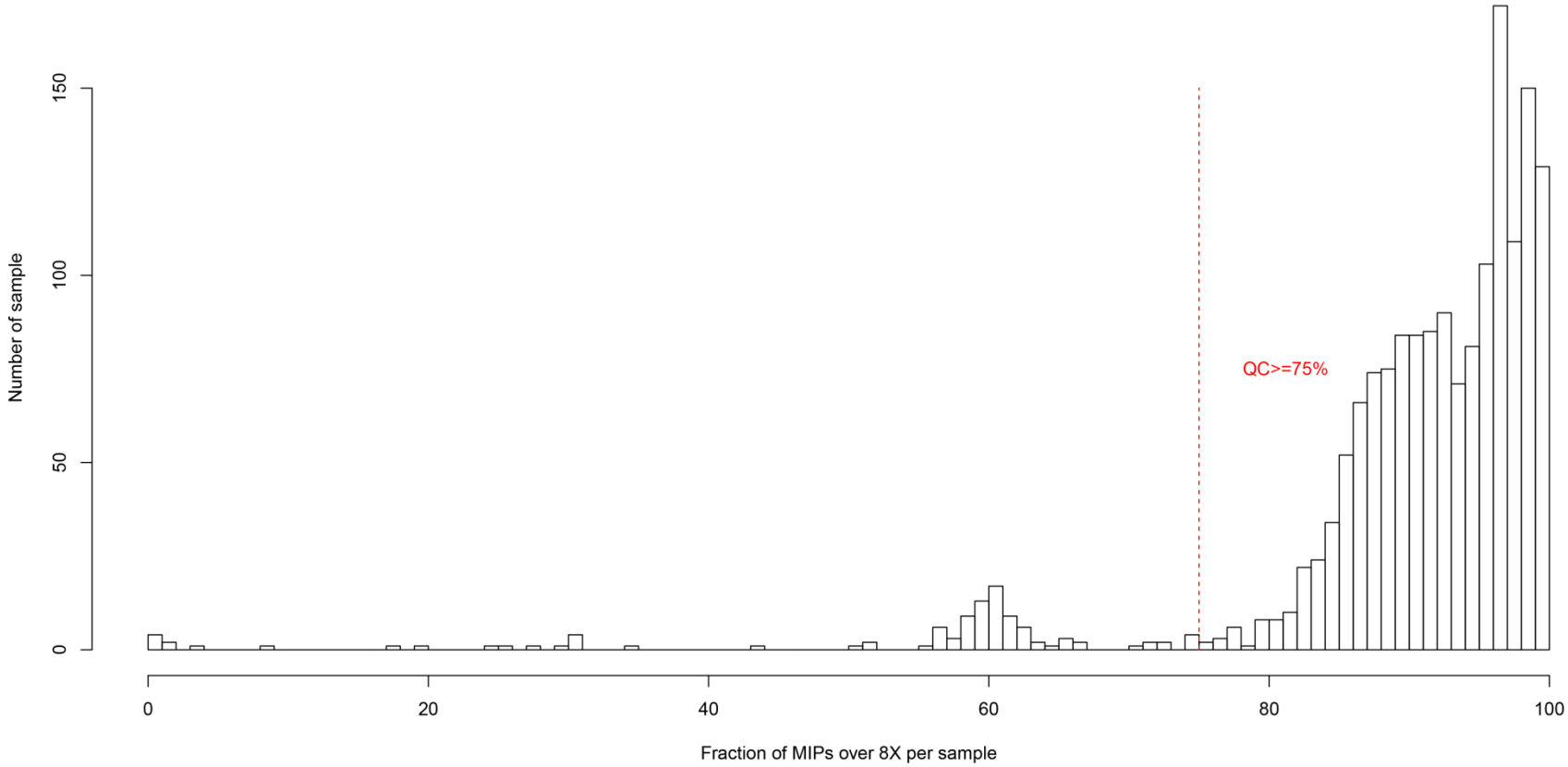
De novo genic mutations among a Chinese autism spectrum disorder cohort

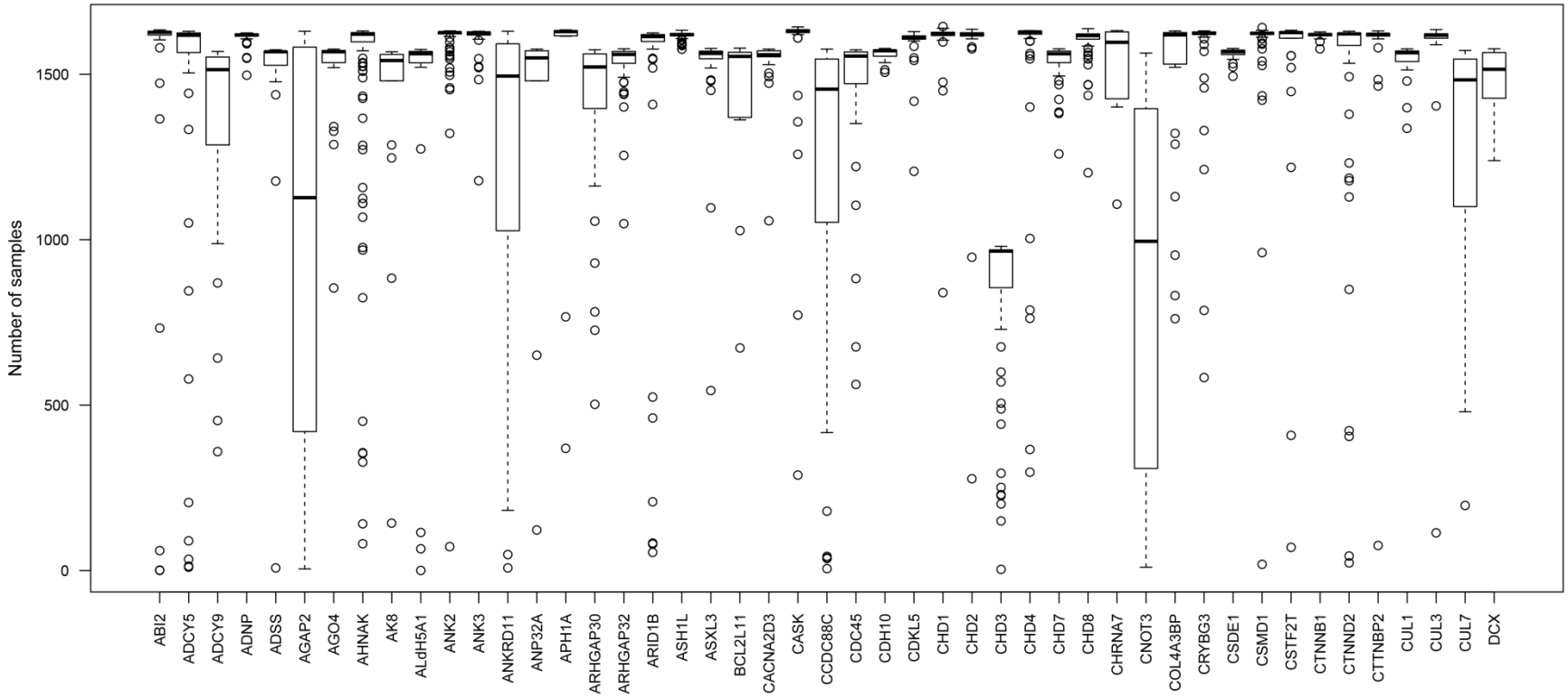
Supplementary Information

Supplementary Figures and Tables

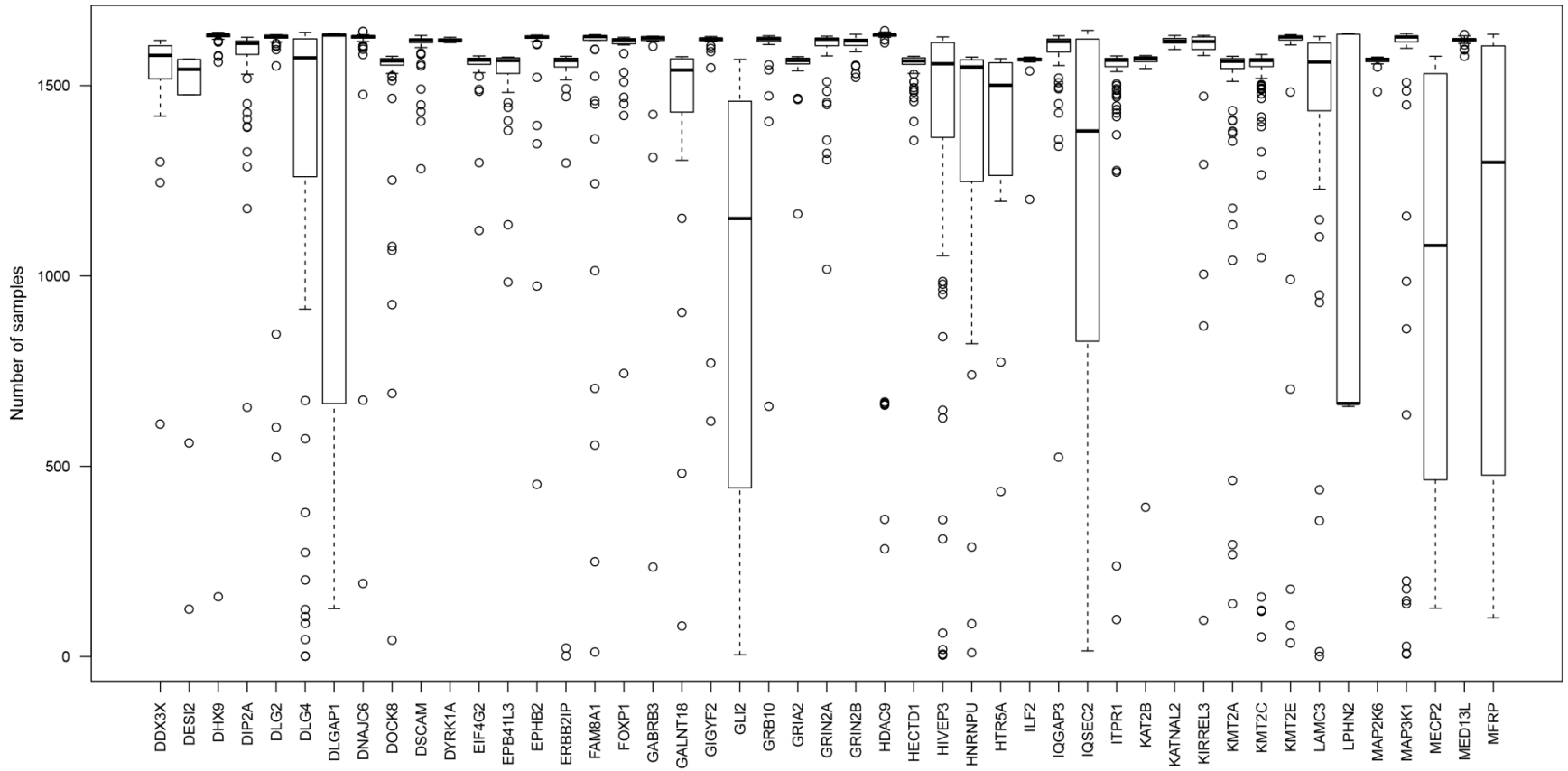


Supplementary Figure 1. Quality control (QC) of MIPs cohort. QC analysis of the percentage of MIPs with at least eight reads per sample.

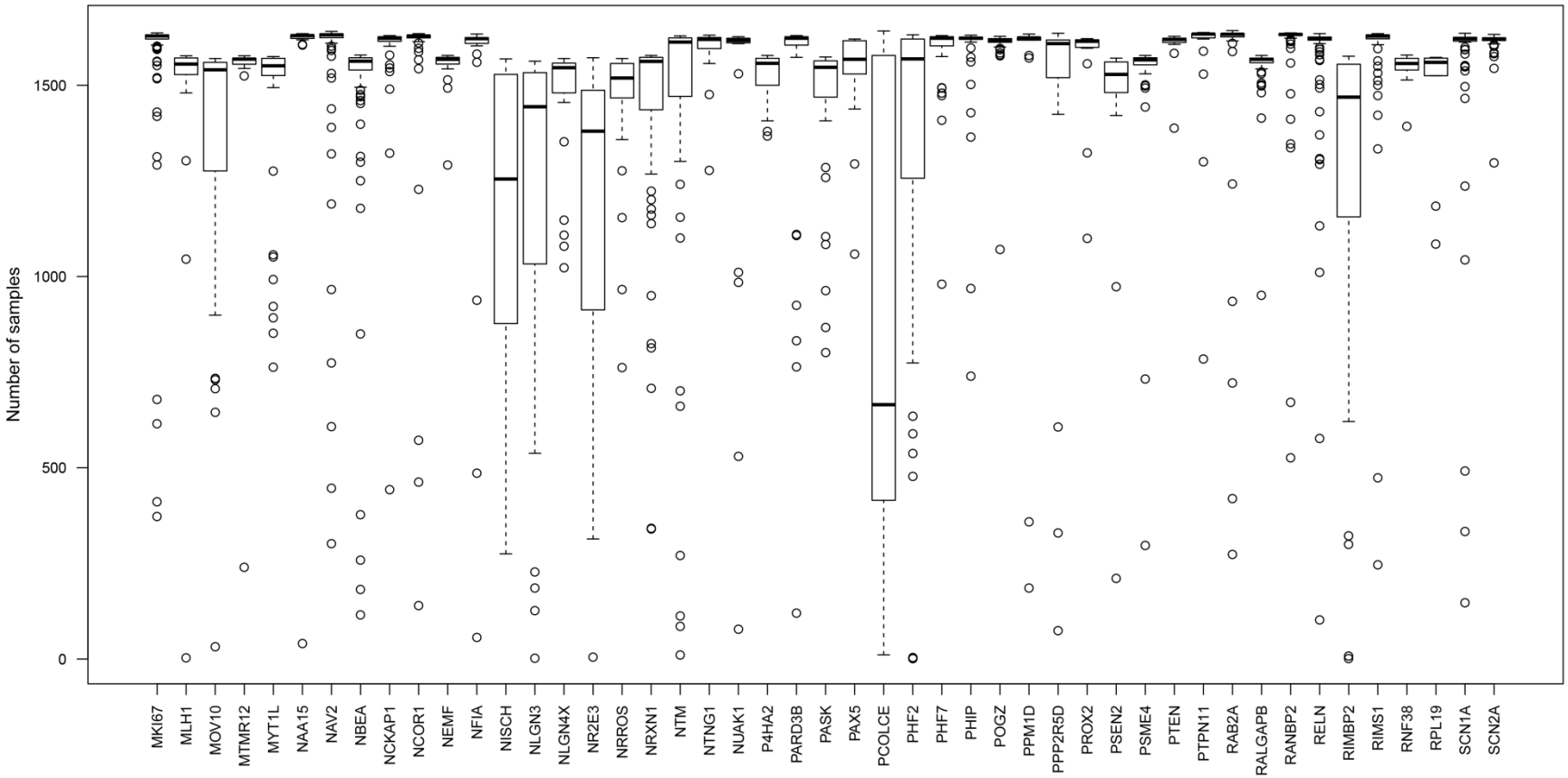
Supplementary Figure 2A



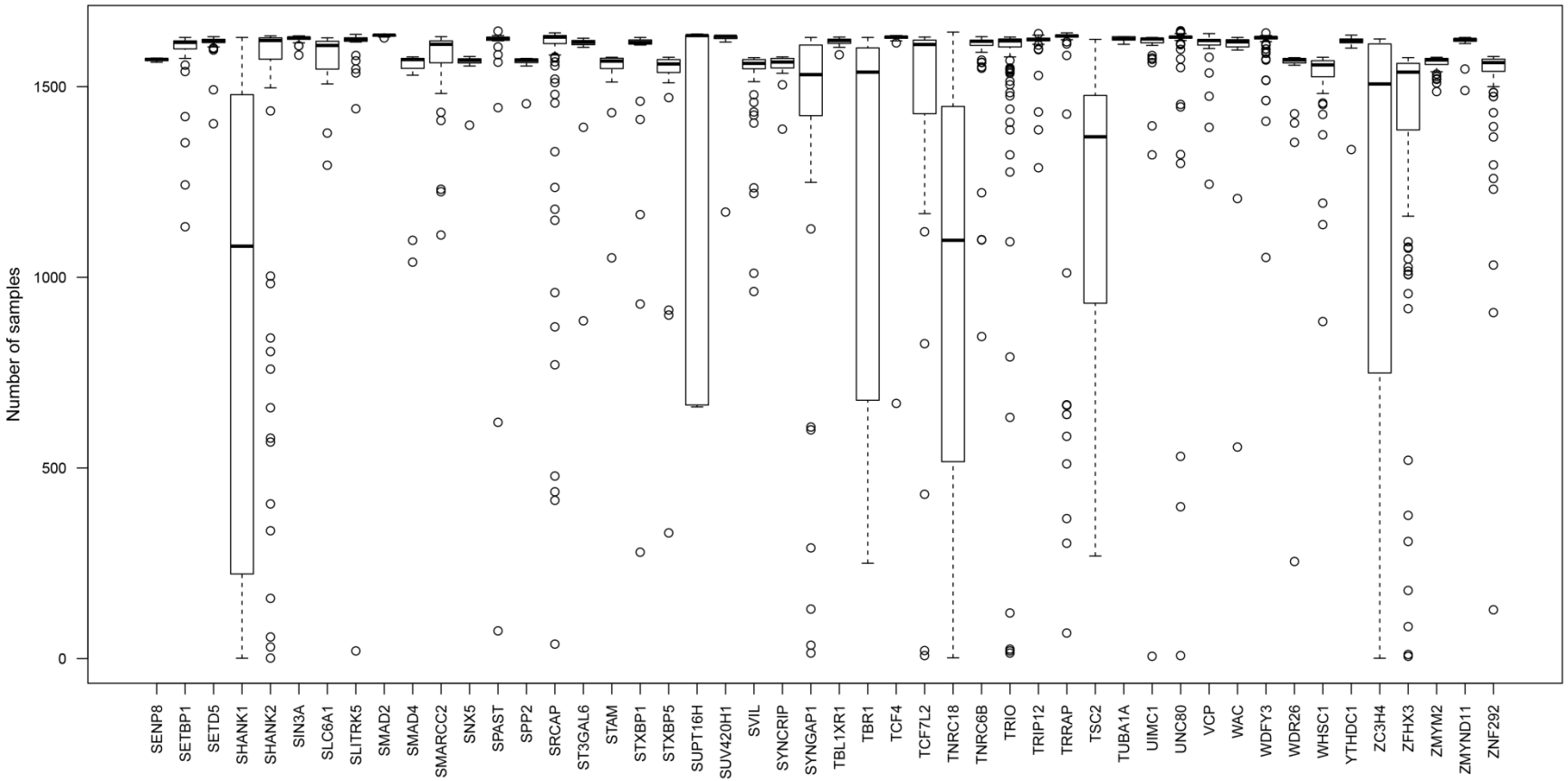
Supplementary Figure 2B



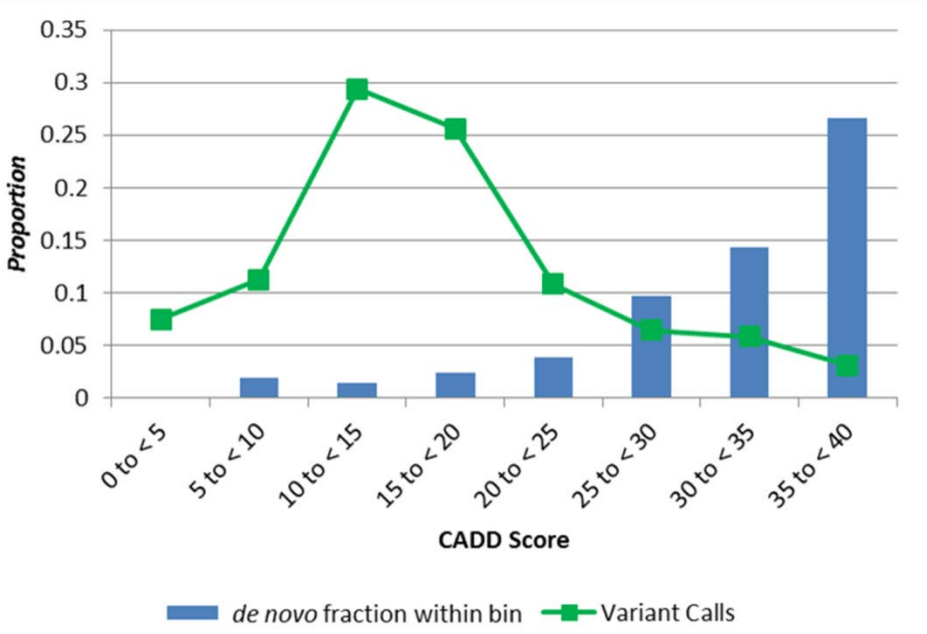
Supplementary Figure 2C



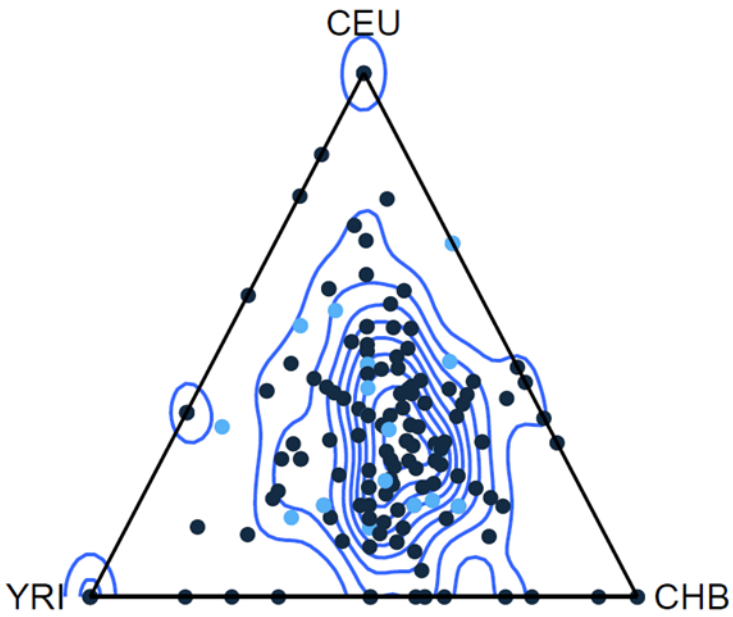
Supplementary Figure 2D



Supplementary Figure 2. Fraction of target based at 8X or greater coverage by gene. (A) Gene 1 to gene 47; **(B)** gene 48 to gene 94; **(C)** gene 95 to gene 141; **(D)** gene 142 to gene 189. Box and whisker plots show the fraction of a sample's target bases at 8X or greater coverage split by gene. All capture samples are included (along with QC failures).



Supplemental Figure 3. Increasing CADD scores predict a higher likelihood of a variant arising *de novo*. Shown are the CADD distributions of all missense variants (n = 480), detected in ASD patients, from the 29 genes with at least one *de novo* (DN) high-impact event (green line) as well as the DN rate at each CADD threshold (blue bars). Most events (n = 264, 55%) are present in the 10–20 CADD score range; concurrently, we observe very low DN yields in this range (1.9%). At increasing CADD scores, we observe both the expected decrease in overall variant frequencies and a striking increase in the DN fraction at each threshold with a peak at 26.7% of variants with a CADD score over 35. Combining all variants with CADD scores ≥ 30, we observe 18.6% of all variants as having arisen *de novo*. CADD score thresholds are therefore a useful predictor of likelihood of DN mutation.



Supplementary Figure 4. The distribution of private variation in the 1000 Genomes Project for the genes targeted by MIPs. The ternary plot compares Northwestern Europeans (CEU; top), Han Chinese from Beijing (CHB; lower right) and Yoruba in Ibadan (YRI; lower left). Each dot, a gene, represents the proportion of private variants contributed by each population; the density of the data is shown as blue contour lines. Dots at the three vertices are genes where private variation is only found in one population. Light blue dots highlight those genes that carried DN LGD mutations in this study (25 genes).

Supplementary Table 1. Detailed DSM-IV criteria for autism diagnosis in ACGC.

A	qualitative impairment in social interaction, as manifested by at least two of the following:		
	1. marked impairments in the use of multiple nonverbal behaviors such as eye-to-eye gaze, facial expression, body posture, and gestures to regulate social interaction	1=YES	2=NO
	2. failure to develop peer relationships appropriate to developmental level	1=YES	2=NO
	3. a lack of spontaneous seeking to share enjoyment, interests, or achievements with other people, (e.g., by a lack of showing, bringing, or pointing out objects of interest to other people)	1=YES	2=NO
	4. lack of social or emotional reciprocity (note: in the description, it gives the following as examples: not actively participating in simple social play or games, preferring solitary activities, or involving others in activities only as tools or "mechanical" aids)	1=YES	2=NO
B	qualitative impairments in communication as manifested by at least one of the following:		
	1. delay in, or total lack of, the development of spoken language (not accompanied by an attempt to compensate through alternative modes of communication such as gesture or mime)	1=YES	2=NO
	2. in individuals with adequate speech, marked impairment in the ability to initiate or sustain a conversation with others	1=YES	2=NO
	3. stereotyped and repetitive use of language or idiosyncratic language	1=YES	2=NO
	4. lack of varied, spontaneous make-believe play or social imitative play appropriate to developmental level	1=YES	2=NO
C	restricted repetitive and stereotyped patterns of behavior, interests and activities, as manifested by at least two of the following:		
	1. encompassing preoccupation with one or more stereotyped and restricted patterns of interest that is abnormal either in intensity or focus	1=YES	2=NO
	2. apparently inflexible adherence to specific, nonfunctional routines or rituals	1=YES	2=NO
	3. stereotyped and repetitive motor mannerisms (e.g., hand or finger flapping or twisting, or complex whole-body movements)	1=YES	2=NO
	4. persistent preoccupation with parts of objects	1=YES	2=NO
I	A total of six (or more) items from (A), (B), and (C), with at least two from (A), and one each from (B) and (C)	1=YES	2=NO
II	Delays or abnormal functioning in at least one of the following areas, with onset prior to age 3 years:	1=YES	2=NO
	a. social interaction; b. language as used in social communication; c. symbolic or imaginative play		
III	The disturbance is not better accounted for by Rett's disorder or Childhood Disintegrative Disorder	1=YES	2=NO
	(I), (II) and (III) are all YES?	1=YES	2=NO

Supplementary Table 2. Regional distribution of ASD samples from ACGC in China.

Region	Number
Shandong	389
Hunan	264
Guangdong	202
Jiangsu	128
Zhejiang	72
Anhui	68
Henan	65
Fujian	54
Hebei	43
Liaoning	40
Sichuan	38
Hubei	36
Jilin	33
Shanxi	32
Jiangxi	29
Heilongjiang	26
Chongqing	18
Inner Mongolia	16
Beijing	15
Shanghai	15
Shanxi(Jin)	14
Guangxi	12
Tianjin	12
Gansu	8
Xinjiang	7
Hainan	5
Guizhou	3
Yunnan	3
Total	**1647**

Supplementary Table 3. Variant counts of the 29 genes with DN mutations identified in ACGC.

Gene	LGD			MIS(CADD>30)			MIS(CADD≤30)		
	DN	Validated in trios	Validated in cases	DN	Validated in trios	Validated in cases	DN	Validated in trios	Validated in cases
SCN2A	7	8	8	1	2	3	4	20	20
CHD8	3	3	3	0	1	1	1	20	20
DSCAM	2	3	3	0	2	4	1	18	20
MECP2	2	2	3	0	0	0	0	3	3
ADNP	1	2	2	0	1	1	1	9	11
ARHGAP32	1	1	1	0	0	0	0	20	22
CDKL5	1	3	3	0	0	0	0	3	3
CUL3	1	1	1	0	0	0	0	0	0
DOCK8	1	3	3	0	7	7	0	32	37
DYRK1A	1	1	1	0	1	1	1	7	7
GIGYF2	1	1	2	0	0	1	1	16	16
GRIN2B	1	1	1	0	0	0	0	9	10
MED13L	1	1	1	0	2	2	0	15	18
MYT1L	1	1	1	0	0	0	0	6	7
NCKAP1	1	1	1	0	0	0	0	3	3
NCOR1	1	1	1	0	1	1	0	25	27
PHIP	1	1	1	0	0	0	0	7	7
POGZ	1	1	1	1	3	3	0	13	14
RIMS1	1	5	8	0	1	1	0	14	18
SHANK1	1	1	1	0	0	0	0	10	11
STXBP1	1	1	2	0	1	1	0	2	2
SYNGAP1	1	1	2	0	0	0	0	2	2
TRIP12	1	1	1	0	1	1	0	11	11
WDFY3	1	2	2	1	4	5	1	20	22
ZNF292	1	1	1	0	0	0	0	29	32
ASH1L	0	0	0	2	5	6	1	21	24
CHD2	0	0	0	1	1	1	0	12	14
ITPR1	0	1	1	1	1	1	0	18	21
TSC2	0	0	0	1	3	3	0	21	27
total	35	48	55	8	37	43	11	386	429

Supplementary Discussion

We considered paternal age at the time of conception (Supplementary Table 1). Fathers of individuals who carry a DN LGD or DN MIS30 mutation in the ACGC are not significantly older than those of individuals who do not carry a predicted high-impact DN variant. In fact, ACGC fathers of DN mutation carriers were significantly younger than fathers of DN mutation carriers from published European cohorts (SSC and The Autism Simplex Collection (TASC) assayed by MIPs; p = 0.016, one-tailed Student's t-test)[1]. Therefore, although increased paternal age at conception has been associated with increased rates of DN mutation[2], this does not appear to explain differences in DN mutation rate in the ACGC compared to other studies. While penetrance and environmental modifiers are a possibility, it is far more likely that the increase in DN *SCN2A* mutations is a reflection of ascertainment and possible technological differences (i.e., exome versus MIPs). Much of our current understanding of European/American autism gene mutation rates has been driven by the SSC (Simons Simplex Collection), which has specifically excluded children with low IQ.

As a surrogate of DN mutation, we tested whether the rates of private variation in our candidate genes differ between ethnic groups. We quantified the distribution of private variation in the

1000 Genomes Project (release: 20130502) by enumerating private variants by gene and population (Western European: CEU, Han Chinese: CHB, and African: YRI). We have provided these data in a ternary plot (Supplementary Figure 4; Supplementary Data 9). Overall, most genes have comparable frequencies of rare variants (CEU:0.3, CHB:0.38, YRI:0.32) supporting the null hypothesis that DN mutations in genes contributing to autism will not differ significantly between global populations. Interestingly, we observed a slight increase in frequency for the CHB (Han Chinese from Beijing); the CHB had the highest proportion of private variants for 95 of the 179 genes we tested, possibly consistent with demographic data that this population shows one of the greatest expansions over the last 5000 years[3]. In the case of *SCN2A*, however, this factor alone is insufficient to account for the increased rate of DN mutations that we observe in the ACGC compared to other published studies.

References

1. O'Roak BJ, Stessman HA, Boyle EA, Witherspoon KT, Martin B, Lee C*, et al.* Recurrent de novo mutations implicate novel genes underlying simplex autism risk. *Nature communications* 2014, **5**: 5595.

2. Kong A, Frigge ML, Masson G, Besenbacher S, Sulem P, Magnusson G*, et al.* Rate of de novo mutations and the importance of father's age to disease risk. *Nature* 2012, **488**(7412)**:** 471-475.

3. Genomes Project C, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM*, et al.* A global reference for human genetic variation. *Nature* 2015, **526**(7571)**:** 68-74.