Neuron, Volume *72*

Supplemental Information

High Frequencies of De Novo CNVs in Bipolar Disorder and Schizophrenia

Dheeraj Malhotra, Shane McCarthy, Jacob J. Michaelson, Vladimir Vacic, Katherine E. Burdick, Seungtai Yoon, Sven Cichon, Aiden Corvin, Sydney Gary, Elliot S. Gershon, Michael Gill, Maria Karayiorgou, John R. Kelsoe, Olga Krastoshevsky, Verena Krause, Ellen Leibenluft, Deborah L. Levy, Vladimir Makarov, Abhishek Bhandari, Anil K. Malhotra, Francis J. McMahon, Markus M. Nöthen, James B. Potash, Marcella Rietschel, Thomas G. Schulze, and Jonathan Sebat

Inventory of Supplemental Information

Table S6: Testing association of de novo CNVs using large case-control rare CNV data set from BiGS bipolar and MGS schizophrenia study.

Table S7: Gene set enrichment analysis of de novo CNVs in controls.

Table S8: De novo CNVs identified in 45 ASD trios.

Table S9: List of Common CNPs. Provided as a separate file online.

Table S10: Evaluation of CNV detection sensitivity between all subjects and their parents by comparing concordance of CNV segmentation calls with CNV genotypes from 47 Common CNPs.

Table S11: Case only genic rare inherited CNVs detected in BD and SCZ and their frequency in BiGS and MGS case-control dataset. Provided as a separate file online.

Supplemental bed file: UCSC bed format custom tracks of all genic rare inherited CNVs detected in BD, SCZ and control subjects. Provided as a separate file online.

Supplemental Experimental Procedures

A) Study subjects: Bipolar Disorder, Schizophrenia, Controls and Autism

B) Microarray Intensity Data Processing.

C) CNV Detection and Quality Control.

D) Mutational Burden Analysis of Genic Rare Inherited CNVs in BD and SCZ.

E) Association of De novo CNVs in Rare CNV Data from Bipolar Genome Study (BiGS) and Molecular Genetic Studies of Schizophrenia (MGS) Samples.

F) Detection and Genotyping of Common CNPs.

G) Selection of CNPs for evaluation of sensitivity.

Supplemental References

Supplemental Figures and Legends

Figure S1. Adaptive Confidence Score (P-value) based filtering of rare CNVs.

The figure describes thresholds for confidence score (P-value) adjusted for filtering various size classes of rare CNVs based on rate of mendelian inconsistency. Rare CNVs (<1 % frequency) from 426 control trios were used for this analysis. The colored lines indicate different size classes of rare cnvs based on number of probes within them. Confidence score was adjusted to achieve a 5% rate of mendelian inconsistency, indicated by horizontal red line in the plot, for all rare CNVs. The P-value thresholds used for filtering various size classes of rare CNVs were: i) 1 X $10^{\wedge -6}$ for 5-10 probe CNVs, ii) 2.5 X $10^{\wedge -6}$ for 11-18 probe CNVs, iii) 2.5 X$10^{\wedge -5}$ for 19-44 probe CNVs, iv) 2.5 x $10^{\wedge -3}$ for 45-100 probe CNVs, v) 1X $10^{\wedge -2}$ for 101-250 probe CNVs and vi) 2.5 X $10^{\wedge -4}$ for >= 251 probe CNVs.

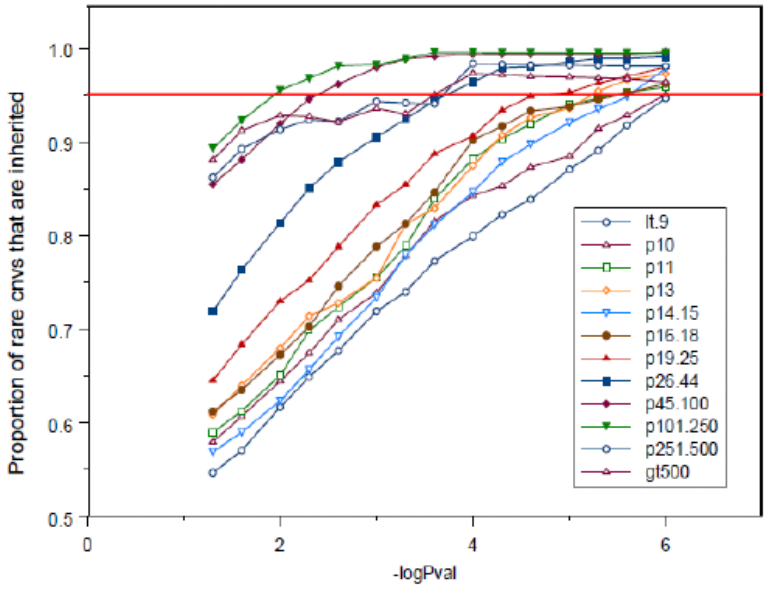


Figure S2. **Parentage test using Glaubitz Relationship Score**. The density plots of Glaubitz Relationship Scores (GRS) in 833 trios with pair wise comparison between all members within a trio are shown. GRS threshold ≥0.37 between subject and both its parents and GRS <0.50 between mother-father was used to define valid parentage.

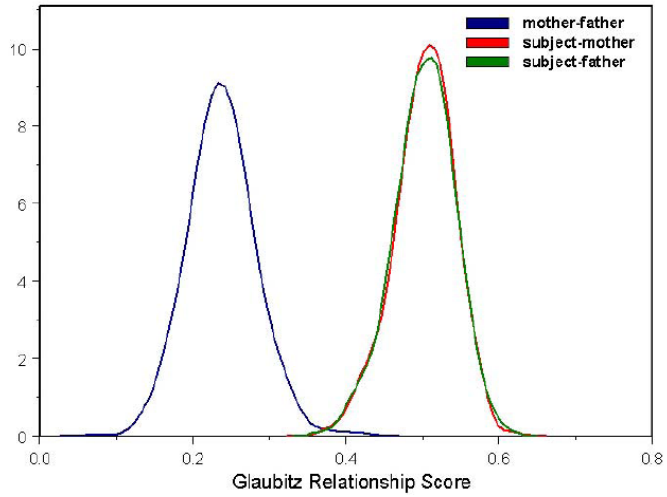
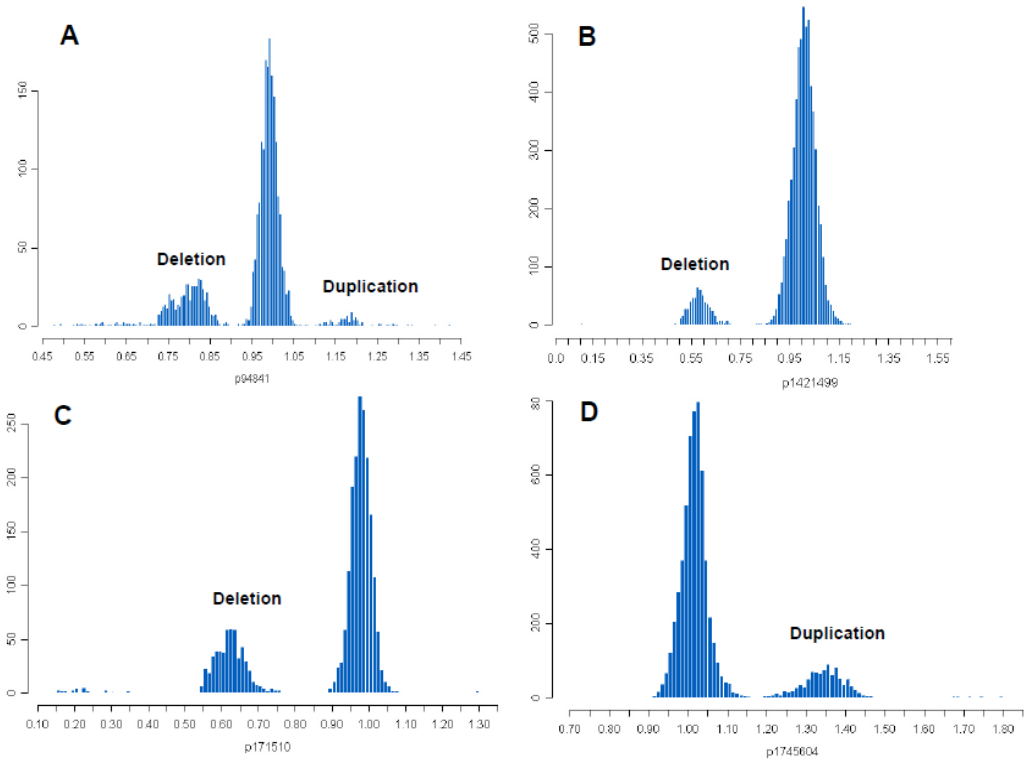


Figure S3. **Evaluating sensitivity of CNV detection in subjects and parents based on concordance of segmentation calls with genotyping calls for 47 common CNPs.** For 47 common cnv regions, clusters of ratios were assigned genotypes as illustrated in four representative examples. Ratio values of upper and lower boundaries of distinct clusters were examined to assign genotypes of “normal”, “deletion” and “duplication” as illustrated. Genotypes obtained in this manner were then compared with segmentation calls, and sensitivity was defined as the average fraction of genotyped deletions or duplications that was detected by segmentation per individual. Using this approach, we did not observe a reduced sensitivity between subjects and parents.



Supplemental Tables

Table S1. **Summary of CNV characteristics in all subjects included in this study**. Columns description is as follows: "All CNVs" refer to autosomal CNVs that passed stringent CNV QC measures. "Rare CNVs" include QC filtered autosomal CNVs that were present in < 1% frequency in CNV data from a reference population of 4081 unrelated individuals. "Rare CNVs filtered" denote rare CNVs that were further filtered based on their confidence scores, segmental duplication(SD) content and overlap with Conrad et al common CNVs. Inherited or putative de novo status of filtered rare CNVs was then determined using confidence score based genotyping.

Table S1

	Children (N=833)					Parents (N=1472)		
	All CNVs	Rare CNVs	Rare CNVs filtered	Rare Inherited CNVs	Rare putative de novo CNVs	All CNVs	Rare CNVs	Rare CNVs filtered
# CNVs	105962	15566	3856	3686	170	185895	26984	7685
Mean/ Median # CNVs/genome	127.2/122	18.7/15	4.6/4	4.4/4	0.2/0	119.6/120.5	17.2/14	4.9/5
Mean/ Median CNV size (kb)	81.2/34	37.6/15.8	83.8/28.3	79.8/28.5	178.4/20.8	80.6/33.8	35.7/15.7	73.2/27.2
% Deletions/ duplications	0.41/0.59	0.48/0.52	0.60/0.40	0.60/0.40	0.60/0.40	0.41/0.59	0.49/0.51	0.62/0.38

Table S2. **Sources of 4081 samples used as reference population for determining CNV frequency**. All samples are genetically unrelated to each other. NimbleGen HD2 microarray data on these samples was processed simultaneously along with 833 trio samples included in the present study. All 4081 samples passed experimental QC filters. We used this population cohort to identify common CNPs using PAM clustering and correlation methods. Filtered set of CNVs from 4081 samples were used to determine rare CNVs (<1% frequency in 4081 samples) in 833 trio samples.

Table S2

Sample Source	# Samples [description]
NIMH (National Institute of Mental Health)	397 [23 SCZ cases, 374 population controls]
NYCP (New York Cancer Project)	353 [all population controls]
HapMap	253 [all population controls]
NYSP (New York State Psychiatric Institute)	239 [all population controls]
Feinstein Institute for Medical Research, New York	230 [all population controls]
Harvard	342 [340 SCZ probands and 2 unaffected parents]
University of Washington	235 [130 SCZ probands and 105 unaffected parents]
Mclean Hospital	207 [101 SCZ probands and 106 parents (4 SCZ, 16 BD or MDD and 86 unaffected)]
Columbia University	174 [all parents (2 SCZ and 172 unaffected)]
Trinity College Dublin	76 [all parents (4 SCZ and 72 unaffected)]
Simons Simplex Collection	978 [all unaffected parents]
GEM (Genetics of Early Onset Mania) Bipolar Study	451 [13 BD probands and 438 parents (125 MDD and 313 unaffected)]
AGRE (Autism Genetic Resource Exchange)	146 [14 ASD probands and 132 unaffected parents]
Total	**4081** [604 SCZ, 14 ASD, 154 BD-MDD and 3309 controls]

Table S3. **Custom tiling arrayCGH validation of putative de novo CNVs**. Provided as an Excel (.xls) file.

145 putative de novo CNVs identified in BD, SCZ and controls were tested for validation by custom Agilent 1 million probe tiling array CGH platform. Five putative de novo CNV regions were validated using Sequenom MassArray based CNV assay due to limited quantity of genomic DNA for tiling array CGH hybridization. The characteristics i.e “Cytoband”, “Start”, “End”, “CNV Size”, “CNV Type” of 145 putative de novo CNVs are listed. For each putative de novo CNV region, we computed median Z score of probe log2 ratios within the region. The Z score values of de novo CNV region in child, mother and father are listed in Columns “Child Z score”, “Mother Zscore” and “Father Zscore” respectively. We used a Zscore cut off of ≥ 2 (for duplication) and ≤ 2 (for deletion) to call an event as de novo (“MeZoD Call”) in child. We then manually examined log 2ratio and Zscore cluster plots of all MeZoD called De novo CNV regions. The value of “1” or “0” in column “Validation by manual inspection” indicates if the call was true or false respectively. Nine putative de novo CNV regions called by MeZoD as “De novo” were invalidated by manual examination and the reasons for their invalidation are detailed in “Comments” column.

Table S4. **Summary of validation rate for putative de novo CNVs**. The rate of validation of 145 putative de novo CNVs distributed across three size categories was evaluated. Validation rate was highest for >100 kb and lowest for < 20 kb putative de novo CNVs.

cnv size category	# putative denovo cnvs	# validated denovo cnvs (proportion)
< 20kb	66	2 (0.03)
20-100kb	56	9 (0.16)
>100kb	23	12 (0.52)

Table S5. Rate of de novo CNVs in BD and SCZ subjects stratified by family history.

Sample	# Trios	# denovo cnvs	# subjects with denovo cnvs	Rate of subjects with denovo cnvs	Fisher's Exact P-value	dels:dups (ratio)
Controls	426	4	4	0.009		3 : 1 (3)
Bipolar	185	10	8	0.043	0.009	5 : 5 (1)
Familial BD	107	7	5	0.047	0.019	
Sporadic BD	78	3	3	0.038	0.078	
Schizophrenia	177	9	8	0.045	0.007	6 : 3 (2.0)
Familial SCZ	44	3	3	0.068	0.026	
Sporadic SCZ	97	5	4	0.042	0.042	
Unknown Family History	36	1	1	0.025	0.334	

Table S6. Testing association of de novo CNVs using large case-control rare CNV data set from BiGS bipolar and MGS schizophrenia studies.

We tested association of all 23 validated de novo CNVs using CNV data from two publicly available data sets, the Bipolar Genome Study (BiGS) and the Molecular Genetics of Schizophrenia (MGS) study. Association of each de novo CNV was evaluated using permutation based approach. Panels A and B describe association results in BiGS and MGS respectively. Chromosome ("chrom"), start ("cnv.start"), end ("cnv.end"), cnv.type ("del" for deletions and "dup" for duplications) and cytoband for each de novo cnv is listed. The Peak.start and Peak.end denote de novo cnv region with at least one overlapping rare cnv. Counts, Odds ratios ("Peak.OR") and P-values ("Peak.Pvalue") are listed for the corresponding peak region.

Table S6

A

					BiGS BIPOLAR					
chrom	**cnv.start**	**cnv.end**	**cnv.type**	**cytoband**	**Peak.start**	**Peak.end**	**Peak.cases**	**Peak.ctrls**	**Peak.OR**	**Peak.Pvalue**
1	755082	1266721	del	1p36.33	758644	875275	2	0	inf [0.50, inf]	0.13119
3	197417247	198249463	del	3q29	197417247	198249463	1	0	inf [0.05, inf]	0.51478
4	187892312	191152733	del	4q35.2	189369106	189435778	3	2	2.16 [0.33, inf]	0.34272
5	17508759	17561308	del	5p15.1	17508759	17534599	1	0	inf [0.05, inf]	0.51478
5	138051454	138254688	dup	5q31.2	138051454	138058292	0	1	0.00 [0.00, inf]	1.00000
6	17639894	17988098	dup	6p22.3	17639894	17749217	0	2	0.00 [0.00, inf]	1.00000
7	151641965	158820241	dup	7q36.1-q36.3	153149244	153157209	4	0	inf [1.05, inf]	0.04488
8	113803018	113843604	del	8q23.3	113803018	113843604	0	1	0.00 [0.00, inf]	1.00000
9	5264155	7119082	dup	9p24.1	6735551	6761267	5	2	3.69 [0.76, inf]	0.10097
9	9855970	9927360	dup	9p23	9855970	9927360	2	1	2.52 [0.19, inf]	0.41743
9	28639802	28696667	del	9p21.1	28639802	28642515	11	19	0.71 [0.35, inf]	0.85899
16	29512728	30124017	dup	16p11.2	29512728	29720497	2	1	2.64 [0.19, inf]	0.41268
16	73021657	73057216	dup	16q22.3	73021657	73057216	0	1	0.00 [0.00, inf]	1.00000
18	6479110	7904147	dup	18p11.31-p11.23	7546209	7691687	3	0	inf [0.61, inf]	0.11752
22	18869860	20006782	dup	22q11.21	19902202	20006782	4	4	1.67 [0.39, inf]	0.35478

Table S6

B

					MGS SCHIZOPHRENIA					
chrom	**cnv.start**	**cnv.end**	**cnv.type**	**cytoband**	**Peak.start**	**Peak.end**	**Peak.cases**	**Peak.ctrls**	**Peak.OR**	**Peak.Pvalue**
1	755082	1266721	del	1p36.33	890962	1018906	3	0	inf [0.41, inf]	0.19993
3	197417247	198249463	del	3q29	197806756	197851785	8	0	inf [1.98, inf]	0.00587
4	187892312	191152733	del	4q35.2	188444875	188467698	4	0	inf [0.64, inf]	0.11683
5	17508759	17561308	del	5p15.1	17508787	17522690	3	0	inf [0.50, inf]	0.15671
5	138051454	138254688	dup	5q31.2	138058292	138060093	1	0	inf [0.05, inf]	0.51925
6	17639894	17988098	dup	6p22.3	17773459	17778478	2	0	inf [0.23, inf]	0.30372
7	151641965	158820241	dup	7q36.1-q36.3	158620398	158647263	13	0	inf [3.39, inf]	0.00028
8	113803018	113843604	del	8q23.3	113803018	113843604	1	2	0.42 [0.01, inf]	0.90409
9	5264155	7119082	dup	9p24.1	6565976	6570982	6	1	5.59 [0.86, inf]	0.07502
9	9855970	9927360	dup	9p23	9855970	9927360	1	3	0.29 [0.01, inf]	0.95326
9	28639802	28696667	del	9p21.1	28639802	28645770	28	41	0.60 [0.39, inf]	0.98678
13	101035498	101078525	del	13q33.1	101058333	101078525	1	0	inf [0.04, inf]	0.58492
16	29512728	30124017	dup	16p11.2	29997384	30011876	18	0	inf [4.94, inf]	0.00001
16	73021657	73057216	dup	16q22.3	73021657	73057216	0	1	0.00 [0.00, inf]	1.00000
18	6479110	7904147	dup	18p11.31-p11.23	6987818	7186935	4	0	inf [0.84, inf]	0.07176
22	18869860	20006782	dup	22q11.21	19308719	19338167	6	12	0.43 [0.16, inf]	0.97645

Table S7. Gene set enrichment analysis of de novo CNVs in controls. For genes within control de novo CNVs, nine functional categories were found to be enriched (P-value <0.05) by primary analysis using DAVID. Two out of the nine categories were observed significant (P-value <0.05, highlighted in bold) by permutation analysis, including “response to drug” and “serine endopeptidase activity”. None of the nine categories were found enriched by PLINK-cnv enrichment test in two independent case-control rare CNV data sets.

Table S7

				De novo CNVs (this study)		Rare CNVs (from Vacic et al)	Rare CNVs (from BiGS)
Category	**Database ID**	**Term**	**CNV Genes**	**DAVID**	**Permutation**	**PLINK-CNV**	**PLINK-CNV**
GOTERM_BP_FAT	GO:0042493	response to drug	PTPRM, CA4, ABCC1, ABCC6	**0.027**	**0.008**	0.096	0.683
GOTERM_MF_FAT	GO:0004252	serine-type endopeptidase activity	AZU1, PRSSL1, PRTN3, CFD	**0.012**	**0.026**	0.408	0.843
GOTERM_CC_FAT	GO:0005576	extracellular region	AZU1, LAMA1, MBL2, DKK1, PRSSL1, PRG2, FSTL3, FGF22, CNTN4, CA2, CFD, ITGBL1	0.023	0.107	0.822	0.018
GOTERM_MF_FAT	GO:0016836	hydro-lyase activity	CA4, ABCC1, CA2	0.008	0.112	0.26	0.642
GOTERM_BP_FAT	GO:0032989	cellular component morphogenesis	LAMA1, NDE1, PTPRM, MYH11, CNTN4	0.031	0.142	0.149	0.494
GOTERM_BP_FAT	GO:0040012	regulation of locomotion	AZU1, LAMA1, KISS1R, PTPRM	0.020	0.151	0.686	0.229
GOTERM_BP_FAT	GO:0016477	cell migration	AZU1, LAMA1, NDE1, PRKG1	0.051	0.262	0.504	0.083
GOTERM_BP_FAT	GO:0006928	cell motion	AZU1, LAMA1, PALM, NDE1, PTPRM, CNTN4, PRKG1	0.003	0.280	0.501	0.498
GOTERM_BP_FAT	GO:0002009	morphogenesis of an epithelium	LAMA1, TBX4, CA2	0.037	0.433	0.079	0.434

Table S8. **De novo CNVs identified in 45 ASD trios.**

Diagnosis	SampleID	Sex	Cytoband	Start	End	CNV Size (bp)	CNV Type	Genes	Breakpoint Sequence
Autism*	01C07329	F	2q24.2	162027747	162199000	171254	deletion	SLC4A10	
Autism*	01C07255	M	2q37.2-q37.3	236227542	242705644	6478103	deletion	76 genes	SegDup
Autism*	02C11069	M	3p14.2	60526210	60727713	201504	deletion	FHIT	
Autism	00C03307	M	14q12	31162911	31200846	37936	deletion	NUBPL	

* De novo CNVs identified in Sebat et al., 2007

Table S9. **List of Common CNPs**. Provided as an Excel (.xls) file.

The table describes characteristics of 493 Common CNPs. “Cnv.id” is a unique identifier for each CNP. Chromosome location (“Cytoband”), start and end genomic coordinates of CNPs for hg18 build are listed. CNP length (“CNV Size”) in base pairs, “CNV Type” (“del” for deletions and “dup” for duplications) and allele frequency (“CNV Allele frequency”) is described. “ CNV Source” refers to study from which CNPs were discovered (“sebat” denotes present study and “conrad” denote Conrad et al study). The upper and lower bound thresholds for ratio values used for genotyping each CNP are listed. 486 CNPs were identified from NimbleGen HD2 microarray data using PAM clustering and correlation approach. Seven CNPs were derived from Conrad et al study. All 47 CNPs greater than or equal to 10 kb in size were used in sensitivity analysis.

Table S10. Evaluation of CNV detection sensitivity between all subjects and their parents by comparing concordance of CNV segmentation calls with CNV genotypes from 47 Common CNPs. Mean and standard deviation (SD) of sensitivity measure for CNV calls in children and parents is listed. Sensitivity was evaluated for four CNV size categories. A two sample T-test pvalue ("Pvalue") was computed and showed no significant difference in sensitivity between children and parents CNV calls across all size categories.

	Children		Parents		
Cnv Size	Mean Sensitivity	SD	Mean Sensitivity	SD	Pvalue
All CNVs	0.96	0.08	0.96	0.09	0.65
≥ 100kb	1.00	0.01	1.00	0.05	0.64
50 -100kb	1.00	0.01	1.00	0.01	0.99
25 - 50kb	0.99	0.01	0.99	0.01	0.56
10-25 kb	0.93	0.11	0.91	0.11	0.63

Table S11. **Case only genic rare inherited CNVs detected in BD and SCZ and their frequency in BiGS and MGS case-control dataset.** Provided as an Excel (.xls) file.

All genic rare inherited CNVs that were observed only in BD or SCZ cases in this study and were not observed in any controls (including our controls and controls from BiGS and MGS dataset) are described.

Supplemental Experimental Procedures

Study Subjects

DNA was derived from whole blood from all Bipolar, Schizophrenia and Control trios included in this study.

a) *Bipolar disorder*

Bipolar disorder samples were collected from four different clinical sites including i) Johns Hopkins University, Maryland, USA, ii) National Institute of Mental Health Intramural Research Program, Bethesda, MD, USA, iii) North Shore Long Island Jewish Hospital, New York, USA and iv) Central Institute of Mental Health, Mannheim, Germany as part of a new collaborative genetic study called GEM (Genetics of Early-onset Mania). 217 complete bipolar trio samples were recruited and microarray data were collected on all samples. We excluded 32 bipolar trios because 11/32 probands did not meet criteria for a bipolar or schizoaffective bipolar (SABP) diagnosis, 8/32 probands failed CNV quality control (QC) and 13/32 trios failed child-parent relationship tests (see parentage testing section). The final bipolar samples included 185 trios of which 173 probands had a diagnosis of bipolar I (BPI), 6 had a diagnosis of bipolar II (BPII) and 6 had a diagnosis of SABP. 176 out of 185 bipolar probands were Caucasian, 8 were Hispanic and 1 was of Asian ancestry. The male/female ratio of bipolar cases was 0.75.

Diagnoses of all probands and their family members were assigned following strict DSM-IV criteria, including episode duration criteria, in both adults and

youth. Subjects were assessed using the SCID-I interview and Diagnostic Interview for Genetic Studies (DIGS) in adults, and the Kiddie Schedule for Affective Disorders, Present and Lifetime Version (KSADS-PL) in youth. There is considerable debate in the literature about the criteria that should be used to assign the diagnosis of bipolar disorder to youth, as well as about how age of onset should be defined in BD. We followed strict DSM-IV guidelines, including duration criteria for manic episodes, to diagnose BD in youth (Leibenluft et al., 2003) as well as in adults.

Age at onset (AAO) was defined by the age at which the subject met full DSM-IV criteria for an affective episode of either polarity, as per SCID interview. Substance induced episodes were excluded. Based on the age distribution of our sample and prior reports (Benazzi, 1999; Schulze et al., 2002; Young and Klerman, 1992), we dichotomized AAO, with early onset defined as an onset prior to or equal to 18 years of age and later onset defined as an onset after the age of 18. A similar AAO threshold was used in the recent GWAS study of BiGS samples (Smith et al., 2009). Here, 58% (107/185) of BD probands had an AAO ≤ 18 yrs (early onset) and 41 % (76/185) had an AAO > 18 yrs. AAO was unknown for two bipolar cases. The mean AAO of bipolar probands in our study was 18.7 yrs (standard deviation (s.d).=7.6, median=18, median absolute deviation (m.a.d).=5.9). Early onset cases included 40 males (37%) and 67 females (63%) and their biological parents. The mean AAO in early onset bipolar cases was 13.6 yrs (s.d. = 4yrs, median=15, m.a.d.=3). Late onset cases

consisted of 38 males (50%) and 38 females (50%) and their biological parents. The mean AAO in late onset bipolar cases was 25.8 yrs (s.d. = 5.6, median= 25, m.a.d.= 5.9).

We used a broad definition for evidence of positive family history (defined as a first degree relative with bipolar I, bipolar II, major depression, SCZ, schizoaffective, autism or intellectual disability) to stratify probands as familial or sporadic.

b) *Schizophrenia*

SCZ samples were collected from three sites: i) Trinity College Dublin,Ireland, ii) Columbia University, New York, USA and iii) McLean Hospital, Belmont,USA. 209 complete SCZ trios were recruited for microarray data collection. 13 probands (trios) failed CNV QC and 19 trios failed a parentage test. The final SCZ samples consisted of 177 trios. 173/177 SCZ probands were Caucasian, 3 were Hispanic and 1 was of African-American ancestry. The male/female ratio in SCZ cases was 1.46 (105 males, 72 females).

Age at onset (AAO) in SCZ was defined as the age at which full DSM-IV criteria for SCZ or schizoaffective disorder were met. The mean AAO of SCZ probands was 21.2 yrs (s.d.=5.3, median= 20, m.a.d.=4.4). Early onset SCZ (defined by AAO <=18) included 40 males (68%) and 19 females (32%) and their parents. The mean AAO in early onset SCZ cases was 16.3 yrs (s.d.= 1.7, median= 17, m.a.d.=1.5). Late onset SCZ cases consisted of 62 males (55%) and 50 females (45%) and their biological parents. The mean AAO in late onset SCZ cases was 23.8 yrs (s.d.=4.7, median=23, m.a.d.= 4.4). AAO was unknown for 6 SCZ cases.

c) *Controls*

426 normal healthy control trios included in the study consisted of unaffected siblings of autism cases from Simon Simplex Collection (SSC) of autism families and their biological parents. DNA from all 426 control trios was derived from whole blood. The male/female ratio in controls was 1 (214 males, 212 females). The mean age of subjects was 13.7 yrs (s.d.=4.4). The mean maternal age at birth was 30.5 yrs (s.d.=4.6) and mean paternal age at birth was 32.8 yrs(s.d.= 5.4).

d) *Autism*

For the purposes of assessing the sensitivity of our methods for detecting de novo events, we included in this study an additional 45 autism trios, all of which had been included in our previous study on autism using low resolution microarray platform (Sebat et al., 2007), and 3 of which carried known de novo CNVs. The details of these samples are described in the previous study.

Microarray Intensity Data Processing

NimbleGen HD2 dual color microarray intensity data were normalized in a two step process: (1) a spatial normalization of probes was performed to adjust for regional differences in intensities across the surface of the array, and (2) the Cy5 and Cy3 intensities were adjusted to a fitting curve by invariant set normalization. Spatial normalization was performed using an R module provided by Kyle Munn (Roche-NimbleGen Inc). Invariant set normalization of intensity data involves

selection of a set of autosomal probes with minimal variability in probe intensities between test and reference samples (Li and Hung Wong, 2001). The test sample intensities of the invariant probe set are then adjusted to the reference distribution. Based on these adjustments, a fitting curve is established to which all other intensities are shifted, preserving the variability in the data. The intensities of X and Y chromosomes were then extrapolated to the fitting curve. The process is repeated with the test and reference samples swapped to simulate a dye-swap experiment. The $\log_2$ ratios were then estimated using the geometric mean of normalized and raw intensity data as follows

$$\log_2(Ratio_i) = \log_2\left(\sqrt{\frac{NormalizedTst_i}{RawRfn_i} \times \frac{RawTst_i}{NormalizedRfn_i}}\right)$$

Where Tst_i *and* Rfn_i are the test and reference intensities for probe *i*.

Finally, $\log_2$ ratios were corrected for genomic wave effects due to regional correlations with GC content based on the fitted linear regression model (Diskin et al., 2008).

CNV Detection and Quality Control

CNVs were identified from processed log2 ratio data of each microarray hybridization using two segmentation algorithms i.e HMMSeg (Day et al., 2007) and GADA (Pique-Regi et al., 2008). CNVs detected in the same genome by both HMMSeg and GADA were merged into a single call set; CNVs of the same type (i.e. deletion or duplication) that were separated by ≤3 probes were merged

into one, and the CNV boundaries were defined as the union of all calls. CNVs detected by only one algorithm were removed.

Filtering of experiments from CNV analysis was based on quality control (QC) measures derived from processed log2 ratio data and filtered CNV properties. Noisy experiments with median absolute deviations (MAD) of log2 ratio values in autosomes >0.23 were removed. In addition, any sample with aneuploidy of autosomal chromosomes was excluded as were experiments with conflicting gender from empirically derived X and Y median probe ratios.

Filtering of CNVs from the data set included removing CNVs that overlapped regions of the genome prone to somatic cell rearrangements. In particular, CNVs intersecting or overlapping T-cell receptor regions (chr7:38,245,705-38,365,141, chr7:141,647,285-142,221,100, chr9:33,608,462-33,652,656, chr14: 21,159,896-22,090,937) and abParts (chr2:88,937,989-89,411,302, chr2:88,966,183-89,377,035, chr2:89,589,457-89,897,555, chr14:105,065,301-106,352,275, chr22:20,715,572-21,595,082) were excluded. We also excluded CNVs with median probe ratios (seg.median) between 0.80 and 1.20, CNVs containing <10 probes and CNVs > 10Mb in size- these likely corresponded to large chromosome abnormalities or cell line artifacts. We removed samples that were outliers with respect to (1) Excess (>400) number of autosomal CNVs per genome-we defined outlier as mean plus three standard deviations; (2) excess

aggregate length of CNV per genome- we used 30Mb (~>1% of the human genome) as cut off for genomic content in combined CNV length.

Mutational Burden Analysis of Genic Rare Inherited CNVs in BD and SCZ.

Rare inherited CNVs were identified using the same data processing and analysis pipeline we used for de novo CNV identification. CNV burden analysis was done only on rare inherited CNVs that impacted one or more genes. We tested the global burden of genic rare inherited CNVs in BD and SCZ compared to healthy controls by measuring the CNV rate, i.e., the number of CNVs per genome, using scripts written in SPlus 8.0. A total of 1907 genic rare inherited CNVs were detected of which 440 were present in BD, 398 in SCZ and 1069 in control subjects (see Supplementary bed file). We defined three comparison groups for burden analysis:

1) All patients vs controls
2) Familial patients vs controls
3) Sporadic patients vs controls.

We tested each of the above three categories for two different class of CNVs, i.e., ≥100 kb and ≥ 500 kb. We did not observe any statistically significant association in our BD (**Table 3**) and SCZ (**Table 3**) samples. In familial cases, we observed 1.8-fold and 2.5-fold enrichment of large ≥500 kb duplications in BD and SCZ respectively. We did not observe the same trend for CNVs >100 kb in size or for large (≥500 kb) deletions.

We further tested the occurrence of all genic rare inherited CNVs in BD and SCZ subjects in rare CNV data from the BiGS BD and the MGS SCZ study. The occurrence of rare inherited CNVs in BiGS and MGS data were determined using ≥50% reciprocal overlap of CNV length..All genic rare inherited CNVs that were detected in only our BD and SCZ samples and were absent in controls including our controls and BiGS and MGS controls are listed in **Table S11.**

Association of De novo CNVs in Rare CNV Data from Bipolar Genome Study (BiGS) and Molecular Genetic Studies of Schizophrenia (MGS) Samples.

We tested statistical association of the regions encompassed by 23 de novo CNVs identified in BD, SCZ and control subjects in this study in CNV data from the BiGS (Smith et al., 2009) and MGS (Levinson et al., 2011) studies. BiGS study subjects included 2777 bipolar cases (2384 Caucasian and 393 African American) and 3508 normal healthy controls (2572 Caucasian and 936 African American). MGS study subjects included 4097 schizophrenia cases (2778 Caucasian and 1319 African-American) and same 3508 controls from BiGS collection. The Affymetrix 6.0 microarray platform was used for data collection on BiGS and MGS samples. The details of Affy 6.0 microarray intensity data processing, CNV discovery and rare CNV detection are described previously (Vacic et al., 2011) in supplementary section 2b-2c, 3b and 5 respectively. The statistical association of 23 de novo CNV regions was tested as follows. We defined each de novo CNV as the region of interest (ROI) and tested statistical

significance of these regions independently in rare CNVs from MGS and BiGS subjects. We maintained the polarity of ROI while testing association, meaning a deletion ROI was tested for association only with rare deletion CNVs and duplication ROI only with rare duplication CNVs. P-value was estimated as follows: all rare CNVs overlapping a ROI were piled up, and the CNV breakpoints which fall within the ROI as well as the region boundaries were used to partition the ROI into a series of non-overlapping segments or bins. Based on the CNV counts within a segment, association was quantified using the Exact Conditional test, with ancestry as covariate. The segment with the lowest one-sided p-value was the peak of association within a ROI. Because segments in different ROIs are driven by the underlying genetic architecture, their numbers and sizes varied widely. Furthermore, numbers of CNVs in nearby segments are highly correlated. To address these issues, we applied a permutation-based p-value correction scheme, where the observed one-sided p-value of the association peak is compared to the distribution of minimal one-sided p-values of any segment within the ROI, computed based on data with case/control labels shuffled at the sample level. In our experiments, p-value was estimated by running 200,000 permutations. We did not observe a statistically significant association for any de novo CNV in BiGS BD samples (**Table S6A**). There were eight de novo CNV regions where we did not observe any overlapping rare CNV of same polarity in BiGS data set. These de novo CNVs are not listed in **Table S6A**. In MGS SCZ data, we found statistically significant associations in three genomic regions namely: i) deletions in 3q29 loci, ii) duplications in 16p11.2 loci, and iii)

duplications at 7q36.3 (**Table S6B**). The three regions have been implicated in SCZ in earlier studies. Seven de novo CNV regions are not listed in **Table S6B** because we did not find any overlapping rare CNV of same polarity in MGS subjects.

Detection and Genotyping of Common CNPs

We used microarray data from QC filtered 4081 samples (**Table S2**) for identifying common CNPs using two automated methods. The first was cluster analysis using Partitioning Around Medoids (PAM) algorithm which examines clustering of ratio values for every autosomal probe on NimbleGen HD2 microarray across 4081 samples. We applied PAM five times by specifying the number of clusters at 2, 3, 4, 5 and 6. PAM evaluates degree of clustering using a measure called "average of silhouette (AOS)" with values ranging between 0 and 1. AOS close to 1 indicates high quality of clustering. The second approach involved searching for genomic regions in which two consecutive autosomal probes showed highly correlated patterns of ratios across 4081 samples. The correlation score for two probes was calculated as the average pair-wise correlation of their measurements across 4081 samples. Using the combination of two automated methods, we defined a candidate CNP probe as the one whose correlation score and AOS's values were in top 0.1% of their respective distributions. The consecutive candidate CNP probes were then merged to define a CNP genomic region. We identified a total of 3229 candidate CNPs regions. We were well aware that both PAM clustering and correlation score methods could identify false positives. Therefore, we used two methods to minimize false

positives CNP regions. The first was by evaluating asymmetry scores for median value of probe ratios of each CNP across 4081 samples. Asymmetry score is calculated as the larger of differences of upper 25% and 10% quantiles from lower 25% and 10% quantiles, respectively. We excluded CNPs with asymmetry score < 0.03 which likely are non polymorphic CNPs. The second approach was to validate our CNP genotypes with CNVs identified by read depth analysis of low coverage sequence data on 120 samples that were common between our microarray data and 1000 genomes pilot project. We used linear regression and applied bonferroni correction for 3229 tests. CNPs with significant regression P-value i.e P $<10^{-6}$ were selected. Using the two approaches, we identified 878 validated CNP regions. Since our microarray hybridizations involved a single reference sample i.e CHP-SKN-1, we wanted to avoid any reference based effects in genotyping CNPs. We examined clustering of all 878 CNP regions and selected 486 CNPs (**Table S9**) for which the reference copy number was diploid and high-quality clustering allowed us to define thresholds of log2 ratios for normal, gain or loss of copy number as shown in **Figure S3** to make genotyping calls for integer copy number states.

Selection of CNPs for evaluation of sensitivity.

Our CNV analysis included only ≥10 kb in size CNVs. Therefore, we evaluated sensitivity of our segmentation algorithms (HMMSeg and GADA) for all CNVs >10kb in size using 486 common CNPs. Furthermore, we wished to test sensitivity in different size bins i.e 10-25 kb, 25-50kb, 50-100kb and >100kb. Only

40 out of 486 common CNPs were ≥10 kb in size and were not uniformly distributed across different size categories. We required at least seven CNPs in each of the four size categories in order to have high confidence in our estimates of sensitivity. We were short of at least one CNP >100kb and four CNPs in 50-100 kb CNV size category. To fill this gap, we used Conrad et al list of common CNVs and genotyped all > 50 kb CNPs from this list in our microarray data of 4081 samples using our targeted genotyping algorithm MeZOD and by manual examination of clusters. We selected seven CNPs (three >100kb and four 50-100kb) which showed high-quality clustering to allow accurate genotyping. We used the combined list of 47 CNPs (**Table S9**) to perform sensitivity analysis. We evaluated the sensitivity of our segmentation-based CNV calling methods (HMMSeg and GADA) by comparing segmentation calls in subjects and parents to genotypes from 47 CNPs. Sensitivity was defined as the average fraction of genotyped gains or losses that were identified by segmentation per individual. In all size classes, we observed high (>90%) sensitivity of our segmentation calls, and sensitivity was similar between subjects and parents (**Table S10**).

Supplemental References

Benazzi, F. (1999). A comparison of the age of onset of bipolar I and bipolar II outpatients. J Affect Disord *54*, 249-253.

Day, N., Hemmaplardh, A., Thurman, R.E., Stamatoyannopoulos, J.A., and Noble, W.S. (2007). Unsupervised segmentation of continuous genomic data. Bioinformatics *23*, 1424-1426.

Diskin, S.J., Li, M., Hou, C., Yang, S., Glessner, J., Hakonarson, H., Bucan, M., Maris, J.M., and Wang, K. (2008). Adjustment of genomic waves in signal intensities from whole-genome SNP genotyping platforms. Nucleic Acids Res *36*, e126.

Leibenluft, E., Charney, D.S., Towbin, K.E., Bhangoo, R.K., and Pine, D.S. (2003). Defining clinical phenotypes of juvenile mania. Am J Psychiatry *160*, 430-437.

Levinson, D.F., Duan, J., Oh, S., Wang, K., Sanders, A.R., Shi, J., Zhang, N., Mowry, B.J., Olincy, A., Amin, F.*, et al.* (2011). Copy number variants in schizophrenia: confirmation of five previous findings and new evidence for 3q29 microdeletions and VIPR2 duplications. Am J Psychiatry *168*, 302-316.

Li, C., and Hung Wong, W. (2001). Model-based analysis of oligonucleotide arrays: model validation, design issues and standard error application. Genome Biol *2*, RESEARCH0032.

Pique-Regi, R., Monso-Varona, J., Ortega, A., Seeger, R.C., Triche, T.J., and Asgharzadeh, S. (2008). Sparse representation and Bayesian detection of genome copy number alterations from microarray data. Bioinformatics *24*, 309-318.

Schulze, T.G., Muller, D.J., Krauss, H., Gross, M., Fangerau-Lefevre, H., Illes, F., Ohlraun, S., Cichon, S., Held, T., Propping, P.*, et al.* (2002). Further evidence for age of onset being an indicator for severity in bipolar disorder. J Affect Disord *68*, 343-345.

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.

Smith, E.N., Bloss, C.S., Badner, J.A., Barrett, T., Belmonte, P.L., Berrettini, W., Byerley, W., Coryell, W., Craig, D., Edenberg, H.J.*, et al.* (2009). Genome-wide association study of bipolar disorder in European American and African American individuals. Mol Psychiatry *14*, 755-763.

Vacic, V., McCarthy, S., Malhotra, D., Murray, F., Chou, H.H., Peoples, A., Makarov, V., Yoon, S., Bhandari, A., Corominas, R.*, et al.* (2011). Duplications of the neuropeptide receptor gene VIPR2 confer significant risk for schizophrenia. Nature *471*, 499-503.

Young, R.C., and Klerman, G.L. (1992). Mania in late life: focus on age at onset. Am J Psychiatry *149*, 867-876.