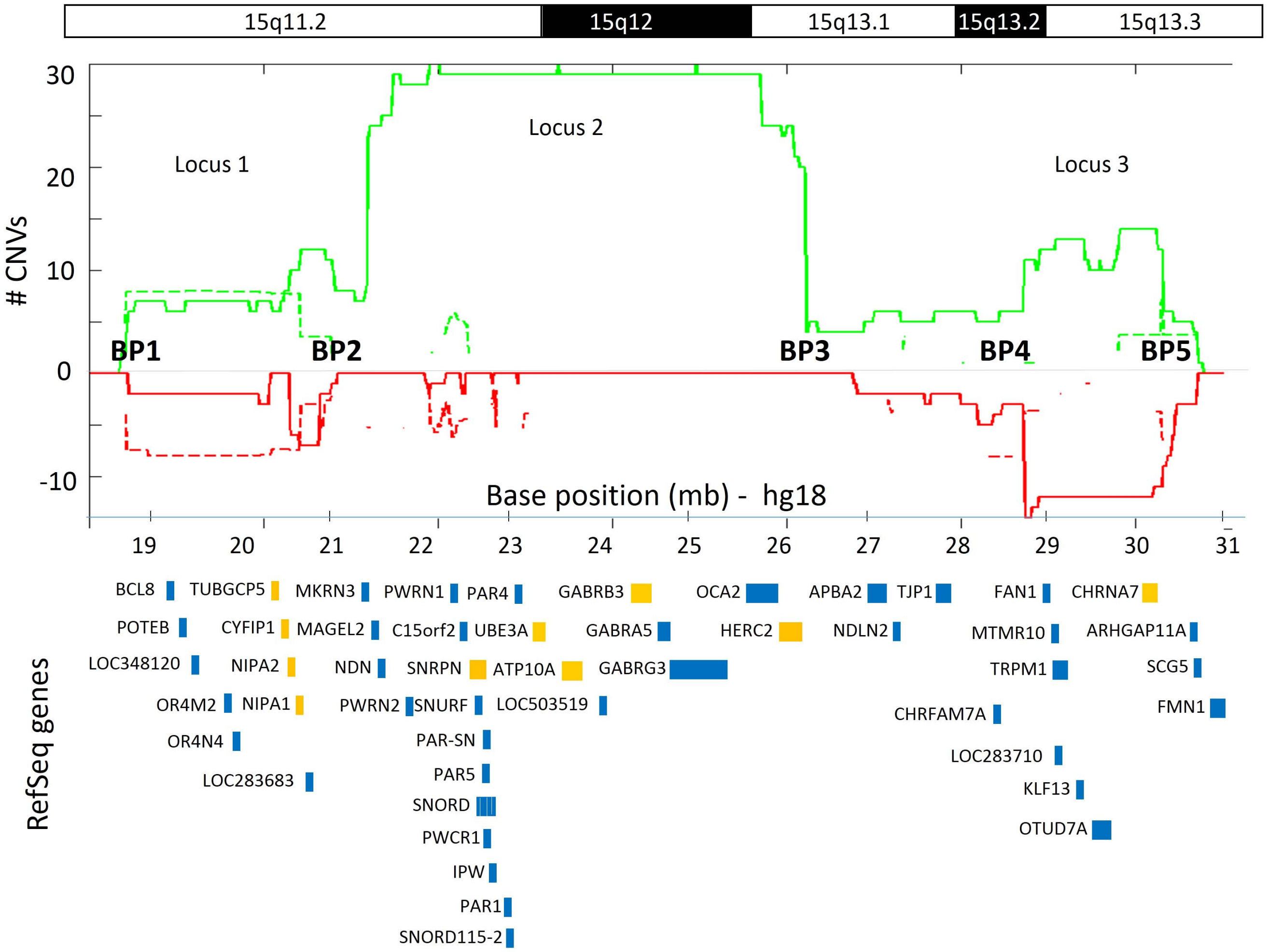

Figure 4. ASD susceptibility CNV loci on human chromosome 15 The number of individuals with duplications (green) and deletions (red) are plotted for both ASD cases (continuous lines) and controls (broken lines, on a $\log_2$ scale) along human chromosome 15q11.2-13.3. RefSeq genes overlapping with these regions are depicted in blue rectangles. Genes that have been associated with ASD according to AutDB [18] are colored in orange. The variation in CNV burden along the 15q11.2 – 13.3 region, suggests three distinct ASD susceptibility loci: Locus 1 within breakpoints (BP) 1-2, Locus 2 within BPs 2-3, and Locus 3 within BPs 4-5.
doi:10.1371/journal.pone.0066707.g004

decreased proliferation of neural progenitor cells and reduced head size, mirroring the microcephaly phenotype commonly seen in cases with 16p11.2 duplication. Alternatively, suppression of *KCTD13* led to increased neural progenitor cell proliferation and increased head size, mirroring the macrocephaly phenotype observed in many cases with 16p11.2 deletion. Likewise, another gene within the 16p11.2 locus, *TAOK2*, was recently shown to influence the formation of basal dendrites in the developing cortex [36]. Therefore, it is likely that these two genes, as well as other genes within the 16p11.2 susceptibility loci, act in concert and contribute to ASD susceptibility.

Our analysis implicated three distinct ASD loci within the 15q11-q13 genomic locus. Duplications within the 15q11.2-q13.1 region, located between chromosomal breakpoints (BPs) BP2-BP3, have long been strongly implicated in ASD pathogenesis [37], whereas deletions of this region are a primary cause of Angelman and Prader-Willi syndromes [38]. The 15q11.2-q13.1 region is flanked by two genomic loci, 15q11.2 and 15q13.2-q13.3 that have demonstrated association with not only autism but also other neurodevelopmental and neuropsychiatric disorders. CNVs within the 15q11.2 genomic loci (BP1-BP2) have been shown to confer risk to epilepsy and developmental delay [39,40], while deletions of the 15q13 region (BP4-BP5) were associated with increased risk for intellectual disability and epilepsy [41]. Notably, the complete overlap of CNVs among controls with the 15q11.2 locus, as seen in other two others CNV loci in this study (3q26.31, and 4q35.2) support the two-hit premise of ASD etiology [42].

Four monogenic genomic loci were identified in our analysis that merit further investigation. The 9p24.3 locus, which contains the *DOCK8* gene overlaps with linkage regions identified in large autism extended pedigrees [43,44], and disruption of *DOCK8* has previously been implicated in two cases of intellectual disability and developmental delay [45]. Therefore, while there is no direct evidence implicating *DOCK8* in ASD, these previous results and our findings argue for a potentially critical role for this gene in ASD susceptibility. The *BCL9* gene resides within a ~1.5–Mb genomic region of the 1q21.1 locus in which both deletions and duplications can result in syndromes associated with numerous phenotypes, including autism [46]. Furthermore, common variants in the *BCL9* gene are associated with schizophrenia, bipolar disorder, and major depressive disorder [47]. *BCL9* is a component of the canonical Wnt signaling pathway, which has

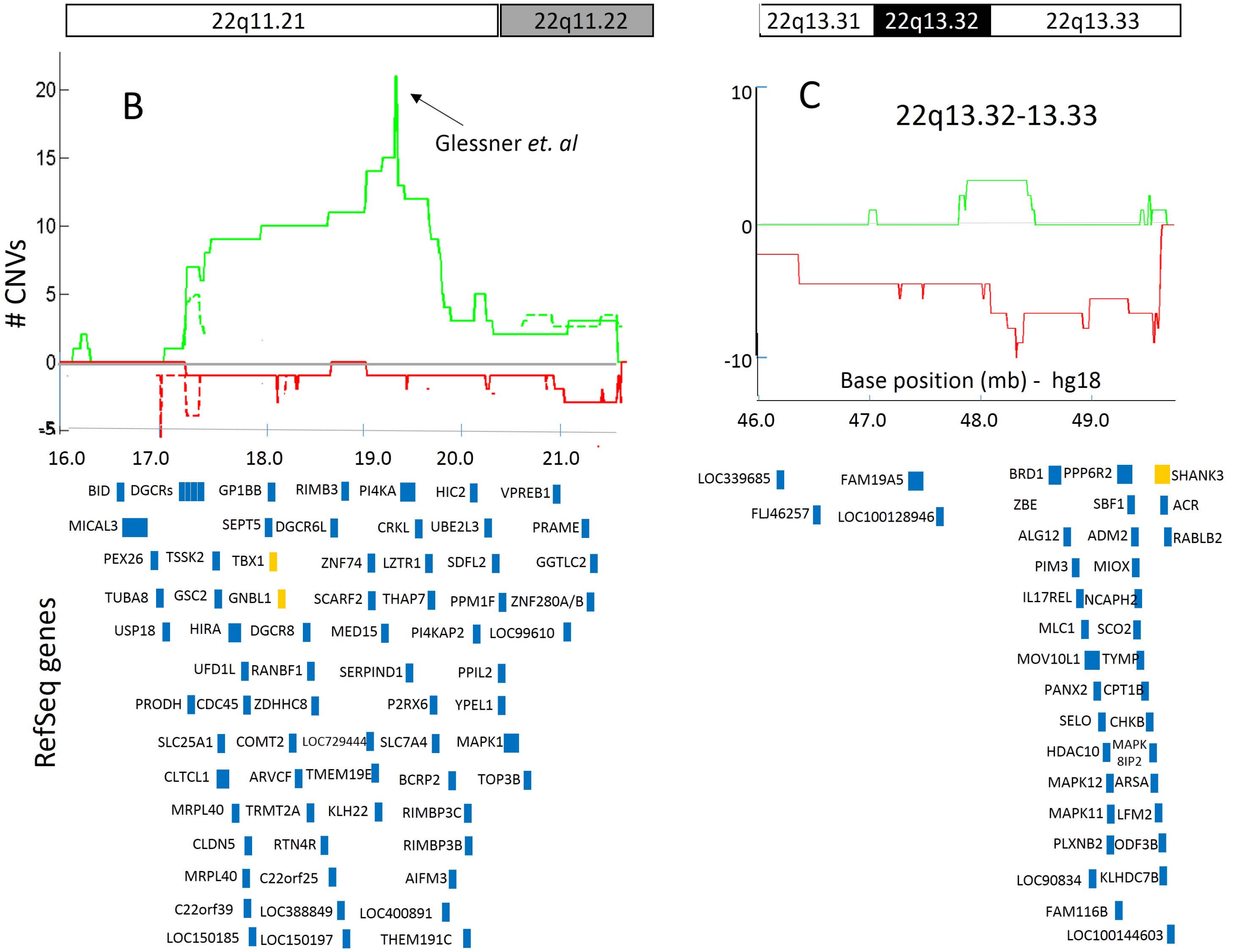

Figure 5. ASD susceptibility CNV loci on human chromosome 22. The number of individuals with duplications (green) and deletions (red) are plotted for both ASD cases (continuous lines) and controls (broken lines, on a $\log_2$ scale) along human chromosome 22q11.21-13.33. RefSeq genes overlapping with these regions are depicted in blue rectangles. Genes that have been associated with ASD according to AutDB [18] are colored in orange. (**A**) The CNV locus on 22q11.21 contains primarily copy number gains (duplications). A black arrow indicates the peak in duplications count due to the CNV data from Glessner *et. al*. (**B**) The CNV locus on 22q13.32-13.33 contains primarily copy number losses (deletions).
doi:10.1371/journal.pone.0066707.g005

been proposed to be affected in ASD [48]. While there is no direct genetic evidence demonstrating that the *BCL9* gene confers genetic risk to ASD susceptibility, our analysis in combination with previous findings strongly implies a potential pathogenic role for this gene. Duplications of the 3q26.31 locus within the *NLGN1* gene were also seen among control individuals, thus questioning the association of this locus with ASD susceptibility [10]. Yet, the functional relevance of *NLGN1* to ASD is supported by a recent report identified a duplication of the *NLGN1* gene in an autistic patient with mild intellectual disability [49], as well as by functional studies demonstrating a role for this gene in neurite outgrowth via interactions with the ASD-associated gene NRXN1 [50]. Finally, the *KPNA3* gene at the 13q14.3 locus, has also been implicated as a potential schizophrenia susceptibility gene [51]. Further investigation will be required to ascertain the relevance of *KPNA3* to these two related psychiatric disorders.

The incorporation of CNV data from both large-scale whole-genome studies and smaller-scale case studies as a framework for ASD-related CNV loci characterization is the major strength of our study. However, this approach has some limitations. First, some of the studies in our dataset included genome-wide CNV data, while other studies focused on the identification of CNVs within specific genomic loci or with a specific mechanism of CNV inheritance. To account for the potential bias arising from this variation, we assigned different weights to studies based on the CNVs/Loci ratio reported in them, and incorporated this weight in the calculation of the CNV burden score (See methods). Another potential source for variation in our data is the differing CNV detection technologies used by the different studies. Accordingly, one might suspect that the higher burden scores observed among larger genomic regions (>500 kb) would be due to the greater likelihood of CNVs within these regions to be detected by all CNV detection methods, whereas smaller CNVs would not be detected by many of the older lower-resolution microarray platforms. While we cannot rule out this possibility, Sanders et. al [14], which employed a high-resolution CNV detection method, reported that the burden of CNVs in their study was remarkably similar to previously published results using lower-resolution CNV detection platforms.

The results of this analysis could have broad clinical and scientific implications. For example, one could use these loci as a guideline for the evaluation of chromosomal microarray (CMA) screening, a procedure that is being increasingly used in genetic evaluation of ASD subjects. Alternatively, the detailed characteristics provided for each of the CNV loci highlighted in the study may be used for further, in-depth exploration of the biological mechanism underlying their role in ASD susceptibility. An intriguing premise is whether the different genetic mechanisms trigger ASD susceptibility in subjects containing CNVs at distinct genomic loci, and whether these genetically diverse individuals present different ASD related phenotypes. We anticipate to increase the resolution of these analyses with continued updates to AutDB and additional control CNV data, which will subsequently provide both scientists and clinicians with a valuable resource for genetics research and clinical diagnostic efforts of ASD.

Supporting Information

Figure S1 Distribution of CNV loci reports. Distribution of 'major' and 'minor' reports across CNV loci in AutDB[18], (A) Venn Diagram, (B) cumulative distribution function (cdf). (C) Top 1% reported CNV loci.
(TIF)

Figure S2 Distribution of CNV sizes. (A) Histogram of the log10 (CNV size) indicate that CNV sizes in our data have a lognormal distribution with a mean = 42.8 kb. (B) CDF plots for the sizes of copy number gains (green), and copy number losses (red).
(TIF)

Figure S3 ASD susceptibility CNV loci on human chromosomes 3 & 4. The number of individuals with duplications (green) and deletions (red) are plotted for both ASD cases (continuous lines) and controls (broken lines) along human chromosomes 3q26.31 & 4q35.2. RefSeq genes overlapping with these regions are depicted in blue rectangles. Genes that have been associated with ASD according to AutDB [18] are colored in orange.
(TIF)

Table S1 Scientific reports of CNVs in ASD individuals. Details of the 48 scientific reports and their CNV data used in this study are listed in the table.
(XLS)

Table S2 CNV loci associated with ASD across different studies. A list of CNV loci associated with ASD from four different studies are depicted in the table. CNV loci highlighted in multiple studies are highlighted in bold.
(DOCX)

Acknowledgments

We thank Catherine C. Swanwick for her review and valuable editorial comments and the IT team of MindSpec for the development of AutDB.

Author Contributions

Conceived and designed the experiments: IM ECL SB. Performed the experiments: ECL. Analyzed the data: IM. Contributed reagents/materials/analysis tools: SB. Wrote the paper: IM ECL SB.

References

1. Iafrate AJ, Feuk L, Rivera MN, Listewnik ML, Donahoe PK, et al. (2004) Detection of large-scale variation in the human genome. Nat Genet 36: 949–951.
2. Sebat J, Lakshmi B, Troge J, Alexander J, Young J, et al. (2004) Large-scale copy number polymorphism in the human genome. Science 305: 525–528.
3. Lee JA, Lupski JR (2006) Genomic rearrangements and gene copy-number alterations as a cause of nervous system disorders. Neuron 52: 103–121.
4. Stankiewicz P, Lupski JR (2010) Structural variation in the human genome and its role in disease. Annu Rev Med 61: 437–455.
5. Lord C, Cook EH, Leventhal BL, Amaral DG (2000) Autism spectrum disorders. Neuron 28: 355–363.
6. Nazeer A, Ghaziuddin M (2012) Autism spectrum disorders: clinical features and diagnosis. Pediatr Clin North Am 59: 19–25, ix.
7. Muhle R, Trentacoste SV, Rapin I (2004) The genetics of autism. Pediatrics 113: e472–486.
8. Hallmayer J, Cleveland S, Torres A, Phillips J, Cohen B, et al. (2011) Genetic Heritability and Shared Environmental Factors Among Twin Pairs With Autism. Arch Gen Psychiatry.
9. Anney R, Klei L, Pinto D, Regan R, Conroy J, et al. (2010) A genome-wide scan for common alleles affecting risk for autism. Hum Mol Genet 19: 4072–4082.
10. Glessner JT, Wang K, Cai G, Korvatska O, Kim CE, et al. (2009) Autism genome-wide copy number variation reveals ubiquitin and neuronal genes. Nature 459: 569–573.
11. Pinto D, Pagnamenta AT, Klei L, Anney R, Merico D, et al. (2010) Functional impact of global rare copy number variation in autism spectrum disorders. Nature 466: 368–372.
12. Weiss LA (2009) Autism genetics: emerging data from genome-wide copy-number and single nucleotide polymorphism scans. Expert Rev Mol Diagn 9: 795–803.
13. Gilman SR, Iossifov I, Levy D, Ronemus M, Wigler M, et al. (2011) Rare de novo variants associated with autism implicate a large functional network of genes involved in formation and function of synapses. Neuron 70: 898–907.
14. Sanders SJ, Ercan-Sencicek AG, Hus V, Luo R, Murtha MT, et al. (2011) Multiple Recurrent De Novo CNVs, Including Duplications of the 7q11.23 Williams Syndrome Region, Are Strongly Associated with Autism. Neuron 70: 863–885.
15. Swanwick CC, Larsen EC, Banerjee-Basu S (2011) Genetic Heterogeneity of Autism Spectrum Disorders. In: Deutsch SI, Urbano MR, editors. Autism Spectrum Disorders: The Role of Genetics in Diagnosis and Treatment: InTech. pp. 65.
16. Miller DT, Adam MP, Aradhya S, Biesecker LG, Brothman AR, et al. (2010) 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.
17. Shen Y, Dies KA, Holm IA, Bridgemohan C, Sobeih MM, et al. (2010) Clinical genetic testing for patients with autism spectrum disorders. Pediatrics 125: e727–735.
18. Basu SN, Kollu R, Banerjee-Basu S (2009) AutDB: a gene reference resource for autism research. Nucleic Acids Res 37: D832–836.
19. Zogopoulos G, Ha KC, Naqib F, Moore S, Kim H, et al. (2007) Germ-line DNA copy number variation frequencies in a large North American population. Hum Genet 122: 345–353.
20. Shaikh TH, Gai X, Perin JC, Glessner JT, Xie H, et al. (2009) High-resolution mapping and analysis of copy number variations in the human genome: a data resource for clinical and research applications. Genome Res 19: 1682–1690.
21. International HapMap C, Altshuler DM, Gibbs RA, Peltonen L, Altshuler DM, et al. (2010) Integrating common and rare genetic variation in diverse human populations. Nature 467: 52–58.
22. Zhang J, Feuk L, Duggan GE, Khaja R, Scherer SW (2006) Development of bioinformatics resources for display and analysis of copy number and other structural variants in the human genome. Cytogenet Genome Res 115: 205–214.
23. Newschaffer CJ, Croen LA, Daniels J, Giarelli E, Grether JK, et al. (2007) The epidemiology of autism spectrum disorders. Annu Rev Public Health 28: 235–258.
24. Pruitt KD, Tatusova T, Brown GR, Maglott DR (2012) NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy. Nucleic Acids Res 40: D130–135.
25. Kumar RA, Marshall CR, Badner JA, Babatz TD, Mukamel Z, et al. (2009) Association and mutation analyses of 16p11.2 autism candidate genes. PLoS One 4: e4582.
26. Schaaf CP, Sabo A, Sakai Y, Crosby J, Muzny D, et al. (2011) Oligogenic heterozygosity in individuals with high-functioning autism spectrum disorders. Hum Mol Genet 20: 3366–3375.
27. Golzio C, Willer J, Talkowski ME, Oh EC, Taniguchi Y, et al. (2012) KCTD13 is a major driver of mirrored neuroanatomical phenotypes of the 16p11.2 copy number variant. Nature 485: 363–367.
28. Itsara A, Cooper GM, Baker C, Girirajan S, Li J, et al. (2009) Population analysis of large copy number variants and hotspots of human genetic disease. Am J Hum Genet 84: 148–161.
29. Kaminsky EB, Kaul V, Paschall J, Church DM, Bunke B, et al. (2011) 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.

30. Marshall CR, Scherer SW (2012) Detection and characterization of copy number variation in autism spectrum disorder. Methods Mol Biol 838: 115–135.
31. Walsh KM, Bracken MB (2011) Copy number variation in the dosage-sensitive 16p11.2 interval accounts for only a small proportion of autism incidence: a systematic review and meta-analysis. Genet Med 13: 377–384.
32. McCarthy SE, Makarov V, Kirov G, Addington AM, McClellan J, et al. (2009) Microduplications of 16p11.2 are associated with schizophrenia. Nat Genet 41: 1223–1227.
33. Shinawi M, Liu P, Kang SH, Shen J, Belmont JW, et al. (2010) Recurrent reciprocal 16p11.2 rearrangements associated with global developmental delay, behavioural problems, dysmorphism, epilepsy, and abnormal head size. J Med Genet 47: 332–341.
34. Degenhardt F, Priebe L, Herms S, Mattheisen M, Muhleisen TW, et al. (2012) Association between copy number variants in 16p11.2 and major depressive disorder in a German case-control sample. Am J Med Genet B Neuropsychiatr Genet 159B: 263–273.
35. Jacquemont S, Reymond A, Zufferey F, Harewood L, Walters RG, et al. (2011) Mirror extreme BMI phenotypes associated with gene dosage at the chromosome 16p11.2 locus. Nature 478: 97–102.
36. de Anda FC, Rosario AL, Durak O, Tran T, Graff J, et al. (2012) Autism spectrum disorder susceptibility gene TAOK2 affects basal dendrite formation in the neocortex. Nat Neurosci 15: 1022–1031.
37. Hogart A, Wu D, LaSalle JM, Schanen NC (2010) The comorbidity of autism with the genomic disorders of chromosome 15q11.2-q13. Neurobiol Dis 38: 181–191.
38. Buiting K (2010) Prader-Willi syndrome and Angelman syndrome. Am J Med Genet C Semin Med Genet 154C: 365–376.
39. Burnside RD, Pasion R, Mikhail FM, Carroll AJ, Robin NH, et al. (2011) Microdeletion/microduplication of proximal 15q11.2 between BP1 and BP2: a susceptibility region for neurological dysfunction including developmental and language delay. Hum Genet 130: 517–528.
40. de Kovel CG, Trucks H, Helbig I, Mefford HC, Baker C, et al. (2010) Recurrent microdeletions at 15q11.2 and 16p13.11 predispose to idiopathic generalized epilepsies. Brain 133: 23–32.
41. van Bon BWM, Mefford HC, de Vries BBA (1993) 15q13.3 Microdeletion.
42. Gau SS, Liao HM, Hong CC, Chien WH, Chen CH (2012) Identification of two inherited copy number variants in a male with autism supports two-hit and compound heterozygosity models of autism. Am J Med Genet B Neuropsychiatr Genet 159B: 710–717.
43. Allen-Brady K, Miller J, Matsunami N, Stevens J, Block H, et al. (2009) A high-density SNP genome-wide linkage scan in a large autism extended pedigree. Mol Psychiatry 14: 590–600.
44. Coon H, Villalobos ME, Robison RJ, Camp NJ, Cannon DS, et al. (2010) Genome-wide linkage using the Social Responsiveness Scale in Utah autism pedigrees. Mol Autism 1: 8.
45. Griggs BL, Ladd S, Saul RA, DuPont BR, Srivastava AK (2008) Dedicator of cytokinesis 8 is disrupted in two patients with mental retardation and developmental disabilities. Genomics 91: 195–202.
46. Brunetti-Pierri N, Berg JS, Scaglia F, Belmont J, Bacino CA, et al. (2008) Recurrent reciprocal 1q21.1 deletions and duplications associated with microcephaly or macrocephaly and developmental and behavioral abnormalities. Nat Genet 40: 1466–1471.
47. Li J, Zhou G, Ji W, Feng G, Zhao Q, et al. (2011) Common variants in the BCL9 gene conferring risk of schizophrenia. Arch Gen Psychiatry 68: 232–240.
48. Kalkman HO (2012) A review of the evidence for the canonical Wnt pathway in autism spectrum disorders. Mol Autism 3: 10.
49. Leblond CS, Heinrich J, Delorme R, Proepper C, Betancur C, et al. (2012) Genetic and functional analyses of SHANK2 mutations suggest a multiple hit model of autism spectrum disorders. PLoS Genet 8: e1002521.
50. Gjorlund MD, Nielsen J, Pankratova S, Li S, Korshunova I, et al. (2012) Neuroligin-1 induces neurite outgrowth through interaction with neurexin-1beta and activation of fibroblast growth factor receptor-1. FASEB J 26: 4174–4186.
51. Morris CP, Baune BT, Domschke K, Arolt V, Swagell CD, et al. (2012) KPNA3 variation is associated with schizophrenia, major depression, opiate dependence and alcohol dependence. Dis Markers 33: 163–170.