Supplementary Information: Detecting copy number variants using exome genotyping arrays in a large Swedish schizophrenia sample

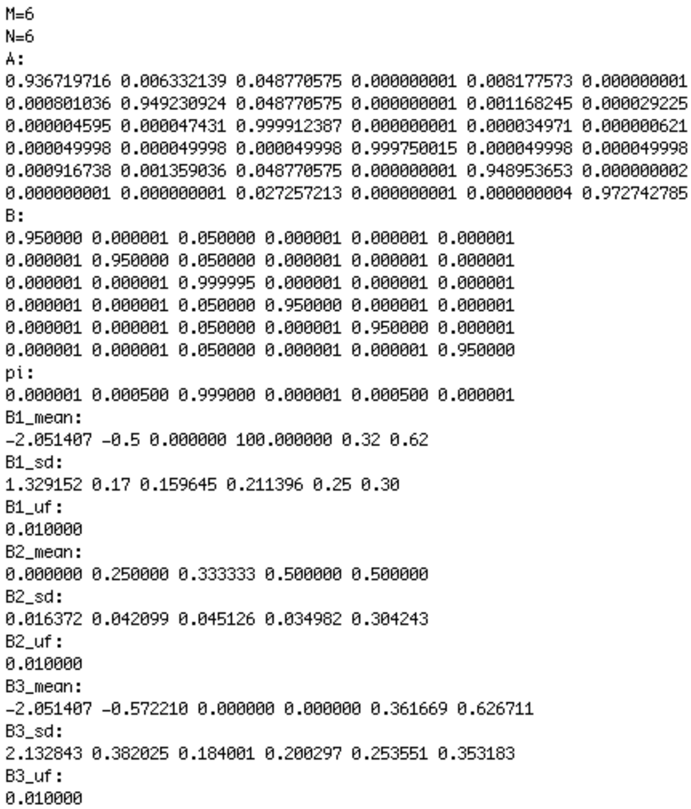
Table of Contents

Supplemental Methods

Obtaining CNV calls from the Illumina exome array

1. Obtain signal intensity values (Log R Ratio and B Allele Frequency) using Illumina's GenomeStudio v2010.3 with the calling algorithm/genotyping module version 1.8.4. A custom cluster file could be created using the GenCall algorithm based on all samples.
2. Obtain and install PennCNV (http://www.openbioinformatics.org/penncnv/penncnv_download.html)
3. Prepare PennCNV signal intensity files from Illumina Report file (Step 1) using "split_illumina_report.pl" script. (http://www.openbioinformatics.org/penncnv/penncnv_input.html)
4. Prepare the GC-model file required by PennCNV:
    a. Download the GC-content file whose build is the same as the signal intensity files (http://hgdownload.cse.ucsc.edu/goldenPath/hg19/database/gc5Base.txt.gz)
    b. Unzip the file and use the linux command "sort -k 2,2 -k 3,3n" to sort by chromosome and position
    c. Run the PennCNV "cal_gc_snp.pl" script to compute the GC model file
5. Prepare PennCNV population frequency of B-allele (PFB) file using EITHER (a) or (b):
    a. Select a large number of representative samples (e.g. 300) and use their intensity files and the "compile_pfb.pl" script in PennCNV to compute the PFB file.
    b. Obtain genotypes of a large number of representative samples using GenomeStudio v2010.3 with the calling algorithm/genotyping module version 1.8.4 with subsequent processing of genotype calling done by the zCall algorithm. Compute the PFB from these genotype calls.
    c. From (a) or (b) above, SNPs with PFB=0 are treated as intensity-only markers by changing to PFB=2.
6. Prepare the HMM model file specialized for the Illumina exome array.
    a. The HMM model file used in this study is provided at the end of this instruction. Additional assistance can be obtained from jin_szatkiewicz@med.unc.edu.
    b. More generally, modify the emission parameters in the default HMM model file "hhall.hmm" provided by PennCNV. The initial emission parameters can be estimated empirically from high-confidence large CNVs, through the BeadStudio Genome Viewer for a set of genotyped individuals and simple linear interpolation. Optimization can be carried out by testing a series of parameter values using a large set of training samples.
7. Obtain CNV calls
    a. Use the PennCNV "detect_cnv.pl" script and with the customized parameter files constructed in Steps 4 through 6 to generate CNV calls for each sample.
    b. Make sure the -conf (to produce confidence score of the CNV calls) and the -log (to document intensity-based sample quality metrics in a log file) options are specified with "detect_cnv.pl".
8. Apply customized quality control to the CNV calls
    a. For example, remove CNV calls with confidence score < 10, or spanning < 10 probes.
    b. Anneal any CNVs that appeared to be artificially split by the PennCNV HMM model using the "combineseg" option of the "clean_cnv.pl" script from PennCNV
9. Apply subject quality control based on CNV metrics
    a. Use the "filter_cnv.pl" script to obtain a summary file from the log file produced in step 7 that contains intensity-based quality control metrics for each sample.
    b. Remove low quality samples with extreme values (e.g., >95th percentile) for LRR_standard deviation or BAF_drift, or were outliers with respect to the total number or total base pairs of CNV calls.

HMM model file used in this study:



```
M=6
N=6
A:
0.936719716 0.006332139 0.048770575 0.000000001 0.008177573 0.000000001
0.000801036 0.949230924 0.048770575 0.000000001 0.001168245 0.000029225
0.000004595 0.000047431 0.999912387 0.000000001 0.000034971 0.000000621
0.000049998 0.000049998 0.000049998 0.999750015 0.000049998 0.000049998
0.000916738 0.001359036 0.048770575 0.000000001 0.948953653 0.000000002
0.000000001 0.000000001 0.027257213 0.000000001 0.000000004 0.972742785
B:
0.950000 0.000001 0.050000 0.000001 0.000001 0.000001
0.000001 0.950000 0.050000 0.000001 0.000001 0.000001
0.000001 0.000001 0.999995 0.000001 0.000001 0.000001
0.000001 0.000001 0.050000 0.950000 0.000001 0.000001
0.000001 0.000001 0.050000 0.000001 0.950000 0.000001
0.000001 0.000001 0.050000 0.000001 0.000001 0.950000
pi:
0.000001 0.000500 0.999000 0.000001 0.000500 0.000001
B1_mean:
-2.051407 -0.5 0.000000 100.000000 0.32 0.62
B1_sd:
1.329152 0.17 0.159645 0.211396 0.25 0.30
B1_uf:
0.010000
B2_mean:
0.000000 0.250000 0.333333 0.500000 0.500000
B2_sd:
0.016372 0.042099 0.045126 0.034982 0.304243
B2_uf:
0.010000
B3_mean:
-2.051407 -0.572210 0.000000 0.000000 0.361669 0.626711
B3_sd:
2.132843 0.382025 0.184001 0.200297 0.253551 0.353183
B3_uf:
0.010000
```

The Swedish Schizophrenia Study

Subjects

All procedures were approved by ethical committees in Sweden and in the US, and all subjects provided written informed consent (or legal guardian consent and subject assent). Data collection for this study took six years (2005-2011). As shown in Table S1, GWAS genotyping was conducted in six separate batches (denoted Sw1-Sw6) using three GWAS arrays (Affymetrix 5.0, Affymetrix 6.0, and Illumina Omni Express). Genotypes were generated as sufficient numbers of samples accumulated from the field work in Sweden. Of the total sample of 11,850 Swedish subjects before QC (5,351 cases with schziphrenia, 6,509 controls), 57.4% (Sw5 and Sw6) have never been reported previously.

Table S1: The six genotyping batches comprising the Swedish sample

Feature	**Sw1**	**Sw2**	**Sw3**	**Sw4**	**Sw5**	**Sw6**
GWAS arrays	affy5	affy6	affy6	affy6	ioexp	ioexp
Subjects (pre-QC)	464	694	1,498	2,388	4,461	2,345

Affy5: Affymetrix 5.0; Affy6: Affymetrix 6.0; ioexp: Illumina Omni Express.

Cases with schizophrenia were identified via the Swedish Hospital Discharge Register (1, 2) which captures all public and private inpatient hospitalizations. The register is complete from 1987 and augmented by psychiatric data from 1973-86. The register contains ICD discharge diagnoses (3-5) made by attending physicians for each hospitalization. (6-9) Case inclusion criteria: ≥2 hospitalizations with a discharge diagnosis of schizophrenia, both parents born in Sweden, and age ≥18 years. Case exclusion criteria: hospital register diagnosis of any medical or psychiatric disorder mitigating a confident diagnosis of schizophrenia as determined by expert review, and included removal of 3.4% of eligible cases due to the primacy of another psychiatric disorder (0.9%) or a general medical condition (0.3%) or uncertainties in the Hospital Discharge Register (e.g., contiguous admissions with brief total duration, 2.2%). The validity of this case definition of schizophrenia is strongly supported.

Controls were selected at random from Swedish population registers. Our goal was to obtain an appropriate control group and to avoid "super-normal" controls. (10) Control inclusion criteria: never hospitalized for schizophrenia or bipolar disorder (given evidence of genetic overlap with schizophrenia), (11-13) both parents born in Sweden, and age ≥18 years.

Genomic characterization and quality control

All genomic locations are given in NCBI build 37/UCSC hg19 coordinates. DNA was extracted from peripheral blood samples using Qiagen technologies at the Karolinska Institutet Biobank.

GWAS array genotyping. As shown in ***Table S1***, samples were genotyped in six batches at the Broad Institute using Affymetrix 5.0 (3.9%), Affymetrix 6.0 (38.6%), and Illumina OmniExpress (57.4%) arrays according to the manufacturers' protocols. Genotype calling and quality control was done in four sets corresponding to data from Affymetrix 5.0 (Sw1), Affymetrix 6.0 (Sw2-4), and the OmniExpress batches (Sw5, Sw6). Genotypes were called using Birdsuite (Affymetrix) or BeadStudio (Illumina).

A multi-step quality control (QC) procedure was carried out. The exclusionary measures of basic quality control were: SNP missingness ≥ 0.05 (before sample removal); subject missingness ≥ 0.02; autosomal heterozygosity deviation; SNP missingness ≥ 0.02 (after sample removal); difference in SNP missingness between cases and controls ≥ 0.02; and deviation from Hardy-Weinberg equilibrium ($P < 10^{-6}$ in controls or $P < 10^{-10}$ in cases). After basic quality control, 77,986 autosomal SNPs directly genotyped on all three GWAS platforms were extracted and pruned to remove SNPs in LD ($r^2 > 0.05$) or with minor allele frequency < 0.05, leaving 39,239 SNPs suitable for robust relatedness testing. Relatedness testing was done with PLINK (16) and pairs of subjects with $\pi > 0.2$ were identified and one member of each relative pair removed at random.

Following quality control, a total of 11,224 subjects (5,001 cases with SCZ and 6,243 controls) remained and were used for subsequent CNV calling and analysis.

Exome array genotyping. DNA samples were sent to the Broad Institute Genetic Analysis Platform (GAP) for genotyping, are placed on 96-well plates for processing using the Illumina Infinium HumanExome BeadChip v1.0. Majority of Exome genotypes were called using GenomeStudio v2010.3 with the calling algorithm/genotyping module version 1.8.4 using the custom cluster file StanCtrExChp_CEPH.egt, subsequent processing of genotype calling was done by zCall(14). The Broad Institute did not filter any SNPs based off of technical quality control metrics. Only samples passing an overall call rate of 98% criteria and standard identity check were released from GAP. Then the 11,224 subjects that passed SNP-based QC filters as described in the previous section "GWAS array genotyping" were extracted for subsequent CNV calling and analysis.

CNV calling and quality control procedures

GWAS array CNV calling and QC

We applied two methods, the Birdseye tool in Birdsuite(17) and the PennCNV software (June 2011 version) (15) to autosomal intensity data from both SNP and CNV probes. All genomic locations are given in NCBI build 37/UCSC hg19 coordinates. Birdseye applies Hidden Markov Model (HMM) to the normalized probe intensities for each allele, using model priors (i.e. allele specific probe responses) generated separately for each type of GWAS arrays. PennCNV applies a HMM to the log R ratios (LRR) and B allele frequencies (BAF). For Ilumina arrays, LRR and BAF were produced by Illumina's GenomeStudio (v2010.3). For Affymetrix arrays, we used the PennCNV procedure (http://www.openbioinformatics.org/penncnv/penncnv_tutorial_affy_gw6.html) to prepare LRR and BAF from Affymetrix .CEL files. For CNV calling, we used PennCNV's default program parameters recommended for each array type (affygw5.hmm, affgw6.hmm, hhall.hmm for Affy5.0, Affy6.0, and Illumina OmniExpress respectively). In addition, the default genomic wave adjustment routine in detect_cnv.pl program were used in generating CNV calls using default model files (affygw5.gcmodel, affygw6.gcmodel, hhall.gcmodel).

For each initial callset (one from Birdseye and the other from PennCNV), a multi-step quality control procedure was applied. First, CNVs were excluded if they were of low confidence CNV (LOD <10, size < 20kb, or spanning < 10 probes). Any CNVs that appeared to be artificially split by the HMM were annealed using an in-house perl script, which recursively joins CNVs as long as the called region is greater or equal to 80% of the entire region to be joined. CNVs were also excluded if they had any overlap with large genomic gaps (downloaded from the UCSC table browser). Next, additional subject quality control was carried out using CNV metrics (***Table S2***). Specifically, subjects were excluded if they had > 40 CNV calls or > 10Mb of CNVs using Birdseye algorithm. (18) Finally, we imposed a 1% frequency threshold, by removing any CNV with greater than 50% of its length spanning a region with CNVs from >1% of total post-QC subjects as implemented in PLINK. (16)

Exome array CNV calling and QC

CNV calling began with raw intensity data processing. A custom cluster file was created using the GenCall algorithm based on all samples. Normalized intensity values were obtained using Illumina's GenomeStudio (v2010.3) with the calling algorithm/genotyping module (v1.8.4). PennCNV (June 2011 version) (15) was applied to the log R ratios (LRR) and B allele frequencies (BAF) calculated from the normalized intensity values. PennCNV implements an hidden Markov model (HMM) that incorporates multiple sources of information, including LRR and BAF at each SNP marker, the distance between neighboring SNPs, and SNP allele frequencies. See section "Obtaining CNV calls from the Illumina eoxme array" for details.

A multi-step quality control procedure was applied to the initial CNVs. First, CNVs were excluded if they were of low-confidence (confidence score < 10, or spanning <10 probes, or confidence:probe ratio < 0.5, or span > 20kb per supporting probe on average). Any CNVs that appeared to be artificially split by the HMM were annealed using the "combineseg" option of the "clean_cnv.pl" script from PennCNV software. CNVs were also excluded if they had any overlap with genomic gaps (downloaded from the UCSC table browser). Next, additional subject quality control was carried out using CNV metrics (***Table S2***). Specifically, subjects were excluded if they had extreme values for LRR_standard deviation (> 0.2, 95th percentile) or BAF_drift (> 0.01, 95th percentile), or were outliers with respect to the total number of CNV calls (>152, 95th percentile). Finally, we imposed a 1% frequency threshold by removing any CNV with > 50% of its length spanning a region with CNVs from >1% of total post-QC subjects as implemented in PLINK. (16) All large CNVs were visually inspected using a custom R script to generate intensity values of the probes involved in CNVs and in the flanking regions (e.g., ***Figure S3***).

CNV datasets for comparison

As summarized in ***Table S2***, for combined analysis, we identified 9,100 subjects (3,962 cases with SCZ and 5,138 controls) passing all quality control procedures. For analysis stratified by GWAS array type, we created CNV subsets for 307 (3.4%) subjects that were genotyped using Affymetrix 5.0; for 3,030 (33.3%) subjects that were genotyped using Affymetrix 6.0; and for 5,736 (63.3%) subjects that were genotyped using Illumina Omni Express.

Table S2. Summary of Subject Quality Control

Genotyping batch	**Sw1**	**Sw2,3,4**	**Sw5,6**	**Total**
By GWAS Array types	Affymetrix 5.0	Affymetrix 6.0	Illumina OmniExpress	
Total subjects genotyped	464	4,580	6,806	11,850
Subjects passing QC I based on GWAS SNP metrics[1]	427	4,261	6,556	11,244
Subjects passing QC II based on CNV metrics for both exome and GWAS arrays (***Analysis dataset***)[2]	307 (3.4%)	3,030 (33.3%)	5,763 (63.3%)	9,100

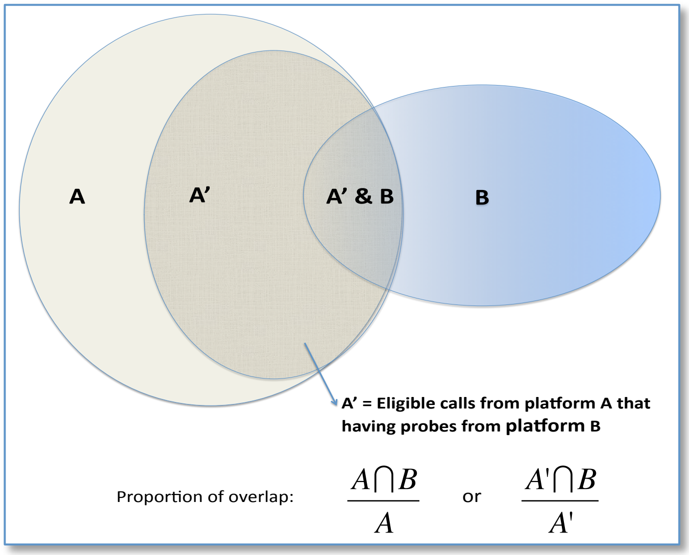
1. Only subjects that passed GWAS QC based on SNP metrics were considered for CNV analysis.

2. Additional subject QC based on CNV metrics was conducted separately for exome array and for GWAS arrays. Finally, subjects that passed all QC steps were identified, comprising the final subjects for combined and stratified comparisons.

Method for comparing CNV datasets

As illustrated in ***Figure 1***, we carried out comparisons with and without restriction on probe overlap between two experimental platforms of interests. In ***Figure S1***, A is defined as all CNVs from dataset A. A’ is defined as a subset of eligible calls comprised of any CNV that have ≥ 3 probes on platform B and on average every such B probes supports <20kb of the CNV. The numerators are any CNV in A or A’ detected by dataset B. A CNV in A is detected when ≥ 50% of its length is overlapped by a CNV in B. The perl scripts accompanying the XHMM package (http://atgu.mgh.harvard.edu/xhmm/) and custom R scripts were used to calculate overlap between two CNV datasets.

Figure S1 strategy for comparing two datasests



A’ = Eligible calls from platform A that having probes from platform B

Proportion of overlap: $\frac{A \cap B}{A}$ or $\frac{A' \cap B}{A'}$

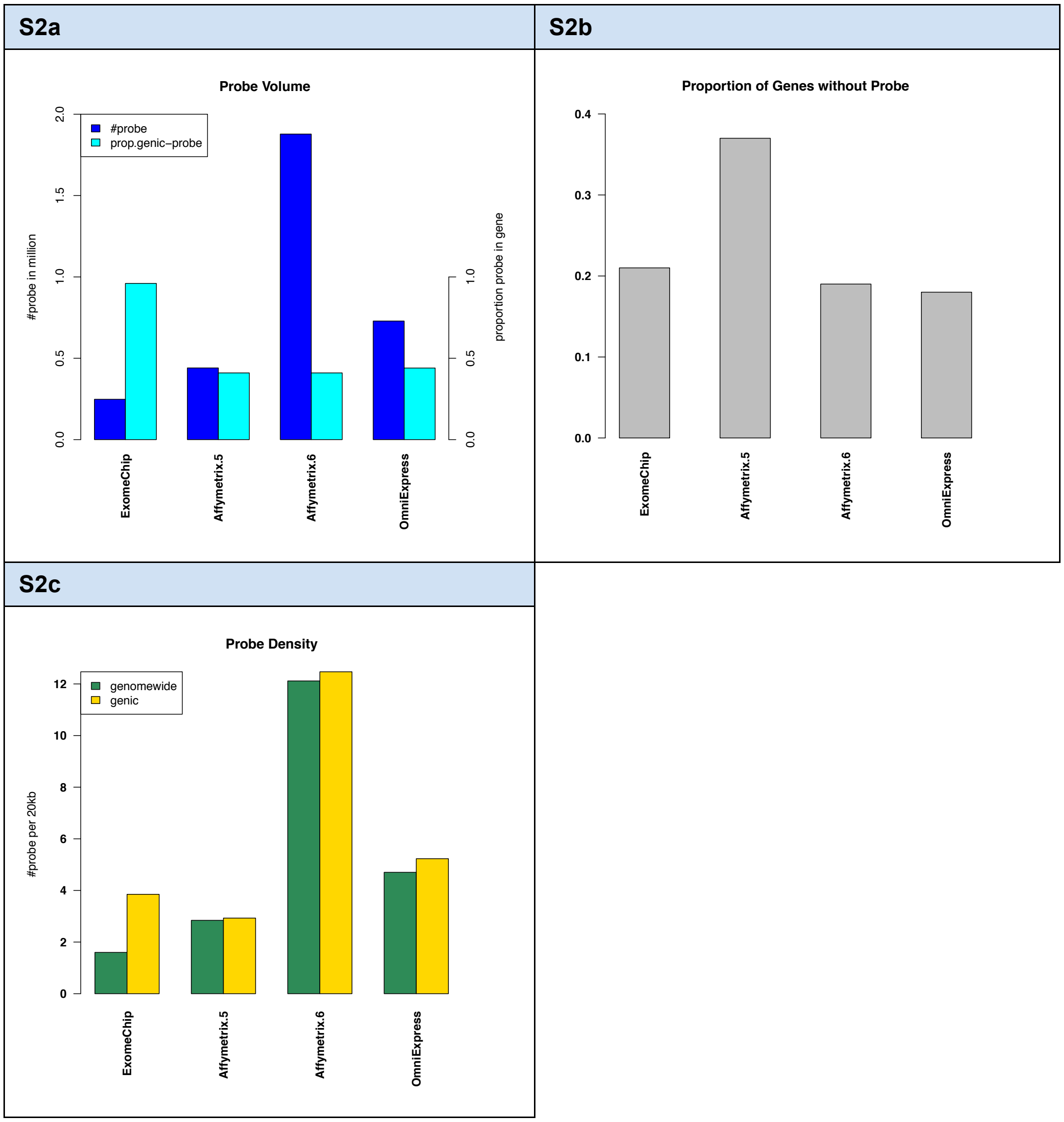
Supplemental Results

Table S3. Probe content comparison genome-wide and in genes

Platform	Genome		Gene[1]				Gene (%) no probe coverage
	Probes	Density[2]	#Probes (% in genes)	Density[2]	Median #probe/gene	Range[3] #probe/gene	
ILMN Exome array	247,870	1.6	238,754 (96%)	3.9	9	1-849	4,858 (21%)
AFFY SNP 5.0	440,638	2.8	181,457 (41%)	2.9	5	1-842	8,645 (37%)
AFFY SNP 6.0	1,877,941	12.1	772,449 (41%)	12.5	14	1-3043	4,394 (19%)
ILMN Omni Express	728,816	4.7	323,731 (44%)	5.2	7	1-1624	4,126 (18%)

1. The gene model is the maximal transcripts for RefSeq mRNA genes resulting in a total of 23,101 genes. 2. Density given as probes/20kb. 3. Range: minimum #probes/gene to maximum #probes/gene

Figure S2. Probe content comparison genome-wide and in genes



We defined a “probe” as any microarray SNP or copy number probe. A genic probe is any probe that overlaps a gene by at least 1bp.

As shown in ***Table S3 and Figures S2***, the exome array has 96% of all probes in genes, and covers 79% of all genes, which similar to the genic coverage of high-density GWAS chips (81% for Affymetrix 6.0 and 82% for Illumina Omni Express). Within genes, the exome array has a higher mean probe density (3.85 probes/20 kb) than Affymetrix 5.0 (2.93 probes per 20kb) but lower than high-density GWAS chips (12.47 probes/20kb for Affymetrix 6.0 and 5.23 probes/20kb for Omin Express).

The genomic distribution of exome array probes is non-uniform. The genomic distance between consecutive probes ranges from 0bp to 22Mb genome-wide and from 0bp to 447kb within genes, with a median inter-probe distance of 171bp.

In terms of allele frequencies, 44% of SNPs on the exome array are not polymorphic and 40% have minor allele frequency < 0.01, based on genotypes generated by zCall (14) from 18,056 chromosomes.

These results suggest that exome array has sufficient genic-probes to interrogate gene-centric CNV although probe features that are specific to the exome array must be taken into account by the computational methods used for detecting CNV.

Table S4. Probe content in regions of interests

Source	Chr	Start	End	#exome.chip.probes	#probes/kb
asd	chr1	10002	5408761	1023	5.28
asd	chr1	146512377	147737376	107	11.45
asd	chr10	4710002	10559994	402	14.55
asd	chr10	81692665	88942014	338	21.45
asd	chr10	128010011	135524747	585	12.85
asd	chr11	2013425	2913424	156	5.77
asd	chr11	31803510	32510988	13	54.42
asd	chr11	43985278	46064560	196	10.61
asd	chr11	115894792	134946516	2318	8.22
asd	chr15	22876633	28557186	297	19.13
asd	chr15	30769996	32701482	115	16.8
asd	chr15	74377175	76162277	439	4.07
asd	chr15	99357971	102521392	444	7.12
asd	chr16	3781465	3861246	19	4.2
asd	chr16	15504455	16284248	148	5.27
asd	chr16	21613957	29042192	800	9.29
asd	chr17	2	2545429	481	5.29
asd	chr17	16706022	20482061	563	6.71
asd	chr17	29162823	30218667	114	9.26
asd	chr17	34907367	36076803	137	8.54
asd	chr17	43632467	44210205	97	5.96
asd	chr2	57741797	61738334	171	23.37
asd	chr2	149216039	149271044	18	3.06
asd	chr2	196925090	205206940	660	12.55
asd	chr2	239954694	243102476	552	5.7
asd	chr22	18546350	22336469	623	6.08
asd	chr22	51045517	51187844	29	4.91
asd	chr3	195672230	197497869	282	6.47
asd	chr4	10002	2073670	502	4.11
asd	chr4	82009852	82963464	24	39.73
asd	chr5	10001	11723854	749	15.64
asd	chr5	88016167	88179024	1	-
asd	chr5	175130403	177456545	447	5.2
asd	chr7	72332744	74616901	229	9.97
asd	chr8	8119296	11765719	452	8.07
asd	chr9	140403364	141153431	107	7.01
asd	chrX	152749901	153390999	166	3.86
decipher	chr1	10002	5408761	1023	5.28
decipher	chr1	145401254	145928123	161	3.27
decipher	chr1	146512931	147737500	107	11.44
decipher	chr11	31803510	32510988	13	54.42
decipher	chr11	43985278	46064560	196	10.61
decipher	chr12	65071920	68645525	179	19.96
decipher	chr15	22876633	28557186	297	19.13
decipher	chr15	30769996	32701482	115	16.8
decipher	chr15	74377175	76162277	439	4.07
decipher	chr15	99357971	102521392	444	7.12
decipher	chr16	60002	834372	554	1.4
decipher	chr16	3781465	3861246	19	4.2
decipher	chr16	15504455	16284248	148	5.27
decipher	chr16	21613957	29042192	800	9.29
decipher	chr16	29501199	30202572	166	4.23
decipher	chr17	2	2545429	481	5.29
decipher	chr17	13968608	15434038	58	25.27
decipher	chr17	16706022	20482061	563	6.71
decipher	chr17	29162823	30218667	114	9.26
decipher	chr17	34907367	36076803	137	8.54
decipher	chr17	43632467	44210205	97	5.96
decipher	chr2	57741797	61738334	171	23.37
decipher	chr2	196925090	205206940	660	12.55
decipher	chr2	239954694	243102476	552	5.7
decipher	chr21	27037957	27548479	35	14.59
decipher	chr22	2	16971860	0	-
decipher	chr22	18546350	23696229	700	7.36
decipher	chr22	51045517	51187844	29	4.91

decipher	chr3	195672230	197497869	282	6.47
decipher	chr4	10002	2073670	502	4.11
decipher	chr5	10002	11723854	749	15.64
decipher	chr5	112101597	112221377	54	2.22
decipher	chr5	126063046	126204952	6	23.65
decipher	chr5	175130403	177456545	447	5.2
decipher	chr7	72332744	74616901	229	9.97
decipher	chr7	95533861	96779486	31	40.18
decipher	chr8	8119296	11765719	452	8.07
decipher	chr9	140403364	141153431	107	7.01
decipher	chrX	460559	867875	2	203.66
decipher	chrX	6441958	8167697	29	59.51
decipher	chrX	102642052	103131767	30	16.32
decipher	chrX	152749901	153390999	166	3.86
devdel	chr1	10001	10077413	1661	6.06
devdel	chr1	145288644	149783376	271	16.59
devdel	chr1	171733378	172333377	20	30
devdel	chr1	242433378	248833377	668	9.58
devdel	chr10	2610001	3210000	77	7.79
devdel	chr10	46929995	48429994	125	12
devdel	chr10	81610021	88910020	335	21.79
devdel	chr10	127760011	135400010	591	12.93
devdel	chr11	310001	3443424	1033	3.03
devdel	chr11	43983425	46063424	196	10.61
devdel	chr11	67753425	71282352	479	7.37
devdel	chr11	128044791	134844790	435	15.63
devdel	chr12	229740	3629739	460	7.39
devdel	chr12	8158734	8358733	39	5.13
devdel	chr12	65073734	68643733	178	20.06
devdel	chr13	19402001	20302000	28	32.14
devdel	chr13	20812001	21012000	5	40
devdel	chr13	113602000	115031898	258	5.54
devdel	chr14	36430250	37230249	37	21.62
devdel	chr14	104480248	106378955	429	4.43
devdel	chr15	22648637	32962708	526	19.61
devdel	chr15	72912947	75792945	511	5.64
devdel	chr15	75972946	78202945	163	13.68
devdel	chr15	83182946	84738996	185	8.41
devdel	chr15	85098997	85798996	143	4.9
devdel	chr15	99362478	102521392	444	7.11
devdel	chr16	160001	5209999	2694	1.87
devdel	chr16	14892500	18292499	210	16.19
devdel	chr16	21352500	30342499	967	9.3
devdel	chr16	83792500	90222499	1419	4.53
devdel	chr17	50001	4153251	927	4.43
devdel	chr17	14069276	15499275	51	28.04
devdel	chr17	16659276	25475873	624	14.13
devdel	chr17	29025875	30215887	133	8.95
devdel	chr17	34725888	36295000	187	8.39
devdel	chr17	43644218	44184217	97	5.57
devdel	chr17	57655219	60305218	244	10.86
devdel	chr17	72088406	81060000	2880	3.12
devdel	chr18	110001	5310000	268	19.4
devdel	chr18	6760001	7360000	126	4.76
devdel	chr18	70949021	77899009	382	18.19
devdel	chr19	199001	8789000	2816	3.05
devdel	chr19	54858189	59058188	1775	2.37
devdel	chr2	110001	1720993	117	13.77
devdel	chr2	3270994	3470993	22	9.09
devdel	chr2	45346497	46046496	42	16.67
devdel	chr2	57746497	61736496	171	23.33
devdel	chr2	96726274	97676273	203	4.68
devdel	chr2	100693569	108443568	406	19.09
devdel	chr2	110822712	110982711	23	6.96
devdel	chr2	111333938	113233529	127	14.96
devdel	chr2	165691755	166391754	40	17.5
devdel	chr2	235735262	243102476	1015	7.26
devdel	chr20	152001	1152000	157	6.37
devdel	chr20	60266606	62829556	959	2.67

devdel	chr21	21028130	21328129	1	-
devdel	chr21	42478131	47975572	1382	3.98
devdel	chr22	17470001	25020000	1185	6.37
devdel	chr22	25370001	26170000	78	10.26
devdel	chr22	44268668	51244566	1137	6.14
devdel	chr3	125001	1425000	64	20.31
devdel	chr3	2125001	9825000	368	20.92
devdel	chr3	195715604	197415603	263	6.46
devdel	chr4	110001	7049099	1325	5.24
devdel	chr4	9840903	10840902	86	11.63
devdel	chr4	81730977	83130976	46	30.43
devdel	chr4	184013007	184513006	45	11.11
devdel	chr4	187263007	187963006	180	3.89
devdel	chr5	47001	1447000	373	3.75
devdel	chr5	3697001	4397000	2	350
devdel	chr5	175517395	177517394	444	4.5
devdel	chr5	180117395	180817394	118	5.93
devdel	chr6	155001	5855001	292	19.52
devdel	chr6	20742022	21142021	4	100
devdel	chr6	92043280	104693307	303	41.75
devdel	chr6	165330011	170908075	481	11.6
devdel	chr7	10239	3833474	483	7.92
devdel	chr7	5733475	6233475	100	5
devdel	chr7	66482566	72272064	66	87.72
devdel	chr7	72662065	74262064	223	7.17
devdel	chr7	74962065	76662064	144	11.81
devdel	chr8	160001	11912591	933	12.6
devdel	chr8	53287448	53887447	33	18.18
devdel	chr8	143252094	145979195	1149	2.37
devdel	chr9	160001	6760000	644	10.25
devdel	chr9	137810180	141080179	1282	2.55
psych	chr1	56280910	56291907	0	-
psych	chr1	145010957	148684147	288	12.75
psych	chr11	88629802	88712013	0	-
psych	chr15	23688945	28422026	237	19.97
psych	chr15	30212709	33212708	161	18.63
psych	chr16	15478051	16302002	154	5.35
psych	chr16	29592500	30301881	166	4.27
psych	chr17	34819671	36203752	186	7.44
psych	chr2	50146497	51490709	20	67.21
psych	chr22	18720001	21870000	498	6.33
psych	chr22	21980001	22450000	93	5.05
psych	chr22	30390001	31970000	406	3.89
psych	chr3	7208954	7222236	0	-
psych	chr3	195715604	197345603	253	6.44
psych	chr5	64992221	65010764	2	9.27
psych	chr6	146615384	146652354	2	18.48
psych	chr7	72773571	74144177	201	6.82
psych	chr7	126737889	126748966	1	-
psych	chr7	153864666	153933894	0	-
psych	chr7	159038641	159117255	0	-
psych	chr9	119260180	119860179	37	16.22

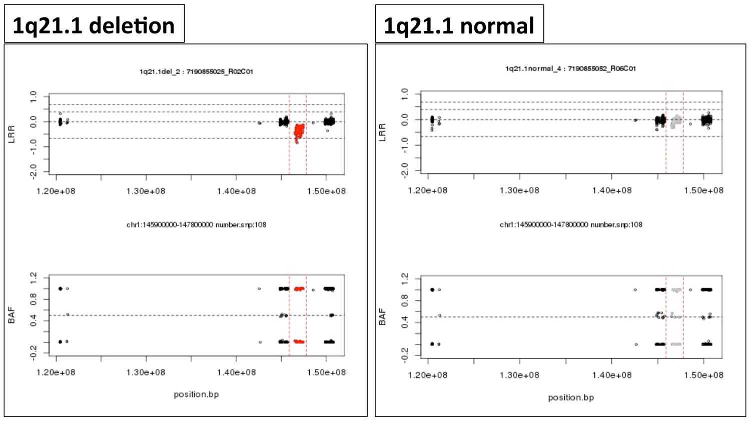
asd: CNVs important to Autism Spectrum Disorder; devdel: CNVs important to developmental delay; psych: CNVs important to psychiatric disorders ; decipher: DECIPHER CNVs.

Examples of CNVs detected by the exome array

In each sub-figure, we compare the intensity data for a specific locus of a sample with the deletion or duplication (left panel) versus a sample with normal copy number (right panel). In each panel, the x-axis indicates genomic position of the probes and y-axis indicates the values of LRR (top) or BAF (bottom). The red vertical lines indicate CNV boundaries predicted from GWAS chips. The red dots indicate exome array probes predicted to be involved in a deletion. The blue dots indicate exome array probes predicted to be involved in a duplication. The gray dots indicate the corresponding probes in the normal sample.

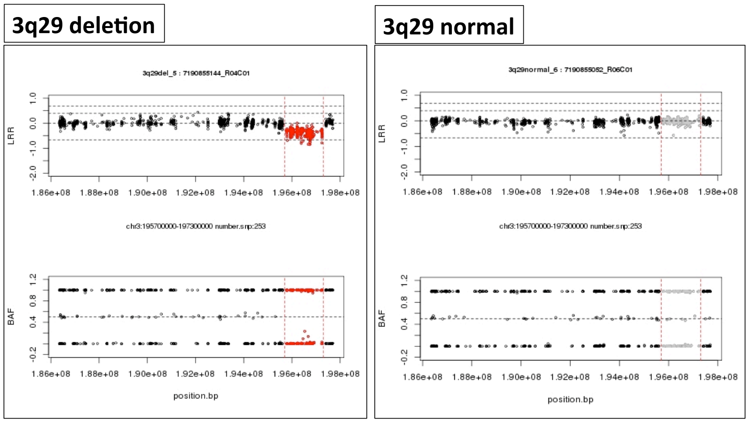
Figure S3. Five examples of large CNVs detected by exome array

Figure S3a: 1q22.1


1q21.1 deletion

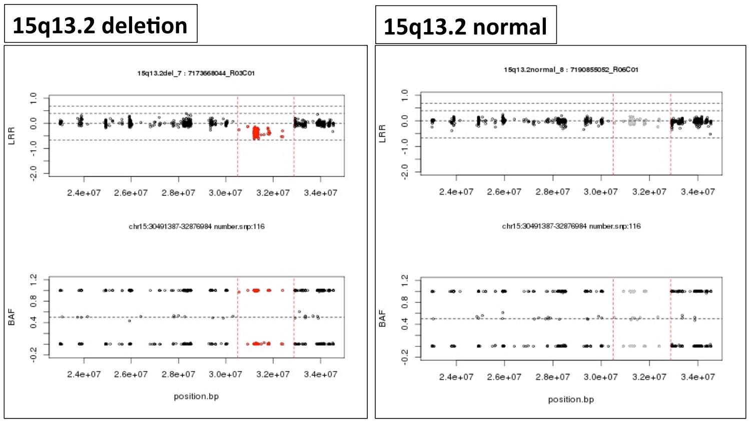
1q21.1 normal

Figure S3b: 3q29


3q29 deletion

3q29 normal

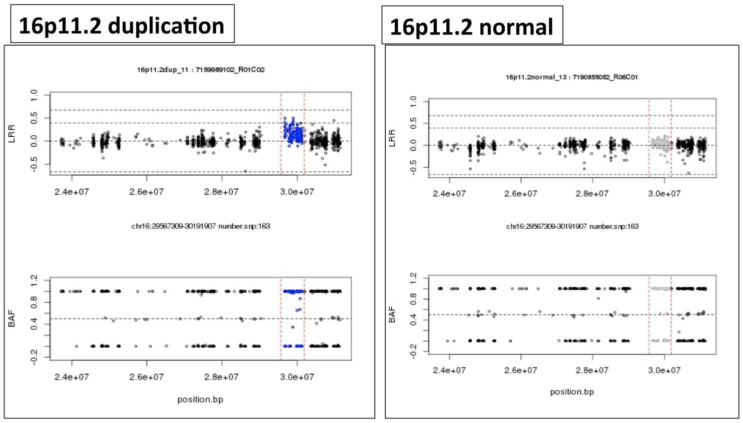
Figure S3c:15q13.2



15q13.2 deletion

15q13.2 normal

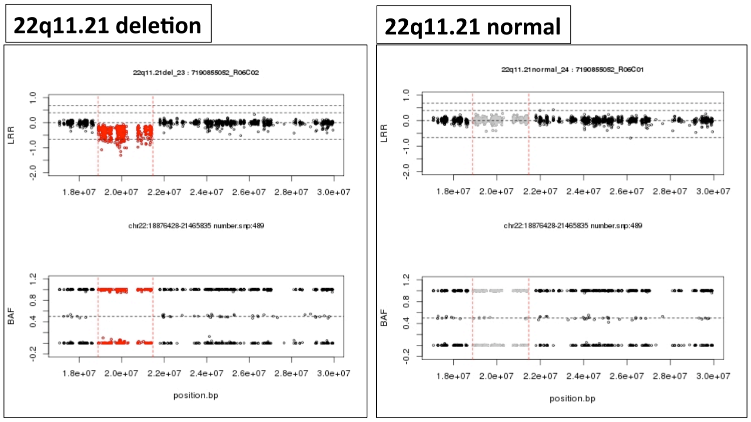
Figure S3d: 16p11.2



16p11.2 duplication

16p11.2 normal

Figure S3e: 22q11.2



22q11.21 deletion

22q11.21 normal

Summary of simulation study

Table S5. ***In sillico*** **sex-mixing results for exome array**

#Probes	Kb	Kb/SNP	Sensitivity
3	0.92	0.31	0.26
3	46.16	15.39	0.20
3	78.35	26.12	0.13
5	3.58	0.71	0.95
5	8.0	1.96	0.91
5	132.58	26.52	0.77
10	3.89	0.39	0.998
10	16.14	1.64	0.999
10	111.5	11.15	1
15	0.99	0.01	1
15	26.15	1.74	1
20	27.16	1.43	1
20	115.89	5.79	1
30	121.48	4.05	1
35	12.9	0.37	1

In sillico sex-mixing experiments: After excluding the pseudo-autosomal regions (in order to remove natural copy number variation), we used a total 5,041 chromosome X probes. Among the 9100 Swedish samples that passed QC, we selected samples with a standard deviation of LRR <0.35 for the chrX probe intensities to form the simulation pool.

1. To estimate specificity, we created 1000 simulated independent samples of normal copy number. In each simulation, a female sample was randomly chosen and its intensities of all chrX probes were randomly permuted to create a "CNV-free" chromosome. After applying the CNV detecting pipeline, any CNV detected from such chromosome was regarded as 'false positive'.
2. To estimate sensitivity, we created 1000 simulated independent samples, where each contains heterozygous deletions spanning N probes at known locations. Each simulation was generated as following: (1) randomly choose a female sample; (2) randomly choose a male sample; (3) randomly permute intensities of all chrX probes to create a "CNV-free" chromosome; (4) Replace consecutive N probes from the female sample with the intensities of the corresponding probes from a male sample to create a virtual sample with pseudo-deletions at known locations. After applying the CNV detecting pipeline, sensitivity was computed by the proportion of "true positive" deletions of all predicted deletions.

in sillico sex-mixing results:

1. The rate of false positive was obtained as 0.03 in unfiltered CNVs, and 0.003 in dataset filtered by confidence score ≥10.
2. In the filtered dataset (confidence score ≥10), we observed:
    a. High sensitivity for deletions spanning ≥10 exome array probes (≥99.8% simulated events were detected).
    b. Reduced sensitivity for CNVs with poor probe coverage (e.g. kb/snp < 20)
    c. Reduced sensitivity for CNVs as sample specific noise as measured in the standard deviation of probe intensities increases (data not shown).

Summary of CNVs from exome & GWAS arrays

Table S6. Comparison of CNV datasets used in Figure 1

Dataset	Size	All CNVs		Deletions		Duplications	
	Range[4]	Total	Mbp	Total (%)	Mbp	Total (%)	Mbp
GWAS[1]	≥20kb	20,665	2,865.5	9,942 (48%)	1,130.5 (39%)	10,723 (52%)	1735 (61%)
GWAS-genic[1,2]	≥20kb	12,553	2,142.8	5,182 (41%)	745.2 (35%)	7,371 (59%)	1,397.6 (65%)
GWAS-genic & detectable[1,3]	≥20kb	5,758	939.9	1,948 (34%)	275.2 (29%)	3,810 (66%)	664.8 (71%)
Exome array	≥0.15kb	26,594	1,298.5	4,707 (18%)	271.3 (21%)	21,887 (82.3%)	1,027.2 (79%)
Exome array	≥20kb	12,503	1,207.5	2,321 (19%)	256.4 (21%)	10,182 (81%)	951 (79%)

[1] GWAS array CNVs were generated using Birdseye. This was the dataset used in Figure 1 of the main text.

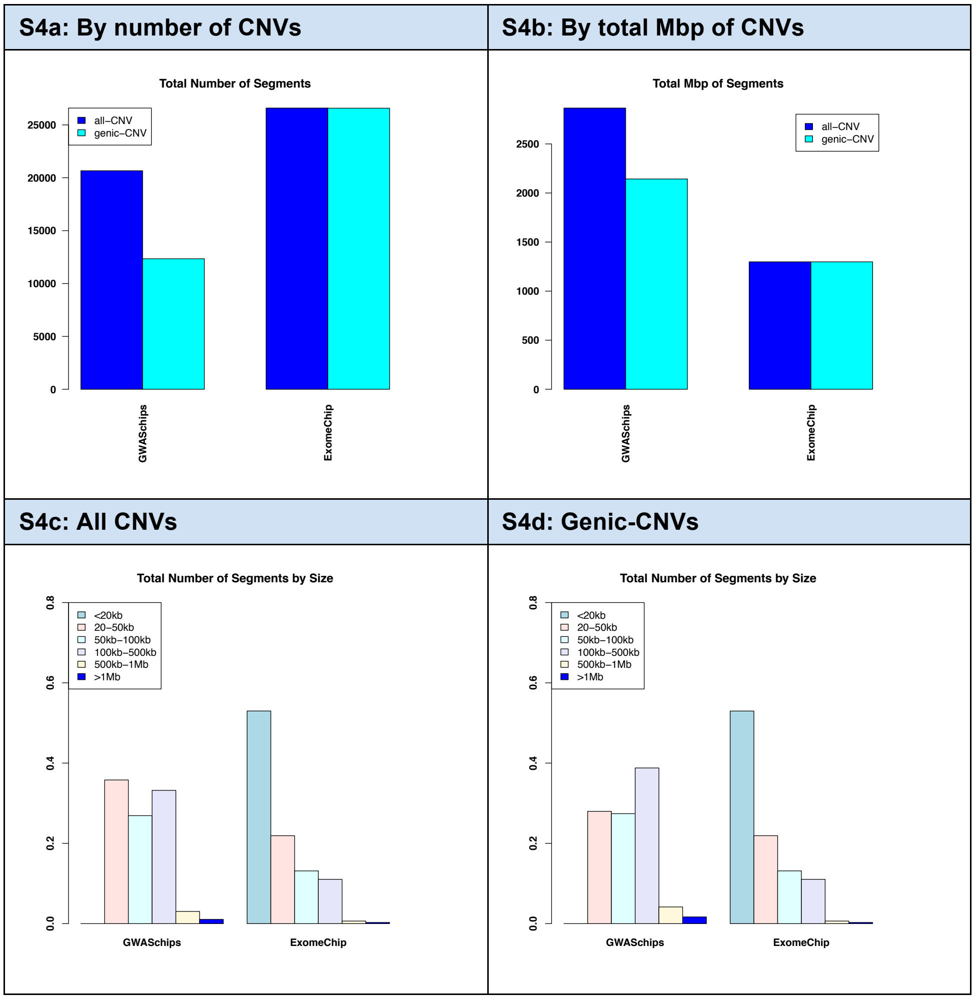
[2] GWAS Birdseye CNVs that intersect ≥1 gene. For exome array dataset, all but 13 CNVs are genic.

*[3] GWAS Birdseye CNVs that intersect ≥1 gene and had ≥1 exome array probe/20kb of its length. This subset can be most reliably detected (**Figure 1** of main text).*

[4] GWAS Birdseye CNVs were filtered to be ≥20kb. The smallest CNVs in the exome array dataset are 0.15kb. The last row of Table S3 restricted the exome array CNVs to those that are ≥20kb for the purpose of comparison.

5. By comparing the results between "GWAS-genic & detectable" and "Exome array", we observed that the relative proportion of duplications was more comparable after controlling for difference in probe design.

Figure S4. Comparison of CNV datasets used in Figure 1



S4a: By number of CNVs

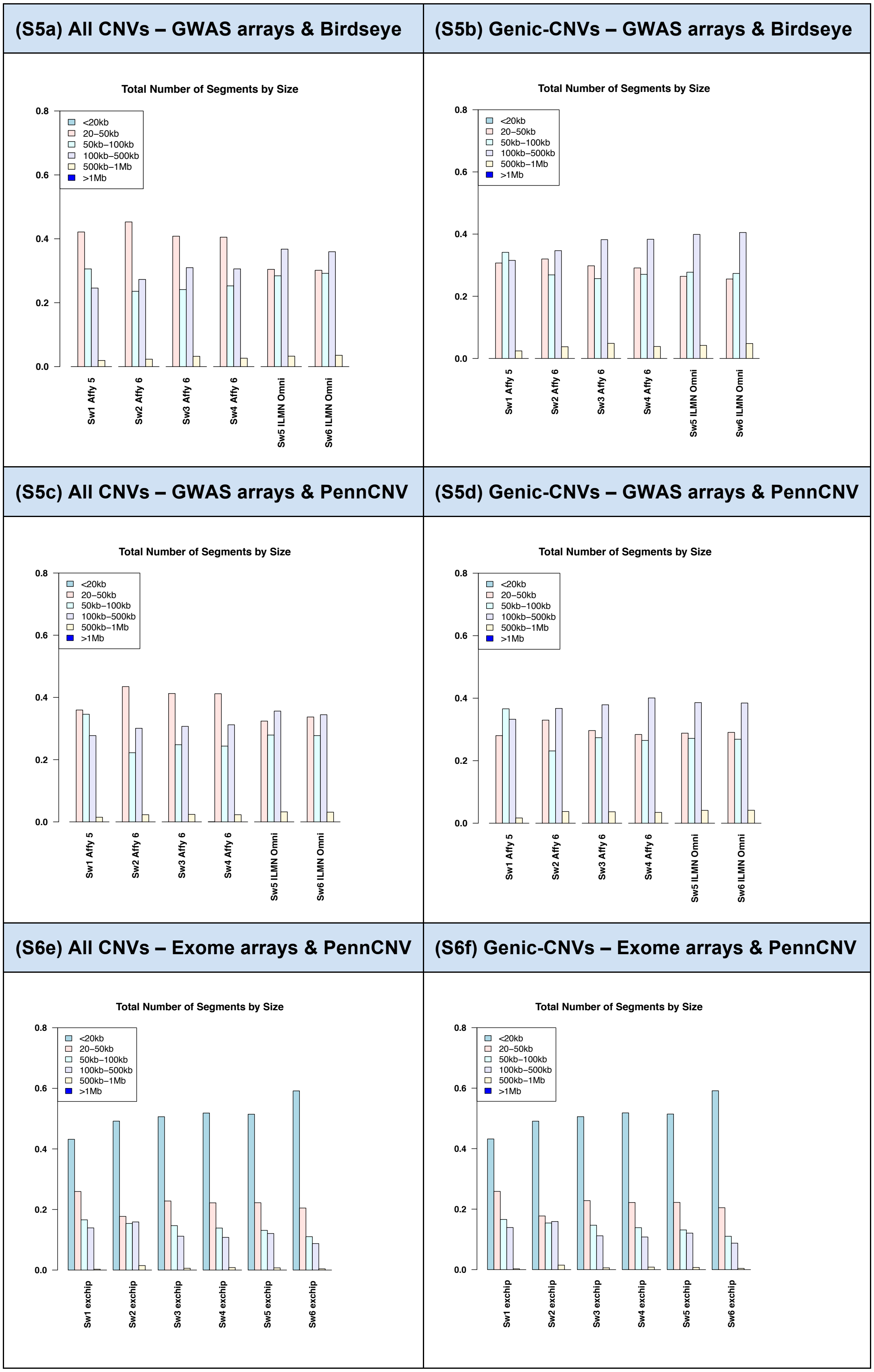
S4b: By total Mbp of CNVs

S4c: All CNVs

S4d: Genic-CNVs

[1] GWAS array CNVs were generated using Birdseye. This was the dataset used in Figure 1 of the main text.

Figure S5. CNV dataset comparison by genotyping batch, array type, and CNV calling algorithm



(S5a) All CNVs – GWAS arrays & Birdseye

(S5b) Genic-CNVs – GWAS arrays & Birdseye

(S5c) All CNVs – GWAS arrays & PennCNV

(S5d) Genic-CNVs – GWAS arrays & PennCNV

(S6e) All CNVs – Exome arrays & PennCNV

(S6f) Genic-CNVs – Exome arrays & PennCNV

Summary results of comparing exome array CNVs to GWAS array CNVs stratified by array type and by CNV calling algorithm.

We contrasted exome array CNVs to GWAS array CNVs stratified by GWAS array type (Affymetrix or Illumina) and by CNV calling algorithm (Birdseye or PennCNV). As discussed above, we considered GWAS array CNVs as the reference for estimating sensitivity and specificity. We estimated sensitivity by computing the proportion of GWAS CNVs captured by the exome array and specificity by the proportion of exome array CNVs overlapping with any GWAS CNV. We did for all CNVs and after stratifying by size and deletion/duplication type. Key results for genic CNVs ≥400 kb are summarized in ***Table S7*** below. ***Figures S6 through S9*** display the full results of stratified analysis using Sw2,3,4 subjects and using Sw5,6 subjects.

Table S7. Summary results of comparing for genic CNVs ≥400kb

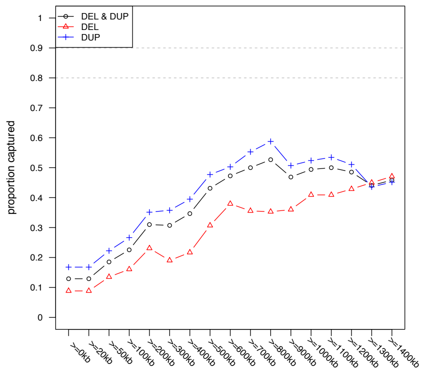
		GWAS arrays		Deletion		Duplication	
Wave	# Subjects (% Total)	Array platform	Calling algorithm	Sensitivity	Specificity	Sensitivity	Specificity
Sw1 [a,b]	307 (3.4%)	Affymetrix 5.0	PennCNV	1.00	1.00	0.86	0.75
Sw1 [a,b]	307 (3.4%)	Affymetrix 5.0	Birdseye	0.50	0	0.67	0.25
Sw2,3,4	3,030 (33.3%)	Affymetrix 6.0	PennCNV	0.94	0.59	0.78	0.84
Sw2,3,4	3,030 (33.3%)	Affymetrix 6.0	Birdseye	0.96	0.60	0.82	0.83
Sw5,6	5,763 (63.3%)	Illumina Omni Express	PennCNV	0.98	0.77	0.86	0.84
Sw5,6	5,763 (63.3%)	Illumina Omni Express	Birdseye	0.99	0.70	0.89	0.77
Sw1-6	9,100 (100%)	Affymetrix 5, 6 Illumina Omni	PennCNV	0.96	0.73	0.83	0.84
Sw1-6 [c]	9,100 (100%)	Affymetrix 5,6 Illumina Omni	Birdseye	0.95	0.68	0.80	0.80

a. *The specificities estimated using Sw1 subjects are based on only 1 eligible deletion and 4 eligible duplications. Thus these estimates are not useful.*
b. *The sensitivities estimated using Sw1 subjects is based on 4 deletions and 6 duplications from Birdseye and 2 deletions and 7 duplications from PennCNV. Thus these estimates are not useful.*
c. *Full results for Sw1-6 using Birdseye are displayed in Figure 1 of the main manuscript.*

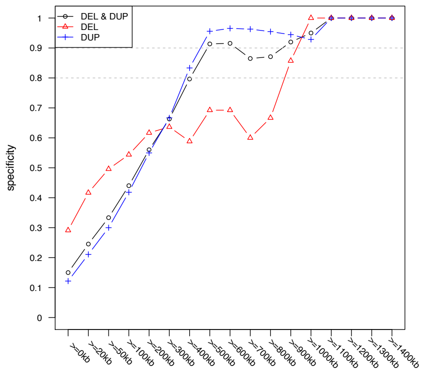
Figure legend. ***Figures S6 through S9*** were created using the same style as ***Figure 2*** of the main manuscript. CNV type is color coded (all CNVs in black, deletions in red, and duplications in blue) and CNV size bin is indicated by the x-axis. Sensitivity and specificity. (a) Sensitivity to detect any GWAS CNVs. (b) Specificity of the exome array CNV dataset to detect any GWAS CNV, estimated by computing the proportion of exome array CNVs overlapping any GWAS CNVs for each size bin of the exome array CNVs. (c) Sensitivity to detect GWAS CNVs limited to genic CNVs and accounting for probe coverage (intersect ≥1 gene and ≥ 1 exome array probe/20kb of its length). (d) Specificity of the exome array CNV dataset compared to genic CNVs from GWAS arrays. Burden tests. The y-axis shows fold changes for CNV burden of cases versus controls, and the x-axes indicate CNV size bins (total numbers of CNVs per bin in parentheses). (e) Burden test using genic CNVs from the GWAS dataset. (f) Burden test using genic CNVs from the exome array dataset. Note that the X-axes stop at the particular bin when the total numbers of CNVs per bin (in parentheses) are comparable between (e) and (f) and hence the total number of bins displayed in (e) and (f) are different.

Figure S6: Comparing subsets of CNV calls: Sw2,3,4; Illumina Exome arrays using PennCNV versus GWAS Affymetrix 6.0 arrays using Birdseye

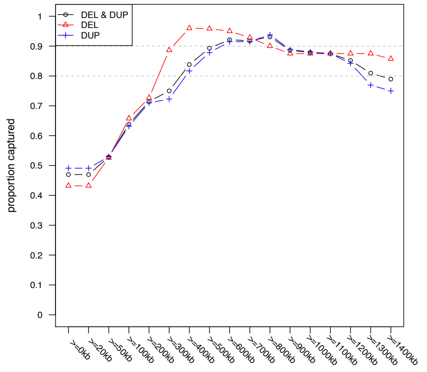
6a: Sensitivity, Genome-wide



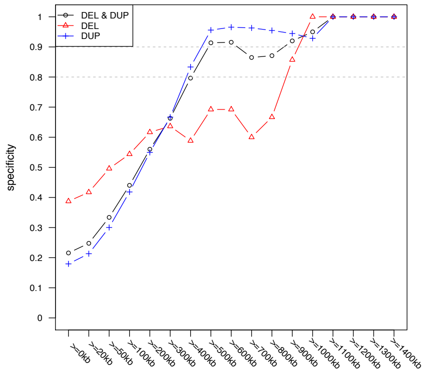
6b: Specificity, Genome-wide



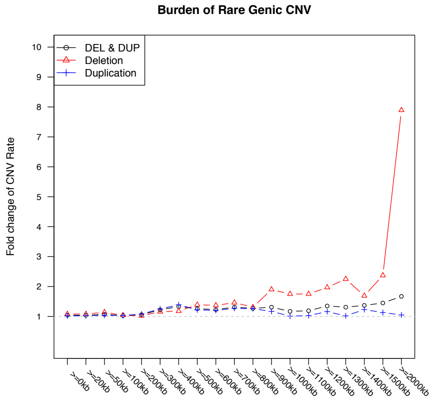
6c: Sensitivity, Genic



6d: Specificity, Genic



6e: CNV burden - GWAS CNVs



6f: CNV burden - exome chip CNVs

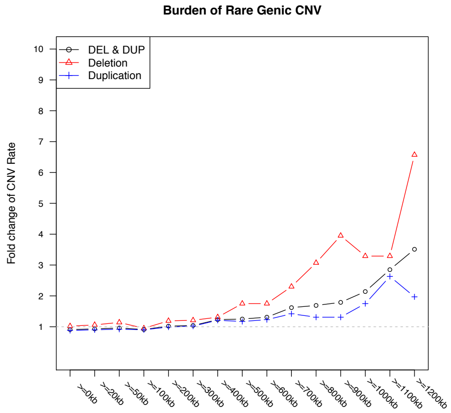
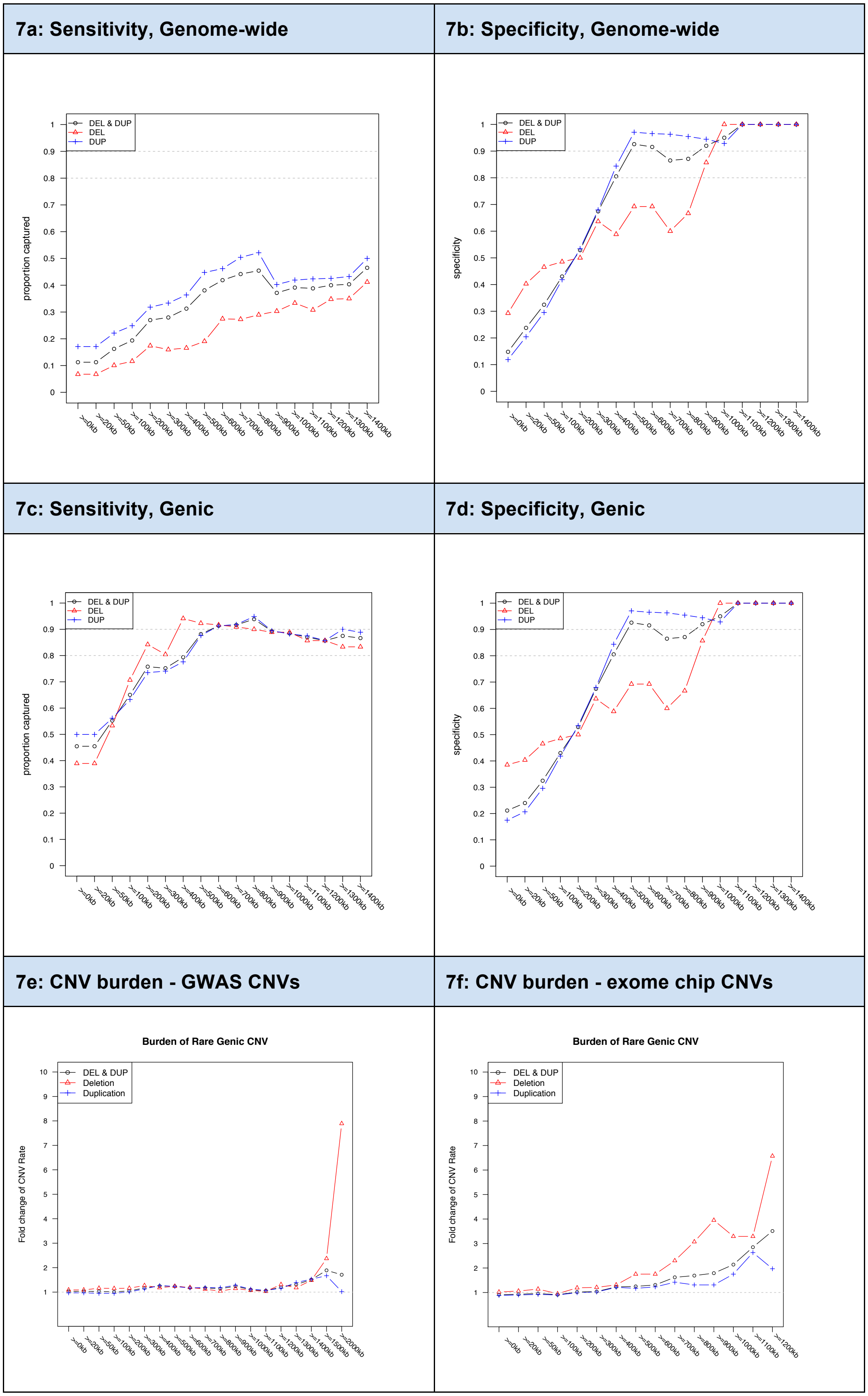


Figure S7: Comparing subsets of CNV calls: Sw2,3,4; Illumina Exome arrays using PennCNV versus GWAS Affymetrix 6.0 arrays using PennCNV



7a: Sensitivity, Genome-wide

7b: Specificity, Genome-wide

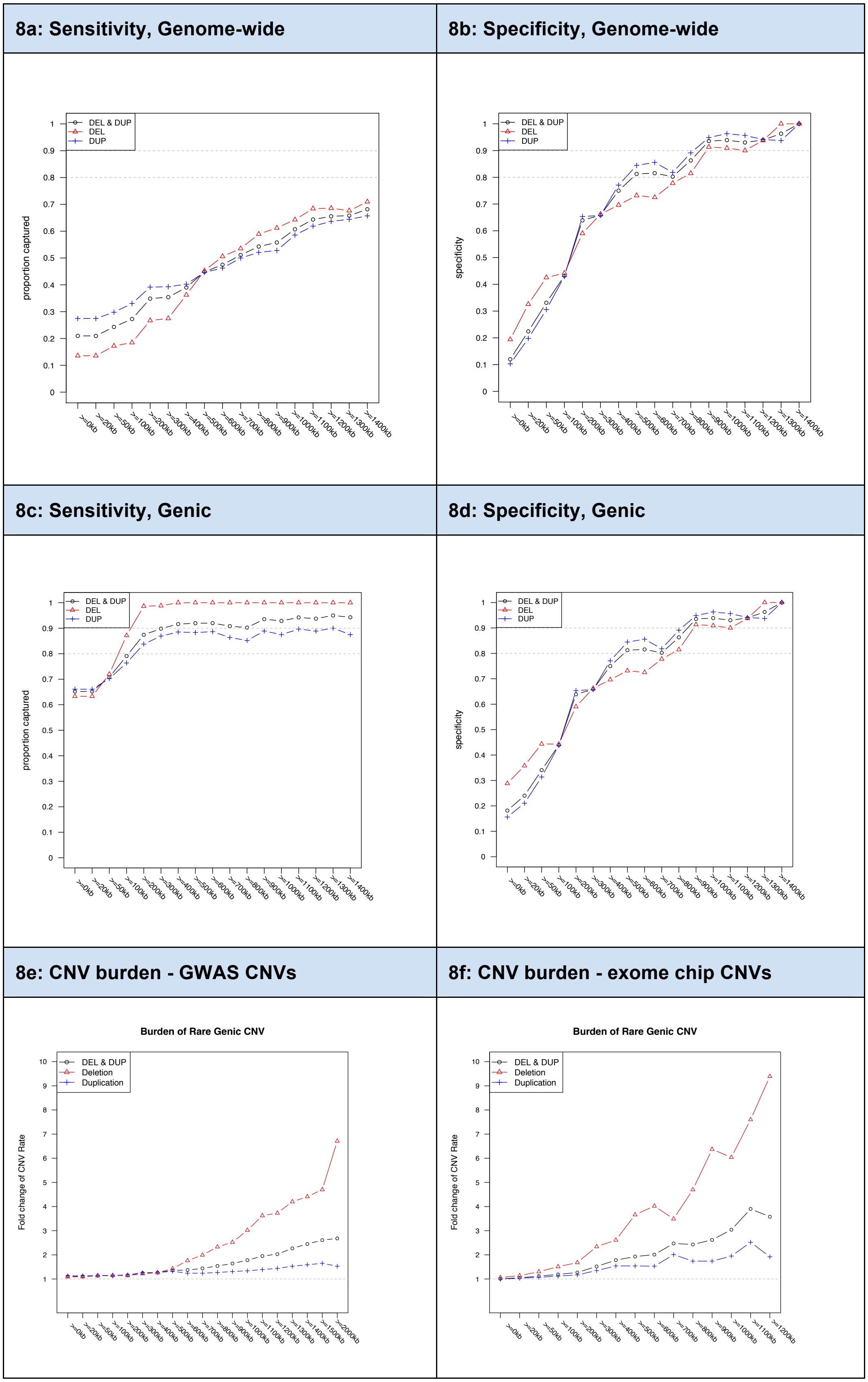
7c: Sensitivity, Genic

7d: Specificity, Genic

7e: CNV burden - GWAS CNVs

7f: CNV burden - exome chip CNVs

Figure S8: Comparing subsets of CNV calls: Sw5,6; Illumina Exome arrays using PennCNV versus GWAS Illumina Omni Express arrays using Birdseye



8a: Sensitivity, Genome-wide

8b: Specificity, Genome-wide

8c: Sensitivity, Genic

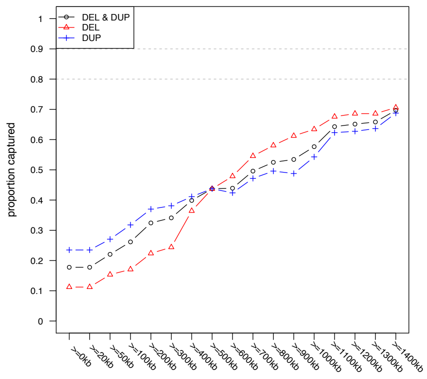
8d: Specificity, Genic

8e: CNV burden - GWAS CNVs

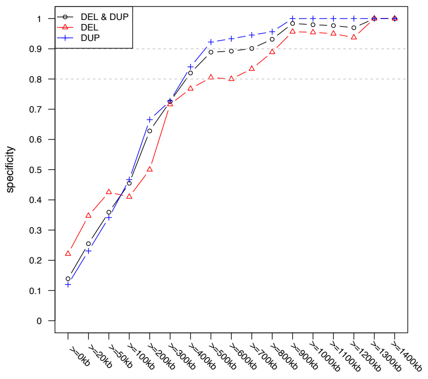
8f: CNV burden - exome chip CNVs

Figure S9: Comparing subsets of CNV calls: Sw5,6; Illumina Exome arrays using PennCNV versus GWAS Illumina Omni Express arrays using PennCNV

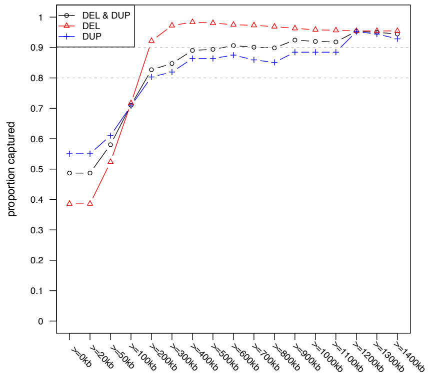
9a: Sensitivity, Genome-wide



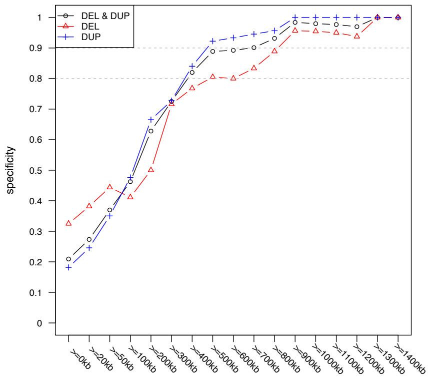
9b: Specificity, Genome-wide



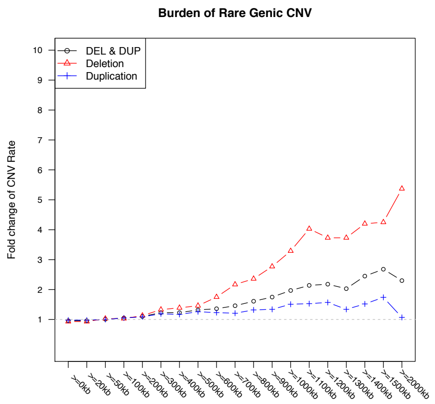
9c: Sensitivity, Genic



9d: Specificity, Genic



9e: CNV burden - GWAS CNVs



9f: CNV burden - exome chip CNVs

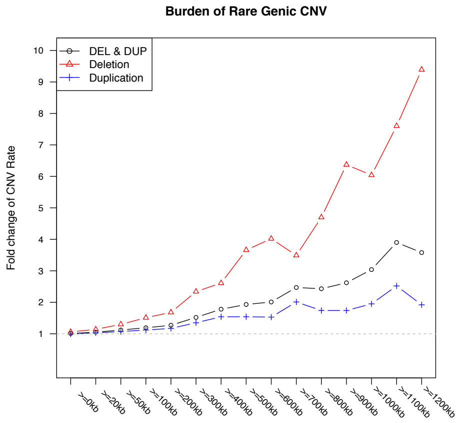
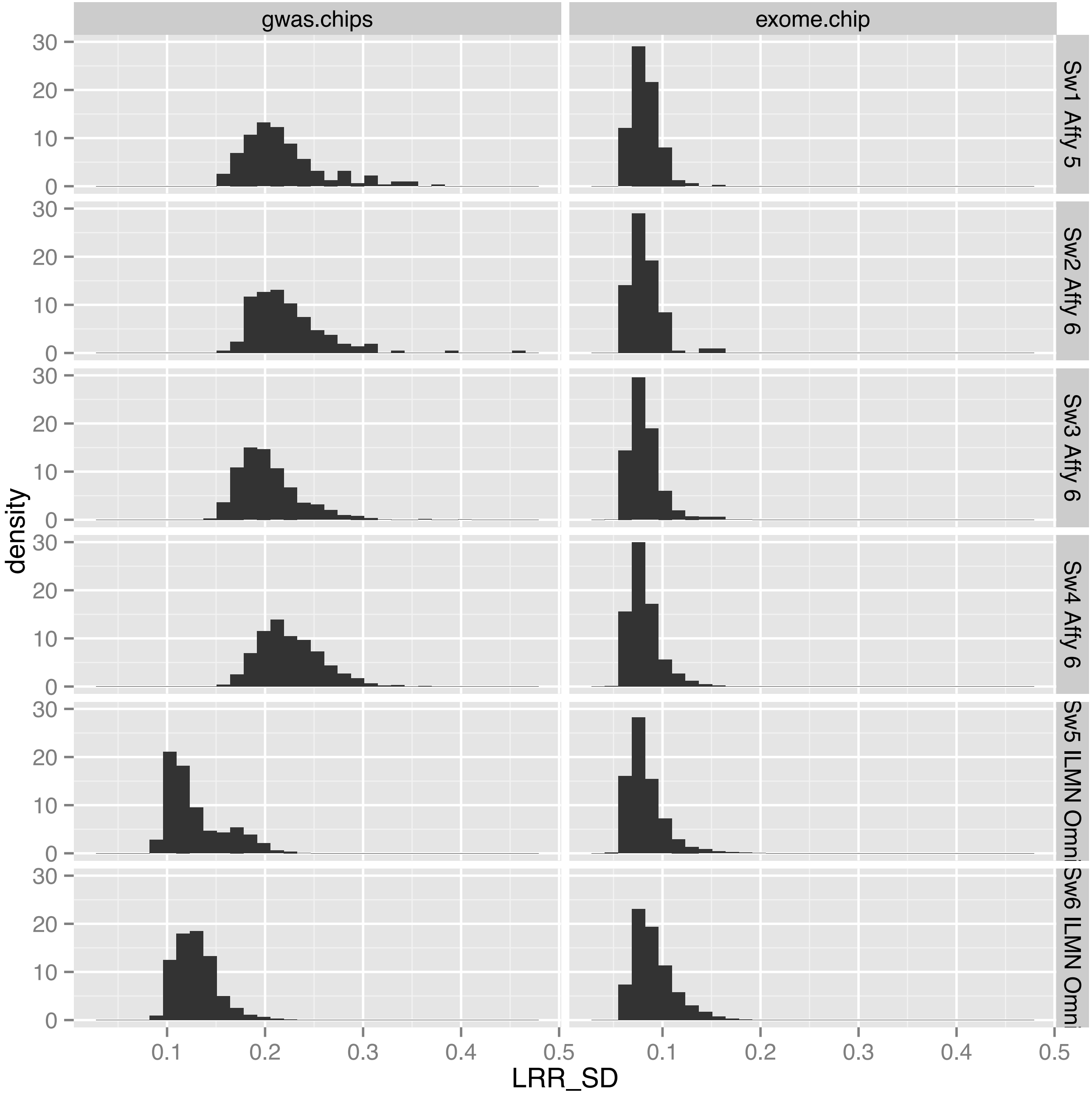


Figure S10: Comparing distributions of LRR_SD by GWAS genotyping batches and array type.



All 9,100 subjects had both exome array genotyping and GWAS array genotyping. For each sample, LRR_SD was computed as the standard deviation of log R Ratio (LRR) and was used as a measure of experimental noise. GWAS arrays (left panels) showed variability between genotyping batches as well as genotyping platforms (Affymetrix vs Illumina). In contrast, exome array genotyping showed more consistency for the entire cohort, as all samples were scanned on a common platform within a relatively short time window thereby minimizing platform and batch variation that can complicate CNV meta-analysis.

Summary of association scan

Using exome array CNVs as input, we scanned the genome using single-marker analysis as implemented in PLINK. A novel nominal association was detected using the exome arrays at 11q12.2 (*P* = 0.0069, multiple testing adjusted *P* = 0.18). Next, we used the –segment-spanning option in PLINK to extract all events at this locus from both exome array data and GWAS array data (Birdseye). ***Table S8*** displays the output. High concordance was observed. Six deletions in cases with SCZ were detected with the smallest common region spanning chr11:60531180-60620982. All six deletions were also detected by GWAS arrays with the smallest common region spanning chr11: 60547604-60624496.

Table S8 PLINK output of CNV events at a nominal novel locus

Exome Arrays									
POOL	**FID**	**IID**	**PHE**	**CHR**	**BP1**	**BP2**	**KB**	**TYPE**	**SCORE**
S1	PT-2M26	1	2	11	60531180	60620982	89.802	DEL	42.307
S1	PT-OQ42	1	2	11	60531180	60620982	89.802	DEL	43.599
S1	PT-ERMN	1	2	11	60525786	60620982	95.196	DEL	104.7
S1	PT-8VXY	1	2	11	60525786	60620982	95.196	DEL	34.772
S1	PT-BP9G	1	2	11	60525786	60620982	95.196	DEL	90.776
S1	PT-L1I6	1	2	11	60525786	60620982	95.196	DEL	73.518
S1	CON	6	6:00	11	60531180	60620982	89.802	NA	NA
S1	UNION	6	6:00	11	60525786	60620982	95.196	NA	NA
GWAS Arrays									
POOL	**FID**	**IID**	**PHE**	**CHR**	**BP1**	**BP2**	**KB**	**TYPE**	**SCORE**
S1	PT-2M26	1	2	11	60526252	60628754	102.502	DEL	74.63
S1	PT-OQ42	1	2	11	60528162	60624496	96.334	DEL	30.93
S1	PT-ERMN	1	2	11	60528162	60624496	96.334	DEL	33.53
S1	PT-8VXY	1	2	11	60526252	60628754	102.502	DEL	72.1
S1	PT-BP9G	1	2	11	60526252	60628754	102.502	DEL	147.26
S1	PT-L1I6	1	2	11	60547604	60624496	76.892	DEL	45.62
S1	CON	6	6:00	11	60547604	60624496	76.892	NA	NA
S1	UNION	6	6:00	11	60526252	60628754	102.502	NA	NA

Figure S11 (page 24) compares probe intensities of representative deletions between exome and GWAS arrays (one from Affymetrix 6 and one from Illumnia OmniExpress). ***Figure S12*** (page 25) displays probe intensities for all 6 deletions detected from exome arrays. In each figure, the x-axis indicates genomic position of the probes and y-axis indicates the values of normalized and transformed intensities (i.e. LRR). The red dots indicate probes predicted to be involved in a deletion. The blue vertical lines indicate the predicted CNV boundary.

Figure S11: Probe intensities from exome and GWAS arrays at chr11q12.2

PT-BP9G was genotyped on Affymetrix 6.0 array. PT-ERMN was genotyped on Illumina Omni Express array.

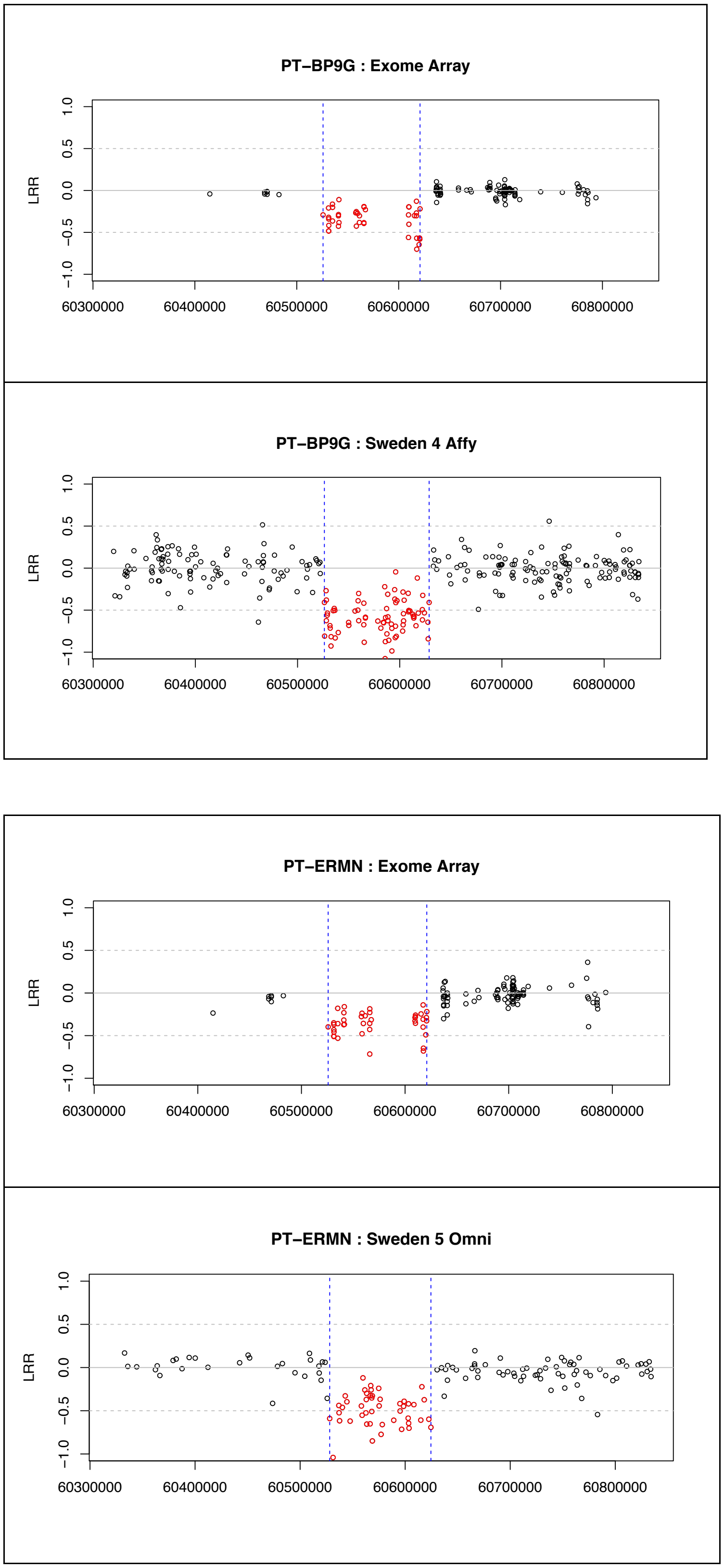
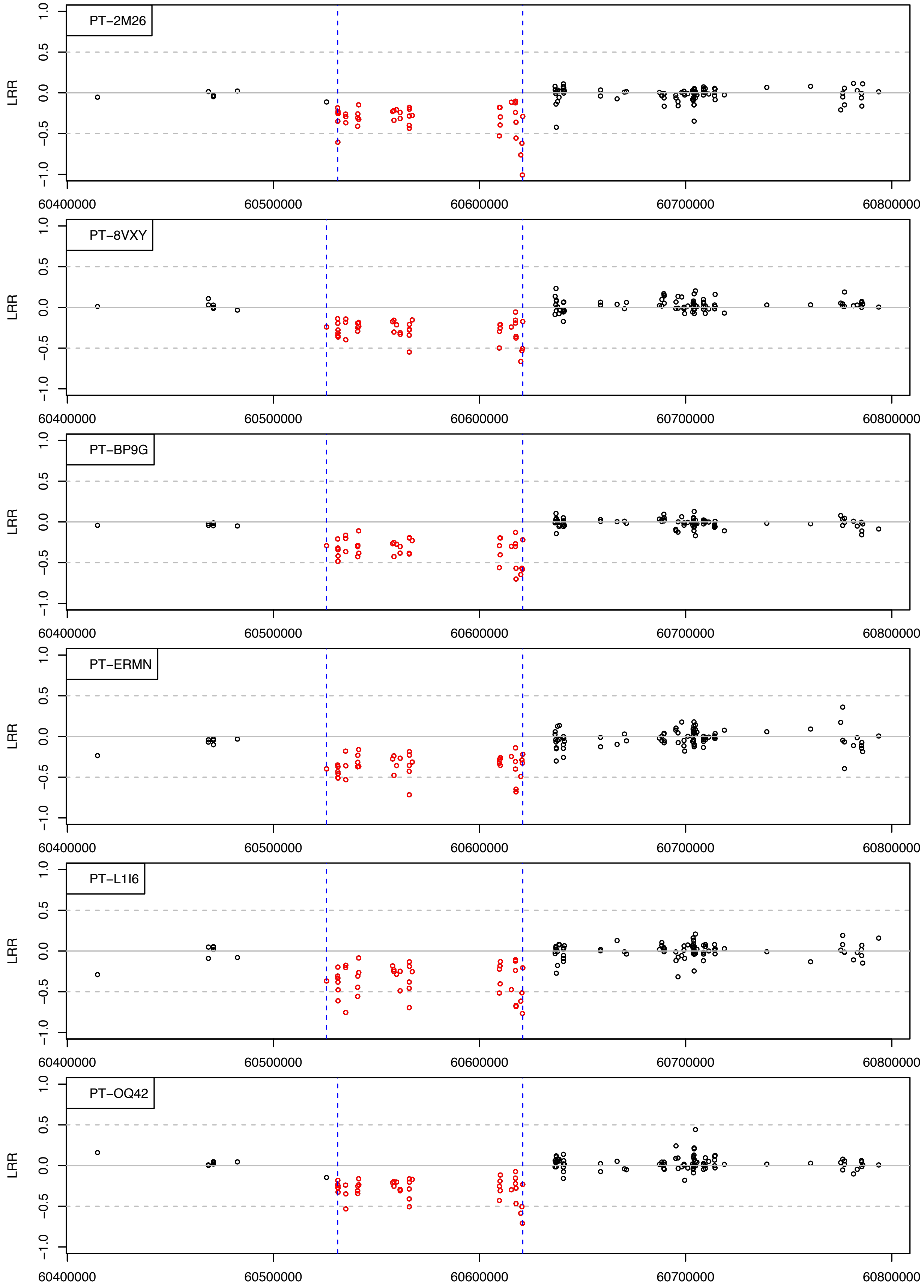


Figure S12: Probe intensities from exome arrays at chr11q12.2



References

1. Kristjansson E, Allebeck P, Wistedt B. Validity of the diagnosis of schizophrenia in a psychiatric inpatient register. *Nordisk Psykiatrik Tidsskrift* 1987; **41:** 229-234.

2. Dalman C, Broms J, Cullberg J, Allebeck P. Young cases of schizophrenia identified in a national inpatient register--are the diagnoses valid? *Social Psychiatry and Psychiatric Epidemiology* 2002 Nov; **37**(11)**:** 527-531.

3. World Health Organization. *International Classification of Diseases*. 8th revised edn. World Health Organization: Geneva, 1967.

4. World Health Organization. *International Classification of Diseases*. 9th revised edn. World Health Organization: Geneva, 1978.

5. World Health Organization. *International Classification of Diseases*. 10th revised edn. World Health Organization: Geneva, 1992.

6. Hultman CM, Sparen P, Takei N, Murray RM, Cnattingius S. Prenatal and perinatal risk factors for schizophrenia, affective psychosis, and reactive psychosis of early onset: case-control study. *Bmj* 1999 Feb 13; **318**(7181)**:** 421-426.

7. Zammit S, Allebeck P, Dalman C, Lundberg I, Hemmingsson T, Lewis G. Investigating the association between cigarette smoking and schizophrenia in a cohort study. *Am J Psychiatry* 2003 Dec; **160**(12)**:** 2216-2221.

8. Andersson RE, Olaison G, Tysk C, Ekbom A. Appendectomy and protection against ulcerative colitis. *N Engl J Med* 2001 Mar 15; **344**(11)**:** 808-814.

9. Hansson LE, Nyren O, Hsing AW, Bergstrom R, Josefsson S, Chow WH *et al.* The risk of stomach cancer in patients with gastric or duodenal ulcer disease. *N Engl J Med* 1996 Jul 25; **335**(4)**:** 242-249.

10. Schwartz S, Susser E. Genome-wide association studies: does only size matter? *Am J Psychiatry* 2010 Jul; **167**(7)**:** 741-744.

11. Craddock N, Owen MJ. The Kraepelinian dichotomy - going, going... but still not gone. *Br J Psychiatry* 2010 Feb; **196**(2)**:** 92-95.

12. International Schizophrenia Consortium. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature* 2009 Jul 1; **460:** 748-752.

13. Lichtenstein P, Yip B, Bjork C, Pawitan Y, Cannon TD, Sullivan PF *et al.* Common genetic influences for schizophrenia and bipolar disorder: A population-based study of 2 million nuclear families. *Lancet* 2009; **373:** 234-239.

14. Goldstein JI, Crenshaw A, Carey J, Grant G, Maguire J, Fromer M *et al.* zCall: a rare variant caller for array-based genotyping. *Bioinformatics* 2012; **28:** 2543-2545.

15. Wang K, Li M, Hadley D, Liu R, Glessner J, Grant SF *et al.* PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome Res* 2007 Nov; **17**(11)**:** 1665-1674.

16. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira M, Bender D *et al.* PLINK: a toolset for whole-genome association and population-based linkage analysis. *American Journal of Human Genetics* 2007; **81:** 559-575.

17. Korn JM, Kuruvilla FG, McCarroll SA, Wysoker A, Nemesh J, Cawley S *et al.* Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs. *Nat Genet* 2008 Oct; **40**(10)**:** 1253-1260.

18. International Schizophrenia Consortium. Rare chromosomal deletions and duplications increase risk of schizophrenia. *Nature* 2008; **455:** 237-241.