Inventory of Supplemental Information

Figures

Figure S1, related to Figure 1: qPCR analysis to optimize transmitted and *de novo* CNV detection.

A) Receiver operating characteristic (ROC) curve of relative quantification qPCR (dilution curve assessed once per oligo set on a different plate, four replicates) accuracy as detection threshold was varied from 1 (normal copy number) to 0 (deletions) or 2 (duplications). True positives were 22 known large deletions and 35 known large duplications; 4,453 assays of pooled DNA controls in regions of rare CNVs were used as true negatives. The closer the plot approaches the top left corner the more accurate the assay. Based on this plot the detection thresholds of 0.7 for deletions and 1.2 for duplications were chosen. **B)** ROC curve of absolute quantification qPCR (dilution curve for every oligo set run alongside the assay, three replicates) accuracy as detection threshold was varied. True positives were 17 known large deletions and 27 known large duplications; 131 assays of pooled DNA controls in regions of rare CNVs were used as true negatives. Based on this plot the detection thresholds of 0.7 for deletions and 1.3 for duplications were chosen. **C)** Confirmation results of rare (<1% DGV), high confidence (PennCNV and QuantiSNP ±GNOSIS, <50% of CNV from one algorithm) transmitted CNV predictions (N=74) from absolute quantification qPCR. No variation in detection accuracy with probe number was seen within this detection range. **D)** Confirmation results of rare, high confidence *de novo* CNV predictions (N=63) from absolute quantification qPCR. The positive predictive value <20 probes is 13% and ≥20 probes is 91% (a further 14 confirmed predictions >100 probes (111-6,013 probes) are not shown). **E)** Prediction accuracy of rare *de novo* CNV predictions (N=403) using relative quantification qPCR; all remaining *de novo* events were also tested by absolute quantification qPCR (N=99) giving 61 confirmed *de novo* events. The positive predictive value <20 probes is 2.6% and ≥20 probes is 53%. Based on this data only *de novo* CNVs with ≥20 probes were considered for further analysis.

Figure S2, related to Figure 2: The burden of *de novo* CNVs and genes mapping within them in 872 probands and 872 matched siblings at different size ranges.

A-C) Total number of rare *de novo* deletions (red) and duplications (blue) in probands and siblings across three size ranges shown above each plot. P-values are displayed above the bars based on a Sign test. **D-F)** Show the corresponding number of Refseq genes within the CNVs of the plot vertically above. A gene is counted if there is any overlap between the CNV and the UTR, exons, or introns. P-values are displayed above the bars based on a Wilcoxon paired test.

Figure S3, related to Figure 3: Maternal and Paternal age at birth with *de novo* CNV status.

A) Boxplots of paternal age at birth are shown for probands (green) and siblings (purple); the notches correspond to 95% confidence intervals of the median. No significant difference was seen between fathers of samples with no identified *de novo* CNVs (N=1,066) and fathers of samples with: *de novo* CNVs, *de novo* duplications, or *de novo* deletions. P-values are displayed above based on a two-tailed equal variance t-test. **B)** The same data is presented as in A, except based on maternal age at birth.

Figure S4, related to Figure 4: Estimated p-value for recurrent *de novo* CNVs.

A) P-values are calculated using the birthday problem logic (Methods) for the 67 *de novo* CNVs confirmed in probands in this paper. The values used were C=232, d=19 (all *de novo* CNVs), d=10 (*de novo* deletions), and d=9 (*de novo* duplications). **B)** P-values for the 246 *de novo* CNVs confirmed in probands in the combination of four studies with this paper (Itsara et al., 2010; Marshall et al., 2008; Pinto et al., 2010; Sebat et al., 2007). The values used were C=232, d=73 (all *de novo* CNVs), d=39 (*de novo* deletions), and d=34 (*de novo* duplications).

Tables

Phenotype	Study	Samples	*De novo* CNV rate			
			All	>100kb	>500kb	>1Mb
Simplex ASD	This paper	872 ASD Quartets	5.8% (n=51)	5.3% (n=46)	4.1% (n=36)	2.5% (n=22)
	This paper	252 ASD Trios	4.8% (n=12)	3.6% (n=9)	2.4% (n=6)	1.6% (n=4)
	Pinto et al., 2010	393 ASD	6.9% (n=27)	4.6% (n=18)	2.5% (n=10)	0.8% (n=3)
	Itsara et al., 2010	60 ASD	5.0% (n=3)	3.3% (n=2)	0.0% (n=0)	0.0% (n=0)
	Marshall et al., 2008	237 ASD	8.4% (n=20)	8.0% (n=19)	6.8% (n=17)	5.9% (n=14)
	Sebat et al., 2007	118 ASD	11.0% (n=13)	9.3% (n=11)	5.1% (n=6)	4.2% (n=5)
	Total	1,932 Simplex ASD	6.6% (n=126)	5.4% (n=105)	3.9% (n=75)	2.5% (n=48)
Multiplex ASD	Pinto et al., 2010	348 ASD	5.4% (n=19)	3.2% (n=11)	1.4% (n=5)	1.1% (n=4)
	Itsara et al., 2010	1,270 ASD	4.3% (n=54)	2.8% (n=36)	1.4% (n=18)	1.0% (n=13)
	Marshall et al., 2008	189 ASD	1.6% (n=3)	1.6% (n=3)	1.6% (n=3)	1.0% (n=2)
	Sebat et al., 2007	77 ASD	2.6% (n=2)	2.6% (n=2)	2.6% (n=2)	2.6% (n=2)
	Total	1,884 Multiplex ASD	4.1% (n=78)	2.8% (n=52)	1.8% (n=33)	1.1% (n=21)
Non-ASD	This paper	872 Unaffected Siblings	1.7% (n=15)	1.5% (n=13)	0.8% (n=7)	0.5% (n=4)
	Itsara et al., 2010	427 Unaffected Siblings	1.2% (n=4)	0.7% (n=2)	0.0% (n=0)	0.0% (n=0)
	Itsara et al., 2010	386 Asthma Cases	2.3% (n=9)	2.1% (n=8)	1.3% (n=5)	1.0% (n=4)
	Sebat et al., 2007	196 Controls	1.0% (n=2)	1.0% (n=2)	1.0% (n=2)	1.0% (n=2)
	Total	1,881 Non-ASD	1.6% (n=30)	1.3% (n=25)	0.7% (n=14)	0.5% (n=10)

Table S1: Rates of *de novo* CNVs in simplex ASD, multiplex ASD, and controls.

Algorithms making the CNV prediction	Type of CNV	Median number of probes	Total number of CNVs tested	Number of CNVs confirmed by qPCR	Number of CNV unconfirmed by qPCR	Positive Predictive Value (PPV)
GN, PN, QT [a]	Del	7	24	23	1	96%
	Dup	15	14	14	0	100%
PN, QT only	Del	6	19	16	3	84%
	Dup	8	17	14	3	82%
PN, QT, ±GN	**All**	**8**	**74**	**67**	**7**	**91%**
GN, PN only	Del	6	4	2	2	50%
	Dup	12	3	2	1	67%
GN, QT only	Del	17	4	2	2	50%
	Dup	NA	0	0	0	0%
PN only	Del	4	6	3	3	50%
	Dup	7	11	6	5	55%
GN only	Del	5	4	1	3	25%
	Dup	6	5	0	0	0%
QT only	Del	7	2	1	1	50%
	Dup	8	2	1	1	50%
Total	**All**	**8**	**115**	**85**	**25**	**75%**

Table S2: Positive predictive value with varying combinations of algorithms.

[a] Three CNV prediction algorithms were used to identify CNVs: PennCNV (PN), QuantiSNP (QT), GNOSIS (GN).

Category		Proband[a] (N=1,124)	Sibling (N=872)	1Mv1 (N=624)	1Mv3 Duo (N=1,372)	Deletion (N=1,996)	Duplication (N=1,996)
All CNVs	Autosomes	82.2 ±1.1	82.3 ±1.4	84.7 ±1.6	81.1 ±1.0	59.7 ±0.8	22.6 ±0.5
	XY	14.1 ±1.1	8.6 ±0.6	9.4 ±0.7	12.7 ±0.9	3.4 ±0.2	8.3 ±0.6
High-confidence CNVs	Autosomes	33.6 ±0.5	32.9 ±0.6	36.8 ±0.7	31.7 ±0.5	24.9 ±0.3	8.5 ±0.2
	XY	0.9 ±0.1	0.9 ±0.1	0.6 ±0.1	1.0 ±0.1	0.4 ±0.0	0.5 ±0.0
Rare, high-confidence CNVs	Autosomes	13.7 ±0.4	13.3 ±0.3	13.2 ±0.4	13.7 ±0.3	9.5 ±0.2	4.3 ±0.1
	XY	0.8 ±0.1	0.8 ±0.1	0.5 ±0.1	1.0 ±0.1	0.3 ±0.0	0.5 ±0.0
De novo, rare, high-confidence CNVs	Autosomes	1.3 ±0.1	1.3 ±0.1	1.1 ±0.1	1.4 ±0.1	1.1 ±0.1	0.3 ±0.0
	XY	0.01 ±0.01	0.01 ±0.01	0.00 ±0.01	0.02 ±0.01	0.01 ±0.00	0.00 ±0.00
≥20 probes, *de novo,* rare, high-confidence CNVs	Autosomes	0.18 ±0.5	0.16 ±0.5	0.17 ±0.6	0.16 ±0.5	0.10 ±0.4	0.07 ±0.3
	XY	0.00 ±0.00	0.00 ±0.00	0.00 ±0.00	0.00 ±0.00	0.00 ±0.00	0.00 ±0.00

Table S3: Number of CNVs per sample at each stage of the experiment.

[a] mean ±95% confidence intervals number of CNVs per sample

Table S4: All confirmed *de novo* CNVs and recurrent regions.

Submitted as a separate excel file.

	16p Deletion (N=8)	16p Deletion Matches (N=40)	16p Duplication (N=6)	16p Duplication Matches (N=30)	7q Duplications (N=4)	7q Duplication Matches (N=20)
DEMOGRAPHICS						
Chronological Age (yrs)	10.9[a, b]	10.5	9	8.6	9.7	9.6
	(2.4)	(3.1)	(4.9)	(4.1)	(4)	(3.6)
COGNITIVE/ADAPTIVE						
Verbal IQ	77.6	81.6	69	77.5	90.3	81.5
	(21.1)	(33)	(25.5)	(32.2)	(16.3)	(35.3)
Verbal Mental Age (yrs)	8.6	8.5	4.6	6.4	8.1	7.6
	(4.1)	(4.6)	(2.1)	(4.4)	(3.5)	(5.1)
Nonverbal IQ	81.1	84.1	82.7	83.5	82.8	82
	(15.9)	(24.5)	(22.6)	(22.7)	(14.2)	(27.8)
Nonverbal Mental Age (yrs)	8.3	8.6	6.1	6.9	7.2	7.5
	(3.1)	(3.8)	(1.9)	(3.6)	(2.6)	(3.7)
VABSII: Adap Beh Comp	71.8	73.5	70	75.3	74.3	70
	(5.8)	(11.5)	(7.6)	(11.1)	(8)	(12.2)
LANGUAGE/COMMUNICATION						
ADI-R Comm Verbal Total	14.1	17	17.5	15.8	14	16.4
	(5.9)	(4.4)	(3.4)	(4.7)	(2.4)	(6.2)
ADI-R Comm Nonverb Total	7.6	9.5	8.5	9.2	7	9.2
	(4)	(3.5)	(4.1)	(3.3)	(1.6)	(4.9)
VABSII: Comm Standard Score	74.5	77.3	77.8	78.5	77.3	73.4
	(7.4)	(13.1)	(10.4)	(15.7)	(9.2)	(11.8)
PPVT4: Standard Score	87.3	86.6	75.2	77.6	91	84
	(18.9)	(29.9)	(16.6)	(31.1)	(24.5)	(35.3)

PPVT4: Ratio IQ	88.7	89.2	64.5	78.1	87	87.3
	(40.3)	(41.1)	(24.6)	(30.3)	(32.9)	(44.6)
CTOPP: Standard Score	6	7.7	8	6.4	7.5	8.5
	(2.9)	(2.7)	(2.6)	(3.5)	(4.9)	(3.4)
SOCIAL						
VABSII: Social Standard Score	71.3	71.3	66.8	72.3	69.3	66.4
	(7.6)	(11.3)	(8.6)	(11.2)	(5.1)	(14.7)
REPETITIVE BEHAVIORS						
RBS-R Total	23.8	27.8	37.3	27.9	26.3	28
	(10.3)	(21.8)	(18.1)	(15.6)	(14.4)	(16)
GENERAL ASD SYMPTOMS						
SRS (Parent Report) T Score	81	78.1	85.7	78.3	85.5	84.6
	(15.9)	(10.3)	(6)	(10.5)	(5.3)	(8)
SRS (Parent Report) Raw Total	99.4	96.2	112.2	93.4	107.8	104.9
	(35.8)	(31.1)	(16.7)	(24.8)	(20.5)	(24)
BEHAVIORAL PROBLEMS						
ABC Stereotypy Subscale	3.4	4.9	6	4.3	6.5	4.4
	(3.7)	(4.2)	(4.3)	(3.6)	(3.3)	(4)
ABC Inapp. Speech Subscale	2.5	3.7	5	3.3	5.5	2.9
	(1.4)	(3.2)	(2.4)	(2.3)	(2.4)	(2.8)
CBCL Internalizing [c]	59.4	58.9	63	60.6	70.3	61.9
	(9.7)	(10.2)	(6.4)	(11.5)	(11.1)	(9.4)
CBCL Externalizing	62.6	53.5	64	56.7	63.8	59.7
	(10.1)	(10.6)	(6.4)	(10.8)	(7.1)	(10.1)
EARLY HISTORY						
Age First Words	2.5	2.1	2.1	2.1	1.6	2

	(1.5)	(1.3)	(0.7)	(1.4)	(1.1)	(1.9)
Age First Phrases	3.4	3	3.3	3.2	2.8	3.4
	(1.7)	(1.3)	(1.2)	(1.6)	(1.2)	(2.5)
GROWTH						
Height (cm) [d]	145.6	143.4	122.8	129.9	141.7	140.6
	(17)	(16.7)	(16.6)	(24.7)	(24.7)	(22.8)
Head Circumference (cm) [e]	56	55	*51.4*	*53.2*	54.8	54.5
	(2.6)	(2.5)	*(1.1)*	*(2.9)*	(2.1)	(3.1)
POSTHOC COMPARISONS (7q11.23 only)						
CBCL Anxious/Depressed					**69.7**	**59.9**
					(17.1)	**(9.4)**
CBCL Withdrawn					71.8	65.5
					(8.5)	(11.7)
CBCL Somatic Complaints					60	59.9
					(7.8)	(7.4)
CBCL Attention Problems					62.5	59.7
					(9)	(9.5)
CBCL Aggressive Behavior					67.3	61.9
					(11.9)	(9.1)
CBCL Affective Problems					65.3	64.7
					(14.9)	(9.5)
CBCL Anxiety Problems					71.5	62.6
					(7)	(10)
CBCL ADD/ADHD					66.5	64.4
					(5.2)	(8.6)
CBCL Total Problems					69	64.7

		(7.5)	(9.2)

Table S5: Exploratory phenotypic comparisons between subjects with 16p11.2 deletions, 16p11.2 duplications, and 7q11.23 duplications and matched probands.

[a] Mean and standard deviation (in parentheses) are shown for each measure.

[b] p-values were calculated using an F-test and are uncorrected; numbers in bold represent a p-value ≤0.05.

[c] All CBCL scores are reported as t-scores.

[d] height is reported having controlled for age

[e] head circumference is reported having controlled for height.

Type	Classification	Rate in Cases			Rate in Controls			Proband/Sibling ratio			P-value (uncorrected)[a]		
		SSC	AGP	SSC no DN	SSC	AGP	SSC no DN	SSC	AGP	SSC no DN	SSC	AGP	SSC no DN
All	**None**	4.38	4.27	3.17	3.34	3.59	3.26	1.31	1.19	0.97	0.09	0.012	0.889
Deletions only	**All**	1.93	1.36	1.26	1.25	1.08	1.24	1.54	1.26	1.02	0.87	0.008	0.197
Duplications only	**All**	2.46	2.91	1.91	2.08	2.51	2.03	1.18	1.16	0.94	0.09	0.072	0.459
CNV frequency													
All	**2-6x**	1.63	1.30	0.93	1.05	1.03	1.00	1.55	1.26	0.93	0.04	0.058	0.303
	1x	1.86	0.85	1.73	1.71	0.83	1.71	1.08	1.03	1.01	0.76	0.375	0.780
Deletions only	**2-6x**	0.74	0.68	0.34	0.29	0.43	0.29	2.54	1.57	1.17	0.13	0.004	0.992
	1x	0.97	0.38	0.72	0.69	0.30	0.68	1.40	1.26	1.05	0.73	0.036	0.298
Duplications only	**2-6x**	0.97	1.22	0.67	0.83	1.05	0.78	1.18	1.16	0.85	0.22	0.203	0.358
	1x	1.27	0.66	1.02	0.97	0.72	0.96	1.31	0.92	1.07	0.07	0.749	0.337
CNV size													
All	**30-500kb**	2.62	2.80	2.58	2.63	2.72	2.60	1.00	1.03	0.99	0.99	0.313	0.905
	>500kb	1.76	1.49	0.59	0.71	0.88	0.66	2.49	1.69	0.89	<0.001	0.005	0.254
Deletions only	**30-500kb**	1.06	1.05	1.05	1.11	0.85	1.10	0.96	1.24	0.95	0.17	0.004	0.105
	>500kb	0.86	0.30	0.22	0.14	0.23	0.13	5.98	1.32	1.63	<0.001	0.209	0.153
Duplications only	**30-500kb**	1.56	1.73	1.54	1.52	1.86	1.50	1.02	0.93	1.03	0.44	0.801	0.509
	>500kb	0.90	1.18	0.37	0.56	0.65	0.52	1.60	1.82	0.70	0.02	0.007	0.697

Table S6: Replication of Autism Genome Project (AGP) findings based on Simons Simplex Collection (SSC) data.

[a] p-values are calculated by Wilcoxon paired test in SSC; recorded as stated for AGP (Pinto et al., 2010)

	Band	Location	Size (kb)	Rarity (1% DGV)	CNVs (Del/Dup)		Fisher's (uncorrected)	Genes
					Proband	Sibling		
Probands	7p14.3	chr7:33,098,254-33,153,804	56	Rare	5 (0/5)	0 (0/0)	0.031	BBS9,RP9
	12q24.3	chr12:130,296,270-130,359,325	63	Rare	11 (11/0)	2 (2/0)	0.011	0
	15q13.3	chr15:29,816,096-30,226,051	410	Common	10 (2/8)	3 (0/3)	0.046	CHRNA7
	16p11.2	chr16:29,563,365-30,107,306	544	Rare	10 (6/4)	0 (0/0)	0.001	Multiple (26)
	17q21.31	chr17:41,532,917-41,710,573	178	Common	97 (0/97)	67 (1/66)	0.009	KIAA1267
	22q11.21	chr22:17,060,279-17,074,487	14	Common	7 (1/6)	1 (0/1)	0.035	0
Siblings	7q36.3	chr7:154,688,930-154,711,482	24	Rare	0 (0/0)	5 (5/0)	0.031	0
	10p11.21	chr10:38,778,547-38,779,644	1	Rare	1 (0/1)	8 (0/8)	0.019	LOC399744
	10q26.3	chr10:135,243,205-135,272,450	29	Common	0 (0/0)	5 (1/4)	0.031	0
	11q11	chr11:54,468,566-54,485,810	17	Common	1 (0/1)	8 (3/5)	0.019	0
	21q11.2	chr21:13,419,718-13,425,083	5	Common	0 (0/0)	5 (0/5)	0.031	0

Table S7: Genome-wide association of recurrent transmitted CNVs.

Table S8: All rare, high-confidence CNVs in samples passing quality control.

Submitted as a separate excel file.

Table S9: Details of all samples passing quality control.

Submitted as a separate excel file.

Supplemental Experimental Procedures

Identity quality control

Once the samples had been genotyped the program Plink was use to check for consistency in reported gender, detect Mendelian inconsistencies and to identify cryptic relatedness using an assessment of inheritance by descent (IBD). The Plink commands used were:

- plink --bfile <Samplefile> --check-sex
- plink --bfile <Samplefile> –-mendel
- plink --bfile <Samplefile> --extract <Hapmap_LD.prune.in> --mind 0.05 --geno 0.1 --maf 0.01 --hwe-all --make-bed --out <Samplefile.indep>
- plink --bfile <Samplefile.indep> --genome --min 0.05 --out <Sample.IBD.Result>

Where 'Hapmap_LD.prune.in' is a pre-defined list of independent SNPs to ensure consistency of results across samples of different sizes. This SNP list was derived from 120 Hapmap individuals with 1Mv1 Illumina data using the command:

- plink --bfile Hapmapfile indep-pairwise 50 5 0.2 --out Hapmap_LD.prune.in

Based on identity quality control 11 families were removed: the mothers did not match the recorded family structure for 2 families (12230, 11134); the mother and father that were swapped for 1 family (13054); one mother 's data was a duplicate of the child sample (11756); one mother's data was a duplicate of a sample from another family (11628); one mother's data was a duplicate of a father sample (11420); five families had samples with incorrectly assigned sex (11794, 12132, 12860, 12217, 12204).

CNV detection

The commands used to detect CNVs were:

- perl detect_cnv.pl -test -hmm lib/hhall.hmm -pfb lib/ho1v1.hg18.pfb -list <List_of_samples.txt> -log <Sample.log.txt> -output <Sample.results.txt>
- quantisnp --config config.dat --output <Sample.results.txt> --sampleid <SampleID> --gender <Gender> --emiters 10 --Lsetting 2000000 --maxcopy 4 --printRS --input-files <Sample.txt> 1M
- perl Combined_CNVv1.73.pl --cnv2 <List_of_samples.txt>

Analysis was automated and the results merged using the in-house script 'CNVision' (www.cnvision.org). QuantiSNP predicted CNVs with a Log Bayes Factor ≥10 were included in the analysis as suggested by the writers; all CNV predictions from PennCNV and GNOSIS were included in the analysis.

CNV Quality Control

Each of the CNV detection algorithms creates a quality control file to identify low quality data. We used the following thresholds to determine a failure:

- PennCNV: LogR standard deviation >0.28, BAF drift >0.01, Waviness factor (WF) deviating from 0 by >0.05.
- QuantiSNP: 90 out of 92 measures per sample being within the following ranges: BAF outliers <0.1, LogR outliers <0.1, BAF standard deviation <0.2, LogR standard deviation <0.4.
- GNOSIS: Quality score >10.

Two further quality control measures were used, Beadstudio call rate of ≥98.5% and an algorithm within CNVision developed to identify two causes of poor CNV prediction identified within the Illumina data: 1) excessively wide/wavy LogR values, 2) excessive numbers of probes with highly negative LogR values. The algorithm counts the number of probes with a logR deviating from 0 by >0.5 and the number of probes with a LogR <-1. Exclusion thresholds were set at 2 standard deviations for the sample population of the specific chip type. These thresholds were verified by visualization of 200 samples including 50 samples with the patterns described above.

A sample was excluded if it failed any of these five CNV quality control measures. 5% of samples were excluded based on these measures. Failing samples were rerun on chips with a 70% success rate of getting good quality data on the second run. If good quality data could not be obtained for the proband and both parents then the family was excluded. 39 families were removed following CNV quality control.

SYBR green qPCR for confirmation of CNVs

Quantitative polymerase chain reaction (qPCR) was used to determine the presence or absence of predicted CNVs in whole-blood DNA (where available, cell-line or saliva DNA used as an alternative in 2% of samples). To ensure consistent and reliable results the qPCR methods were defined and tested as outlined below.

SYBR green qPCR primer design

Repeat-masked and SNP-masked sequence was generated for each half of the predicted CNV with the aim of designing two non-overlapping amplicons per CNV. Primers were designed by Sigma to strict parameters to minimize primer dimerization and variation in amplification efficiency:

- Half of the CNV region, using only the 75% of sequence nearest the middle of the CNV
- ≤3 consecutive identical nucleotides
- ≤2 G or C nucleotides in the 3' terminal five nucleotides
- GC content 20-80%
- Amplicon size 110-150bp
- Primer length 20-25bp (optimal 23bp)
- Primer Tm 58-61°C (optimal 60°C)
- Amplicon Tm 75-85°C (optimal 80°C)

If primers could not be designed to these specifications a second round of design was undertaken with the same criteria as above except that:

- Half of the CNV region, using all the sequence
- Amplicon size 105-160bp
- Primer Tm 56-61°C (optimal 58°C)
- Attempt to design a single probe using the whole prediction as the input
- Add 160bp (the maximum length of the qPCR product) either side of the prediction

Finally if a primer could not be designed with these criteria then the entire length of the CNV plus 160bp on either end (the maximal length of the amplicon) was submitted for design keeping all the other parameters the same.

Primer pairs were excluded if:

- ≥1 PCR product identified using in-silico PCR
- Either primer aligned to ≥1 position within 2,000bp of the expected amplicon using BLAST

Two control primers were designed within 'house-keeping genes' in which no CNVs have been reported in the DGV or literature. In addition both control primers had been used for multiple SYBR green qPCR experiments for previous experiments and no CNVs had been identified. The controls primers are designed within the genes ZNF423 ('ZNF423Primer') and HMBS ('HMBSPrimer'). The sequences used are:

- ZNF423 – Forward 5'-AGATGATCGGAGATGGTTGTG-3'
- ZNF423 – Reverse 5'-GATCTGCTCGTGCCTCTTCAA-3'
- HMBS – Forward 5'-GGCTTCAGAAAAGGAGAGTGTCTGGT-3'
- HMBS – Reverse 5'-CCCTCCCTCCCCCAGCCATT-3'

All primers were synthesized by Sigma-Aldrich (www.sigmaaldrich.com). The pairs of synthesized primers were mixed at equimolar concentrations in a single 2D tube then this was divided into eight aliquots in 2D barcoded tubes using a Biomek® FX^P Laboratory Automation Workstation (Beckman Coulter). The control primers were synthesized alongside all other primers, however they were divided into 450 aliquots in 2D tubes.

SYBR green qPCR primer validation

To ensure that the primer pairs in the experiment have similar amplification efficiencies to the two sets of control primers the amplification efficiency of every primer was examined against pooled female control DNA (Promega; G152A). The correlation coefficient was calculated for each primer pair with control DNA dilutions of 1:4, 1:16, 1:64, and 1:256. PCR efficiency was calculated by plotting the Threshold cycle (CT) as a function of Log10 concentration of the template used. Primer pairs were eliminated if:

- PCR efficiency <90%
- Slope <-3.7 or >-3.1

- R2 < 99%
- Product Tm <75°C

SYBR green qPCR accuracy with relative quantification

Relative quantification is when the dilutional curve was run once per primer pair prior to the experiment during the primer validation step as opposed to absolute quantification in which the dilutional curve is run alongside each primer pair used on the same plate as the experiment. The advantage of relative quantification is that the time and cost of the experiment are reduced by a factor of three.

57 regions (22 deletions and 35 duplications) that had been previously confirmed using absolute quantification (see below) were used as true positives. All the CNVs identified for confirmation were rare in the general population (defined as ≤90% of the CNVs length overlapping a list of regions at >1% of the DGV), therefore the assumption was made that any change in estimated copy number seen in a pooled control (CT_Female, CT_Male or CT_Both) was a false positive; accordingly the number of true negatives (no change in copy number in these controls) was 4,452.

Based on these results the true positive rate and false positive rate were calculated based on the relative quantification data. These values were plotted as receiver operating characteristic curves (Figure S1) as the threshold for identifying a deletion or duplication was reduced or increased respectively from 1 (the expected value of estimated copy number for 2 copies). Based on these plots the threshold for detection of deletions was set as 0.7 and for duplications 1.2. At these thresholds the sensitivity was 86.4% for deletions and 77.1% for duplications; specificity was 92.9% for deletions and 77.5% for duplications.

SYBR green qPCR accuracy with absolute quantification

44 regions (17 deletions and 27 duplications) showing large (≥100 probes, ≥250kbp) with highly convincing plots on visualization were used as true positives. All the CNVs identified for confirmation were rare in the general population (defined as ≤90% of the CNVs length overlapping a list of regions at >1% of the DGV), therefore the assumption was made that any change in estimated copy number seen in a pooled control (CT_Female, CT_Male or CT_Both) was a false positive; accordingly the number of true negatives (no change in copy number in these controls) was 131.

The true positive rate and false positive rate were calculated based on this absolute quantification data. These values were plotted as receiver operating characteristic curves (Figure S1) as the threshold for identifying a deletion or duplication was reduced or increased respectively from 1 (the expected value of estimated copy number for 2 copies). Based on these plots the threshold for detection of deletions was set as 0.7 and for duplications 1.3. At these thresholds the sensitivity was 100% for deletions and 100% for duplications; specificity was 97.9% for deletions and 100% for duplications.

Estimating CNV prediction accuracy in transmitted CNVs

120 autosomal CNVs predicted by ≤50 Illumina probes and with ≤10% of their length overlapping a predefined list of regions in the Database of Genomic Variation (DGV) with a population frequency of >1% were selected at random from the list of total predictions in the initial 585 families (Figure 1) ensuring that all combinations of prediction algorithms were represented. The presence of the CNVs in the sample was assessed using absolute quantification SYBR green qPCR with triplicate reactions. Interpretable results were obtained for 115 and these are presented in Table S1.

Defining high-confidence CNVs

Based on the estimated accuracy of detection a CNV was determined as being 'high confidence' if it was detected by PennCNV, QuantiSNP, and GNOSIS with ≥50% of its length being present in two or more algorithms or if it was detected by PennCNV and QuantiSNP with ≥50% of its length being present in two or more algorithms. CNVs present in the X transposed region (XTR, defined as chrX:88,343,459-92,429,752) were removed since the similarity between chrX and chrY in this region causes a false positive duplication in all male samples and the region is common in the DGV. Using this high confidence threshold the positive predictive value was found to be 91% (Table S1).

Variation in CNV prediction with DNA source

7 samples (4 probands, 3 siblings) were genotyped using saliva DNA and 60 samples (29 parents, 20 probands, 11 siblings) were genotyped using cell-line DNA because whole blood DNA was not available. Cell-line and saliva DNA resulted in more autosomal CNV predictions than whole-blood DNA: 82.0 ±0.9 (mean ±95%

confidence intervals) CNVs with whole blood, 95.5 ±9.3 CNVs with cell-line, and 135.6 ±35.0 with saliva. However, this difference was minimised by using multiple algorithms to identify high-confidence (33.2 ±0.4, 35.0 ±4.2, and 39.7 ±6.5 respectively) and rare-high confidence predictions (13.5 ±0.2, 14.4 ±2.6, and 19.9 ±6.5 respectively). Ultimately no *de novo* events were confirmed in non-whole-blood DNA.

De novo CNV prediction

De novo CNVs were detected using an algorithm within CNVision to calculate the parental values of: 1) mean LogR of all the probes within the predicted proband CNV, 2) the percentage of probes with a LogR deviating from 0 by ≥50% in the same direction as the proband's probe, and 3) the number of probes within the CNV that are homozygous, heterozygous, consistent with a duplication or not in one of these categories. The *de novo* detection algorithm was trained using large confirmed *de novo* CNVs to determine appropriate detection thresholds.

Estimating CNV prediction accuracy in *de novo* CNVs

65 autosomal CNVs with ≤10% of their length overlapping a predefined list of regions in the Database of Genomic Variation (DGV) with a population frequency of >1% were selected at random from the list of 897 *de novo* CNV predictions in the initial 585 families (Figure 1). If qPCR probes could not be designed for the region an alternative region was selected at random. The presence of the CNVs in the sample was assessed using absolute quantification SYBR green qPCR with triplicate reactions. Interpretable results were obtained for 63 and these are shown in Figure S1. The positive predictive value (PPV) was found to be 13% <20 probes and 91% ≥20 probes.

High-throughput SYBR green qPCR *de novo* confirmation

While the PPV of *de novo* CNVs was only 13% <20 probes the number of predicted *de novo* CNVs beneath this threshold was high. To identify the true positive events a high-throughput qPCR experiment using relative quantification was performed to interrogate all 897 *de novo* CNV predictions.

High-throughput SYBR green qPCR *de novo* confirmation: samples

Whole blood DNA for the samples was obtained from the Biological Response Indicators Facility Core (The State University of New Jersey). The DNA was divided into five aliquots in 2D barcoded tubes using a Biomek® FX^{P} Laboratory Automation Workstation (Beckman Coulter). Samples of distilled water were put in 195 2D tubes.

High-throughput SYBR green qPCR *de novo* confirmation: controls

Laboratory control DNA for two samples (one male, one female) was run on Illumina 1Mv3 Duo microarrays and CNVs were predicted using CNVision as described earlier. Large, rare CNVs were identified and qPCR primers were designed using the same methods described earlier. A deletion CNV in the male sample (DNA: 'PosDelDNA', Primer: 'PosDelPrimer') and duplication CNV in the female sample (DNA: 'PosDupDNA', Primer: 'PosDupPrimer') were confirmed; neither CNV was present in the other sample. These samples and primers were used as positive and negative controls on every plate. The primers were synthesized by Sigma and each primer and sample was divided into 45 aliquots in 2D tubes.

Three further control samples were used: pooled female ('CT_Female', Promega G152A), pooled male ('CT_Male', Promega G147A) and pooled male and female ('CT_Both', Promega G304A). These control samples were run once against every primer pair used on each plate.

The following 18 controls were run on every plate to ensure accurate results and identify possible contamination (shown as Sample:Primer – reason for control):

- PosDelDNA:PosDelPrimer – Positive control (deletion)
- PosDelDNA:PosDupPrimer – Negative control (duplication)
- PosDelDNA:HMBSPrimer – Reference1 for positive control (deletion)
- PosDelDNA:ZNF423Primer – Reference2 for positive control (deletion)
- PosDupDNA:PosDelPrimer – Negative control (deletion)
- PosDupDNA:PosDupPrimer – Positive control (duplication)
- PosDupDNA:HMBSPrimer – Reference1 for positive control (duplication)
- PosDupDNA:ZNF423Primer – Reference2 for positive control (duplication)
- CT_Female:HMBSPrimer – Reference1 for CT_Female
- CT_Female:ZNF423Primer – Reference1 for CT_Female

- CT_Male:HMBSPrimer – Reference2 for CT_Male
- CT_Male:ZNF423Primer – Reference2 for CT_Male
- CT_Both:HMBSPrimer – Reference1 for CT_Both
- CT_Both:ZNF423Primer – Reference2 for CT_Both
- Water:PosDelPrimer – Negative control for PosDelPrimer
- Water:PosDupPrimer – Negative control for PosDupPrimer
- Water:HMBSPrimer – Negative control for HMBSPrimer
- Water:ZNF423Primer – Negative control for ZNF423Primer

A further four controls were run once for every experimental primer pair (ExpPrimer1 in this example, shown as Sample:Primer – reason for control):

- CT_Female:ExpPrimer1 – Female control
- CT_Male:ExpPrimer1 – Male control
- CT_Both:ExpPrimer1 – Pooled control
- Water:ExpPrimer1 – Negative control for ExpPrimer1

Each sample (ExpSample1 in this example) was run against the two reference controls and the experimental primer (ExpPrimer1 in this example, shown as Sample:Primer – reason for control):

- ExpSample1:ExpPrimer1 – Reaction at site of predicted CNV
- ExpSample1:HMBSPrimer – Reference1 for ExpSample1
- ExpSample1:ZNF423Primer – Reference2 for ExpSample1

High-throughput SYBR green qPCR *de novo* confirmation: data generation

26,206 2D barcoded tubes containing primers, DNA samples or water were generated and stored on barcoded plates. 10-15 plates of 96 sample tubes were generated each day; samples were taken from all 96 2D tubes simultaneously using a Biomek® FX^P Laboratory Automation Workstation (Beckman Coulter) with a 96-well head. The 96 samples were dispended onto a 384-well plate four times (with the top left well being A1, A2, B1, B2 in turn) so that four identical sets of wells were generated from each tube. This process was then repeated with the primer plate. This resulted in every reaction being conducted in quadruplicate.

Up to seven combinations of sample and primer were analyzed on each 384-well plate. All members of the family were analyzed together to ensure consistency of result.

A perl script was written to allocate samples and primers into the optimal configuration on 96-tube plates. This script monitored the location and barcode of all samples and primers within the project to ensure that no tube was required more than once at the same time. Once the optimal plate arrangement was calculated for all 96-well plates being processed that day, a series of XL20 input files were generated. These files were used to control the BioMicroLab XL20 Tube Handler to dispense the samples and primers into the correct configuration; the barcode of every tube was checked.

Once the 96-well sample plate and 96-well primer plates had been dispensed in quadruplicate onto the 384-well qPCR plate the following reaction conditions were achieved: 10 µl total volume, 1.75ng of whole-blood derived genomic DNA, 1X Power SYBR Green PCR master mix (Applied Biosystems, Foster City, California, USA), 400nM of the forward and reverse primers. The plates were run using the ABI Prism 7900 high-throughput sequence detection system; Applied Biosystems' standard thermal cycling parameters were used throughout.

All fluid-handling, sample rearraying, and qPCR reactions were conducted by from the Biological Response Indicators Facility Core (The State University of New Jersey). The sample and primer allocation to each plate was conducted remotely at State Lab, Yale University. The data generated were sent to State Lab for analysis in the form of SDS files uploaded to 'YouSendIt'.

High-throughput SYBR green qPCR *de novo* confirmation: data analysis

The qPCR data were analyzed using the relative quantification method: the sample and primer of interest are compared to the calibrator (the same sample with a control primer i.e. HMBS or ZNF423) to compare the number cycles taken to achieve a specific intensity of SYBR green signal. The difference in PCR efficiency

between primers is normalized using the controls (CT_Female, CT_Male and CT_Both) which were run against both sets of primers.

Data were initially analyzed using SDS 2.3 (Applied Biosystems). The specificity of the PCR products was confirmed by visual inspection of the melting curve. Reactions with melting peaks at annealing temperatures other than expected for the reaction were removed from the analysis; this includes reactions with more than one melting peak. The threshold cycle (C_T) values of each sample were exported into text file for further analysis.

A custom perl script was written to merge the plate layouts generated earlier (when allocating the samples and primers and making XL20 input files) and the text file containing the C_T values from the experiment. Having identified the four reactions (unless any had been removed following visual inspection) with identical samples and primers the script calculated the standard deviation of the C_T values. If the standard deviation of the CT values was >0.3 then the reaction with a C_T value the furthest from the mean of the C_T values was removed. This outlier removal process was continued until the standard deviation was ≤0.3 or only one C_T value was left. The output was printed to a text file for further analysis.

A further custom perl script was used to calculate the estimated copy number from the C_T values. The text file from the merging of the plate layout and the C_T values was used as the input file. Copy number was calculated using the comparative C_T method ($\Delta\Delta C_T$) (Livak and Schmittgen, 2001). The formula used was given below:

Estimated copy number = $2^{(-\Delta\Delta CT)}$

Where:

- $\Delta\Delta C_T$ = (C_T Region:Sample – C_T Ref:Sample) – (C_T Region:Control – C_T Ref:Control)
- C_T Region:Sample = mean C_T values for the region of interest and sample of interest (e.g. ExpPrimer1 and ExpSample1)
- C_T Ref:Sample = mean C_T values for the reference region and sample of interest (e.g. HMBSPrimer and ExpSample1)
- C_T Region:Control = mean C_T values for the region of interest and the control sample (e.g. ExpPrimer1 and C_T_Female)
- C_T Ref:Control = mean C_T values for the reference region and the control sample (e.g. HMBSPrimer and C_T_Female)

Using the equation above the expected copy number estimates are:

- 0 for a homozygous deletion (0 copies)
- 0.5 for a hemizyous deletion (1 copy)
- 1 for regions of normal copy number (2 copies)
- 1.5 for a duplication (3 copies)

In view of the comparatively low estimated specificity of qPCR with relative quantification and the large number of prediction we elected to confirm all 86 CNVs identified as *de novo* with absolute quantification qPCR. 58% (N=50) of regions were ultimately confirmed in 46 samples.

High-throughput SYBR green qPCR *de novo* confirmation: results

5.5% of probands (n=32 out of 585) were found to have at least one rare *de novo* CNV compared with 2.4% of designated siblings (n=14 out of 585) giving an odds ratio of 2.2 (p=0.005, Fisher's exact test). 53% of *de novo* predictions based on ≥20 probes (N=94) were confirmed compared with 2.6% with <20 probes (N=430). 82% of failures were false-positive predictions in offspring, 18% were false-negatives in parents. The comparatively low PPV <probes compared with the value identified previously (2.6% vs. 13%) was attributed to the larger proportion of very small *de novo* CNVs in the high-throughput experiment (Figure S1).

Refining *de novo* CNV confirmation in light of high-throughput qPCR results

Given a large number of predictions, and the low yield of true positives with small numbers of probes (Figure S1), we elected to restrict further analysis of *de novo* events to those encompassing ≥20 probes. Furthermore the results were used to improve the prediction thresholds of the *de novo* algorithm within CNVision, specifically: exclusion of all regions with a probe density (size / number of probes) >5,000; more stringent LogR threshold for deletions and duplications <100 probes; addition of standard deviation of LogR threshold for deletions and duplications <100 probes; more stringent threshold for number of probes showing a LogR

deviation similar to the proband. All *de novo* events reported in this study were then confirmed by qPCR using absolute quantification.

Parent-of-origin for *de novo* CNVs

The parental chromosome from which the *de novo* CNV arose was calculated using the B allele Frequency to identify informative SNPs. The parent-of-origin could be determined in 78 out of the 83 *de novo* CNVs; the other CNVs lacked informative SNPs. Of the 76 autosomal *de novo* CNVs in whom parent-of-origin was determined 35 (46%) were of paternal origin and 41 (54%) were of maternal origin; this is not significantly different from the expected value of 50% from each parent (p=0.57, binomial test, two-tailed). This is consistent with previous findings in which no difference in parent-of-origin was seen for 47 *de novo* CNVs in multiplex autism families (Itsara et al., 2010).

Mechanism of *de novo* CNVs

The mechanism of origin was determined by comparing the ends of the CNV to regions of segmental duplication. In probands 36% (24 out of 67) were most likely a result of non-allelic homologous recombination (NAHR) compared with 38% (6 out of 16) in siblings. One *de novo* CNV was caused by an isodicentric chr15 and two CNVs in one proband were caused by an unbalanced translocation.

Individuals with more than one *de novo* CNV

We detected four individuals (three probands, one sibling) with more than one *de novo* CNV. The proband of family 11435, has a 1.3Mb telomeric deletion at 16p13.3 and a 3.2Mb telomeric duplication at 9p24.2-p24.3, suggestive of an unbalanced translocation; this was confirmed by the presence of a derivative chromosome 16 on FISH analysis. The proband of family 13036 carries a 5.3Mb duplication of chromosome 20q directly adjacent to the centromere. This finding suggests the presence of an additional derivative chromosome derived from chromosome 20; confirmation by FISH analysis is in progress. This proband also has two pericentromeric duplications involving the p and q arm of chromosome 6. These regions may reflect a single *de novo* duplication spanning the centromere in which the probe density is poor. The proband of family 12330 and the sibling of family 12331 (the consecutive family IDs is co-incidental) each individual carry two *de novo* CNVs in close proximity on the same chromosome, show the same copy number change, and originate on the same parental chromosome. In view of these factors they are likely to represent a single complex rearrangement in both families. Of note the proband of family 12430 has two very similar overlapping duplications to those seen in 12330 suggesting a similar complex rearrangement. Overall of the four individuals with more than two observed *de novo* CNVs only one (13036) is likely to represent two independent *de novo* events. This conclusion does not change the results in the paper since the analyses are based on the number of samples with at least one CNV, the number of genes within CNVs, or the distribution of events in siblings.

Quantification of CNVs using Digital Array

The observation of 14 CNVs at 16p11.2 in probands and 0 in siblings (Figure 4) is striking. In order to verify that 16p11.2 CNVs were not being missed in siblings we used The Fluidigm® 48,770 Digital Array Integrated Fluidic Circuit (IFC) in 1,498 samples: 650 matched probands and siblings plus an additional 198 probands from trio families (a subset of the samples described in this study). Two other regions with recurrent *de novo* CNVs were also analyzed in a similar manner using the same sample set: *CDH13* and 16p13.2. These regions were chosen to be representative of the wider set of *de novo* CNVs, in that they included deletions and duplications, CNVs across a wide size range (34kb to 5.3Mb), and sufficient true-positives to inform experimental design. The region of interest was represented by three probes for each region and ordered from pre-designed Taqman copy number assays (ABI):

- *CDH13* exon2: Hs01319842_cn (chr16:82,891,971)
- *CDH13* exon5: Hs01582788_cn (chr16:83,250,968)
- *CDH13* exon13: Hs01802449_cn (chr16:83,816,948)
- 16p11.2: Hs02301463_cn (chr16:29,845,140)
- 16p11.2: Hs01712568_cn (chr16:29,922,447)
- 16p11.2: Hs03006732_cn (chr16:30,128,276)
- 16p13.2: Hs01167560_cn (chr16:9,017,087)
- 16p13.2: Hs00799207_cn (chr16:9,197,064)
- 16p13.2: Hs02242470_cn (chr16:8,962,123)

Briefly, a 4 μL aliquot of master mix containing rehydrated DNA was transferred into each sample inlet on the digital array. The final reaction mixture (4 μL) consisted of 20 ng DNA, 1X Taqman gene expression master mix (ABI), 1X RNaseP Cy5-Taqman®, 1X Taqman assay for the gene of interest (900 nM primers and 250 nM probe) using both FAM- TaqMan® and VIC- TaqMan®, and 1X GE sample loading reagent (Fluidigm). Probands and siblings from the same family were analysed on the same array in neighbouring inlets all cases.

Each of the mixtures was injected into separate inlets on the chip. Digital Array Integrated Fluidic Circuits are placed on the NanoFlex Integrated Fluidic Circuit Controller to load the sample mixture into reaction chambers. Then, thermal cycling and fluorescence detection was generated on the BioMark™ Real Time QPCR system using default thermocycling conditions: 95°C, 10 min hot start, followed by 40 cycles of two-step PCR (15 s at 95°C for denaturing and 1 min at 60°C for annealing and extension).

The BioMark™ Digital PCR Analysis Software is used to analyze the number of reaction chambers that are positive for the gene or genes of interest by counting amplifications from single molecules. The positive chambers are then partitioned into bins based on their intensity and mean DNA concentration is estimated using Poisson probabilistic analysis. The relative copy number of the gene of interest in the sample was calculated by the ratio of the signals from the target gene and the reference gene (Qin et al., 2008).

All expected CNVs were observed (N=19) while no additional CNVs were identified. This matches the hypothesis that the Illumina 1M microarrays are extremely sensitive at detecting large *de novo* and transmitted CNVs.

Population structure of recurrent *de novo* CNVs

To evaluate ancestry, all parents are projected onto a five-dimensional ancestry map using eigenvector decomposition (Crossett et al., 2010; Lee et al., 2009). Then Euclidean distances were measured for the parent-of-origin of 16p11.2 or 7q11.23 *de novo* or transmitted events. The mean and median distance between these pairs of parents were calculated. To determine if these parents were unusually close in the ancestry space, we randomly sampled sets of N parents 500 times from all parents (e.g., N=16 for all 16p11.2 *de novo* events and N = 4 for 7q11.23 duplications), and calculated the average and median distances for each sampling. In this way a bootstrap distribution under random sampling is developed, and the summary statistics for the true sample can be compared to the bootstrap distribution to obtain an approximate p-value.

For all ancestries and all 16p11.2 *de novo* CNVs, we found approximate p-value for mean is 0.988 and for the median it is 0.998. We conclude that these 16 parents do not show unusually similar ancestry. Next, to focus on a more homogeneous sample, we limited our analysis to parents of European ancestry (N = 13 transmitting parents). These parents also show no evidence of unusual ancestral clustering. Again their average distance was almost always larger than a random sample (p-value for mean, 0.836; for median, 0.586). Similar results were obtained when we examined the data separately for parents transmitting 16p11.2 deletions and duplications. The parents transmitting 7q11.23 *de novo* duplications also did not show unusual clustering with respect to ancestry (p-value for mean is 0.51 and for the median it is 0.63); all were of European ancestry.

Genotype-Phenotype analysis of 16p11.2 and 7q11.23

For each subject with a 16p11.2 deletion (N=8), 16p11.2 duplication (N=6) and 7q11.23 duplication (N=4), five other probands were selected based on matching criteria. Matching was performed in a hierarchical manner starting with age (within 2-years of index case if under 8 years of age, within 3 years if aged 8-12 and within 4 years if older than 12 years of age), then gender, genetic distance (based on the five-dimensional ancestry map), site of recruitment, and finally whether the sample was from a quad or trio family. All samples with *de novo* CNVs or CNVs in regions previously associated with ASD were removed prior to the matching; each proband was only allowed to act as a matched control for one sample. Of the 90 samples selected as matched probands the extent of the matching was as follows: 100% could be matched by age and gender; 89% could be matched based on a pre-specified neighborhood of genetic similarity; 50% could be matched by site; and 39% could be matched to the quad/trio origin.

For continuous variables, matching was taken into account by treating each stratum of a "case" proband matched to 5 "control" probands as a block, and the data analyzed as a randomized block design by using analysis of covariance. Thus mean values were allowed to vary across blocks and to be altered by case-control status; the test of interest, difference due to de novo status (yes or no), is an F-test with N, M degrees-of-freedom (N is the number of de novo events of interest and M is the residual degrees-of-freedom after accounting for model terms) . Because IQ is known to affect many behavioral measures associated with ASD, it was treated as a covariate in models for outcomes besides itself and Body Mass Index (BMI). For diagnostic status, matching was taken into account by using a conditional logit model.

Studies of subjects with 16p11.2 deletions, 16p11.2 duplications and 7q11.23 duplications have reported a wide range of phenotypes (Bochukova et al., 2010; Hanson et al., 2010; Rosenfeld et al., 2010; Shinawi et al., 2010). Our primary analyses of probands with a 16p or 7q event target four features (Table 2), namely full-scale IQ, severity of autism, diagnosis and body mass index (BMI), the latter motivated by observations that 16p deletions (Bijlsma et al., 2009; Walters et al., 2010) and duplications (Reymond et al., 2010) tend to have opposite impact on BMI. In addition to these primary analyses, we conducted wider, exploratory analyses (Table 2) that must be interpreted in light of more extensive multiple testing; in total, 33 exploratory analyses were performed, 10 of which were reported in Table 2 and the remainder in Table S5, which also contains results from 9 post-hoc analyses of phenotypes for 7q duplication carriers.

For each case proband carrying one of the three distinct *de novo* events, five other probands (controls) were selected based on hierarchical matching criteria: first age, then gender, genetic distance (based on 'five dimensional' ancestry map), ascertainment site, and whether the sample was from a quartet or trio. Case probands did not differ significantly from the control probands for three of the four target features (Table 2). Carriers of 7q11.23 duplications tended to be have a less severe presentation, as measured by the ADOS-CSS, while a somewhat greater proportion of carriers of 16p11.2 duplications met CPEA diagnosis criteria for Asperger Syndrome rather than with strict autism. Results for BMI were interesting in the sense that case probands with 16p11.2 deletions tended to have higher than average BMI, but their means were not significantly different from that of their matched controls. In contrast, BMIs of case probands with 16p11.2 duplications were significantly lower than their matched controls. When we treated copy number of the 16p11.2 region as an ordinal variable (1, 2, and 3 copies), and used the matched controls proband as the diploid sample, BMI significantly diminished as copy number increased ($r = -0.24$, $p = 0.03$).

A few features stood out in our exploratory analyses (Table 2). Case probands carrying 16p11.2 deletions differed notably (i.e., $p < 0.05$) from matched controls – without correction for multiple testing – only for age at first concern, with parents of these case probands tending to recall later ages of first concern than parents of control probands. The analyses also suggest that case probands with 16p11.2 deletions had somewhat less severe social impairments and fewer restricted and repetitive behaviors (as indicated by lower scores on the ADI-R Social Interaction and ADOS RRB domains, respectively). In comparison to matched control probands, case probands carrying 16p11.2 duplications had greater behavioral problems, as indicated by notably higher scores on the Aberrant Behavior Checklist (ABC) Hyperactivity subscale and somewhat higher ABC total scores. Case probands with 16p duplications also tended to demonstrate more social/communication impairment during a standardized observational assessment (ADOS Social + Communication and ADOS Social Affect.)

Case probands carrying 7q11.23 duplications also had more behavioral problems than their matched control probands, as evidenced by notably higher ABC total and ABC Irritability subscale scores and somewhat higher scores on the ABC Lethargy/Social Withdrawal subscale. In contrast, case probands with 7q11.23 also had notably lower ADOS Social + Communication totals, reflecting less severe social and communication impairments. They also tended to demonstrate fewer restrictive and repetitive behaviors (ADOS RRB) and less severe ASD-related behaviors, as indicated by the ADOS CSS.

Like BMI, some features from Table 2 showed a correlation with copy number at 16p11.2. Thus we explored the correlations between features in Table 2 with copy number, after accounting for IQ. Repetitive and Restrictive Behaviors, whether measured by the ADI-R (0.31, $p = 0.01$) or ADOS (0.45, $p = 0.009$) showed substantial correlation with copy number at 16p11.2, with increases in RRBs observed with increasing copy number. Scores on the ABC Hyperactivity subscale also tended to increase with copy number ($r=0.35$, $p = 0.08$).

Recurrent rare structural variations at 15q11.2-13

While CNVs are common within the segmental duplication, CNVs spanning the breakpoints are rare (with the exception of BP1-BP2). Several recurrent, rare CNVs in this region have specific names (Figure 5):

- **Class 1 interstitial deletion:** 6.5 Mb deletion at BP1-BP3 associated with Angelman syndrome (maternal deletion, must involve UBE3A (Knoll et al., 1989)) and Prader-Willi syndrome (PWS, paternal deletion, unclear whether a single gene is causative or if several paternally imprinted genes are required: MKRN3, MAGEL2, NDN, SNURF-SNRPN and snoRNAs (Hogart et al., 2010)).
- **Class 2 interstitial deletion:** 5 Mb deletions at BP2-BP3 also associated with Angelman syndrome and PWS. Of note the phenotype in PWS is less severe than in class I interstitial deletions (Butler et al., 2004) suggesting that the four genes in BP1-BP2 (TUBGCP5, CYFIP1, NIPA2, NIPA1) may play a role.

- **Class 1 interstitial duplication:** 6.5 Mb duplication at BP1-BP3. Maternal duplications are strongly associated with a neurodevelopmental phenotype including ASD in >85% of cases (Cook et al., 1997). The phenotype in paternally derived duplications is more variable, often with no neurodevelopmental abnormality. Cases of ASD have been reported with paternally derived duplications but may be co-incidental (Hogart et al., 2010). An atypical duplication within this region in a patient with ASD implicated the genes ATP10A and GABRB3 (Weiss et al., 2008). Numerous linkage studies support the role of GABRB3 (Grafodatskaya et al., 2010).
- **Class 2 interstitial duplication:** 5 Mb duplication at BP2-BP3 with a similar phenotype to Class 2 duplications.
- **Class 3B isodicentric chr15:** derivative chromosome 15 with two centromeres and two extra copies of BP1-BP3 (tetrapoloidy). It has a similar phenotype to Class 1 interstitial duplications but tends to be more severe consistent with the higher number of copies (Battaglia et al., 1997).
- **Class 5A isodicentric chr15:** derivative chromosome 15 with two centromeres and two extra copies of BP1-BP4 (tetrapoloidy) and a single extra copy of BP4-BP5 (trisomy). The phenotype is similar to Class 3B isodicentric chr15.
- **15q13.3 deletion:** 1.5 Mb deletion at BP4-BP5 (including CHRNA7) which may be associated with Intellectual disability (ID). Initially described in 6 unrelated ID samples out of 1,797 cases (0.3%) vs. 1 out of 960 controls (0.1%) (Sharp et al., 2008). In a case series of 18 deletions, 16 were associated with a degree of ID and one also had autisitic behaviour; of note at least 10 deletions were inherited from parents with no evidence of ID (van Bon et al., 2009). Three studies have looked at deletions in clinical samples sent for microarray: (Miller et al., 2009) reported 5 deletions out of 1,445 unrelated samples (0.35%); 2 showed autistic features (1 ASD, 1 PDD-NOS) and the other three had ID; (Ben-Shachar et al., 2009) reported 14 children and 6 parents with deletions out of 8,200 samples (0.24%), 12/14 had ID and 6/14 had ASD. (Shen et al., 2010) reported 2 deletions out of 993 ASD samples (0.20%); both were diagnosed with PDD-NOS. About a third of deletions in which the inheritance is determined are found to be *de novo*.
- **15q13.3 duplication:** 1.5 Mb duplication at BP4-BP5 (including CHRNA7) which may be associated with ID and/or ASD. A case series of 4 duplications found ID in all subjects; 2 had autistic behaviour (1 PDD-NOS, one ASD-like) (van Bon et al., 2009). (Miller et al., 2009) reported 1 *de novo* duplication and 1 maternally inherited duplication out of 1,445 unrelated samples submitted for clinical testing for aCGH; one had autism, the other language delay. In 751 AGRE families they found a further *de novo* duplication (autism) and an inherited atypical duplication (not mediated by NAHR and involving on the proximal end: MTMR15, MTMR10, TRPM1) in two siblings patients with autism. (Shen et al., 2010) reported 2 *de novo* duplications out of 993 ASD samples (both diagnosed as autism).
- **15q13.3 microdeletion:** 680 kb deletion between BP5 and the segmental duplication between BP4 and BP5 (including CHRNA7). Deletions in this region are present at about 0.5% in the DGV, however no such deletions were seen in 3,699 controls and at a rate of 0.03% in samples sent for aCGH for clinical purposes. A degree of ID was present in 9 out of 10 samples in four families. No *de novo* deletions were detected (Shinawi et al., 2009).
- **15q13.3 microduplication:** 360-680 kb duplication between BP5 and the segmental duplication between BP4 and BP5 (including CHRNA7). They are assigned to classes 1-5 depending on the location and extent of the duplication and reciprocal deletions/duplication (Szafranski et al., 2010). Duplications in this region are present at about 1% in the DGV and were found at a rate of 0.6% in samples sent for aCGH for clinical purposes. 55 such duplications were seen; of 11 that were phenotypically characterised a range of developmental delay, ASD, and speech delay was reported. No *de novo* duplications were detected (Szafranski et al., 2010).

15q11.2-3 CNVs and Schizophrenia:

- 15q13.3 deletion (BP4-BP5): identified as a risk factor for schizophrenia with 7 out of 7,951 cases, 8 out of 33,250 controls; 0.09% vs. 0.02% (Stefansson et al., 2008), and 9 out of 3,391 cases and 0 out of 3,181 ancestry-matched controls; 0.27% vs. 0% (ISC, 2008).
- BP1-BP2 deletion: identified as a risk factor for schizophrenia (26 out of 7,951 cases, 79 out of 33,250 controls; 0.33% vs. 0.24%) (Stefansson et al., 2008); deletions seen at 0.5% in DGV.

- CHRNA7: alongside the 15q13.3 deletions described above, 11 studies have shown linkage intervals in this region and two have shown association (Leonard and Freedman, 2006); there are several negative linkage and association studies for this region too.

15q11.2-3 CNVs and Epilepsy:

- 15q13.3 deletion (BP4-BP5): identified as a risk factor for idiopathic generalized epilepsy with 12 out of 1,223 cases, 0 out of 3,699 ancestry-matched controls; 0.98% vs. 0% (Helbig et al., 2009); 3 of the cases had a degree of ID, none had ASD. In a separate study 7 deletions were seen out of 539 cases while none were found in 3,777 controls; 1.3% vs. 0%; no cases had ASD. (Dibbens et al., 2009).
- CHRNA7: Linkage studies in epilepsy (Elmslie et al., 1997; Neubauer et al., 1998) show a linkage interval near CHRNA7. A knockout mouse model that has an abnormal EEG (Orr-Urtreger et al., 1997).

Transmitted CNV Burden

CNV burden in transmitted CNVs was assessed for regions previously associated with ASD (Figure 5). The regions used were those identified as 'ASD implicated' in Supplementary Table 9 of Pinto et al., 2010. The lists for intellectual disability (ID) and ASD candidates were also tested; the results were not significant. Brain-expressed genes were determined as those regions with ≥10 overlapping reads of RNASeq data derived from adult temporal lobe (unpublished data); this definition included 84% of RefSeq genes.

To limit the risk of an arbitrary threshold for rare CNVs influencing the result burden analyses were routinely performed using all combinations of four populations to identify common CNVs (DGV, all parents, fathers, mothers) and multiple thresholds (1%, 0.5%, 0.1%). Furthermore rarity was also calculated separately for deletions and duplications in each of these populations and thresholds. The results were robust to the definition of rarity.

AGP burden analysis

The SSC data were treated in the same manner as the AGP data (Pinto et al., 2010) to determine if the same trends were present. The list of high confidence CNVs was restricted to those with ≥5 probes and ≥30kb length. All CNVs >50% overlap with a segmental duplication region were removed. From the remaining CNVs, a list of common regions showing CNVs in ≥1% of the population was defined. A CNV was defined as rare if it ≥50% of its length was not present in this common regions.

The rare CNVs were annotated to define the CNV frequency at that locus. A CNV was defined as a single occurrence if ≥50% of its length was not seen in any other cases or controls. A CNV was defined as being seen at 2-6x frequency if ≥50% of its length was seen in a set of regions present at a frequency of 2-6x in the cases and controls. This annotation process was performed separately for all CNVs together, deletions, and duplications. Genes were defined by the coding transcription start and stop sites from the RefSeq gene list plus 10kb on either end. A CNV was counted as containing the gene if there was any overlap between the CNV and the gene. The results are shown in Table S6.

Genome-wide association of recurrent transmitted CNVs.

All high-confidence autosomal CNVs in 872 matched probands and siblings were examined to identify regions of enrichment**.** The CNVs were not filtered by frequency (common CNVs were included) or inheritance (both transmitted and *de novo* CNVs were included). The results were restricted to 3,677 regions that were that represented ≥10 SNPs on the Illumina 1Mv1 array to limit the detection of enrichment as a product of inaccurate CNV boundary detection. A Fisher's exact test was performed on these regions to identify regions of genome-wide enrichment. 11 regions had a p-value ≤0.05 prior to correction for multiple comparisons; however none of these regions are significant after correction (Table S7). The regions identified in probands tended to be larger (p=0.03, Wilcoxon test) and have more genes (p=0.13, Wilcoxon test). The regions 16p11.2 and 15q13.3 are also identified by the recurrent *de novo* analysis. The region of enrichment at 7p14.3 includes the gene *Bardet-Biedl Syndrome 9* (*BBS9*); mutations in this gene can contribute to Bardet-Biedel syndrome (retinal dystrophy, polydactyly, mental retardation, and mild obesity) which exhibits recessive inheritance with a modifier of penetrance (Burghes et al., 2001).

Calculations for association of rare, recurrent, *de novo* CNVs.

Synopsis: We observe a much higher rate of *de novo* events in probands (51/872) than siblings (15/872), making it reasonable to conclude that the probands carry events affecting risk. However given that we also see *de novo* events in unaffected siblings, the *de novo* events in probands must be a mixture of risk and neutral CNVs. A reasonable estimate of the fraction affecting risk is (51-15)/51 or 71%, but which ones are the risk

CNVs? The distribution of *de novo* CNVs in siblings gives us a way to disentangle risk from neutral CNVs in probands because it tells us how the neutral CNV recur in the genome. We use this distribution to estimate the number of effective CNVR, or eCNVR, the number of sites available for placement of *de novo* CNVs. Then we use the estimated rate of neutral events, the number of probands evaluated, and the number of eCNVR to predict the nature and number of neutral multiplicities. In other words, how often are neutral CNVs recurrent and how many occur at each eCNVR. Based on that distribution we obtain a cut-off beyond which a larger number of recurrent *de novo* events found in probands almost surely comprise a set of risk CNVs. Notice that the threshold established is a genome-wide threshold, akin to the genome-wide significant threshold for linkage analysis (Lander and Kruglyak, 1995).

Overlapping samples in literature-based analysis of recurrent *de novo* CNVs

To identify recurrent *de novo* CNVs in idiopathic ASD we searched for large-scale CNV investigations meeting four criteria: standardized diagnosis, genome-wide detection, confirmed *de novo* structural variations, and sufficient information to permit the identification of duplicate samples. Four such studies were identified (Itsara et al., 2010; Marshall et al., 2008; Pinto et al., 2010; Sebat et al., 2007). Between these studies there were six CNVs identified by more than one study (Table S4), assessed on the basis of location, sample ID, and descriptions in the supplementary materials.

While it was possible to identify samples with *de novo* CNVs present in multiple studies, it was not possible to identify if the samples without *de novo* CNVs were similarly present in multiple studies because non-ambiguous sample identifiers were not available for all studies.

In Table S1 CNV frequencies were based on the total sample size reported for these studies, without considering whether a sample was present in more than one of the studies. The total estimate of samples assessed (1,932 simplex and 1,884 multiplex, giving 3,816 in total) therefore included six samples with *de novo* CNVs and an unknown number of samples without *de novo* CNVs that were present in more than one study; i.e. the total number of samples we report could be a slightly inflated estimate.

In Table S4, and for the analysis of recurrent *de novo* events, *de novo* CNVs that were reported in more than one sample – but were plausibly the same sample – have been listed only once. However, the estimate of the total number of samples assessed included the unknown number of samples without *de novo* CNVs that were reported on in more than one study. This results in a more conservative estimate of recurrent significance because the increased sample size produces a high estimate of "d" (see birthday problem), specifically d=67 for all *de novo* CNVs. Given that 6 out of 143 (4.2%) of *de novo* CNVs in other studies were reported more than once, it is reasonable to assume that about 4.2% of the samples without *de novo* events were also reported in more than one study, or equivalently about 107 samples out of the 2,549 were duplicates. This gives a revised total estimate of samples of 3,709. Using this estimate leads to a corrected value of d=64, but it is not a meaningful difference in the estimate.

Unseen species problem

To determine if it is unusual to find multiple *de novo* CNVs at the same location we first estimated how many likely positions were available for placement of the observed *de novo* CNVs, i.e. the number of eCNVRs. To estimate this quantity we use methods from the so-called "unseen species problem". This approach uses the frequency and number of observed CNV types (or species) to infer how many species are present in the population, including those yet to be observed. Based on the observed *de novo* CNVs in the sibling group, we apply a formula for calculating the number of species (C). $C = c/u + g^2*d*(1-u)/u$, in which: c = the total number of distinct species observed; c_1 = the number of singleton species; d = total number of CNVs observed; g = the coefficient of variation of the fractions of CNVs of each type, and $u = 1 – c1/d$ (Bunge and Fitzpatrick, 1993). In the calculations presented in this manuscript we assume that g equals 1 due to the small number of observations. For the de novo events in siblings, c1=14, c=15, d=16 and C=232. This estimate is bolstered by additional data, specifically the *de novo* events in the asthma trios and the unaffected siblings in the AGRE sample (Itsara et al., 2010). Looking at these data independently we find c_1=12, c=13, d=14 and C=150. Moreover, combining the sibling data with these data yields an estimate that is quite similar to the original point estimate: c_1=24, c=26, d=29 and C=290 (1 tripleton, 1 doubleton, and 24 singletons). Finally adding the data for all other controls described in this paper (Sebat et al., 2007) also gives a similar estimate: c_1=25, c=28, d=32 and C=242.

The estimate of eCNVR sites are similar to recent results that estimate mutation rates for over 4,000 known CNVRs (Fu et al., 2010). 104 CNVRs are estimated to have high mutation rates (~1 in 1,000 transmissions) and hundreds more having intermediate mutation rates (~1 in 10,000 transmissions). The remainder is estimated

to have mutation rates of 1 in 100,000 or less. A probability calculation shows that, while most CNVRs are extremely rare, most observed *de novo* events will occur at the CNVRs with high mutations rates (≥1 in 10,000 transmissions). In light of our approach of screening out CNVs that correspond to common regions of structural variation in the genome, Fu et al.'s findings suggest that the effective number of CNVR sites is of the same order of magnitude as we obtained above and our estimate of "C" (232) is conservative.

Birthday problem

The birthday problem answers the following question: what is the probability of observing at least one pair of matching birthdays in a group. This problem translates to our setting by considering CNVRs as analogous to days and location of *de novo* events as analogous to birthdays. The bigger C, the more surprising it is to find matching *de novo* events. To determine how many *de novo* events are to be distributed among eCNVRs we note that we observe 15 subjects carrying *de novo* events out of 872 siblings, so the *de novo* rate per sibling is 0.017 or 1.7% per subject. Extrapolating to the larger sample of 1,124 probands we expect d = 1.7% x 1,124 ≈ 19 events, where we define d as the *de novo* rate in siblings times the number of probands. For d=19 the probability of observing at least one matching pair is 0.53. It is worth noting that substituting an even more conservative estimate of C=104, which is the number of high mutation rate CNVRs detected thus far in the human genome (Fu et al. 2010), does not substantively alter the results presented in the body of the text.

In its original formulation the birthday problem is restricted to the probability of observing at least one pair of matching birthdays. Our interest extends to how unusual it is to see a group of m>2 matching events under the null hypothesis of no association with ASD. This calculation is performed empirically by distributing d events at random among C eCNVRs and then counting the maximum number of CNVs falling in the same location. Repeating this experiment many times, we obtain an estimate of the probability of finding m or more counts for at least one eCNVR under the null hypothesis. We performed the same calculations for estimating the probability of eCNVR multiplicities for the larger sample (3,816 ASD subjects) (Figure S3).

Unseen species and birthday problem for *de novo* CNVs mediated by NAHR only

De novo CNVs mediated by non-allelic homologous recombination (NAHR) occur at a higher frequency than *de novo* CNVs mediated by other mechanisms. The majority of recurrent events noted in this paper are mediated by NAHR, including 16p11.2 and 7q11.23. Thus it might be argued that the values of C and d should be estimated from NAHR mediated CNVs only to derive a highly conservative estimate of significance for recurrent NAHR *de novo* CNVs.

Within the SSC siblings there are four *de novo* CNVs mediated by NAHR, two of these are recurrent (Table S4). Therefore c_1=2, c=3, d=4 and C is 10. This estimate was based on 4 siblings with NAHR mediated CNVs out of 872 so the rate of NAHR *de novo* CNVs per subject is 0.46%. Scaling this up to 1,124 probands gives an expected count of d = 0.46% x 1,124 ≈ 5. For d=5 and C=10 the probability of observing at least one matching pair is 0.69. The likelihood of seeing four recurrent CNVs (e.g. 7q11.23) by chance remains very low at 0.004.

Therefore even when NAHR *de novo* CNVs are considered alone the main findings of this paper remain unchanged and strongly support an association with ASD.

Estimating the significance of loci based on single locus calculations

As noted above, our approach to testing yields a genome-wide threshold to evaluate significance. An alternative approach would be to assess individual loci, contrasting the de novo rate in probands with the rate observed in control families. This rate for control families is not available for loci of interest here, but we can bound the de novo rate from above by an estimate of the population frequency of the CNV of interest. For 16p11.2 deletions and duplications we were able to obtain such estimates by gleaning rates for multiple studies (Bochukova et al., 2010; de Kovel et al., 2010; Fernandez et al., 2010; Glessner et al., 2009; Marshall et al., 2008; McCarthy et al., 2009; Rosenfeld et al., 2010; Weiss et al., 2008). What we find is that 16p11.2 deletions occur at a rate of 3 per 10,000 approximately, and likewise duplications occur at rate 4 per 10,000 control subjects. These "control subjects" were variously characterized, and some could still be affected by a psychiatric disorder. Literature estimates for 7q11.23 duplications are also of similar magnitude to these estimates (Molina et al., 2011).

Let us conservatively take the de novo rate for these loci as 5 per 10,000. Then, using binomial theory, we can estimate the p-value of 0.003 approximately for ≥4 *de novo* events (e.g. 7q11.23), a p-value of 2×10^{-6} for ≥7 *de novo* events (e.g. 16p11.2 deletions), and p-value of 3×10^{-11} for ≥11 *de novo* events (e.g. all 16p11.2 *de novo* CNVs) out of a total sample of 1,124 probands.

Predicting the CNV-mediated risk loci for autism based on the SSC data

Using the unseen species formulation we estimated eCNVRs for *non-risk* variants to be 232 from the distribution of *de novo* CNVs found in siblings. We applied the same methodology to estimate the number of autism *risk* eCNVRs by analyzing the distribution of *de novo* CNVs found in probands. This requires that we infer the number of *de novo* events that are true risk loci. While this cannot be done definitively, we consider 11 16p11.2, 4 7q11.23 CNVs and each of the recurrent *de novo* regions restricted to probands as the most likely candidates. In addition, based on the presumption that the majority of *de novo* CNVs in siblings are neutral, we calculate that approximately 75% of *de novo* events in probands are risk variants (67 *de novo* proband CNVs - 16 *de novo* CNVs in siblings/67 *de novo* events). Based on this, we consider 27 of the 44 single occurrence *de novo* CNVs in probands as risk variants and c_1=27, c=33, d=51 (where c= total number of species observed; c_1= the number of singleton species; d = the expected number of risk CNVs occurring in the new sample, which is expected to be of equal or greater size to the set reported herein). Using the formula given above, we calculate *de novo risk* eCNVR = 130.

Predicting future finds for risk loci in the SSC

Next to determine how many *de novo* risk variants are likely to be confirmed in a future SSC sample, we accounted for the uneven distribution of recurrent events. Using the observed frequency of presumed risk variants in our sample, we approximate the distribution of risk eCNVs using a beta distribution (method of moments applied to the frequency distribution of presumed risk variants). The estimated parameters of the beta distribution were then used to assign relative probabilities for each of the 130 eCNVRs.

With this estimated probability framework, we then performed the following simulation experiment: sample another 67 *de novo* risk variants in phase II of the project and combining these CNVs with the 67 already observed, and we performed this experiment many times. We than evaluated the number of CNVs showing 4 or more recurrences at a single locus over the set of experiments. For this larger sample 4 matching variants are required to reach the significance threshold. We find that our previous results (16p11.2 and 7q11.23) are confirmed and typically 2-3 new risk variants are identified. The newly identified risk variants are often, but not always, confirmations of those *de novo* CNV intervals seen twice in the current sample.

Predicting the CNV-mediated risk loci for autism based on the wider set of *de novo* events

Using the 219 rare de novo CNVs predicted in 204 probands from this paper combined with the four other papers (Itsara et al., 2010; Marshall et al., 2008; Pinto et al., 2010; Sebat et al., 2007) gives an estimate of de novo burden of 5.3% in probands (219 out of 3,816). In controls (where present) the burden is 1.6% (31 out of 1,881). This gives an estimate of 71% for the percentage of *de novo* events that contribute to risk. Using this estimate to calculate the number of risk eCNVRs gives an estimate of 234 (c_1=59, c=88, d=158).

References

Battaglia, A., Gurrieri, F., Bertini, E., Bellacosa, A., Pomponi, M.G., Paravatou-Petsotas, M., Mazza, S., and Neri, G. (1997). The inv dup(15) syndrome: a clinically recognizable syndrome with altered behavior, mental retardation, and epilepsy. Neurology *48*, 1081-1086.

Ben-Shachar, S., Lanpher, B., German, J.R., Qasaymeh, M., Potocki, L., Nagamani, S.C., Franco, L.M., Malphrus, A., Bottenfield, G.W., Spence, J.E*., et al.* (2009). Microdeletion 15q13.3: a locus with incomplete penetrance for autism, mental retardation, and psychiatric disorders. J Med Genet *46*, 382-388.

Bijlsma, E.K., Gijsbers, A.C., Schuurs-Hoeijmakers, J.H., van Haeringen, A., Fransen van de Putte, D.E., Anderlid, B.M., Lundin, J., Lapunzina, P., Pérez Jurado, L.A., Delle Chiaie, B*., et al.* (2009). Extending the phenotype of recurrent rearrangements of 16p11.2: deletions in mentally retarded patients without autism and in normal individuals. Eur J Med Genet *52*, 77-87.

Bochukova, E.G., Huang, N., Keogh, J., Henning, E., Purmann, C., Blaszczyk, K., Saeed, S., Hamilton-Shield, J., Clayton-Smith, J., O'Rahilly, S*., et al.* (2010). Large, rare chromosomal deletions associated with severe early-onset obesity. Nature *463*, 666-670.

Bunge, J., and Fitzpatrick, M. (1993). Estimating the Number of Species: A Review. Journal of the American Statistical Association *88*, 364-373.

Burghes, A.H., Vaessin, H.E., and de La Chapelle, A. (2001). Genetics. The land between Mendelian and multifactorial inheritance. Science *293*, 2213-2214.

Butler, M.G., Bittel, D.C., Kibiryeva, N., Talebizadeh, Z., and Thompson, T. (2004). Behavioral differences among subjects with Prader-Willi syndrome and type I or type II deletion and maternal disomy. Pediatrics *113*, 565-573.

Cook, E.H., Lindgren, V., Leventhal, B.L., Courchesne, R., Lincoln, A., Shulman, C., Lord, C., and Courchesne, E. (1997). Autism or atypical autism in maternally but not paternally derived proximal 15q duplication. Am J Hum Genet *60*, 928-934.

Crossett, A., Kent, B.P., Klei, L., Ringquist, S., Trucco, M., Roeder, K., and Devlin, B. (2010). Using ancestry matching to combine family-based and unrelated samples for genome-wide association studies. Stat Med *29*, 2932-2945.

de Kovel, C.G., Trucks, H., Helbig, I., Mefford, H.C., Baker, C., Leu, C., Kluck, C., Muhle, H., von Spiczak, S., Ostertag, P*., et al.* (2010). Recurrent microdeletions at 15q11.2 and 16p13.11 predispose to idiopathic generalized epilepsies. Brain *133*, 23-32.

Dibbens, L.M., Mullen, S., Helbig, I., Mefford, H.C., Bayly, M.A., Bellows, S., Leu, C., Trucks, H., Obermeier, T., Wittig, M*., et al.* (2009). Familial and sporadic 15q13.3 microdeletions in idiopathic generalized epilepsy: precedent for disorders with complex inheritance. Hum Mol Genet *18*, 3626-3631.

Elmslie, F.V., Rees, M., Williamson, M.P., Kerr, M., Kjeldsen, M.J., Pang, K.A., Sundqvist, A., Friis, M.L., Chadwick, D., Richens, A*., et al.* (1997). Genetic mapping of a major susceptibility locus for juvenile myoclonic epilepsy on chromosome 15q. Hum Mol Genet *6*, 1329-1334.

Fernandez, B.A., Roberts, W., Chung, B., Weksberg, R., Meyn, S., Szatmari, P., Joseph-George, A.M., Mackay, S., Whitten, K., Noble, B*., et al.* (2010). Phenotypic spectrum associated with de novo and inherited deletions and duplications at 16p11.2 in individuals ascertained for diagnosis of autism spectrum disorder. J Med Genet *47*, 195-203.

Fu, W., Zhang, F., Wang, Y., Gu, X., and Jin, L. (2010). Identification of copy number variation hotspots in human populations. Am J Hum Genet *87*, 494-504.

Glessner, J.T., Wang, K., Cai, G., Korvatska, O., Kim, C.E., Wood, S., Zhang, H., Estes, A., Brune, C.W., Bradfield, J.P*., et al.* (2009). Autism genome-wide copy number variation reveals ubiquitin and neuronal genes. Nature *459*, 569-573.

Grafodatskaya, D., Chung, B., Szatmari, P., and Weksberg, R. (2010). Autism spectrum disorders and epigenetics. J Am Acad Child Adolesc Psychiatry *49*, 794-809.

Hanson, E., Nasir, R.H., Fong, A., Lian, A., Hundley, R., Shen, Y., Wu, B.L., Holm, I.A., Miller, D.T., and Clinicians, p.S.G. (2010). Cognitive and behavioral characterization of 16p11.2 deletion syndrome. J Dev Behav Pediatr *31*, 649-657.

Helbig, I., Mefford, H.C., Sharp, A.J., Guipponi, M., Fichera, M., Franke, A., Muhle, H., de Kovel, C., Baker, C., von Spiczak, S*., et al.* (2009). 15q13.3 microdeletions increase risk of idiopathic generalized epilepsy. Nat Genet *41*, 160-162.

Hogart, A., Wu, D., LaSalle, J.M., and Schanen, N.C. (2010). The comorbidity of autism with the genomic disorders of chromosome 15q11.2-q13. Neurobiol Dis *38*, 181-191.

ISC, I.S.C. (2008). Rare chromosomal deletions and duplications increase risk of schizophrenia. Nature *455*, 237-241.

Itsara, A., Wu, H., Smith, J.D., Nickerson, D.A., Romieu, I., London, S.J., and Eichler, E.E. (2010). De novo rates and selection of large copy number variation. Genome Res *20*, 1469-1481.

Knoll, J.H., Nicholls, R.D., Magenis, R.E., Graham, J.M., Lalande, M., and Latt, S.A. (1989). Angelman and Prader-Willi syndromes share a common chromosome 15 deletion but differ in parental origin of the deletion. Am J Med Genet *32*, 285-290.

Lander, E., and Kruglyak, L. (1995). Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. Nat Genet *11*, 241-247.

Lee, C., Abdool, A., and Huang, C.H. (2009). PCA-based population structure inference with generic clustering algorithms. BMC Bioinformatics *10 Suppl 1*, S73.

Leonard, S., and Freedman, R. (2006). Genetics of chromosome 15q13-q14 in schizophrenia. Biol Psychiatry *60*, 115-122.

Livak, K.J., and Schmittgen, T.D. (2001). Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) Method. Methods *25*, 402-408.

Marshall, C.R., Noor, A., Vincent, J.B., Lionel, A.C., Feuk, L., Skaug, J., Shago, M., Moessner, R., Pinto, D., Ren, Y.*, et al.* (2008). Structural variation of chromosomes in autism spectrum disorder. Am J Hum Genet *82*, 477-488.

McCarthy, S.E., Makarov, V., Kirov, G., Addington, A.M., McClellan, J., Yoon, S., Perkins, D.O., Dickel, D.E., Kusenda, M., Krastoshevsky, O.*, et al.* (2009). Microduplications of 16p11.2 are associated with schizophrenia. Nat Genet *41*, 1223-1227.

Miller, D.T., Shen, Y., Weiss, L.A., Korn, J., Anselm, I., Bridgemohan, C., Cox, G.F., Dickinson, H., Gentile, J., Harris, D.J.*, et al.* (2009). Microdeletion/duplication at 15q13.2q13.3 among individuals with features of autism and other neuropsychiatric disorders. J Med Genet *46*, 242-248.

Molina, O., Anton, E., Vidal, F., and Blanco, J. (2011). Sperm rates of 7q11.23, 15q11q13 and 22q11.2 deletions and duplications: a FISH approach. Hum Genet *129*, 35-44.

Neubauer, B.A., Fiedler, B., Himmelein, B., Kämpfer, F., Lässker, U., Schwabe, G., Spanier, I., Tams, D., Bretscher, C., Moldenhauer, K.*, et al.* (1998). Centrotemporal spikes in families with rolandic epilepsy: linkage to chromosome 15q14. Neurology *51*, 1608-1612.

Orr-Urtreger, A., Göldner, F.M., Saeki, M., Lorenzo, I., Goldberg, L., De Biasi, M., Dani, J.A., Patrick, J.W., and Beaudet, A.L. (1997). Mice deficient in the alpha7 neuronal nicotinic acetylcholine receptor lack alpha-bungarotoxin binding sites and hippocampal fast nicotinic currents. J Neurosci *17*, 9165-9171.

Pinto, D., Pagnamenta, A.T., Klei, L., Anney, R., Merico, D., Regan, R., Conroy, J., Magalhaes, T.R., Correia, C., Abrahams, B.S.*, et al.* (2010). Functional impact of global rare copy number variation in autism spectrum disorders. Nature *466*, 368-372.

Qin, J., Jones, R.C., and Ramakrishnan, R. (2008). Studying copy number variations using a nanofluidic platform. Nucleic Acids Res *36*, e116.

Reymond, A., Zufferey, F., Harewood, L., Kutalik, Z., Martinet, D., Chrast, J., Walters, R.G., Bouquillon, S., Valsesia, A., Hippolyte, L.*, et al.* (2010). Gene dosage at the 16p11.2 locus controls body mass index. In American Society of Human Genetics (Washington DC).

Rosenfeld, J.A., Coppinger, J., Bejjani, B.A., Girirajan, S., Eichler, E.E., Shaffer, L.G., and Ballif, B.C. (2010). Speech delays and behavioral problems are the predominant features in individuals with developmental delays and 16p11.2 microdeletions and microduplications. J Neurodevelop Disord, 26–38.

Sebat, J., Lakshmi, B., Malhotra, D., Troge, J., Lese-Martin, C., Walsh, T., Yamrom, B., Yoon, S., Krasnitz, A., Kendall, J.*, et al.* (2007). Strong association of de novo copy number mutations with autism. Science *316*, 445-449.

Sharp, A.J., Mefford, H.C., Li, K., Baker, C., Skinner, C., Stevenson, R.E., Schroer, R.J., Novara, F., De Gregori, M., Ciccone, R.*, et al.* (2008). A recurrent 15q13.3 microdeletion syndrome associated with mental retardation and seizures. Nat Genet *40*, 322-328.

Shen, Y., Dies, K.A., Holm, I.A., Bridgemohan, C., Sobeih, M.M., Caronna, E.B., Miller, K.J., Frazier, J.A., Silverstein, I., Picker, J.*, et al.* (2010). Clinical genetic testing for patients with autism spectrum disorders. Pediatrics *125*, e727-735.

Shinawi, M., Liu, P., Kang, S.H., Shen, J., Belmont, J.W., Scott, D.A., Probst, F.J., Craigen, W.J., Graham, B.H., Pursley, A.*, 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.

Shinawi, M., Schaaf, C.P., Bhatt, S.S., Xia, Z., Patel, A., Cheung, S.W., Lanpher, B., Nagl, S., Herding, H.S., Nevinny-Stickel, C.*, et al.* (2009). A small recurrent deletion within 15q13.3 is associated with a range of neurodevelopmental phenotypes. Nat Genet *41*, 1269-1271.

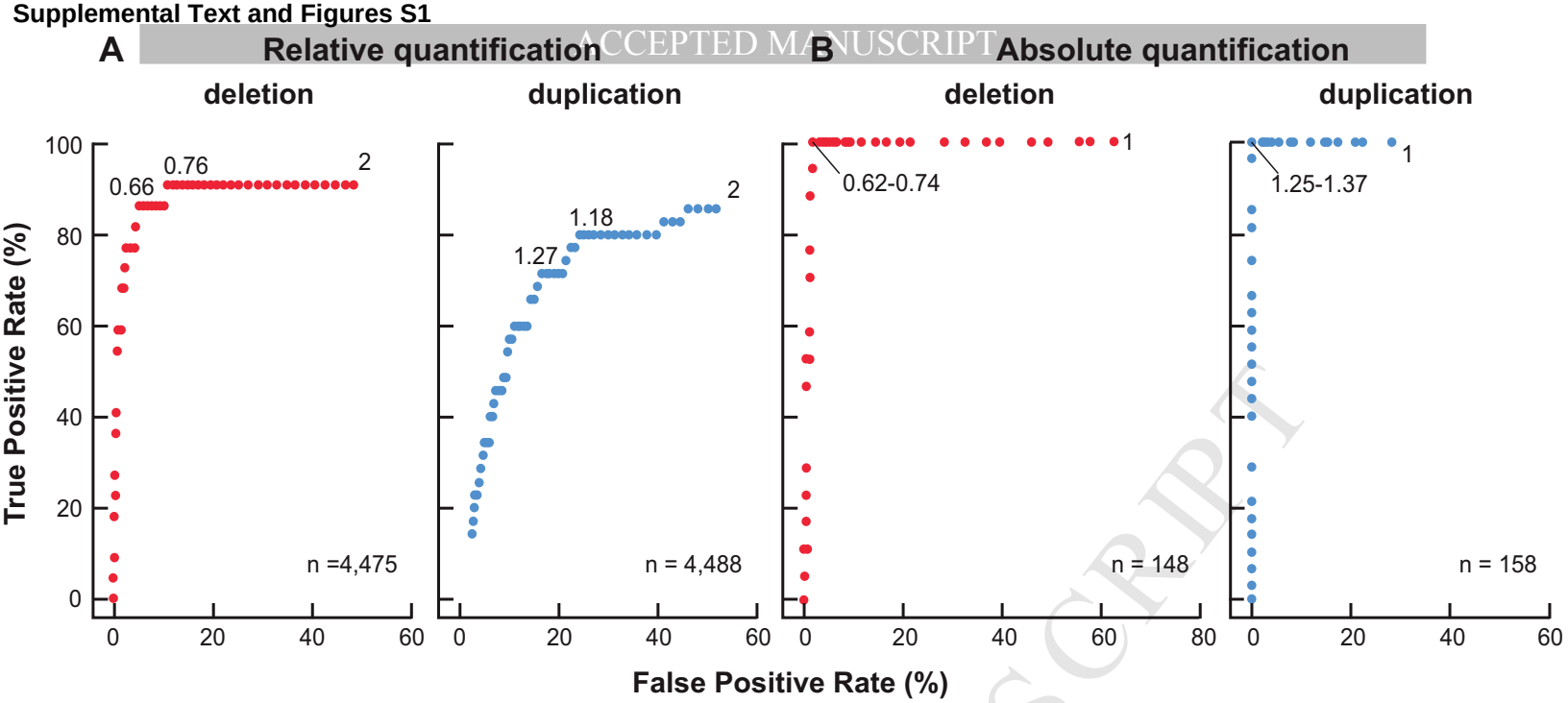
Stefansson, H., Rujescu, D., Cichon, S., Pietiläinen, O.P., Ingason, A., Steinberg, S., Fossdal, R., Sigurdsson, E., Sigmundsson, T., Buizer-Voskamp, J.E.*, et al.* (2008). Large recurrent microdeletions associated with schizophrenia. Nature *455*, 232-236.

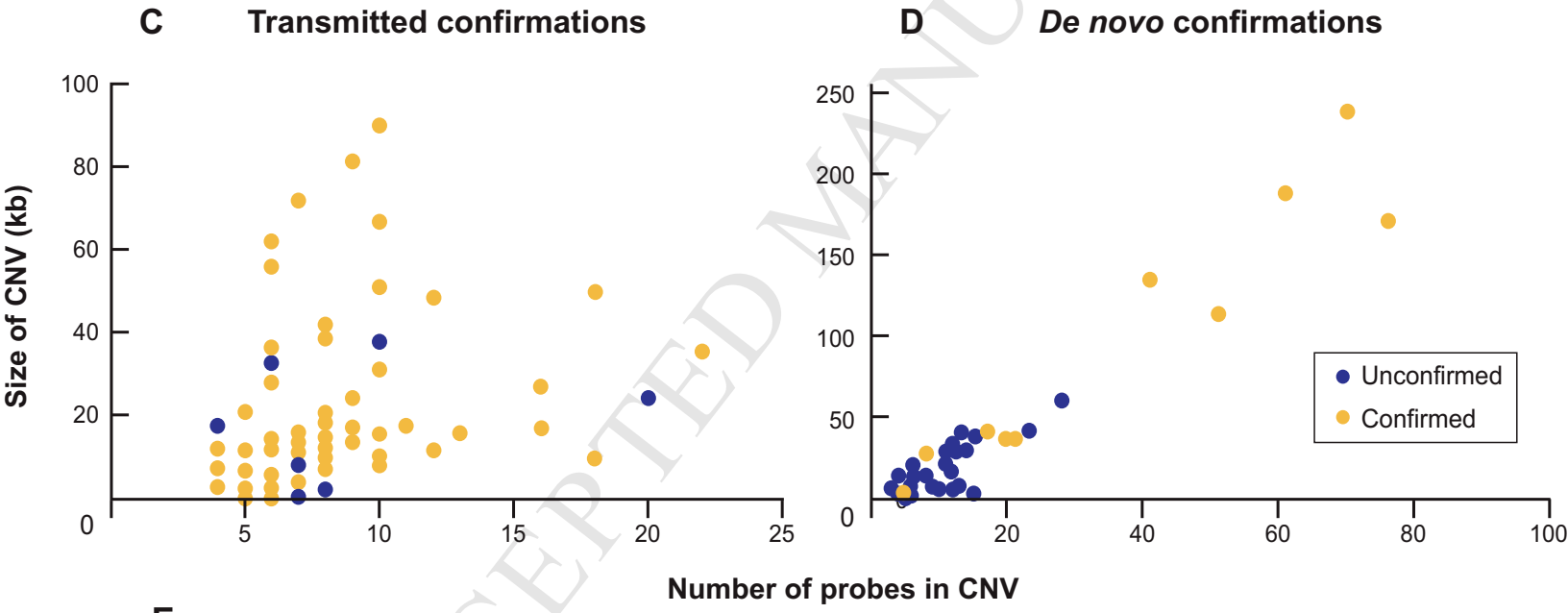
Szafranski, P., Schaaf, C.P., Person, R.E., Gibson, I.B., Xia, Z., Mahadevan, S., Wiszniewska, J., Bacino, C.A., Lalani, S., Potocki, L.*, et al.* (2010). Structures and molecular mechanisms for common 15q13.3 microduplications involving CHRNA7: benign or pathological? Hum Mutat *31*, 840-850.

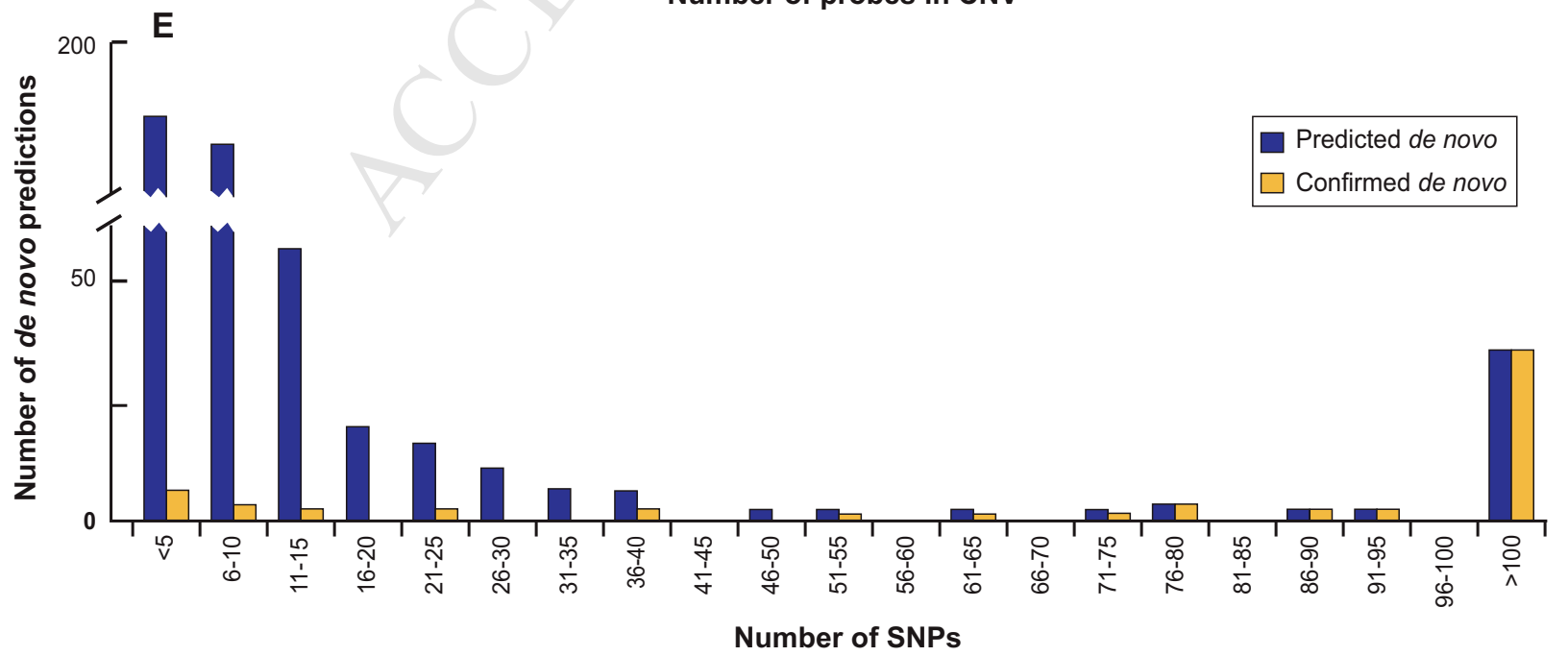
van Bon, B.W., Mefford, H.C., Menten, B., Koolen, D.A., Sharp, A.J., Nillesen, W.M., Innis, J.W., de Ravel, T.J., Mercer, C.L., Fichera, M.*, et al.* (2009). Further delineation of the 15q13 microdeletion and duplication syndromes: a clinical spectrum varying from non-pathogenic to a severe outcome. J Med Genet *46*, 511-523.

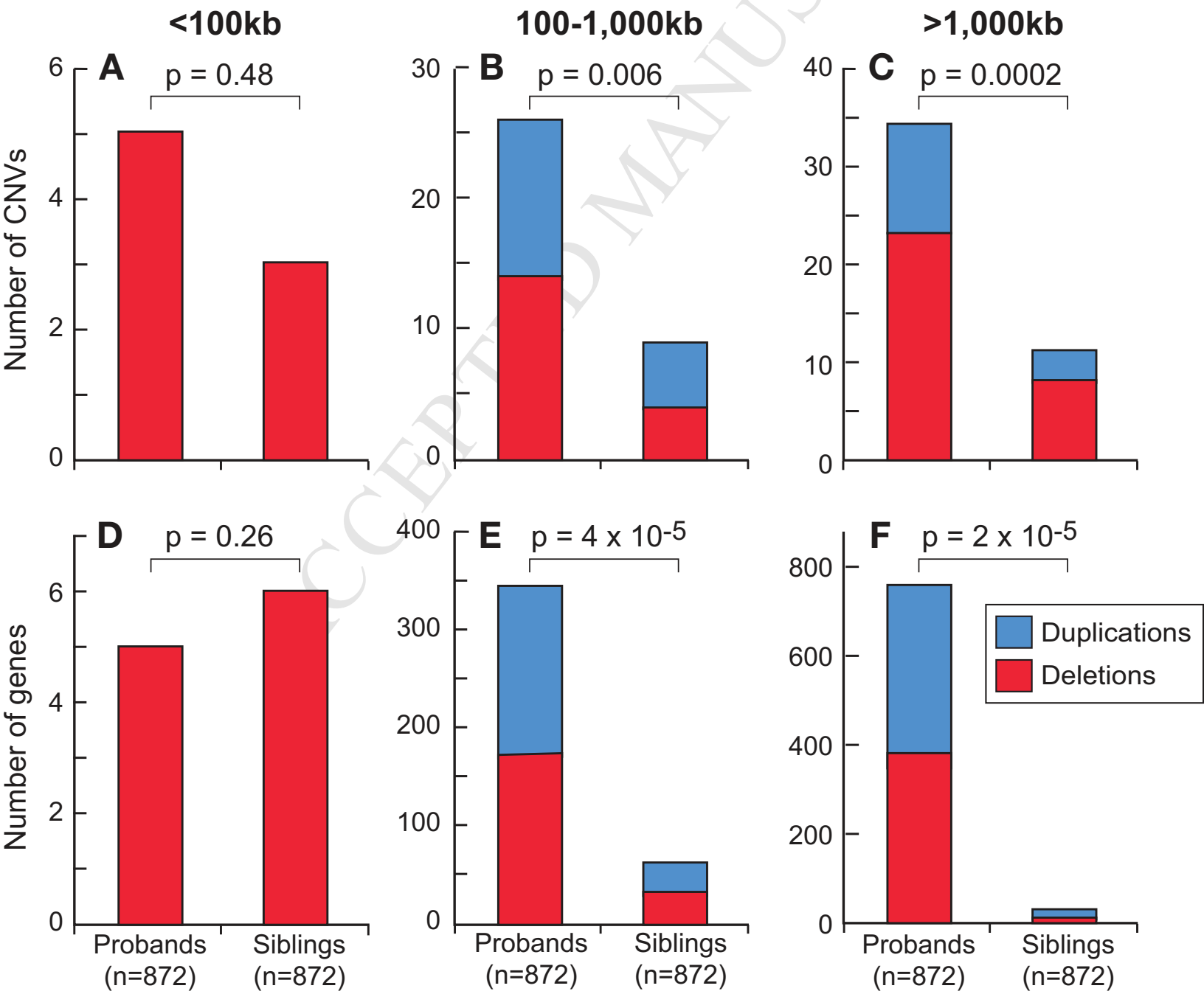
Walters, R., Jacquemont, S., Valsesia, A., de Smith, A., Martinet, D., Andersson, J., Falchi, M., Chen, F., Andrieux, J., Lobbens, S.*, et al.* (2010). A new highly penetrant form of obesity due to deletions on chromosome 16p11.2. Nature *463*, 671-675.

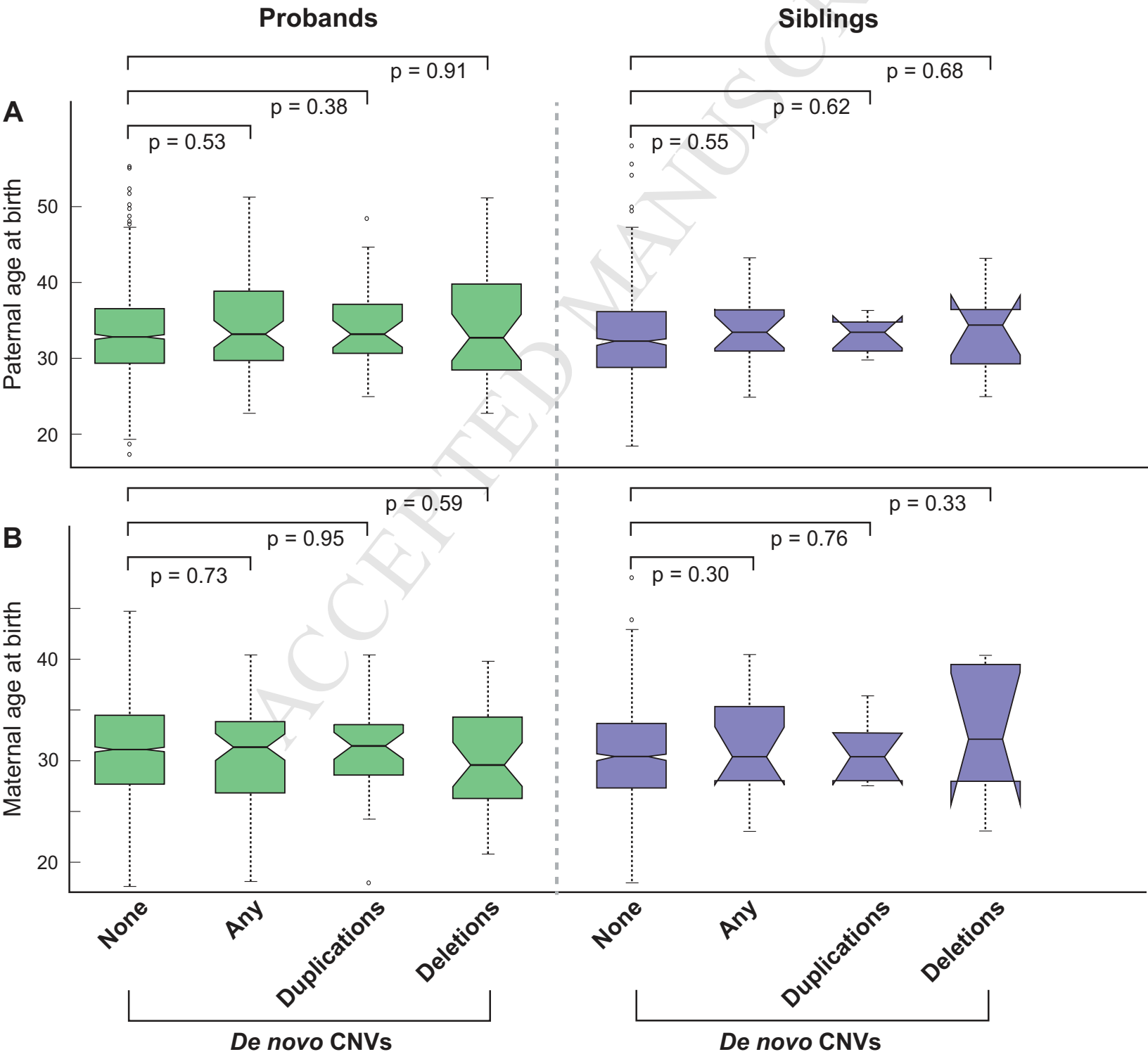
Weiss, L.A., Shen, Y., Korn, J.M., Arking, D.E., Miller, D.T., Fossdal, R., Saemundsen, E., Stefansson, H., Ferreira, M.A., Green, T.*, et al.* (2008). Association between microdeletion and microduplication at 16p11.2 and autism. N Engl J Med *358*, 667-675.

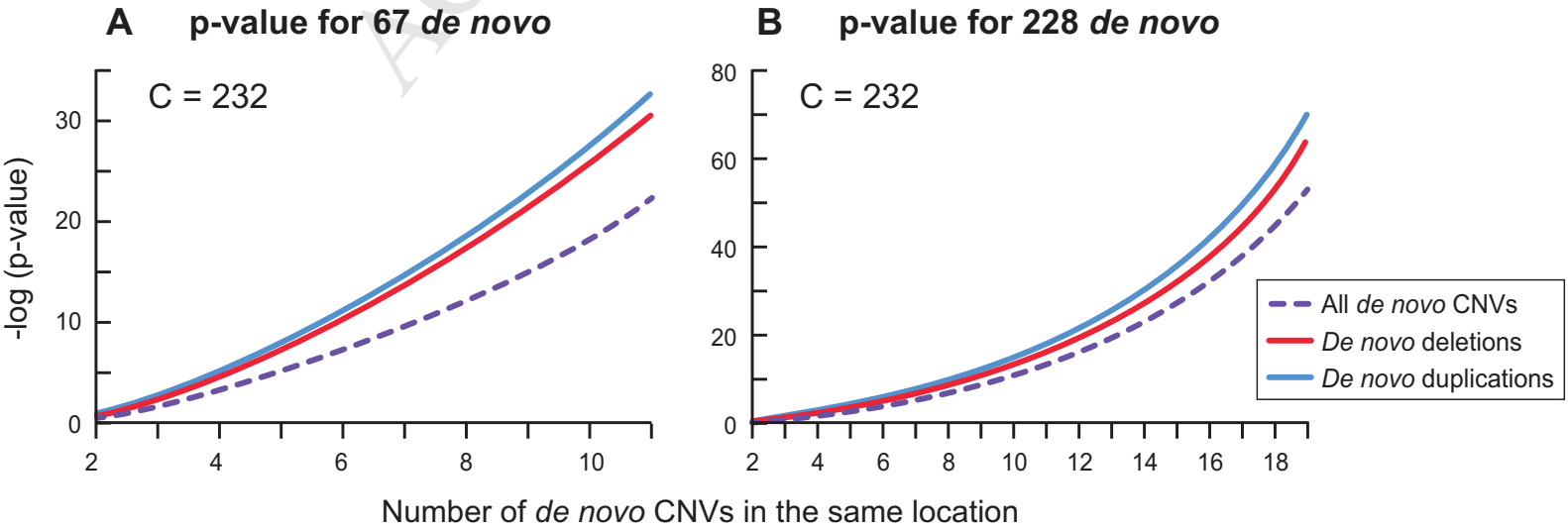












The study evaluates copy number variants (CNVs) in 1,174 families with a single child diagnosed with Autism Spectrum Disorder (ASD). Strong association is found between *de novo* duplications at 7q11.23 and ASD. Reciprocal deletions at 7q11.23 cause Williams-Beuren Syndrome featuring a highly social personality, in contrast to core deficits seen in autism. Association with ASD is replicated at chromosomes 16p11.2, 15q11.2-13 and the gene *Neurexin 1*; new potential ASD risk regions are identified involving the genes *CDH13*, *USP7*, and *C16orf72*.

Emily L. Crawford11, Lea Davis15, Nicole R. Davis Wright2, Rahul M. Dhodapkar2, Michael DiCola9, Nicholas M. DiLullo2, Thomas V. Fernandez2, Vikram Fielding-Singh16, Daniel O. Fishman17, Stephanie Frahm9, Rouben Garagaloyan18, Gerald S. Goh4, Sindhuja Kammela2, Lambertus Klei19, Jennifer K. Lowe20, Sabata C. Lund5, Anna D. McGrew11, Kyle A. Meyer21, William J. Moffat2, John D. Murdoch4, Brian J. O'Roak22, Gordon T. Ober2, Rebecca S. Pottenger23, Melanie J. Raubeson2, Youeun Song2, Qi Wang9, Brian L. Yaspan11, Timothy W. Yu24, Ilana R. Yurkiewicz2, Arthur L. Beaudet14, Rita M. Cantor6,25, Martin Curland18, Dorothy E. Grice26, Murat Günel1,4,13, Richard P. Lifton4,27, Shrikant M. Mane28, Donna M. Martin29, Chad A. Shaw14, Michael Sheldon30, Jay A. Tischfield30, Christopher A. Walsh31, Eric M. Morrow32, David H. Ledbetter33, Eric Fombonne34, Catherine Lord5,35, Christa Lese Martin7, Andrew I. Brooks9, James S. Sutcliffe11, Edwin H. Cook, Jr. 15,36, Daniel Geschwind20,36, Kathryn Roeder8, Bernie Devlin19, Matthew W. State1-4*

Your Name: Stephan Sanders

Date: 18/04/11