Supplementary text and tables for

De Novo Rates and Selection of Large Copy Number Variation

Andy Itsara, Hao Wu, Joshua D. Smith, Deborah A. Nickerson, Isabelle Romieu, Stephanie J London, Evan E. Eichler

Supplemental Methods:

CNV Discovery: The HMM analyzed each chromosome of each sample separately. HMM state assignments were merged into segments according to the following criteria: consecutive probes of the same state less than 50kb apart were merged, and if two segments of the same state were separated by an intervening sequence of ≤5 probes and ≤10kb, both segments and intervening sequence were called as a single variant. Before further analysis, samples were eliminated if the hybridization did not have genome-wide LogR standard deviation ≤0.25, absolute value of the average LogR ≤ 0.1, and average b-deviation < 0.05. To decrease the false discovery rate, putative CNVs calls were then subject to additional filtering. Putative CNVs were divided into two categories: "large" CNV calls >100 probes or >1Mb and "small" CNVs <100 probes and <1Mb. Large CNVs were manually curated. Manual curation was used to exclude potential false positives, whole chromosome aneuploidies, and potential cell line mosaicism and artifacts. Small CNVs were subject to the following automated filtering criteria: homozygous deletions were required to have ≥3 probes, median LogR Z-score ≤ -4, and mean b-deviation ≥ 0.1 or ≥ 3 probes and median LogR Z-score ≤ -8; hemizygous deletions were required to span ≥10 probes, have LogR Z-score ≤ -1.5, and less than 10% of probes called as heterozygous; for duplications we required ≥ 10 probes, LogR Z-score ≥1.5, and b-deviation among heterozygote probes ≥ 0.075. Using these parameters, this CNV discovery technique was previously estimated to have a false discovery rate of 14-23% with a sensitivity of ~60% with an effective resolution of ~30kb.

In order to decrease overfragmentation by the HMM, CNV calls <1Mb within the same sample were manually inspected and merged if they were found to represent the same CNV. Finally, samples were removed if they were outliers with respect to the number of CNVs, false positives found during manual inspection or possible artifacts during merging of HMM calls.

QC Parameters Used in CNV Calling

Study	Illumina Platform	Max Number of CNVs	Max Number of large CNV false positives	Max number of possible artifacts during CNV merging
Asthma	HumanHap550	25	2	1
HapMap	1M Duo	75	5	2
AGRE	HumanHap550	25	15	2

***de novo* CNV Identification:** Parent-child relationships within a trio were considered validated if >98% of successfully genotyped SNPs were concordant with Mendelian inheritance. As a negative control, false trios consisting of three randomly chosen individuals were found to display on average ~80% of SNPs concordant with Mendelian inheritance.

To assess the ability of manual curation to exclude inherited CNVs during manual inspection of trio data, we generated copy number genotypes in 269 HapMap samples through manual curation at previously reported copy number polymorphisms (CNP) (McCarroll et al. 2008).

Illumina 1MDuo genotype data was obtained for 269 HapMap samples (GEO Accessions GSE16894, GSE16895, and GSE16896). As >99% of probes on the Illumina 550K platform are present on the Illumina 1MDuo platform, we additionally generated the equivalent of Illumina 550K data by subsampling Illumina 1MDuo data, allowing us to gauge the performance of manual curation on both platforms.

For each of the Illumina 1MDuo and Illumina 550K platforms, we chose 10 loci for copy number genotyping by manual curation in 269 HapMap samples. We chose 10 random CNPs from those with ≥10 probes on the platform being assessed. The ≥10 probe criteria was applied because the CNVs identified in the CNV discovery phase (and hence those that would undergo manual curation in *de novo* CNV identification) were similarly required to include ≥10 probes.

Copy number genotypes reported by McCarroll et al. were used to assess the performance of manual curation (Supplemental Table 18, Supplemental Table 19). Two of the chosen CNPs, CNP 2082 on Illumina 550K and CNP 1434 on Illumina 1MDuo, were entirely contained within segmental duplications (SDs). Although manual curation was nearly perfect in genotyping CNP 2082 (Supplemental Table 18), it performed poorly on CNP 1434 (Supplemental Table 19). The variable performance of manual curation in genotyping CNPs within segmental duplications was expected given previously described difficulties in ascertaining CNVs within these regions (Cooper et al. 2008; Conrad et al. 2009).

The remaining CNPs each had the majority of their lengths outside of SDs. We defined sensitivity and specificity with respect to the ability of manual curation to flag a sample as copy number variant (copy number not equal to 2). Under this metric of performance, manual curation of Illumina 550K data had 100% sensitivity (42/42) and >99.9% (2372/2373) specificity for identifying copy number variants. Manual curation of Illumina 1MDuo had 94.7% (36/38) sensitivity and 99.8% (2375/2380) specificity.

In summary, manual curation outside of SDs has high sensitivity to detect CNVs given a defined locus. Therefore outside of SDs, we expect it to be an effective method of excluding inherited CNVs erroneously flagged as *de novo* due to undercalling of CNVs in parents. Finally, it should be noted that although our analysis did not explicitly remove candidate *de novo* CNVs within segmental duplications, all *de novo* CNVs identified in this study had >50% of their length outside of SDs.

Calculation of Selection Coefficient using Mutation-Selection Balance

We calculate the selection coefficient based on a slight modification of the classic mutation-selection model assuming either linked mutations with no recombination or unlinked mutations within a haploid genome. We assume an infinite, random mating diploid population (i.e. ignoring genetic drift) and consider the frequency of gametes with a given number of mutations (i.e. large CNVs). In the limit that the mutation rate (μ) and equilibrium frequency of mutation-bearing genomes (q) is small, we observe that both models converge to the classical approximation, $s = \mu/q$.

Linked mutations in a haploid genome with no recombination

Under this model, we ignore back mutation and haploid genomes acquire mutations at a rate μ. The mutations are linked with no recombination so that a given haploid genome simply collects mutations that never segregate away from one another with each generation. For simplicity, we assume that the relative fitness of a diploid genome is 1 if it has no mutation, and 1-s for one or more mutations. If p_j is the frequency of a haploid genome with j mutations, then we have the following:

$$\begin{array}{ll} \text{Allele frequencies:} & p_0 \underset{\mu}{\rightarrow} p_1 \underset{\mu}{\rightarrow} p_2 \underset{\mu}{\rightarrow} \ldots \\ \text{Relative fitnesses:} & 1 \quad\; 1 - s \quad 1 - s \quad \ldots \end{array}$$

After selection on diploids from an earlier generation, the resulting fraction of gametes that will be of haplotype p_0 in the next generation will be

$$p_0^* = \frac{2p_0^2 + (1-s)2\sum_{j=1}^{\infty} p_0 p_j}{2p_0^2 + (1-s)\left[\sum_{\substack{i,j=0,\\ i\neq j}}^{\infty} 2p_i p_j + 2\sum_{i=1}^{\infty} p_i^2\right]}(1-\mu)$$

. Canceling out the factors of 2 and using the fact that the sum of allele frequencies and genotypes are separately equal to1, this equation is greatly simplified to

$$p_0^* = \frac{p_0^2 + (1-s)p_0(1-p_0)}{p_0^2 + (1-s)(1-p_0^2)}(1-\mu) \quad \text{(Equation 1)}$$

At equilibrium, $p_0^* = p_0$. After some algebra, we have that $p_0 = 0$ or

$$s = \frac{\mu}{(1+\mu)q - q^2}$$

where $q = 1 - p_0$. For small μ and q, this simplifies to the classic equation $s = \mu/q$.

Unlinked mutations

Similar to the previous model, we ignore back mutation with a given haploid genome acquiring mutations at rate μ and relative fitnesses of 1 for a diploid genome without mutations, and 1 – s otherwise. However, this model assumes all mutations segregate independently. For a diploid genome inheriting *i* mutations from one gamete and *j* mutations from the other, it will generate gametes with up to *i*+*j* mutations following a binomial distribution with probability of success 0.5. Given haploid genomes with *j* mutations at frequency p_j the frequency of gametes with *k* mutations in the following generation will then be

$$p_k^* = (1-\mu)\sum_{l=0}^{\infty}\left[\binom{k+l}{k}\left(\tfrac{1}{2}\right)^{k+l}\left(\sum_{j=0}^{k+l} p_j p_{l+k-j}\right)(1-s_{l+k})\right] + \mu \cdot p_{k-1}^*$$

The large summation considers the contributions from all possible diploid genotypes that can create a gamete with *k* mutations. The value s_{l+k} is 0 if $l+k=0$ and *s* otherwise. For p_0, the fraction of gametes with no mutations, the formula simplifies considerably:

$$p_0^* = (1-\mu)\left[\sum_{l=0}^{\infty}\left(\tfrac{1}{2}\right)^{l}\left(\sum_{j=0}^{l} p_j p_l\right)(1-s_l)\right]$$

$$p_0^* = (1-\mu)\left\{p_0^2 + \sum_{l=1}^{\infty}\left(\left(\tfrac{1}{2}\right)^{l}\sum_{j=0}^{l} p_j p_{l-j}\right)(1-s)\right\} \qquad \text{(Equation 2)}$$

Under assumption that p_j for $j>0$ will be small, we can drop quadratic terms in Equation 2 that do not have p_0.

$$p_0^* = (1-\mu)\left[p_0^2 + \sum_{l=1}^{\infty}\left(\tfrac{1}{2}\right)^{l-1} p_0 p_l (1-s)\right]$$

Finally, if the frequency of multiple mutations in an individual is small, we can drop p_l for $l \geq 2$ yielding

$$p_0^* = (1-\mu)\left[p_0^2 + (1-s)p_0(1-p_0)\right] \qquad \text{(Equation 3)}$$

Finally, Equation 3 must be normalized by the sum of all p_i^* so that sum of allele frequencies is 1 in the next generation. Using the fact that all diploid genomes with one or more mutations have relative fitness (1-*s*),

$$p_0^* = \frac{p_0^2 + (1-s)p_0(1-p_0)}{p_0^2 + (1-p_0^2)(1-s)}(1-\mu)$$

As this is identical to Equation 1, solving for equilibrium frequency will again yield the classic equation $s = \mu/q$.

Calculation of Confidence Intervals for *s*: The variances for the mutation rate (μ) and frequency of CNVs >500kb (q) were estimated as $p(1-p) / n$ where p is the mutation rate or frequency of large CNVs and n is the number of transmissions or allele frequency. The variance in the selection coefficient *s* was calculated assuming no covariance between estimates of μ and q using the first-order approximation

$$\frac{\sigma_s^2}{s^2} = \frac{\sigma_\mu^2}{\mu^2} + \frac{\sigma_q^2}{q^2}$$

95% Confidence Intervals were calculated as $s \pm 1.96 \cdot \sigma_S$.

Supplemental Figure Legends

Supplemental Figure 1. Selected Extended CEPH pedigrees used for *de novo* CNV validation. Two of the three extended CEPH pedigrees in which we attempted to validate *de novo* CNVs using array CGH are shown. Individuals carrying a putative *de novo* CNV (blue circle), HapMap trios (red boxes), and individuals tested (green boxes) have been highlighted.

Supplemental Figure 2. Array CGH confirms a putative CNV predicted to be present in HapMap trio children but neither parent. Plots of SNP array data (a, c) and array CGH data (b, d) for a duplication and deletion (a,b) and deletion (c,d) predicted in the child, but neither parent.

Supplemental Figure 3. A predicted *de novo* duplication in CEPH individual NA12707 fails to transmit to any of eight children. Array CGH data at hg18, chr13:103,202,760-103,228,137 (highlighted with gray background, blue vertical bars) along with 50 kb of flanking sequence is shown for NA12707 (arrow), NA12708, and eight children in extended CEPH pedigree 1358. SegMNT mean signal is indicated by red lines.

Supplemental Figure 4. A predicted *de novo* deletion in CEPH individual NA10831 fails to transmit to any of eight children. Array CGH data at hg18, chr7:84,122,104-84,384,907 (highlighted with gray background, blue vertical bars) along with 100 kb of flanking sequence is shown for NA10831 (arrow), NA10830, and eight children in extended CEPH pedigree 1408. SegMNT mean signal is indicated by red lines.

Supplemental Figure 5-6. Segregation analysis of flanking SNPs confirms predicted *de novo* CNVs in the HapMap are cell line artifacts. Labeled, extended CEPH pedigrees are shown with phased genotypes of nearby SNPs and microsatellites printed vertically underneath each individual. In the second and third generations, a red line indicates the relative position of the CNV. Haplotypes in the individual of interest (indicated by an arrow) and the composition of transmitted haplotypes has been highlighted in green or yellow. A local map of physical and genetic distances along with the positions of the markers and predicted CNV is shown below the pedigree. In all pedigrees, nearby markers suggest that each chromosome homologue is transmitted at least once. Thus, failure to observe inheritance of a putative *de novo* CNV is unlikely to be due to lack of transmission of one of the two chromosome homologues.

Supplemental Figure 7. SNP array data of all candidate *de novo* CNVs from asthma trios. For each candidate *de novo* CNV (see Table 3, main text), SNP array data from the father, mother, and child are displayed as indicated by the pedigree in the lower-right corner of each panel. Each plot shows LogR ratio (vertical bars), B-allele frequency (solid points), and segmental duplications (green locks) with genomic coordinates on the x-axis and a common scale on the y-axis. The predicted CNV in the child is highlighted by a gray background and contrasting LogR ratio (red) and B-allele frequency (blue). The corresponding region in each parent is indicated by a dotted box.

Supplemental Figure 8. Observed frequency of *de novo* CNVs as a function of minimum CNV size across several studies.

Supplemental Figure 9. Performance of *de novo* CNV identification in this study compared to 6 *de novo* CNVs previously reported by Sebat et al. A subset (~150) of samples from AGRE pedigrees used in this study were previously analyzed using ROMA and reported to have six *de novo* CNVs (Sebat et al. 2007). More recently generated SNP array data (Bucan et al. 2009) of these events is shown. LogR Ratio (black vertical bars), B-allele frequency (blue dots), and segmental duplications (green blocks), and CNV boundaries determined by (Sebat et al. 2007) with ROMA (lifted from hg17, dotted rectangle) have been plotted along genomic coordinates (hg18, Build 36). Using independently generated data and analyses, we identified three of the events as *de novo* CNVs (tan rectangles; b, d, e). Due to aberrant B-allele frequencies inconsistent with a hemizygous deletion, two events were intentionally excluded as potential cell line or somatic artifacts (a, c). The remaining event, a previously reported 5Mb duplication detected with ROMA (f), did not display signal indicative of a duplication using SNP arrays and was not called using our HMM-based approach. Owing to the strength of association, it is important to note that the likely false positive CNVs we report above do not alter the previously reported conclusion that simplex autism is enriched for *de novo* CNVs (Sebat et al. 2007).

Supplemental Tables

Supplemental Table 1. HapMap total CNV counts

	N	CNVs	CNVs per Sample	p-value vs. child
father*	57	1453	25.49	0.9
mother*	54	1440	26.67	0.29
child*	55	1393	25.33	-
parents*	111	2893	26.06	0.49

father versus mother 0.35

*includes data from incomplete trios

Supplemental Table 2. HapMap classification of CNVs identified in probands

Classification	count
maternal inheritance	374
paternal inheritance	333
both transmitted (homozygous deletions)	287
de novo	32
unclear transmission, but inherited*	192
unclear parental CNV genotypes	22
Likely false positive in proband	126
Total CNVs in complete trios	1366
incomplete trio data	27
total transmitted	994
total de novo	32
total assigned	1026

*reflects a situation in which a CNV is inherited, but could have been transmitted from either parent

Supplemental Table 3. Putative *de novo* CNVs identified in the HapMap

Sample	chrom	start(hg18)	size	type	Overlap with other HapMap CNV calls	frequency in controls (N=2339)	exclude mother	exclude father	SD frac	Population	Notes
NA12707	chr13	103202760	25377	gain	1	0	Y	Y	0	CEU	
NA10831	chr14	78107903	94684	loss	1	0	Y	Y	0	CEU	
NA10831	chr7	84122104	262803	loss	1	0	Y	Y	0	CEU	
NA12878	chr7	1821039	30902	loss	1	0	ND	ND	0.87	CEU	contains 27kb SD block
NA12865	chr20	52043114	27829	loss	1	0	Y	Y	0	CEU	
NA18500	chr12	131661753	16465	gain	1	0	ND	ND	0	YRI	
NA18500	chr2	216033519	84284	loss	1	0	ND	ND	0	YRI	

Supplemental Table 4. Asthma total CNV counts

	N	CNVs	CNVs per Sample	p-value vs. probands
fathers	395	1925	4.87	0.59
mothers	392	1999	5.1	0.18
probands	411	2025	4.93	-
parents	787	3924	4.99	0.65
father versus mother				0.07

Supplemental Table 5. Classification of CNVs identified in probands with asthma

Classification	count
maternal inheritance	522
paternal inheritance	490
both transmitted (homozygous deletions)	264
putative *de novo*	11
unclear transmission, but inherited*	82
questionable proband CNV call	50
unclear parental CNV genotypes	68
likely false positive in proband	408
incomplete trio data	130
total	2025
total transmitted	1358
total *de novo*	11
total assigned	1369

*reflects a situation in which a CNV is inherited, but could have been transmitted from either parent

Supplemental Table 6. p-values in validation of *de novo* asthma CNVs by custom array CGH

Sample	chrom	Start(hg18)	Size	type	probes	p-value*, child	p-value, mother	p-value, father
de novo CNVs								
10871	chr1	106371568	9955627	gain	20171	$<1.2x10^{-5}$	0.85	0.93
10942	chr12	98433426	147234	loss	305	$6.3x10^{-4}$	0.56	ND
11020	chr16	15387380	809653	gain	1558	$<8.4x10^{-6}$	0.76	ND
10653	chr16	54000488	351314	gain	724	$9.3x10^{-3}$	0.61	0.79
10054	chr18	45282024	1912345	gain	ND	ND	ND	ND
10186	chr2	60591731	158513	gain	ND	ND	ND	ND
10421	chr22	17295347	2497006	loss	4891	$<8.7x10^{-6}$	0.99	0.87
2648	chr22	17295963	2486274	loss	4870	$<8.6x10^{-6}$	0.98	0.93
10846	chr4	179040624	61669	gain	253	0.0052	0.72	ND
excluded putative *de novo events*								
723	chr11	38249818	33141	loss	349	0.94	0.86	0.63
593**	chr1	195089653	74058	gain	295	0.039	0.61	0.69

*one-tailed empirical p-values. For a CNV of N probes, the null distribution for a given hybridization and sample was created using the mean signal in a sliding window of N probes across the entire array excluding CNVs predicted in initial CNV discovery.

**Overlaps a previously reported copy number polymorphism

Supplemental Table 7. Summary of CNV rates across different studies

						Counts			Fractions			
Study	N	*de novo* events	Median Size (kb)	Mean Size (kb)	mu	SD med	SD assoc	no SD	SD med	SD assoc	no SD	*de novo* freq p-value
Asthma	386	9	810	2042	1.17E-02	3	1	5	0.33	0.11	0.56	-
AGRE	1638	60	156	693	1.83E-02	12	5	43	0.2	0.08	0.72	0.22
Sebat	196	2	4051	4051	5.10E-03	0	0	2	0	0	1	0.35
Xu	159	2	2804	2804	1.69E-02	0	2	0	0	1	0	0.52
Total	2379	73	182	947	1.53E-02	15	8	50	0.21	0.11	0.68	-

Supplemental Table 8. Selection coefficient estimates using different data sets

Data Set	Number of Transmissions	Estimated Mutation Rate	95% CI	estimate of s
Asthma	772	6.5E-03	(0.0008-0.0121)	0.16 (0.02-0.31)
AGRE	3514	5.1E-03	(0.0028-0.0075)	0.13 (0.06-0.19)
Sebat et al.	392	5.1E-03	(0-0.0122)	0.13 (0-0.31)
Stefansson et al.*	9878	3.2E-03	(0.0021-0.0044)	0.08 (0.05-0.11)
Xu et al.	159	6.3E-03	(0-0.0150)	0.16 (0-0.38)
asthma, AGRE, Sebat et al.	4678	5.3E-03	(0.0033-0.0074)	0.13 (0.08-0.19)
asthma, Sebat et al. Stefansson et al.	11042	3.5E-03	(0.0024-0.0046)	0.09 (0.06-0.12)
All Studies	14556	3.9E-03	(0.0029-0.0049)	0.10 (0.07-0.13)

*Mutation rates are systematically underestimated for (Stefansson et al. 2008) as 5558 of 9878 transmissions were parent-child pairs for which no duplications and only a subset of deletions could be ascertained

Supplemental Table 9. Summary of CNV counts in the AGRE collection

	total CNV calls	autosomal CNV calls	N	CNVs per Sample
father	4626	4600	778	5.912596401
mother	5034	5025	838	5.996420048
unaffected child	3970	3969	664	5.977409639
affected child	10043	10005	1688	5.927132701

Wilcoxon signed-rank p-values comparing CNVs per sample

	father	mother	unaffected	affected
father	x	0.4939	0.9178	0.7702
mother	x	x	0.6135	0.8605
unaffected	x	x	x	0.7059
affected	x	x	x	x

Supplemental Table 10. Classification of CNVs identified in probands for AGRE collection

Classification	count
maternal inheritance	3103
paternal inheritance	3059
both transmitted (homozygous deletions)	1124
putative *de novo*	209
unclear transmission, but inherited^	330
questionable proband CNV call	67
unclear parental CNV genotypes	61
false positive in proband	2735
immune somatic rearrangement*	151
cell line artifacts, not at immune system loci**	
Total	10839
transmitted	7618
putative *de novo*	209
Assigned	7827

*see Supplemental Table 20

**see Supplemental Table 21

^reflects a situation in which a CNV is inherited, but could have been transmitted from either parent

Supplemental Table 11. Summary of previous CNV analyses of the AGRE collection

	N (AGRE) total	N (AGRE) affected	systematic screen for *de novo* events	Platform
(Sebat et al. 2007)	148	117	y	85K probe ROMA
(Kumar et al. 2008)	410	410	n	19K probe BAC array
(Szatmari et al. 2007)	2213	Not Reported	n	Affymetrix 10K SNP array
(Weiss et al. 2008)	2861	1441	n	Affymetrix 5.0 SNP array
(Bucan et al. 2009)	3832	1673	n	Illumina 550K SNP array
this study*	3896	1688	y	Illumina 550K SNP array

*represents the same underlying data as Bucan et al. with a different analysis

Supplemental Table 12. Comparison to previously reported *de novo* CNVs in AGRE

			CN	reported location	called CNV,	*de novo*,	
Study	Sample	chr	change	(hg18 or cyto)	this study	this study	comments
Sebat et al.	HI0120	chr6	loss	13997280-15261931	no	ND	excluded as mosaic deletion (see Supplemental Figure 9)
Sebat et al.	HI0120	chr13	loss	44199441-46143178	yes	yes	X
Sebat et al.	HI1392	chr7	loss	15353403-15505283	no	ND	excluded as mosaic deletion (see Supplemental Figure 9)
Sebat et al.	HI0101	chr10	gain	50562149-61478511	yes	yes	X
Sebat et al.	HI1704	chr16	loss	5992836-6200816	yes	yes	X
Sebat et al.	HI1910	chr20	gain	111824-5428110	no	ND	no evidence of CN change in data, this study (see Supplemental Figure 9)
Bucan et al.*	HI3079	chr1	loss	1q21.1	yes	yes	x
Bucan et al.*	HI3692	chr1	gain	1q21.1	yes	no	father of HI3687, HI3690, HI3690, no parental information
Bucan et al.*	HI3688	chr1	gain	1q21.1	yes	no	paternal inheritance, from HI3692
Bucan et al.*	HI3690	chr1	gain	1q21.1	yes	no	paternal inheritance, from HI3692
Bucan et al.*	HI3689	chr1	gain	1q21.1	yes	no	paternal inheritance, from HI3692
Bucan et al.	HI4971	chr8	loss	8q21.2-21.3	no	ND	sample mix-up? See 10q24.2 samples below
Bucan et al.	HI2802	chr8	loss	8q21.2-21.3	no	ND	sample mix-up? See 10q24.2 samples below
Bucan et al.	HI1414	chr8	loss	8q21.2-21.3	no	ND	sample mix-up? See 10q24.2 samples below
Bucan et al.	HI2828	chr10	loss	10q24.2	no	ND	typo in figure? has large *de novo* 8q21.2 deletion
Bucan et al.	HI2401	chr10	loss	10q24.2	no	ND	typo in figure? has small 8q21.2 deletion
Bucan et al.	HI2402	chr10	loss	10q24.2	no	ND	typo in figure? has small 8q21.2 inherited deletion from HI2401
Kumar et al., Weiss et al.	HI0646	chr16	loss	16p11.2	no	ND	false negative -- manually removed in this study
Kumar et al., Weiss et al.	HI0624	chr16	loss	16p11.2	yes	yes	x
Kumar et al., Weiss et al.	HI2467	chr16	loss	16p11.2	yes	yes	in agreement with Kumar, mosaic with HI2466
Kumar et al., Weiss et al.	HI2466	chr16	loss	16p11.2	yes	yes	in agreement with Kumar, mosaic with HI2467
Kumar et al., Weiss et al.	HI2997	chr16	loss	16p11.2	yes	yes	x
Szatmari et al.	HI0128	chr7	loss	121543000-122291000	yes	no	inherited from mother, HI0126
Szatmari et al.	HI0298	chr13	gain	47048100-47569100	yes	ND	no SNP data for father HI0297
Szatmari et al.	HI0299	chr13	gain	47048100-47569100	yes	ND	no SNP data for father HI0297
Szatmari et al.	HI2741	chr8	gain	3909530-3909710	no	ND	small CNV, no probe coverage on platform
Szatmari et al.	HI1404	chr17	gain	14304400-15237700	yes	no	inherited from father, HI1408

*inclusion of parents and children for 1q21 duplication suggest it was not the authors' intent to report these CNVs as *de novo*

Supplemental Table 13. Comparison of *de novo* CNV rates in simplex autism

	Simplex autism excluding AGRE samples, Sebat et al.	AGRE simplex autism, this study	Asthma trios, this study
Simplex autism excluding AGRE samples, Sebat et al.	10, N=78*	0.07	3.2×10^{-4}
AGRE simplex autism, this study	x	2, N=60	0.65
Asthma trios, this study	x	X	9, N=386

p-values, two-sided Fisher's exact comparing *de novo* CNVs per transmission

*diagonals entries indicate number of *de novo* events identified in N samples.

Supplemental Table 14. Relative rates of *de novo* CNVs in multiplex autism pedigrees by phenotype

	de novo CNVs	N	events / transmission	relative enrichment	p, two-sided Fisher's exact
All Affected	56	1270	$2.2x10^{-2}$	4.1	$1.6x10^{-3}$
Autism	53	1113	$2.4x10^{-2}$	4.4	$9.2x10^{-4}$
Spectrum, NQA*	3	157	9.6x10-3	1.8	0.43
Unaffected	4	368	5.4x10-3	1	-

*Spectrum = Broad Spectrum, NQA = Not Quite Autism

Supplemental Table 15. Relative rates of *de novo* CNVs in multiplex autism pedigrees by phenotype and CNV size

<500kb

	de novo CNVs	N	events / transmission	relative enrichment	p, two-sided Fisher's exact
All Affected	40	1270	1.6x10-2	2.9	0.03
Autism	38	1113	1.7x10-2	3.1	0.02
Spectrum, NQA*	2	157	6.4x10-3	1.2	1
Unaffected	4	368	5.4x10-3	1	-

>500kb

	de novo CNVs	N	events / transmission	relative enrichment	p, two-sided Fisher's exact
All Affected	16	1270	5.9x10-3	-	0.03
Autism	15	1113	6.7x10-3	-	0.03
Spectrum, NQA*	1	157	3.2x10-3	-	0.3
Unaffected	0	368	0	-	-

Supplemental Table 16. Parental origin of *de novo* CNVs

SD class	maternal	paternal	p-value	undetermined
mediated	7	6	1	2
associated	2	3	1	1
no SD	12	17	0.4583	23
all	21	26	0.5601	26

Study	maternal	paternal	p-value	undetermined
asthma	7	2	0.1797	0
AGRE	14	24	0.1433	26
combined	21	26	0.5601	26

Supplemental Table 17. Selection coefficients for various human diseases

Disease	Mode of Inheritance	Selection coefficient, *s*	Reference
Porphyria Variegata	Auto. Dominant	0.02-0.07	(Stine and Smith 1990)
Lipoid Proteinosis	Auto. Recessive	0.07	(Stine and Smith 1990)
BRCA1 mutations	Auto. Dominant	0.04-0.08	(Pavard and Metcalf 2007)
Huntington Disease	Auto. Dominant	0.34	(Stine and Smith 1990)
Achondroplasia	Auto. Dominant	0.8	(Møorch and Andersen 1941), (Crow 1986)

Supplemental Table 18. Estimation of manual curation error rates for Illumina 550K SNP arrays

All Curated Loci Excluding CNP 2082*, 550K Probes

		McCarroll et al., 2008				
		CN ≠ 2	CN = 2	Total		
Manual	CN ≠ 2	42	1	43	Sensitivity:	1
Curation	CN = 2	0	2372	2372	Specificity:	>0.999
	Total	42	2373	2415		

***CNP 2082 is entirely within segmental duplication. The table above therefore represents an estimate of the performance of manual curation outside of SDs.

Individual Inspected Loci

CNP 2082** hg18, chr15:32487975-32617680

		CN, McCarroll et al., 2008			
		0	1	2	3
CN, Manual	0	4	0	0	0
Curation	1	0	48	0	1
	2	0	0	207	0
	3	0	0	2	7

**Locus is entirely within segmental duplication

CNP 12657 chr19:46143504-46205185

		CN, McCarroll et al., 2008	
		1	2
CN, Manual	1	4	0
Curation	2	0	265

CNP 10684 hg18, chr4:28251431-28339922

		CN, McCarroll et al., 2008	
		1	2
CN, Manual	1	2	0
Curation	2	0	266

CNP 11361 hg18, chr8:4598305-4697836

		CN, McCarroll et al., 2008		
		1	2	3
CN, Manual	1	2	0	0
Curation	2	0	264	0
	3	0	0	3

CNP 12054 hg18, chr12:130466075-130524698

		CN, McCarroll et al., 2008		
		1	2	3
CN, Manual	1	3	0	0
Curation	2	0	263	0
	3	0	1	1

CNP 2174 hg18, chr16:34324072-34614568

		CN, McCarroll et al., 2008		
		2	3	4
CN, Manual	2	251	0	0
Curation	3	0	16	2
	4	0	0	0

CNP 12167 hg18, chr14:44895806-45085468

		CN, McCarroll et al., 2008	
		1	2
CN, Manual	1	3	0
Curation	2	0	265

CNP 10791 hg18, chr4:132165824-132577643

		CN, McCarroll et al., 2008	
		2	3
CN, Manual	2	266	0
Curation	3	0	2

CNP 11185 hg18, chr7:9093698-9196410

		CN, McCarroll et al., 2008	
		1	2
CN, Manual	1	2	0
Curation	2	0	266

CNP 11200 hg18, chr7:19379124-19511836

		CN, McCarroll et al., 2008	
		2	3
CN, Manual	2	266	0
Curation	3	0	2

Supplemental Table 19. Estimation of manual curation error rates for Illumina 1MDuo SNP arrays.

All Loci Excluding CNP 1434*, Illumina 1MDuo Probes

		McCarroll et al., 2008				
		CN ≠ 2	CN = 2	Total		
Manual	CN ≠ 2	36	5	41	Sensitivity	0.95
Curation	CN = 2	2	2375	2377	Specificity	>0.99
	Total	38	2380	2418		

*****Manual curation performed poorly on CNP 1434, a locus entirely within segmental duplication. The table above therefore represents an estimate of the performance of manual curation outside of SDs.

Individual Inspected Loci

CNP 1434** hg18, chr9:43255666-43735571

		CN, McCarroll et al., 2008			
		0	1	2	3
CN, Manual	0	19	2	1	0
Curation	1	0	23	2	0
	2	3	60	124	0
	3	0	0	29	0

**Locus is entirely within segmental duplication

CNP 11939 hg18, chr12:303069-414108

		CN, McCarroll et al., 2008	
		2	3
CN, Manual	2	266	0
Curation	3	0	2

CNP 12360 hg18, chr15:91656315-91667141

		CN, McCarroll et al., 2008	
		1	2
CN, Manual	1	1	1
Curation	2	1	265

CNP 10168 hg18, chr1:176926933-176940811

		CN, McCarroll et al., 2008		
		1	2	3
CN, Manual	1	3	3	0
Curation	2	0	262	0
	3	0	1	0

CNP 12670 hg18, chr19:50542545-50583764

		CN, McCarroll et al., 2008	
		2	3
CN, Manual	2	266	0
Curation	3	0	3

CNP 2200 hg18, chr16:74115584-74133500

		CN, McCarroll et al., 2008	
		2	3
CN, Manual	2	254	1
Curation	3	0	14

CNP 11361 hg18, chr8:4598305-4697836

		CN, McCarroll et al., 2008		
		1	2	3
CN, Manual	1	2	0	0
Curation	2	0	264	0
	3	0	0	3

CNP 11816 hg18, chr11:4466713-4518969

		CN, McCarroll et al., 2008	
		1	2
CN, Manual	1	2	0
Curation	2	0	267

CNP 12515 hg18, chr17:36920703-36936394

		CN, McCarroll et al., 2008	
		2	3
CN, Manual	2	266	0
Curation	3	0	3

CNP 10674 hg18, chr4:18697657-18733331

		CN, McCarroll et al., 2008	
		1	2
CN, Manual	1	3	0
Curation	2	0	265

Supplemental Table 20. Loci excluded as sites of immune somatic rearrangement

chrom	start (hg18)	end (hg18)	locusName	comments
chr14	21214600	22095500	TCRalpha	encodes alpha chain of T-cell receptor; undergoes VJ-recombination
chr7	141636000	142225000	TCRbeta	encodes beta chain of T-cell receptor; undergoes V(D)J-recombination
chr14	105046000	106368000	IgHeavy	encodes immunoglobulin heavy chain
chr22	20675000	21620000	IgLambda	encodes immunoglobulin lambda light chain
chr2	88935000	89418000	IgKappa	encodes immunoglobulin kappa light chain
chr6	29775000	33225000	HLA*	HLA locus*

*excluded due to high variability in structure and sequence, making interpretation of array data difficult

Supplemental Table 21. CNV calls in AGRE pedigrees outside of immune loci marked as potential cell line artifacts

chrom	start (hg18)	end (hg18)	sample(s)	comments
chr3	15234465	15323827	HI0120	possible mosaic deletion
chr7	38297796	38310481	HI2591	possible mosaic deletion
chr7	15445998	15503875	HI1392	possible mosaic deletion
chr11	1830648	1868015	HI5581	possible mosaic deletion / false positive
chr14	104225150	104462050	HI0120,HI3581, HI2862, HI3855	ambiguous LogR normalization in region - mosaic deletion?
chr14	104686693	104850350	HI2862	possible mosaic deletion / false positive
chr19	20717774	20972627	HI0507	possible mosaic deletion

Supplemental References

Bucan, M., Abrahams, B.S., Wang, K., Glessner, J.T., Herman, E.I., Sonnenblick, L.I., Alvarez Retuerto, A.I., Imielinski, M., Hadley, D., Bradfield, J.P. et al. 2009. Genome-wide analyses of exonic copy number variants in a family-based study point to novel autism susceptibility genes. *PLoS Genet* **5:** e1000536.

Conrad, D.F., Pinto, D., Redon, R., Feuk, L., Gokcumen, O., Zhang, Y., Aerts, J., Andrews, T.D., Barnes, C., Campbell, P. et al. 2009. Origins and functional impact of copy number variation in the human genome. *Nature*.

Cooper, G.M., Zerr, T., Kidd, J.M., Eichler, E.E., and Nickerson, D.A. 2008. Systematic assessment of copy number variant detection via genome-wide SNP genotyping. *Nat Genet*.

Crow, J.F. 1986. *Basic concepts in population, quantitative, and evolutionary genetics*. W.H. Freeman, New York.

Kumar, R.A., KaraMohamed, S., Sudi, J., Conrad, D.F., Brune, C., Badner, J.A., Gilliam, T.C., Nowak, N.J., Cook, E.H., Jr., Dobyns, W.B. et al. 2008. Recurrent 16p11.2 microdeletions in autism. *Hum Mol Genet* **17:** 628-638.

McCarroll, S.A., Kuruvilla, F.G., Korn, J.M., Cawley, S., Nemesh, J., Wysoker, A., Shapero, M.H., de Bakker, P.I., Maller, J.B., Kirby, A. et al. 2008. Integrated detection and population-genetic analysis of SNPs and copy number variation. *Nat Genet* **40:** 1166-1174.

Møorch, E.T. and Andersen, H. 1941. *Chondrodystrophic dwarfs in Denmark*. E. Munksgaard, Copenhagen,.

Pavard, S. and Metcalf, C.J. 2007. Negative selection on BRCA1 susceptibility alleles sheds light on the population genetics of late-onset diseases and aging theory. *PLoS One* **2:** e1206.

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.

Stefansson, H., Rujescu, D., Cichon, S., Pietilainen, 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*.

Stine, O.C. and Smith, K.D. 1990. The estimation of selection coefficients in Afrikaners: Huntington disease, porphyria variegata, and lipoid proteinosis. *Am J Hum Genet* **46:** 452-458.

Szatmari, P. Paterson, A.D. Zwaigenbaum, L. Roberts, W. Brian, J. Liu, X.Q. Vincent, J.B. Skaug, J.L. Thompson, A.P. Senman, L. et al. 2007. Mapping autism risk loci using genetic linkage and chromosomal rearrangements. *Nat Genet* **39:** 319-328.

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.