Supplemental Data

List of Figures

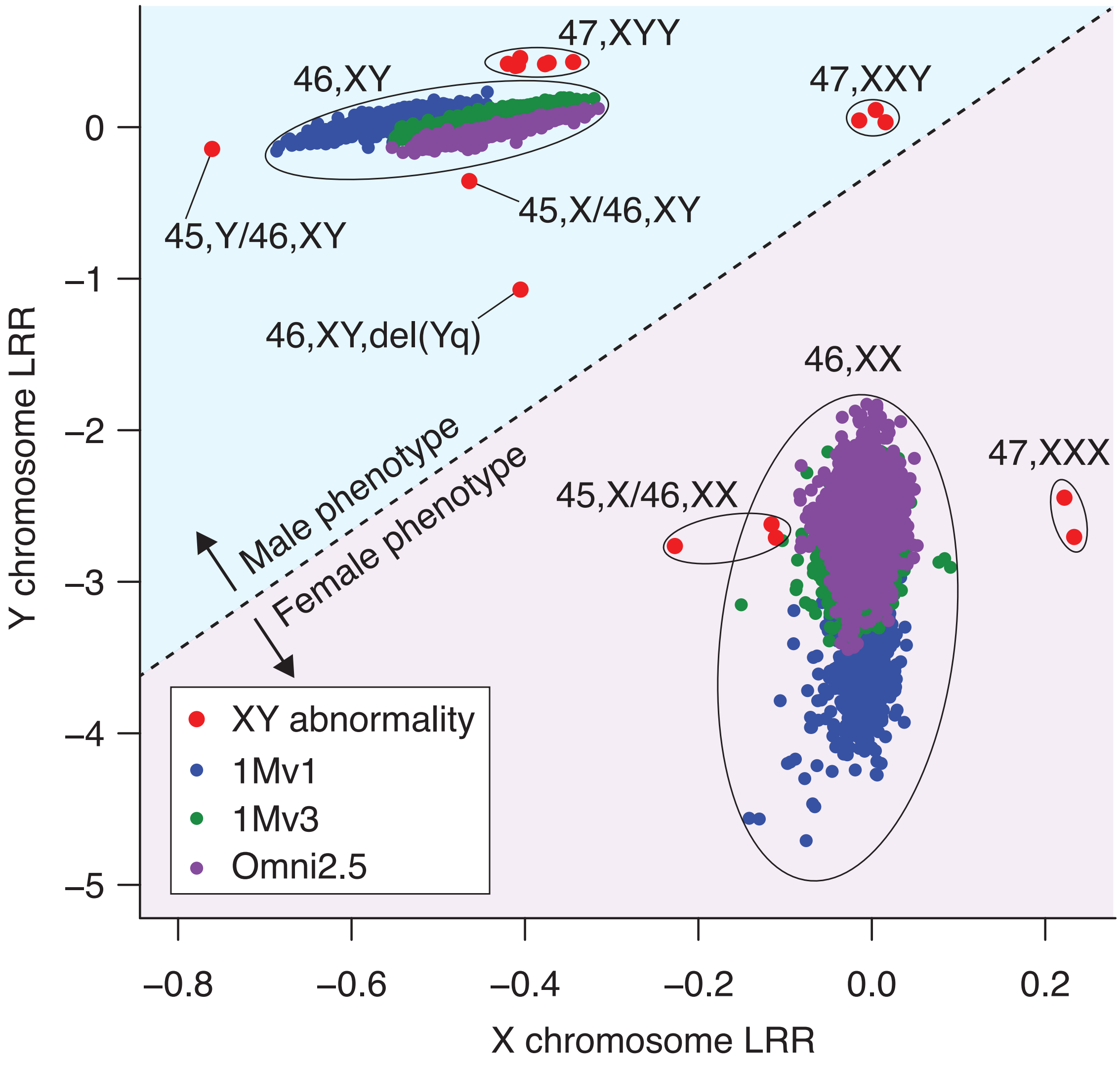

Figure S1. Sex determination by analysis of LRR. Related to Figure 1.

Genotypic sex was determined through comparison of LRR on the X chromosome (x-axis) and LRR on the Y chromosome (y-axis) based on Illumina SNP genotyping data. The majority of samples from all three BeadChip types, 1Mv1 (blue), 1Mv3 (green), and Omni2.5 (purple) formed two distinct clusters corresponding to females (46,XX) and males (46,XY) with a normal complement of sex chromosomes. The raw SNP genotyping data for all outliers were visualized to assess whether a chromosomal aneuploidy was present; samples with an abnormal complement of sex chromosomes are shown in red with the corresponding cytogenetic description. Samples that appear as outliers, but are not colored in red, have no evidence of chromosome aneuploidy on visualization. Predicted phenotypic sex is shown by the pink (female) and blue (male) background.

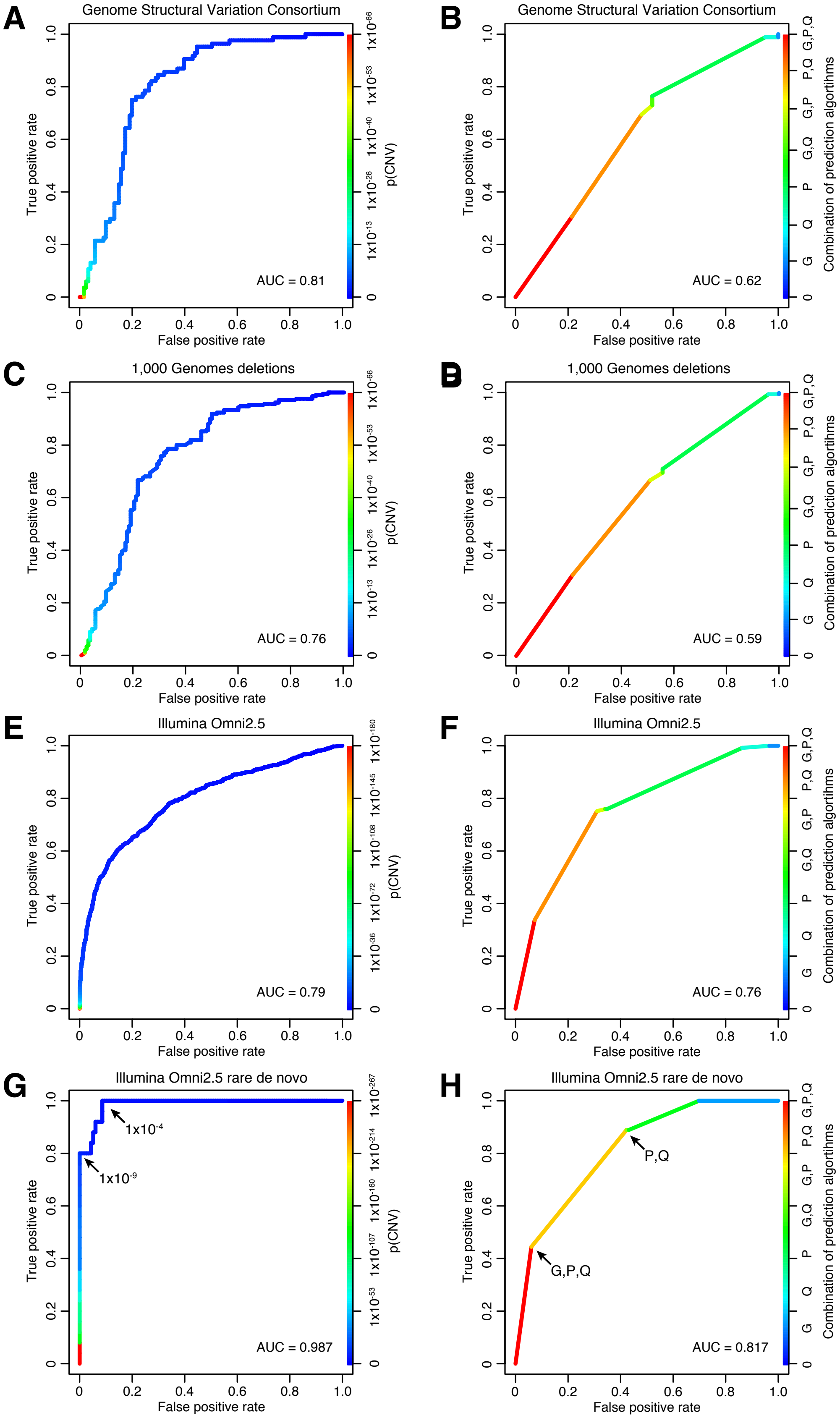

Figure S2. Receiver operating characteristic (ROC) curves for CNV detection. Related to Figure 2.

Figure S2. Receiver operating characteristic (ROC) curves for CNV detection. Related to Figure 2.

A) The ability of the p_{CNV} metric to distinguish CNVs detected and validated by the Genome Structural Variation (GSV) Consortium (Conrad et al., 2010) from those not detected/validated was assessed for three HapMap samples analyzed with a high density CGH array by the GSV consortium and on an Illumina Omni2.5 SNP genotyping array to estimate p_{CNV}. The results are assessed using a Receiver Operating Characteristic Curve (ROC) which achieves an area under the curve (AUC) of 0.81. **B)** The analysis described in 'A' is repeated using a combination of CNV prediction algorithms (PennCNV (Wang et al., 2007), QuantiSNP (Colella et al., 2007), and GNOSIS (Sanders et al., 2011)); a lower AUC is achieved than with the p_{CNV} metric. **C** and **D)** The analysis is repeated for the same samples, but using deletions predicted by the 1,000 genomes project. A higher AUC is achieved using p_{CNV} than algorithm overlap. **E** and **F)** The analysis is repeated using CNVs detected on the Illumina 1Mv3 SNP genotyping array and assessing the ability the p_{CNV} metric and the combination of CNV prediction algorithms to distinguish CNVs also detected on the Illumina Omni2.5 SNP genotyping array. The p_{CNV} metric achieves a higher AUC. **G** and **H)** The analysis is repeated for qPCR validated *de novo* CNVs. Again the p_{CNV} metric achieves a higher AUC with a threshold of p_{CNV} 1x10^{-9} achieving 100% specificity and 80% sensitivity.

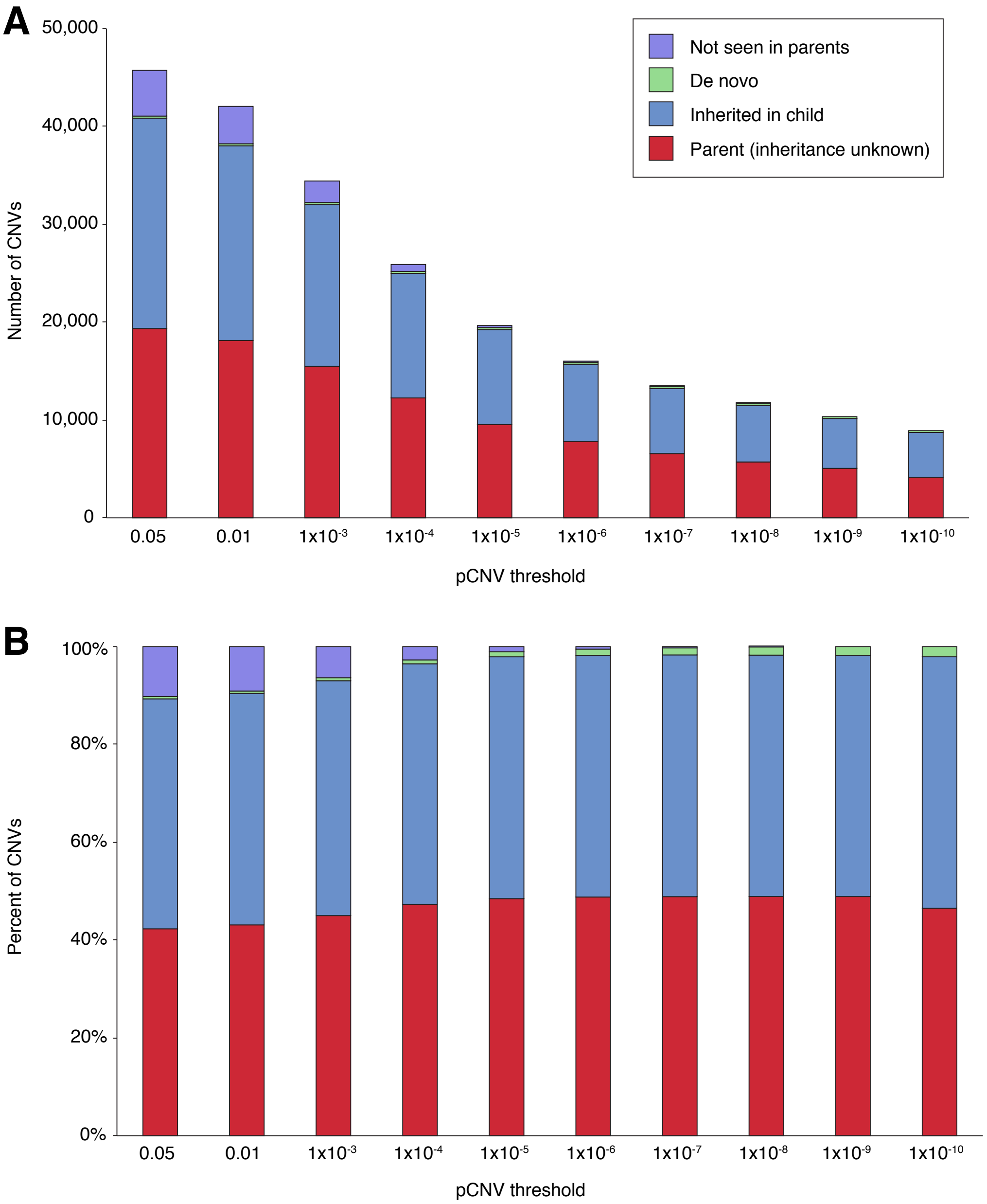

Figure S3. CNV inheritance by pCNV threshold. Related to Figure 2.

A) 52,249 rare (population frequency ≤0.1%) autosomal CNVs were detected in the 2,591 families. Based on the detection of a corresponding CNV in the parent, the inheritance in the child was estimated to be 'inherited' (blue), 'De novo' (green, meeting the criteria for *de novo* detection, see section 13.1), or 'Not seen in parents' (purple). For comparison the CNVs detected in the parents are also shown (red). The 'Not seen in parents' category is likely to be highly enriched for false positive calls and we would expect a similar proportion of false positive calls in the parents, in whom we cannot assign inheritance. As the p_{CNV} threshold was reduced, leading to more stringent detection, the number of detected CNVs decreases, with the 'Not seen in parents' reducing most rapidly. A complete list of these CNVs are available in Table S3. **B)** This figure shows the same data as in 'A', however the percentage of CNVs in each of the categories is shown rather than the absolute number. Based on this image a p_{CNV} threshold of 1x10^{-4} was used for the detection of rare inherited CNVs 14.1 and parental CNV burden 14.6.

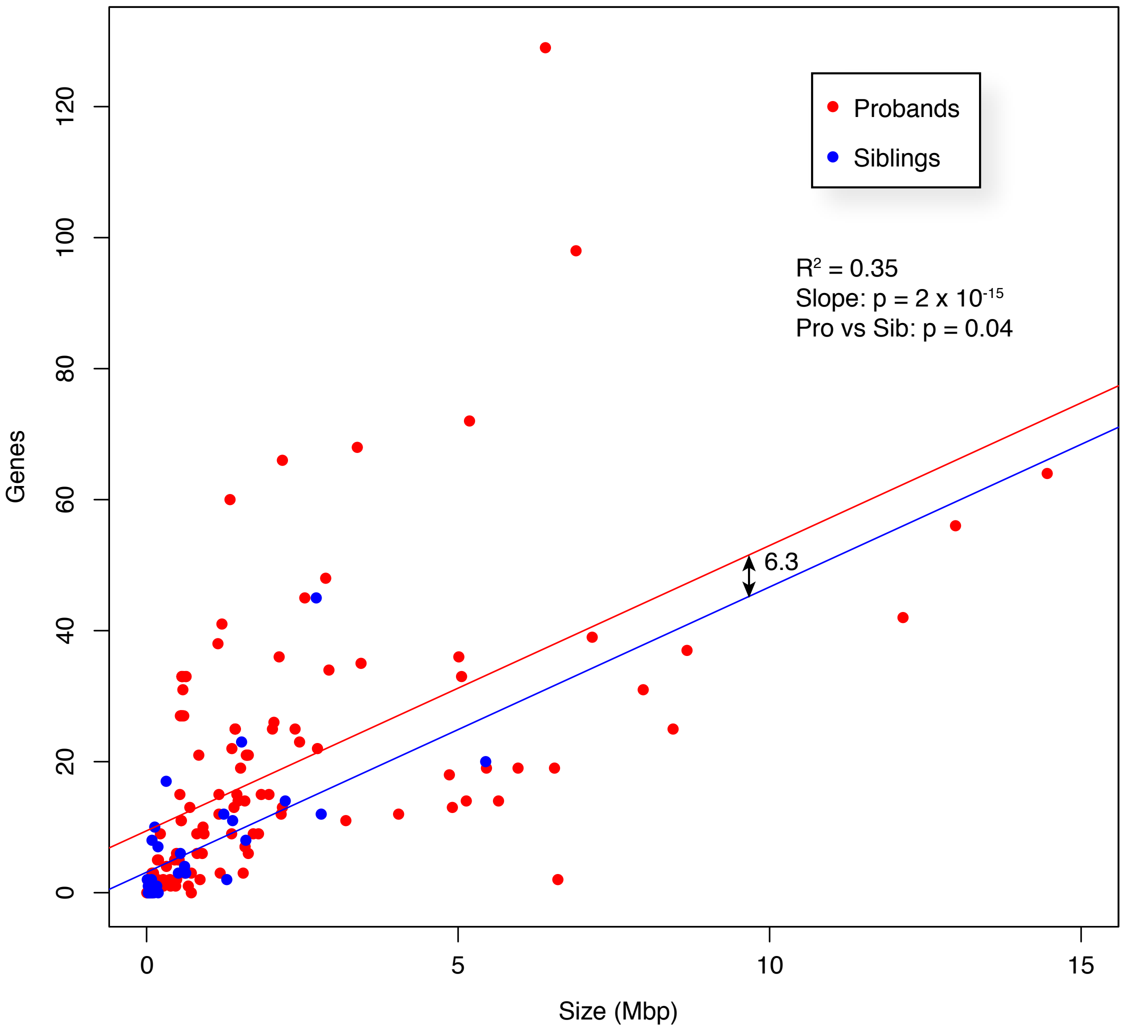

Figure S4. Relationship between *de novo* CNV size and gene count in probands and siblings. Related to Figure 2. A linear model was used to explore the relationship between phenotype, CNV size, and the number of genes within CNVs for all filtered *de novo* CNVs from trio and quartet families (Table S2). Size and gene number were correlated (gene count $\sim$ size; $R^2 = 0.34$; $\beta = 4.7\text{x}10^{-6}$; $p < 2\text{x}10^{-16}$). Adding phenotype to the model showed 6.3 more genes per *de novo* CNV in probands compared to sibling (gene count $\sim$ size $+$ phenotype; $R^2 = 0.36$; $p = 0.04$ for phenotype); this is represented by the higher y-intercept for the line in probands than siblings. No interaction was observed between size and phenotype therefore the slope does not differ between probands and siblings (gene count $\sim$ size * phenotype; $R^2 = 0.36$; $p = 0.61$ for interaction term).

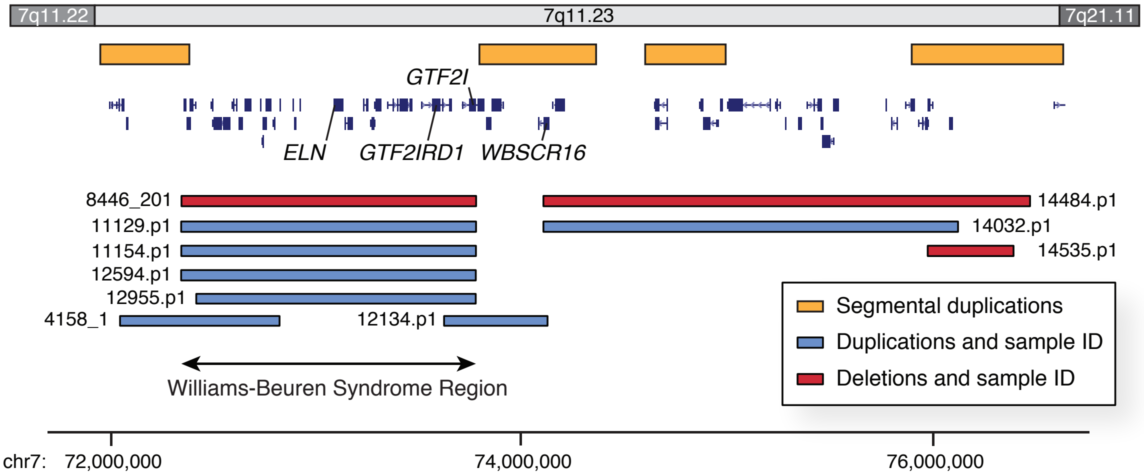

Figure S5. Multiple overlapping dnCNVs in the 7q11.23 region

Ten dnCNVs (seven duplications and three deletions) were observed in the 7q11.23 region. Six duplications and one deletion intersect with the Williams-Beuren Syndrome (WBS) region, however two of these are atypical duplications that only intersects with one end of the WBS region. The larger of these atypical duplications intersects with 10 genes, while the smaller one only intersects with the general transcription factor genes *GTF2IRD1* and *GTF2I*. The smaller atypical duplication overlaps with a deletion and duplication over the gene *WBSCR16*. No dnLoF mutations were identified throughout this region.

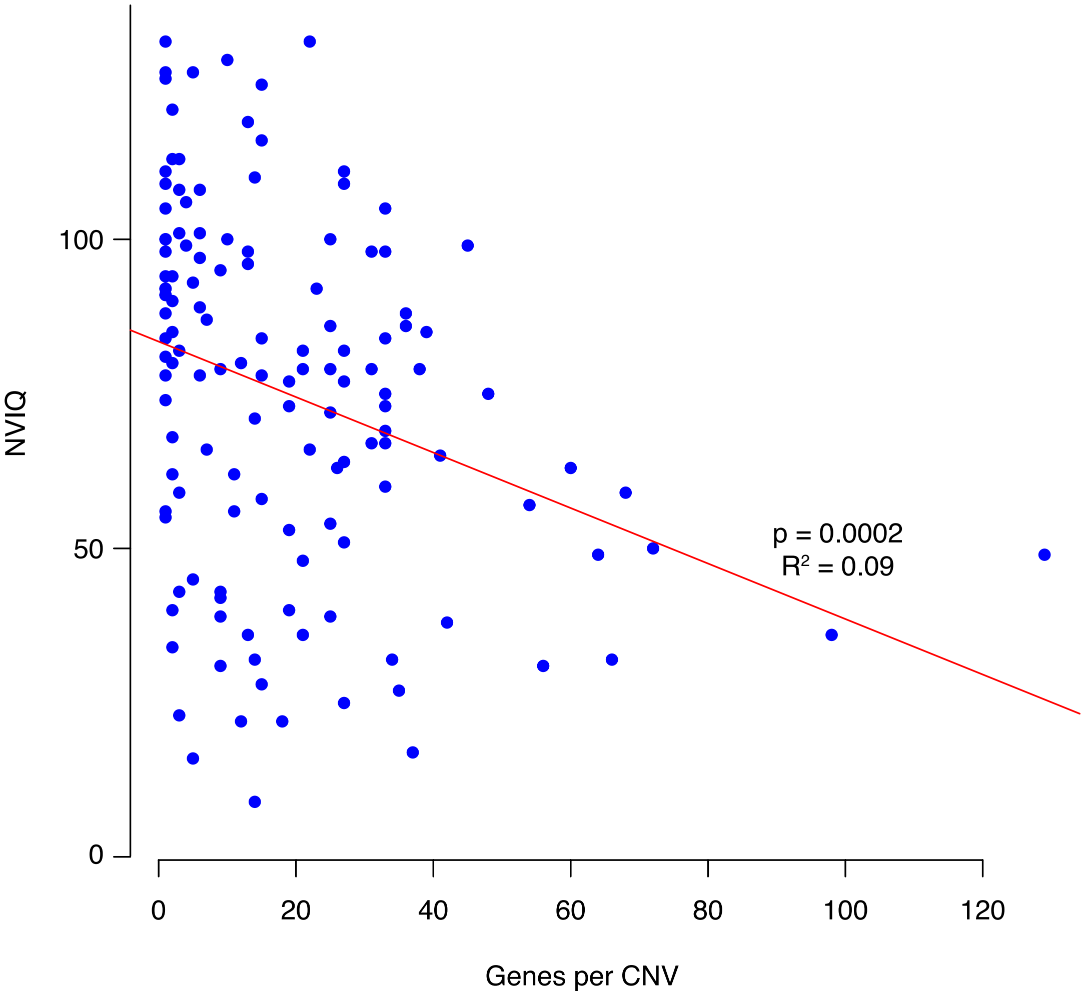

Figure S6. Relationship between CNV gene count and NVIQ. Related to Figure 3.
The number of genes within probands dnCNVs in comparison to non-verbal IQ (NVIQ). Each point represents a proband from the SSC with a dnCNV. The NVIQ decreases as the number of genes increases (p=0.0002, linear regression).

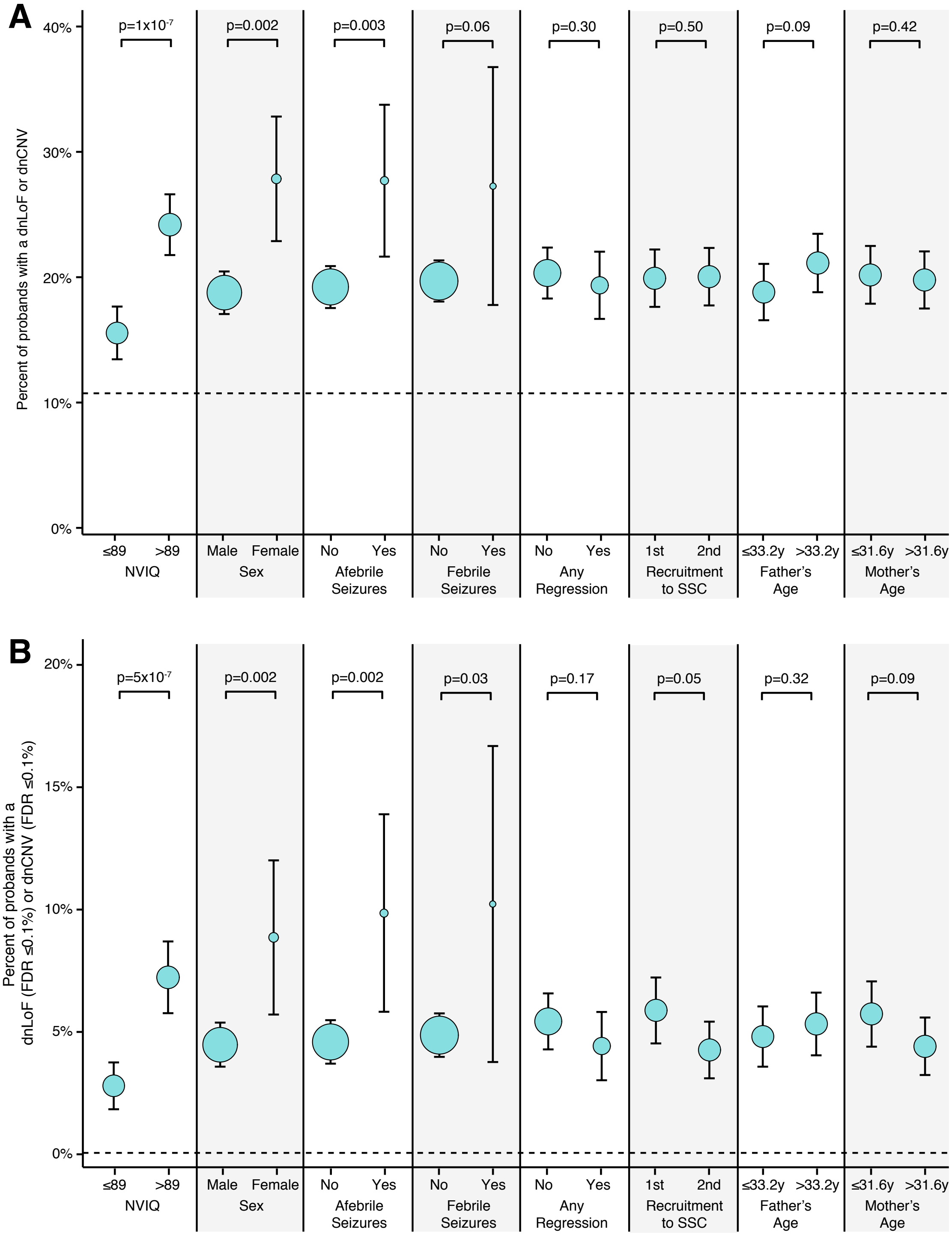

Figure S7. Relationship of phenotypic factors and *de novo* rate. Related to Figure 3.
A) The rate of *de novo* mutations in probands is shown for a range of phenotypic factors for the 2,591 probands in the SSC. For non-verbal IQ (NVIQ), father's age, and mother's age the median of the proband group is used to divide the group in half. Recruitment to the SSC is divided into two groups based on the family ID assigned (above or below 13183). P-values were calculated using a Fisher's exact test comparing the two groups of probands. The whiskers show the 95% confidence intervals estimated with the odds ratio. The size of the circles represents the number of samples in each group. **B)** The analysis in 'A' is repeated considering only *de novo* mutations at loci with an FDR $\leq$0.1%.

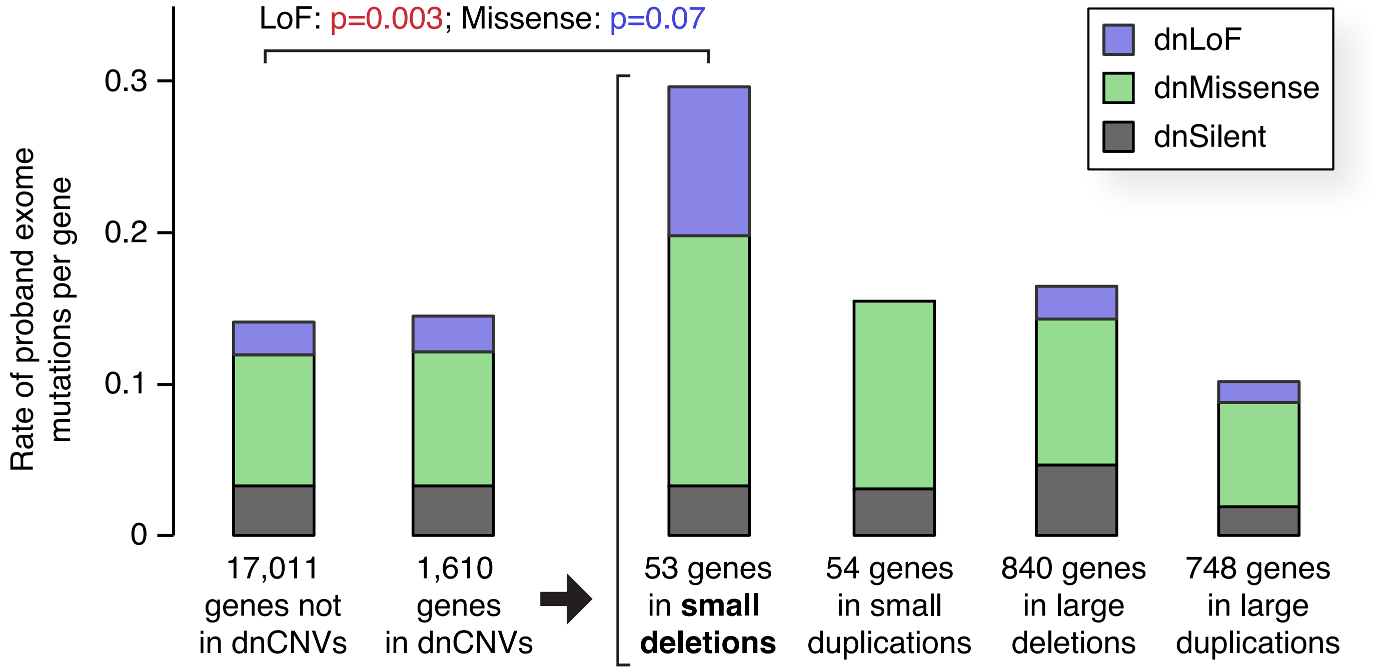

Figure S8. Enrichment for *de novo* LoF mutations in small *de novo* deletions in the SSC. Related to Figure 5. The number of *de novo* mutations per gene observed with exome sequencing of the SSC are shown in different groups of genes based on SSC CNV overlap. Mutation rates are normalized for gene mutability based on gene size and GC content. Exome mutations are divided into silent (grey), missense (green), and LoF (purple). No excess of exome mutations is observed in the 1,610 genes within dnCNV regions compared to the 17,011 genes outside of dnCNVs. Dividing the dnCNV regions by size ($\leq$7 genes vs. $>$7 genes) and type (deletion vs. duplication) reveals strong enrichment for dnLoF (p=0.003, Fisher Exact Test) in small *de novo* deletions only.

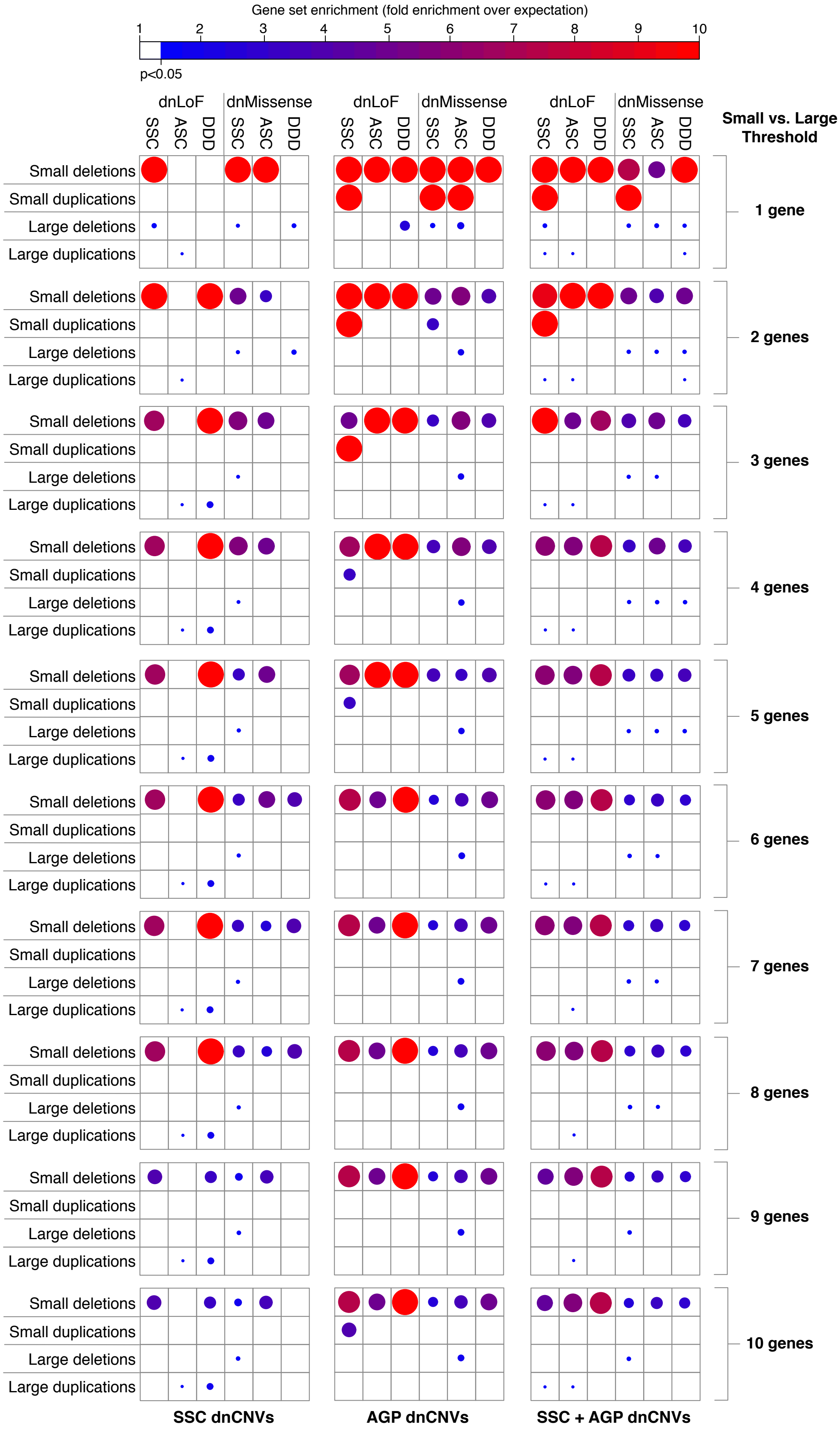

Figure S9. Enrichment for *de novo* across size thresholds. Related to Figure 5.

Figure S9. Enrichment for *de novo* across size thresholds. Related to Figure 5.
The enrichment of genes within dnCNVs is shown by the size and shade of the circle (red and large = high degree of enrichment; blue and small = modest degree of enrichment); only results reaching nominal significance (assessed with a hypergeometric test) are shown. Small *de novo* deletions show consistent enrichment for dnLoF and dnMissense mutations across three cohorts: SSC (Iossifov et al., 2014), Autism Sequencing Consortium (ASC) (De Rubeis et al., 2014), and Deciphering Developmental Disorders (DDD) (Deciphering Developmental Disorders, 2015). This result is observed for dnCNVs detected in the SSC and Autism Genome Project (AGP) (Pinto et al., 2014) independently and in combination. Furthermore, by varying the threshold defining small and large dnCNVs between and 1 and 10 genes (7 is used in the main manuscript) we show that the results are robust across a range of thresholds.

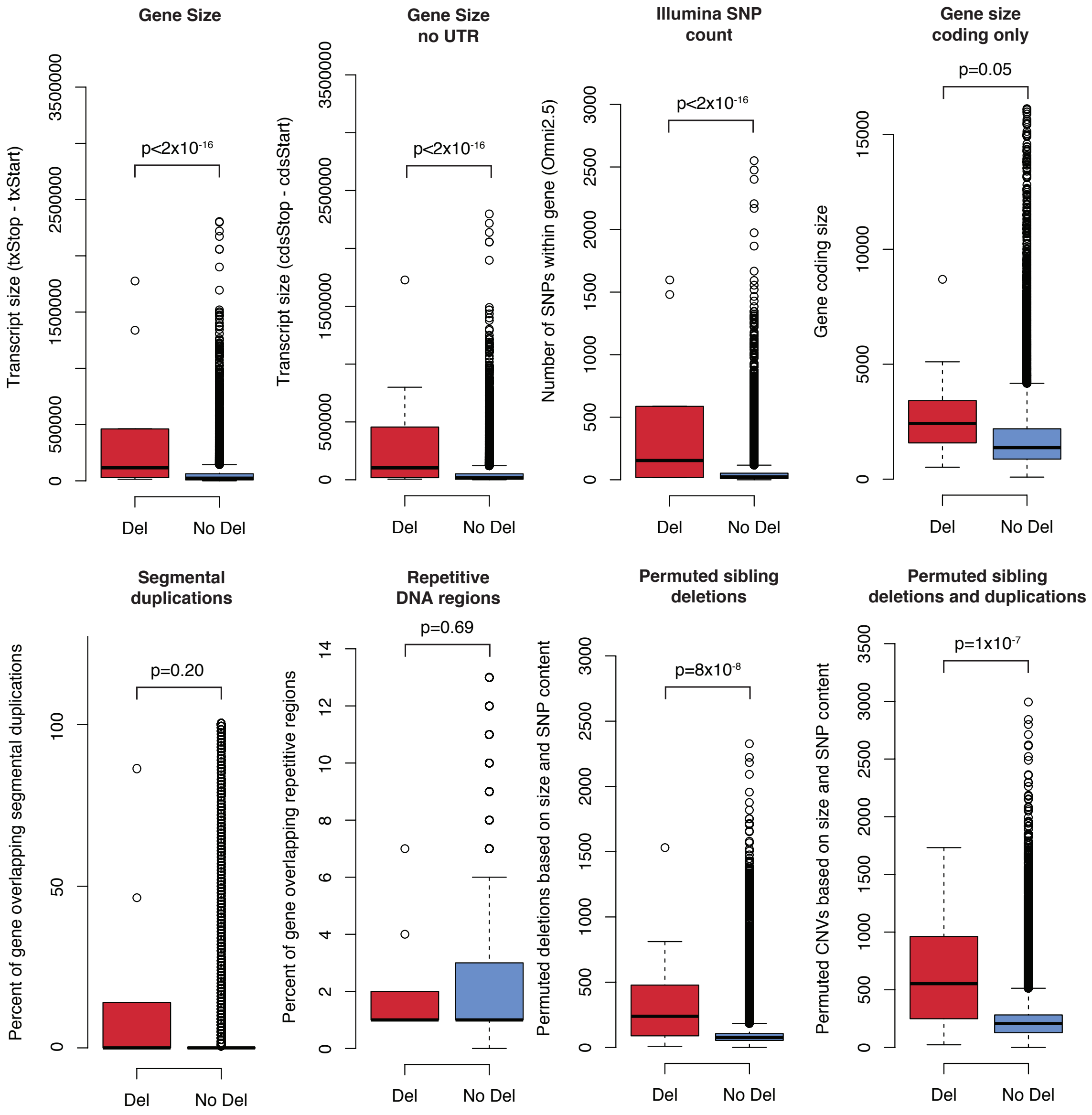

Figure S10. Factors influencing the rate of small *de novo* deletions in siblings. Related to Figure 6. Boxplots show the distribution of eight factors within genes in small *de novo* deletions in siblings (red) and all other RefSeq genes (blue). P-values are estimated using linear regression in R. The first three factors relate to gene transcript length and are show significantly higher values in genes within deletions. The fourth factor considered is the gene size counting only coding regions (exons) within the gene and it is nominally significant. The first two plots on the second line show the distribution of segmental duplications and repetitive regions of DNA, neither of which are significantly associated with genes within deletions. The last two plots show permuted CNVs based on CNV size and Illumina Omni2.5 SNP content for small *de novo* deletions in siblings and small *de novo* deletions and duplications in siblings.

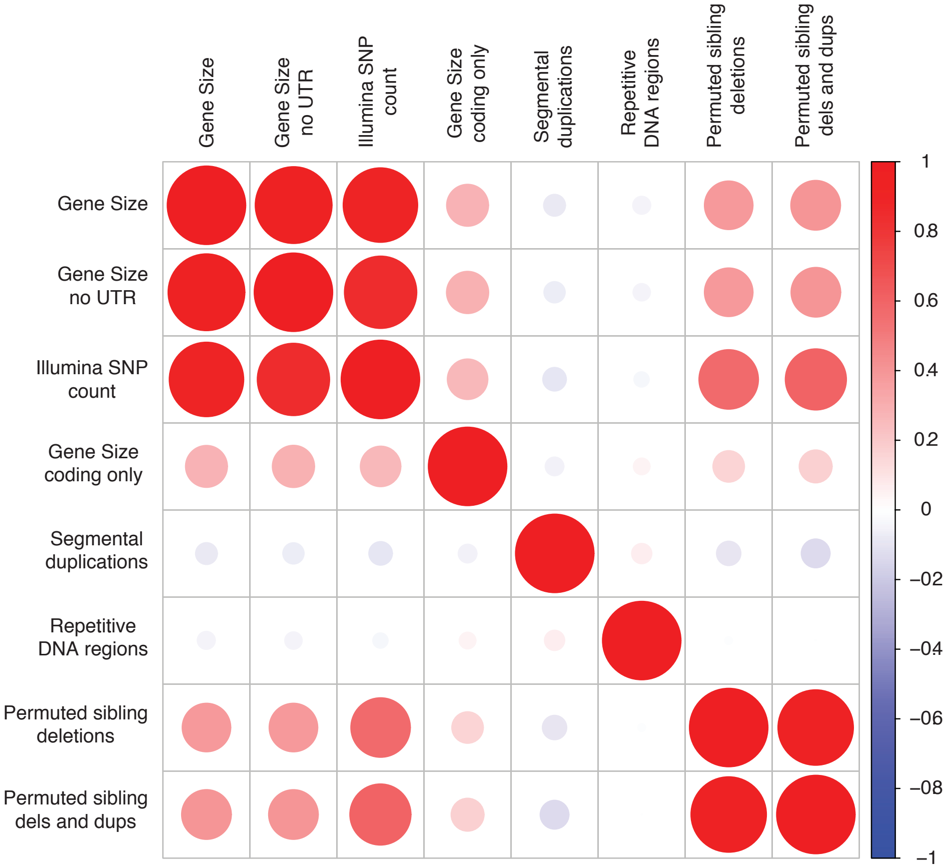

Figure S11. Correlation of factors influencing small *de novo* deletions. Related to Figure 6.
The correlation of the eight factors shown in figure S10 with each other are shown. The degree of correlation is shown by the area of the circle and the depth of color, while the color (red vs. blue) shows whether this correlation is positive (red) or negative (blue). Gene size and the permuted CNVs are highly correlated with each other, while the segmental duplications and repetitive regions of DNA are not.

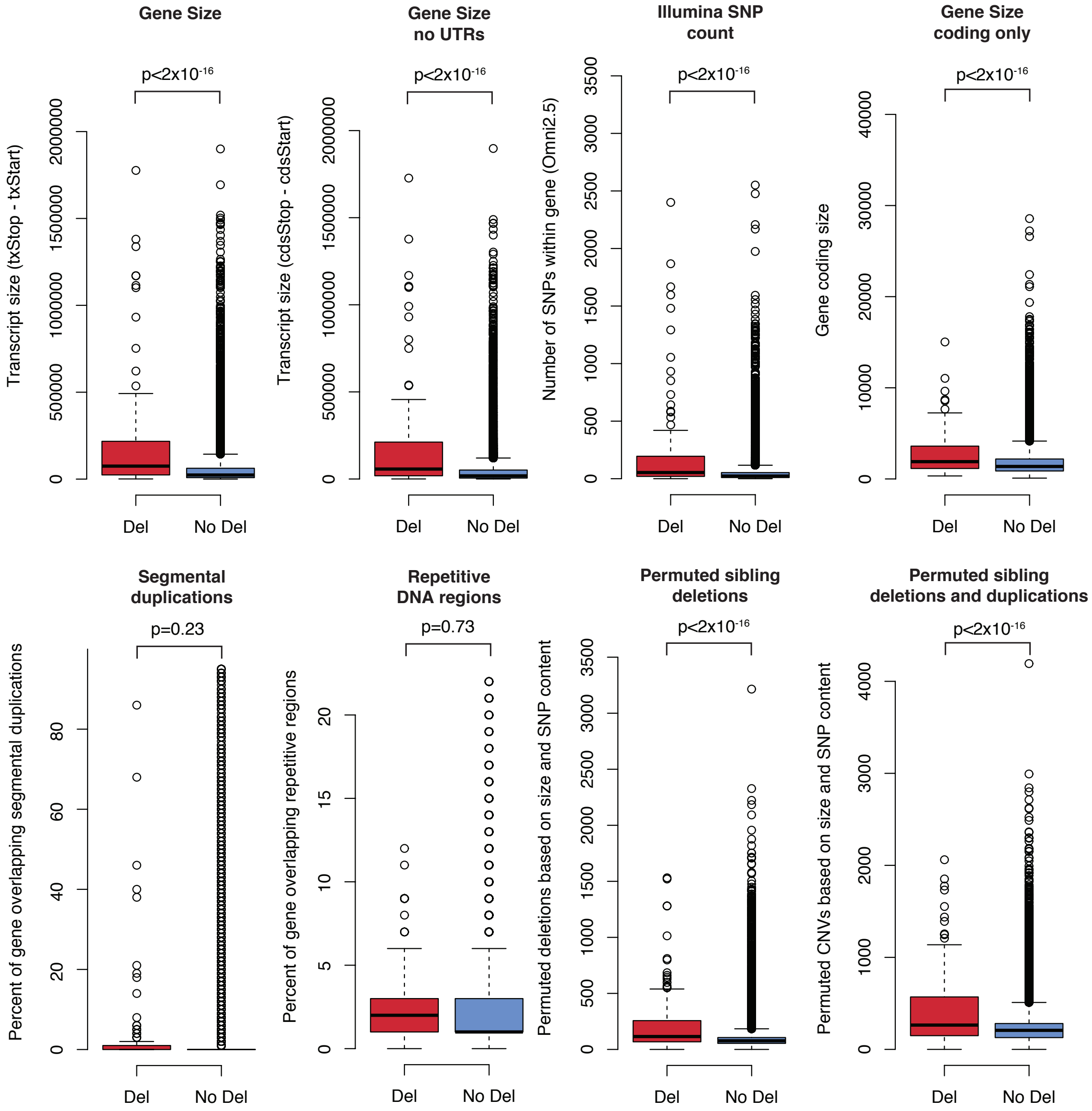

Figure S12. Factors influencing the rate of small *de novo* deletions in probands and siblings. Related to Figure 6. The analysis presented in figure S10 was repeated with the inclusion of proband, alongside sibling, small *de novo* deletions to increase power. The factors significantly associated with deletions remain unchanged from the analysis based on sibling deletions alone.

List of Tables

Table S1. List of samples included in the CNV and exome analysis (.xlsx file). Related to Figure 1.

The following columns are present:

1. **Cohort** The cohort the sample is from
2. **Family** The familyID of the sample
3. **Proband** The sampleID of the proband
4. **Father** The sampleID of the father
5. **Mother** The sampleID of the mother
6. **Sibling** The sampleID of the sibling ('.' indicates there was no sibling)
7. **ProbandSex** The sex of the proband
8. **SiblingSex** The sex of the sibling
9. **Exome_ThisManuscriptSsc** Whether the sample was included the SSC exome data in this manuscript
10. **Exome_ThisManuscriptSscAsc** Whether the sample was included the SSC and ASC exome data in this manuscript
11. **CNV_ThisManuscriptSsc** Whether the sample was included the SSC CNV data in this manuscript
12. **CNV_ThisManuscriptSscAgp** Whether the sample was included the SSC and AGP CNV data in this manuscript
13. **CNV_Levy2011** Whether the sample was included the Levy *et al.* 2011 SSC CNV analysis (Levy et al., 2011)
14. **CNV_Sanders2011** Whether the sample was included the Sanders *et al.* 2011 SSC CNV analysis (Sanders et al., 2011)
15. **CNV_Pinto2014** Whether the sample was included the Pinto *et al.* 2014 AGP CNV analysis (Pinto et al., 2014)
16. **Exome_Iossifov2012** Whether the sample was included the Iossifov *et al.* 2012 SSC exome analysis (Iossifov et al., 2012); these sample IDs were estimated from the downloaded data, rather than the manuscript
17. **Exome_Oroak2012** Whether the sample was included the O'Roak *et al.* 2012 SSC exome analysis (O'Roak et al., 2012)
18. **Exome_Neale2012** Whether the sample was included the Neale *et al.* 2012 ARRA exome analysis (Sanders et al., 2012)
19. **Exome_Sanders2012** Whether the sample was included the Sanders *et al.* 2012 SSC exome analysis (Sanders et al., 2012)
20. **Exome_Willsey2013** Whether the sample was included the Willsey *et al.* 2013 exome analysis (Willsey et al., 2013)
21. **Exome_Dong2014** Whether the sample was included the Dong *et al.* 2014 SSC exome analysis (Dong et al., 2014)
22. **Exome_Iossifov2014** Whether the sample was included the Iossifov *et al.* 2014 SSC exome analysis (Iossifov et al., 2014)

Table S2. List of *de novo* CNVs in cases and controls (.xlsx file). Related to Figure 2.

This table shows all the *de novo* CNVs detected in this manuscript along with those described in (Sebat et al., 2007), (Marshall et al., 2008), (Itsara et al., 2010), (Sanders et al., 2011), (Levy et al., 2011)), and (Pinto et al., 2014). There are two sheets (Cases and Controls) and the following columns are present:

1. **PatientID** The sample ID of the case or control
2. **Band** The chromosomal band of the CNV
3. **Chr** The chromosome of the CNV
4. **Start(hg18)** The start co-ordinates of the CNV in hg18
5. **Stop(hg18)** The stop co-ordinates of the CNV in hg18
6. **Start(hg19)** The start co-ordinates of the CNV in hg19
7. **Stop(hg19)** The stop co-ordinates of the CNV in hg19
8. **Size** The size of the CNV based on hg18 co-ordinates
9. **Del/Dup** Whether the CNV is a hemizygous deletion (HemiDel), heterozygous deletion (Del) or duplication (Dup). If the CNV is present in more than one child in the family it is marked '_Germline', if the CNV had a clear mosaic appearance in one parent it is marked '_Parent_Mosaic', if the CNV had a clear mosaic appearance in child it is marked '_Mosaic'.
10. **Mechanism** Whether the CNV location is most consistent with non-allelic homologous recombination (NAHR), homologous recombination (HR), non-homologous recombination (NHR), a micro homologous hotspot (Micro_HS) (Girirajan et al., 2013), Meiotic non-dysjunction, or Unknown.
11. **Simplex/Multiplex** Whether the family is simplex (only one affected individual), multiplex (multiple affected individuals), or unknown.
12. **Phenotype** Whether the sample has ASD or is Unaffected
13. **Gender** Whether the sample is Male, Female, or 'Not_stated'
14. **Parent-of-origin** Whether the CNV is present on the Father's allele, the Mother's allele, Unknown, or 'Not_stated'
15. **Study** The analysis or publication that identified the CNV. 'Sebat_2007' (Sebat et al., 2007), 'Marshall_2008' (Marshall et al., 2008), 'Itsara_2010' (Itsara et al., 2010), 'Levy_2011' (Levy et al., 2011)), 'Pinto_2014' (Pinto et al., 2014), 'This_study' (all CNVs identified in this analysis including those from (Sanders et al., 2011)), 'SSC_Below_pCNV_threshold' (*de novo* CNVs detected but not meeting the p_{CNV} threshold of 1x10^{-9}), 'SSC_not_2591' (*de novo* CNVs detected, but not in the 2,591 samples in the SSC or STC, see Table S1).
16. **dnCNV_SSC_(del/dup)** The maximum number of overlapping *de novo* CNVs in the SSC samples included in the analysis. In parentheses this number is shown for deletions and duplications separately. If the CNV has a population frequency greater than 0.1% it is marked 'Common'.
17. **dnCNV_SSCAGP_(del/dup)** The maximum number of overlapping *de novo* CNVs in the SSC and AGP samples included in the analysis. In parentheses this number is shown for deletions and duplications separately. If the CNV has a population frequency greater than 0.1% it is marked 'Common'.
18. **RefSeqGenes** The official gene symbol of coding RefSeq genes within the CNV interval sorted alphabetically
19. **RefSeqAccension** The NCBI accession number of coding RefSeq isoforms within the CNV interval
20. **RefSeqGeneCount** The number of unique coding RefSeq genes within the CNV interval,
21. **Location** Whether the CNV impacts Exons, Introns, UTRs, or is Intergenic
22. **Freq** The maximal population frequency of the estimate from the 5,182 SSC parents and the DGV (Macdonald et al., 2014); this estimate was used as a filtering criteria to distinguish rare ($\leq$0.1%) CNVs.

23. **Rare0.1** Whether the CNV met the rare criteria (Y/N)
24. **Validated** Whether the CNV has been validated (Yes, No_primer, Not_attempted, Unknown)
25. **SSC_P** The genome-wide corrected p-value for recurrence based on the SSC data only
26. **SSC_Q** The genome-wide corrected q-value/FDR for recurrence based on the SSC data only; this metric was used as a filtering criteria for Table 1 in the main manuscript.
27. **SSCAGP_P** The genome-wide corrected p-value for recurrence based on the SSC and AGP data
28. **SSCAGP_Q** The genome-wide corrected q-value/FDR for recurrence based on the SSC and AGP data; this metric was used as a filtering criteria for Table 2 in the main manuscript.
29. **Main_analysis** Whether the CNV was included in the main analysis of *de novo* CNVs detected in the SSC samples using the Illumina SNP genotyping data (SSC), or in the combined SSC and AGP analysis (SSC_Levy_Pinto), or not included in any of the analyses described in the paper (N).
30. **Trio/Quartet** Whether the sample is from an SSC Trio family (both parents and an ASD case), an SSC Quartet family (both parents, an ASD case, and an unaffected sibling), or neither (NA).

Table S3. List of rare CNVs in cases and controls (.xlsx file). Related to Figure 2.

This table shows all the rare CNVs detected in this manuscript. The following columns are present:

1. **familyID** The family ID of the sample
2. **patientID** The sample ID of the sample
3. **chipID** The Illumina chip ID of the sample
4. **Band** The chromosomal band of the CNV
5. **Chr** The chromosome of the CNV
6. **Start(hg18)** The start co-ordinates of the CNV in hg18
7. **Stop(hg18)** The stop co-ordinates of the CNV in hg18
8. **Start(hg19)** The start co-ordinates of the CNV in hg19
9. **Stop(hg19)** The stop co-ordinates of the CNV in hg19
10. **Size** The size of the CNV based on hg18 co-ordinates
11. **SNPs** The number of Illumina probes within the CNV
12. **Del/Dup** Whether the CNV is a heterozygous deletion (Del) or duplication (Dup).
13. **Mechanism** Whether the CNV location is most consistent with non-allelic homologous recombination (NAHR), an NAHR hotspot (NAHR_HS), homologous recombination (HR), non-homologous recombination (NHR), a mini homologous hotspot (Mini_HS), a micro homologous hotspot (Micro_HS), or an Alu hotspot (Alu_HS). Definitions from (Girirajan et al., 2013).
14. **Phenotype** Whether the sample is a 'Proband', 'Sibling', 'Father', or 'Mother'
15. **Gender** Whether the sample is Male or Female
16. **Inheritance** Whether the CNV is inherited from the 'Father', 'Mother', 'BothParents', 'DeNovo', or 'NotSeenInParent'. Parental CNVs are marked 'Parent'
17. **RefSeqGenes** The official gene symbol of coding RefSeq genes within the CNV interval sorted alphabetically
18. **RefSeqAccension** The NCBI accession number of coding RefSeq isoforms within the CNV interval
19. **RefSeqGeneCount** The number of unique coding RefSeq genes within the CNV interval
20. **Location** Whether the CNV impacts Exons, Introns, UTRs, or Intergenic
21. **pCNV** The p_{CNV} value estimated by CNVision
22. **Freq** The maximal population frequency of the estimate from the SSC and the DGV; this estimate was used as a filtering criteria to distinguish common ($>$0.1%) and rare ($\leq$0.1%) CNVs.
23. **Trio/Quartet** Whether the sample is from an SSC Trio family (both parents and an ASD case) or an SSC Quartet family (both parents, an ASD case, and an unaffected sibling).
24. **Phase** Whether the family was included in the prior analysis '1' (Sanders et al., 2011), or is new to this analysis '2'.
25. **DeNovoFile** Whether the CNV is listed in the *de novo* CNV table (Table S2)

Table S4. Estimates of *de novo* and inherited burden (.xlsx file). Related to Figure 2.

This table shows the burden estimate for CNVs and LoF between probands and siblings (sheets 'Mutations', 'Samples', and 'Genes') or males and females (sheets 'MutationSex', 'SampleSex', and 'GeneSex'). The columns are described below (in the later three sheets proband and sibling is replaced by female and male):

1. **Inheritance** Whether the variants were '*Denovo*' or 'RareInherited'
2. **Rarity** The population frequency filter used: '≤0.1%' or '≤1%'
3. **Type** The type of variant: 'Del', 'Dup', 'All CNV' (Del + Dup), or 'LoF'
4. **SampleSet** Whether the analysis is on 'All' samples, 'Phase1' (Sanders et al., 2011), or 'Phase2'; see Table S1
5. **Size** The size of variants considered: 'All', '≥1 gene', '1 gene', '2-3 genes', '4-10 genes', '11-20 genes', '>20 genes'
6. **WRST** The p-value from a one-sided Wilcoxon Signed Rank Test; paired in the first three sheets, unpaired in the sex comparison three sheets
7. **Sign** The p-value from a one-sided Sign; in the three sex comparison sheets this is replaced by a Fisher's exact test due to the different sample sizes
8. **NPro** The number of probands in the analysis
9. **sampPro** The estimate in the probands (totalPro / NPro)
10. **sdPro** The standard deviation of the estimate in the probands
11. **totalPro** The count in the probands
12. **sePro** The standard error of the estimate in the probands
13. **ciPro** The 95% confidence intervals of the estimate in the probands
14. **NSib** The number of siblings in the analysis
15. **sampSib** The estimate in the siblings (totalSib / NSib)
16. **sdSib** The standard deviation of the estimate in the siblings
17. **totalSib** The count in the siblings
18. **seSib** The standard error of the estimate in the siblings
19. **ciSib** The 95% confidence intervals of the estimate in the siblings
20. **Figure** The figure in which the estimate was used (if any)

Table S5. List of *de novo* exome mutations in cases and controls (.xlsx file). Related to Figure 4.

This table shows all exome mutations from SSC and ASC families included in three publications: (Iossifov et al., 2014), (Dong et al., 2014), and (De Rubeis et al., 2014). The following columns are present:

1. **patientID** The sample ID of the sample
2. **Chr** The chromosome of the mutation
3. **Pos(hg19)** The start co-ordinate of the mutation (VCF format)
4. **rsId** The rsID of the variant is present in dbSNP
5. **Ref** The reference allele of the mutation (VCF format)
6. **Alt** The alternate allele of the mutation (VCF format)
7. **Type** The type of mutation (SNV or Indel)
8. **Phenotype** If the sample is a case (2) or control (1)
9. **Sex** If the sample is a female (2) or male (1)
10. **Study** The publication in which the mutation was recently (but not necessarily originally) described: 'SSC_Iossifov' relates to (Iossifov et al., 2014), 'SSC_Dong' relates to (Dong et al., 2014), and 'ASC_DeRubeis' relates to (De Rubeis et al., 2014)
11. **Gene** The official gene symbol of the gene at which the mutation occurs (0 if none)
12. **Accession** The RefSeq accession ID of the gene at which the mutation occurs (0 if none)
13. **Effect** The predicted impact of the mutation on the gene, options are: 3UTR, 5UTR, InFrame, Intergenic, Intron, LoF_3splice, LoF_5splice, LoF_frameshift, LoF_nonsense, Missense, Near_3splice, Near_5splice, Promoter, Silent, StartCodonDisrupted, and StopCodonDisrupted
14. **AA_Change** The predicted amino acid change (original, amino acid number, modified)
15. **BP_Change** The nucleotide change (original, gene nucleotide number, modified)
16. **PPH2output** The output of Polyphen2 prediction for missense variants: benign_Mis1, possiblyDamaging_Mis2, probablyDamaging_Mis3; variants that are not annotated by PolyPhen2 are marked as '.'
17. **Confidence** The confidence of the prediction: validated (by PCR followed by Sanger Sequencing or MiSeq), highConf, mediumConf, and lowConf; all four categories were included in the analysis
18. **ESPVarFreq** The frequency of the variant in the 6,500 individuals in the ESP database `evs.gs.washington.edu/EVS/`)
19. **ESPHetFreq** The percentage of individuals with a heterozygous variant in the 6,500 individuals in the ESP database
20. **ESPHomoFreq** The percentage of individuals with a homozygous variant in the 6,500 individuals in the ESP database
21. **SSCBurden** Whether the sample was included in the SSC burden analysis (Figure 4 and Table S4): 'Y' or '.'

Table S6. List of genes with TADA FDR values (.xlsx file). Related to Figure 5.

This table shows the input data and TADA FDR/q-values for 18,665 RefSeq genes. 'mis3' is defined as missense variants that are estimated to be 'probably_damaging' by the PolyPhen2 algorithm (Adzhubei et al., 2010). The *de novo* exome data includes 4,036 ASD cases from the ASC and SSC (Iossifov et al., 2014) (De Rubeis et al., 2014). The case-control data includes 1,601 ASD cases and 5,397 population controls described in the ASC (Liu et al., 2013) (De Rubeis et al., 2014). The transmitted-non-transmitted data is from 1,445 trio families in the ASC (De Rubeis et al., 2014). The small *de novo* deletion data is from 2,591 ASD cases from the SSC (this manuscript, (Sanders et al., 2011), and (Levy et al., 2011)) and 2,096 ASD cases from the AGP (Pinto et al., 2014). The following columns are present:

1. **RefSeqGeneName** The official gene symbol based on RefSeq definitions.
2. **TadaGeneName** The official gene symbol copied from the original TADA table (He et al., 2013).
3. **ExomeMutRate** The mutation rate for *de novo* exome variants copied from the original TADA table (He et al., 2013).
4. **dnLoF** The number of *de novo* loss of function (LoF) mutations in 4,036 ASD cases.
5. **caseLoF** The number of LoF variants in 1,601 ASD cases.
6. **ctrlLoF** The number of LoF variants in 5,397 population controls.
7. **transLoF** The number of transmitted LoF variants in 1,445 ASD cases.
8. **ntransLoF** The number of non-transmitted LoF variants in 1,445 ASD cases.
9. **dnMis3** The number of *de novo* mis3 mutations in 4,036 ASD cases.
10. **caseMis3** The number of mis3 variants in 1,601 ASD cases.
11. **ctrlMis3** The number of mis3 variants in 5,397 population controls.
12. **transMis3** The number of transmitted mis3 variants in 1,445 ASD cases.
13. **ntransMis3** The number of non-transmitted mis3 variants in 1,445 ASD cases.
14. **SmallDelMutRate** The mutation rate for small *de novo* deletions, see section 25.
15. **dnSmallDel** The number of small *de novo* deletions in 4,687 ASD cases.
16. **tadaFdrAscSscExomeSscAgpSmallDel** TADA FDR for the combined analysis of exome data and small *de novo* deletions; this represents the most thorough identification of ASD genes in this manuscript.
17. **65genes_tadaFdrAscSscExomeSscAgpSmallDel** For the 65 genes with a FDR≤0.1 in the combined analysis, the dataset that enabled the specific ASD gene to meet this threshold is listed. Datasets were considered in the following order: SSC exome, SSC small *de novo* deletions (this manuscript), ASC exome, the combination of SSC and ASC exome data using the TADA metric, SSC and ASC exome combined with small *de novo* deletions from the SSC and AGP.
18. **tadaFdrAscSscExome** TADA FDR for the combined analysis of exome data only. This metric was used for Figure 6A in the main manuscript.
19. **59genes_tadaFdrAscSscExome** For the 59 genes with a FDR≤0.1 in the exome analysis, the dataset that enabled the specific ASD gene to meet this threshold is listed. Each gene could be discovered by: SSC only (SSC), ASC only (ASC), or both the SSC and ASC (SSC_ASC). Note that two genes (*RAPGEF4* and *CACNA2D3*) out of the 59 below the FDR≤0.1 threshold are not in the list of 65 (column 17) because the sample size increased while no further evidence of ASD genetic association was added.
20. **tadaFdrAscExome** TADA FDR for the ASC exome data only.
21. **tadaFdrSscExome** TADA FDR for the SSC exome data only.

Supplemental Experimental Procedures

Contents

1. The Simons Simplex Collection (SSC)

The Simons Simplex Collection (SSC) (Fischbach and Lord, 2010) was created to identify cases of sporadic ASD, which are more likely to involve a *de novo* mutation than multiplex families. Cases aged between 4 and 18 years were collected across 12 sites in the USA using standardized criteria, including: ASD in the proband, defined as clinical consensus informed by Autism Diagnostic Interview - Revised (ADI-R) (Lord et al., 1994) and the Autism Diagnostic Observation Schedule (ADOS)(Lord et al., 2000); a mental age of greater than 18 months in the proband; and the absence of ASD in either parent or the siblings, based on clinical consensus and a threshold on the Social Responsiveness Scale (SRS) (Constantino, 2002).

The total collection identifies 11,921 samples from 2,951 families and is split into four sub collections:

1. **Simons Simplex Collection (SSC)** 2,644 families that meet the full inclusion and exclusion criteria.
2. **Simons Twin Collection (STC)** 15 families that meet the full inclusion and exclusion criteria for the SSC and, in addition, the proband has a twin (4 monozygotic, 11 dizygotic).
3. **Simons Ancillary Collection (SSC)** 256 families that do not meet the full inclusion and exclusion criteria for the SSC, usually due to the diagnosis of ASD in another family member.
4. **Excluded** 36 families that are excluded from the collection, usually due to the proband not meeting the diagnostic criteria for ASD.

A blood sample was taken from each proband, both parents, and, if available, unaffected siblings. The blood samples were sent to the Rutgers University Cell and DNA Repository (RUCDR) where DNA was extracted and lymphoblast cell lines (LCL) were generated; DNA was also extracted from the LCLs. In addition, the samples were checked for fragile X syndrome mutations (an exclusion criteria). In the course of the collection the RUCDR instituted a genotyping test that used 96 SNPs with high fixation index values (F_{ST}, a measure of population structure with a high value for a SNP that is unique to specific populations) as a rapid test of sample identity and to confirm familial relationships. A total of 6,979 samples in the collection were assessed using this panel (59%).

The whole-blood (WB) and LCL DNA were sent to Yale University where the samples were analyzed using the Illumina SNP genotyping arrays (1Mv1, 1Mv3 Duo, or Omni2.5M) and Cold Spring Harbor Laboratory where they underwent analysis using a customized 2.1M NimbleGen hybridization array (Levy et al., 2011). The same samples were also analyzed using exome sequencing split between three sites: Yale University, University of Washington, and Cold Spring Harbor Laboratory.

1.1. Clinically ascertained siblings

Within the 2,100 unaffected siblings there are two children with DiGeorge's Syndrome (22q11.2 deletions; 14468.s1 and 13653.s1) and one child with Down's syndrome (Trisomy 21; 11824.s1). Discussion with the recruiting team reveled that at least one of the children with 22q11.2 deletions (14468.s1) and the child with trisomy 21 (11824.s1) were specifically recruited due to the presence of a known ASD associated genotype in an unaffected child and the potential scientific interest of this occurrence. However, this ascertainment deviates from the ideal of a cohort of simplex idiopathic ASD and poses a complication for genomic analyses.

Using the p_{CNV} metric (section 7) the trisomy 21 did not meet the detection threshold due to the per chromosome LRR normalization step in data processing (see seciton 7.1). Since the sample was excluded in the previous analysis (Sanders et al., 2011) and the CNV detection was performed equally across the probands and siblings in the cohort, the sub-threshold detection was used as justification to exclude this 'CNV' but include the family. Careful examination of the SNP genotyping data showed there were no further cases of whole chromosome aneuploidies on the autosomes. Aneuploidies of the sex chromosomes are described in section 3.4 and Figure S1.

Both the 22q11.2 deletions were detected in the dataset. The decision was taken to include both families and CNVs in the burden analysis, despite the slight bias against discovering ASD association that this would create (for context these deletions contain 45 RefSeq genes; the only other *de novo* CNV with more than

20 genes in the 2,100 siblings contains 23 genes). The clinically ascertained 22q11.2 deletion (14468.s1) was excluded from the analysis of CNV recurrence (section 15) and total CNV loci (section 16), since this recurrent CNV at a known ASD risk loci had a dramatic effect on the outcome of the analysis. In all other analyses, including burden, the CNV is included.

2. Illumina Infinium Bead Array

To identify single nucleotide polymorphism (SNP) genotypes, the Illumina Infinium beadarray uses a silicon wafer with over a million microwells, each of which analyzes a single SNP which can be identified by the 2D location on the wafer. Each microwell contains a 3μm silica bead that is covered with hundreds of thousands of identical 50-mer oligonucleotide hybridization probes that are specific to the genomic region immediately upstream or downstream of the particular SNP. This array is called a BeadChip.

To analyze a sample, 750ng of genomic whole-blood DNA was whole-genome amplified, fragmented, applied to the BeadChip, and allowed to hybridize overnight (16 hours). The oligonucleotide probes were designed so that the end of the probe is one nucleotide away from SNP to be analyzed; through a single nucleotide extension reaction, using fluorescently tagged nucleotides (A and T in red, C and G in green), the genotype of the SNP was revealed as a fluorescent color. Using the Illumina iScan or HiScan instrument, the red and green fluorescence of each bead was recorded by a high-resolution camera, these images were used to generated red and green '.idat' files that were used as the input for data analysis.

The '.idat' files were uploaded into Illumina's GenomeStudio, along with the hg18 SNP manifest ('.bpm'), which describes the location and identity of the SNPs on the BeadChip, and a cluster file ('.egt') which describes the expected distribution of data for normalized controls. The default cluster files provided by Illumina are based on 120 HapMap individuals, however, to reduce variation, a new cluster file was generated for each BeadChip type from 100 fathers and 100 mothers from the SSC with high quality data. Two values were then calculated: θ which is the ratio of red intensity to green intensity and R which is the sum of red intensity and green intensity.

Using the cluster file, the θ value was transformed into the B allele frequency (BAF) which represents the genotype so that 0 corresponds to a homozygous A allele (AA), 0.5 corresponds to a heterozygous A and B allele (AB), and 1 corresponds to a homozygous B allele (BB). The R value was transformed into the log R ratio (LRR) which represents the total quantity of DNA present; it is a ratio between the signal intensity from DNA in the sample and the signal intensity from DNA in the multiple controls in the cluster file, expressed as a logarithm in base 2 so that 0 represents normal copy number (2 copies), a positive value is observed with duplications ($\geq$3 copies), while a negative value is observed with deletions ($\leq$1 copy). The data generated from this analysis were printed out as a 'FinalReport' text file for each sample, that contains one line per SNP and includes the LRR and BAF data.

3. Confirming sample identity

Following this analysis each sample had genotypic data available from one of three SNP genotyping arrays (1M, 1Mv3, Omni2.5) and exome sequencing; furthermore 59% of samples had a 96-SNP genotypic identity panel generated as the sample was received at RUCDR. Sample collection is a complicated process that has the potential to introduce errors as the samples are handled. Broadly there were five main processing steps at which a sample error may occur:

1. Collection site
2. RUCDR
3. LCL generation
4. SNP genotyping array analysis
5. Whole-exome sequencing analysis

While methods have been developed to test the identity of SNP genotyping data(Purcell et al., 2007) there are no simple methods to check identity in exome sequencing data or between SNP genotyping and exome sequencing data. To perform a thorough check of sample identity in the SSC a novel tool was developed for cross-platform identity analysis and organization of genomic samples.

3.1. SNP identity barcode

The 96-SNP panel used by RUCDR gives a quick and reliable method of checking sample identity, however less than a third of the SNPs are represented on the Illumina SNP genotyping arrays and less than 10% are reliably tested by exome sequencing. Furthermore, the SNPs in this panel have a high F_{ST} making them ideal for determining ancestry, but sub-optimal for checking family relationships. We therefore selected SNPs that were well represented in both SNP genotyping and exome data, had a minor allele frequency (MAF) of at least 35%, and that were separated by at least 5 million nucleotides. In total 289 SNPs met these criteria.

Each SNP was assessed in genomic data to determine if it was heterozygous, recorded as 'C', or homozygous, recorded as 'A', and an identity barcode was generated by sorting these results by their genomic location in hg19. While this method does not make full use of the genomic data, for example 'AA' and 'TT' would both be recorded as homozygous, it ensures that the analysis is independent of the method used to record the SNP in the genotyping data, so that 'top strand' and 'forward strand' data will not alter the identity barcode. Furthermore with 289 SNPs there are 1 x 10^{86} possible combinations, substantially more than the 7 x 10^{9} humans, even after considering linkage disequilibrium and missing data.

An identity barcode was generated from SNP genotyping data by assessing the genotype calls within each FinalReport file; no calls were recorded in the barcode as an 'N' and these were not used for identity assessment. For exome data the identity barcode was generated from a BAM file; a SNP was identified as homozygous if $\geq$90% of reads showed the same nucleotide and heterozygous if between 25% and 75% of reads showed the nucleotide with the highest frequency at that position. Only reads with a nucleotide quality score $\geq$25 were considered. If neither frequency criteria was met, the read depth was $<$20, or $\geq$20% of nucleotides were recorded as 'N's in the BAM file then the SNP was recorded in the identity barcode as an 'N' which was not used for identity assessment.

The 289 SNPs used as an input were listed in hg18 and hg19 allowing both genome builds to be indexed. To facilitate rapid indexing the script was written to allow multi-threading. All assessed files were recorded in a common database file in FASTA format using the sample ID and file name as the sequence identifier. In addition, a separate database was generated using the 32 SNPs in common between the RUCDR 96-SNP panel and the Illumina SNP genotyping data to index FinalReports; for BAM files from exome data only 10 SNPs were shared, which is insufficient for identity analysis.

3.2. Assessing SNP barcodes

Two methods were developed to assess the 289-SNP identity barcodes between samples. The first method uses BLAT (Kent, 2002) to perform a rapid assessment between all the samples within an index, or between two indexes, and reports samples with $\geq$80% homology.

The second method assesses the exact number of matches between the barcodes of any two files and prints out a list of all samples with $\geq$85% homology. While this method is more accurate in the face of poor quality genomic data, it is considerably slower, especially with larger indexes. This slower method is useful for comparisons involving the 32 SNP index generated from RUCDR data or for confirming the presumed identity of samples.

3.3. Identity check using SNP genotyping data

While the SNP barcode allows rapid and simple identification of matching samples across differing data types, it is less suitable for determining family relationships. To make this assessment the SNP genotyping data were used to generate a plink file (Purcell et al., 2007), which was assessed against the expected pedigree file for Mendelian errors and identity by descent (IBD). Families that contained $\geq$1% of SNPs with Mendelian errors, or IBD metrics that did not match the expectation from the pedigree file, were examined manually to determine the cause of the error and if an alternative arrangement of samples within the family would resolve the errors. Furthermore samples were compared across families to identify duplicates or samples recorded in the wrong family.

To ensure that the modification of the pedigree file was correct, all samples were compared against the RUCDR 96-SNP panels. In families with errors that had not already had been analyzed with the RUCDR 96-SNP panels all family members were run through the SNP panel retrospectively.

3.4. Sex check using SNP genotyping data

To check that the genetic sex matched the recorded sex in the phenotype information from the collecting site, the percent of heterozygosity on the X chromosome was determined for all samples. Those with $\leq$10% heterozygosity were determined to be male, while samples with $>$10% heterozygosity were determined to be female. On examination of samples with discordant genotypic and recorded phenotypic sex it was noted that several samples had sex chromosome aneuploidies or runs of homozygosity on the X chromosome, therefore a second assessment of sex was performed by plotting the mean LRR for the X chromosome against the mean LRR for the Y chromosome (Figure S1). The raw LRR and BAF data were visualized for both sex chromosomes for all samples that were outliers on this analysis, or that had discordant genotypic and phenotypic sex. If a clear explanation was identified, such as sex chromosome aneuploidy, then the result was recorded and the sample retained. If a difference between genotypic and phenotypic sex remained then the sample was checked against the RUCDR 96-SNP panel which was run retrospectively if the data were not available. Eighteen samples had abnormalities of the sex chromosomes, while five samples had discordant genotypic and phenotypic sex. The five discordant samples were: 12217.p1, recorded female but genotypically male, trio family; 14472.s1, recorded male but genotypically female; 14472.s2, recorded female but genotypically male; 14487.s1, recorded male but genotypically female; and 14639.s1, recorded female but genotypically male. Since the RUCDR IDs were correct, and there was no evidence to suggest a proband and sibling has been switched, all five samples were retained.

3.5. Resolving sample identity

All samples with potential identity issues were assessed using the combination of data from the SNP barcode analysis, Mendel/IBD check, sex check, and RUCDR 96-SNP panel. If a clear explanation for the identity error could be identified and resolved, such as parents mislabeled at the genotyping facility, then the family was retained and the identity corrected. All families with identity issues that could not be resolved, for example non-paternity of the proband; were excluded.

The data from all samples in the families that passed the identity check were used to make a final barcode database of the SSC. This SSC barcode database was then used to check the identity of all the BAM files generated by exome sequencing.

4. SNP genotyping data quality

Having resolved the identity issues, the genotyping data were assessed for quality. A sample was excluded if it failed one or more of six measures:

1. **Beadstudio** Genotyping call rate of $\geq$98.5%
2. **Wide LRR** $\geq$3.5% of SNPs with a LRR between -0.5 and 0.5
3. **Extreme LRR** $\geq$0.5% of SNPs with a LRR $>$-1
4. **PennCNV** LRR standard deviation $>$0.28, BAF drift $>$0.01, Waviness factor (WF) deviating from 0 by $>$0.05
5. **QuantiSNP** For each sample four measures were assessed for each autosomal chromosome: BAF outliers $<$0.1, LRR outliers $<$0.1, BAF standard deviation $<$0.2, and LRR standard deviation $<$0.4. If 86 out of 88 measures per sample were outside these ranges then the sample failed quality control.
6. **GNOSIS** Quality score $>$10

4.1. Contaminated samples

In the course of checking samples that failed quality control it was noticed that a subset had unusual patterns of B allele frequency (BAF) across all chromosomes, consistent with four copies despite a LRR consistent with two copies. This pattern probably represents contamination, for example a mix of DNA from two samples. Having identified this pattern of genotyping error the entire dataset was assessed for contaminated samples by identifying outliers on two metrics: 1) the ratio of heterozygous and homozygous to total SNPs; and 2) the ratio of 'duplicate' SNPs (BAF of 0.25-0.4 or 0.6-0.75) to total SNPs. Samples with evidence of contamination were excluded from the analysis.

5. Identification of copy number variants (CNVs)

Copy number variants (CNVs) were detected using CNVision, as described previously (Sanders et al., 2011). The results of three CNV detection algorithms, PennCNV (Wang et al., 2007), QuantiSNP (Colella et al., 2007), and GNOSIS (Sanders et al., 2011), run with default parameters, were combined to form contiguous CNV loci. CNVision was updated to work with QuantiSNPv2.3 and to run in a UNIX environment to facilitate rapid data processing through parallelization.

6. Joining neighboring CNVs

CNV prediction algorithms frequently mis-classify large CNVs as multiple small neighboring CNVs. The extent of this breakage varies between algorithms, for example trisomy 21 (the duplication of all of chromosome 21) is broken into 4 sections by QuantiSNP, 121 sections by PennCNV, and 134 sections by GNOSIS. The use of multi-algorithm CNV prediction improves this problem considerably, however merging the results for trisomy 21 for these three algorithms still yields 4 sections. The breaks often occur at regions with poor data quality (*e.g.* high variance in LRR or uninformative SNPs with low MAF) or in regions with few SNPs (*e.g.* centromeres, segmental duplications). Such 'broken' CNVs interfere with interpretation since genes with variant copy number may be missed or a large CNV might be counted twice in a statistical analysis.

To resolve these artificial breaks an algorithm was developed to join CNVs that were separated by <50% of the number of SNPs in the larger of two neighboring CNVs, but only if the neighboring CNVs were on the same chromosome and not separated by a CNV with differing copy number. This rule was applied recurrently to the data until no further CNVs could be joined. To check the performance of this algorithm, 60 CNVs (20 from each of the three Illumina platforms) that had been joined were visualized. Of the 60 joined CNVs, 57 (95%) were correctly joined, while 3 were not, one due to a small intervening homozygous deletion, that had not been predicted, between two heterozygous deletions and two due to a small region with normal copy number. Since this represented a substantial improvement over the un-joined CNVs the algorithm was applied to the entire dataset.

7. p_{CNV} Calculation

While many methods exist to predict CNVs in SNP genotyping data, they are prone to large numbers of false positives and false negatives. These inaccuracies are amplified when trying to identify extremely rare variants such as *de novo* CNVs. Previously we had used predetermined thresholds of LRR and SNP count, alongside visualization of the raw genotyping data to improve the accuracy. However, these approaches risked missing important *de novo* events that did not meet the imposed thresholds and results that were dependent on the researcher who performed the visualization. Furthermore, visualization could not practically be applied to the entire dataset of inherited CNVs. To ameliorate these concerns a new statistical approach was developed to access the accuracy of each predicted CNV. This method estimates a p-value for the null-hypothesis that a there is no deviation from the expected distribution of data in the SNPs within a predicted CNV. This p-value is estimated by assessing the ratio of the likelihood of the observed deviation in LRR or BAF for each

SNP in a CNV in a region with two copies versus a region with one or three copies, depending the on type of CNV. The metric has been optimized to identify rare and potentially *de novo* CNVs; an overview of the methodology is shown in Figure S2.

SNP quality control: A SNP was excluded from p_{CNV} analysis if there was no measure of LRR or if the no call rate for that SNP was greater than or equal to the mean no call rate for all SNPs plus two standard deviations. SNPs with no genotype were included (if below this threshold) since they might represent true homozygous deletions. Filtering SNPs with high no call rates is likely to decrease the ability to detect common deletions at the expense of improving the detection of rare or *de novo* CNVs.

7.1. $p_{CNV,LRR}$ Calculation

LRR Metrics: Estimates of the expected LRR deviation in deletions and duplications were obtained by determining the median LRR for all SNPs within large, qPCR confirmed, autosomal CNVs in reclustered Illumina FinalReports. The values determined were -0.440 for deletions and 0.209 for duplications; both metrics were obtained from over 20,000 SNPs. A third metric was also determined: the number of SNPs with a LRR less than -1 on autosomal chromosomes (0.022%).

LRR Normalization: For each sample the mean LRR for that particular chromosome was measured. Though the mean deviation from 0 (expected) is small (*e.g.*-0.003 across chr1 for 4,748 samples on Illumina 1Mv3) the deviation per sample can be quite large (standard deviation 0.020 for same measure). To account for these differences, the LRR value for each SNP within a potential CNV was normalized by subtracting the observed LRR deviation in that sample across the same chromosome. This normalized LRR is denoted *y*. This normalization technique has the potential to obscure large CNVs, especially on smaller chromosomes, since the mean LRR would be expected to be somewhat deviated. However, in practice the p_{CNV} metric accumulates sufficient evidence by assessing every SNP in large CNVs that this is only a concern for chromosomal aneuploidy (*e.g.* trisomy 21) and in these cases the BAF assessment still accumulates sufficient evidence to give a low combined p_{CNV} metric.

Homozygous deletions: SNPs with a LRR <-1 were excluded from consideration of heterozygous deletions and duplications, since a highly negative SNP exerts disproportionate influence on these more subtle changes in copy number. However, the number of these low LRR SNPs was used to assess homozygous deletions by comparison with the observed rate of 0.022% of such SNPs in regions of normal copy number. A likelihood was estimated under two hypotheses:

1. Likelihood of no CNV (L_{null}):

$$L_{null} = (1 - 0.00022)^o.(0.00022)^u \tag{1}$$

2. Likelihood of a homozygous deletion ($L_{homodel}$):

$$L_{homodel} = (1 - 0.00022)^u.(0.00022)^o \tag{2}$$

Where: o is the number of SNPs with a LRR $\geq$-1; and u is the number of SNPs with a LRR <-1.
The likelihood ratio statistic, LR, is then calculated for a homozygous deletion:

1. Likelihood ratio for a homozygous deletion based on LRR ($LR_{homodel,LRR}$):

$$LR_{homodel,LRR} = 2(log(L_{homodel}) - log(L_{null})) \tag{3}$$

To obtained and estimate of $p_{homodel,LRR}$ the chi-squared distribution was used with j degrees of freedom, where j is the total number of SNPs passing quality control within the CNV.

Deletions, duplications, and no CNVs: For each SNP that passed the quality filters and had a LRR $\geq$-1, the likelihood of the observed, normalized LRR (*y*, see 'LRR Normalization' above) was calculated using the probability density function, under three hypotheses:

1. Likelihood of no CNV (L_{null}):

$$L_{null} = \phi\left(\frac{y - 0}{s}\right) + 0.01 \tag{4}$$

2. Likelihood of deletion CNV (L_{del}):

$$L_{del} = \phi\left(\frac{y - (-0.440)}{s}\right) + 0.01 \tag{5}$$

3. Likelihood of duplication CNV (L_{dup}):

$$L_{dup} = \phi\left(\frac{y - 0.209}{s}\right) + 0.01 \tag{6}$$

Where: ϕ is the probability density function; y is the normalized LRR; the number subtracted from y is the expected LRR under this hypothesis (see 'LRR Metrics' above); s is the LRR standard deviation for all SNPs on this chromosome in this sample; and 0.01 is added for stability, specifically to limit the impact a single SNP can carry within the estimate. The probability density function makes the assumption of a normal distribution of LRR values: analysis of random SNPs shows this is a reasonable assumption outside of regions with common CNVs.

The likelihood is computed for all SNPs within the CNV from 1 to j, where j is the total number of SNPs passing quality control within the CNV. The log values of the likelihood per SNP are summed to calculate the log likelihood:

1. log likelihood of no CNV (l_{null}):

$$l_{null} = \sum_{j} log\left(L_{null,j}\right) \tag{7}$$

2. log likelihood of deletion CNV (l_{del}):

$$l_{del} = \sum_{j} log\left(L_{del,j}\right) \tag{8}$$

3. log likelihood of duplication CNV (l_{dup}):

$$l_{dup} = \sum_{j} log\left(L_{dup,j}\right) \tag{9}$$

The likelihood ratio statistic LR is then calculated for a deletion and duplication separately:

1. Likelihood ratio for a deletion based on LRR ($LR_{del,LRR}$):

$$LR_{del,LRR} = 2log\left(l_{del} - l_{null}\right) \tag{10}$$

2. Likelihood ratio for a duplication based on LRR ($LR_{dup,LRR}$):

$$LR_{dup,LRR} = 2log\left(l_{dup} - l_{null}\right) \tag{11}$$

To obtained and estimate of $p_{del,LRR}$ the chi-squared distribution was used with j degrees of freedom, where j is the total number of SNPs passing quality control within the CNV. The same approach was used to estimate of $p_{dup,LRR}$.

7.2. $p_{CNV,BAF}$ Calculation

BAF Metrics: To assess the contribution of BAF to CNV detection, the BAF was used to measure the number of SNPs that appeared homozygous (≤0.1 or ≥0.9), heterozygous (≥0.42 and ≤ 0.58), consistent with a duplication ((≥0.25 and ≤0.4) or (≥0.6 and ≤0.75)), or 'other' (all other values). These thresholds were determined by examining the distribution of SNPs within deletions (homozygous values), duplications (duplication values), and regions of normal copy number (heterozygous values). To estimate the likelihood of observing these categories within different types of CNV, the following metrics were estimated from regions without predicted CNVs, regions with confirmed large deletions, and regions with confirmed large duplications:

1. Within regions without predicted CNVs:
   - Homozygous ($homo$): 73.7%
   - Heterozygous (het): 26.0%
   - Duplication (dup): 0.1%
2. Within deletions:
   - Homozygous ($homo$): 99.8%
   - Heterozygous (het) or duplication (dup): 0.2%
3. Within duplications:
   - Homozygous ($homo$): 62.7%
   - Heterozygous (het): 0.9%
   - Duplication (dup): 33.3%

Homozygous deletions: The calculation of $p_{CNV,BAF}$ is not performed for homozygous deletions.

Deletions: To estimate the likelihood of a deletion and the likelihood of no CNV the following equations were used:

1. Likelihood of no CNV based on BAF ($L_{null,BAF}$):

$$L_{null,BAF} = (0.261)^{het+dup}.(0.737)^{homo} \tag{12}$$

2. Likelihood of a deletion based on BAF ($L_{del,BAF}$):

$$L_{del,BAF} = (0.002)^{het+dup}.(0.998)^{homo} \tag{13}$$

These likelihoods were used to calculate the likelihood ratio statistic for a deletion, $LR_{del:BAF}$:

1. Likelihood ratio of a deletion based on BAF ($LR_{del,BAF}$):

$$LR_{del,BAF} = 2(log(L_{del,BAF}) - log(L_{null,BAF})) \tag{14}$$

Duplications: To estimate the likelihood of a duplication and the likelihood of no CNV the following equations were used:

1. Likelihood of no CNV based on BAF ($L_{null,BAF}$):

$$L_{null,BAF} = (0.260)^{het}.(0.737)^{homo}.(0.001)^{dup} \tag{15}$$

2. Likelihood of a duplication based on BAF ($L_{dup,BAF}$):

$$L_{dup,BAF} = (0.009)^{het}.(0.627)^{homo}.(0.333)^{dup} \tag{16}$$

These likelihoods were used to calculate the likelihood ratio statistic for a duplication, $LR_{dup:BAF}$:

1. Likelihood ratio of a duplication based on BAF ($LR_{dup,BAF}$):

$$LR_{dup,BAF} = 2(log(L_{dup,BAF}) - log(L_{null,BAF})) \tag{17}$$

7.3. Combining $p_{CNV,LRR}$ and $p_{CNV,BAF}$ to make p_{CNV}

The type of CNV is determined prior to considering BAF by choosing the lowest of the three values: $p_{homodel,LRR}$, $p_{del,LRR}$, and $p_{dup,LRR}$. The final estimate of p_{CNV} is determined as follows:

1. **Homozygous deletions:** Homozygous deletions are assessed on the basis on LRR values alone (*i.e.* the measure is unchanged by BAF). The final estimate of $p_{homodel,LRR}$ determined using the chi-squared distribution is selected.
2. **Deletions:** The measures of $LR_{del,LRR}$ and $LR_{del,BAF}$ are summed to give the combined likelihood ratio statistic LR_{del}. The chi-squared distribution is used to estimate p_{del} with $2j + 1$ degrees of freedom, where j is the total number of SNPs assessed within the CNV.
3. **Duplications:** The measures of $LR_{dup,LRR}$ and $LR_{dup,BAF}$ are summed to give a combined likelihood ratio statistic LR_{dup}. The chi-squared distribution is used to estimate p_{dup} with $2j + 2$ degrees of freedom, where j is the total number of SNPs assessed within the CNV.

8. Validating p_{CNV}

To test the efficacy of the p_{CNV} metric a comparison was performed against three independent data sets:

1. Deletions and duplications predicted in HapMap samples by the Genome Structural Variation (GSV) Consortium (Conrad et al., 2010)
2. Deletions predicted in HapMap samples by the 1000 Genomes (1KG) Project (Genomes Project, 2010)
3. Deletions and duplications predicted in 30 SSC trios on the Illumina 1Mv3 and replicated on the higher resolution Illumina Omni2.5
4. *De novo* deletions and duplications predicted in 30 SSC trios on the Illumina 1Mv3 and replicated on the higher resolution Illumina Omni2.5

8.1. p_{CNV} and GSV CNVs

Deletions and duplications predicted by the GSV consortium were obtained from Table S4 of the GSV paper (Conrad et al., 2010). Data were extracted for three independent HapMap samples for whom Illumina Omni2.5 data had been generated: NA12891, NA12892, NA19238. In total, there were 1,501 autosomal CNVs (1,190 deletions with a copy number of 1 and 311 duplications with a copy number of 3). Of these, 72 (69 deletions and 3 duplications) were detected by at least one algorithm in CNVision with the same copy number and at least 50% reciprocal overlap in predicted location and these were labeled as true CNVs. A further 100 autosomal CNVs (64 duplications and 36 deletions) showed no evidence of any overlap with GSV findings, these were labeled as false CNVs. To assess the performance of p_{CNV}, a receiver operating characteristic (ROC) curve was generated using four approaches to distinguishing real CNVs: p_{CNV}, combination of detection algorithms, number of SNPs, and size of CNV.

To generate the ROC curve the following definitions were used:

- **True positive** Omni2.5 CNV overlapping with GSV CNV of same copy number with 50% reciprocal overlap
- **False positive** Omni2.5 CNV not overlapping with any GSV CNV
- **True negative** Omni2.5 CNV not overlapping with any GSV CNV and not meeting the detection threshold
- **False negative** Omni2.5 CNV overlapping with GSV CNV of same copy number with 50% reciprocal overlap and not meeting the detection threshold

Figure S2A-D show the ROC curves generated for p_{CNV}, combination of CNV detection algorithms, number of SNPs, and size of CNV respectively. The best performance was obtained with p_{CNV} with an area under the curve (AUC) of 0.81, followed by the combination of CNV detection algorithms which had an AUC of 0.62. Both number of SNPs and size of CNV performed poorly.

While p_{CNV} demonstrated a good ability to distinguish real CNVs, there overall overlap between the CNVs detected on the Illumina Omni2.5 and the CNVs detected by the GSV, using a high density comparative genomic hybridization (CGH) array, was low (5%). The majority of CNVs that were not detected either had small numbers of probes or were in regions with a high population frequency of copy number variation leading to minimal changes in the LRR following normalization against 200 parents.

8.2. p_{CNV} and 1000 Genomes deletions

Deletions predicted by the 1000 Genomes project pilot paper (Genomes Project, 2010) were downloaded from the project FTP site: union.2010_06.deletions.genotypes.vcf.gz. Data were extracted for three independent HapMap samples for whom Illumina Omni2.5 data had been generated: NA12891, NA12892, NA19238. In total there were 4,824 validated autosomal deletions with a copy number of 1, however many of the calls were in close proximity (suggesting a single CNV fragmented into numerous loci) or were overlapping. Of autosomal deletions predicted by CNVision from the Illumina Omni2.5, 126 overlapped with a 1000 Genome deletion (373 of the 1000 Genome loci showed a degree of overlap with the Omni2.5 data) and these were labeled as true deletions. A further 28 autosomal deletions predicted by CNVision showed no overlap with the 1000 Genome deletions and these were labeled as false deletions. Using the same approach to making an ROC curve as described above for the GSV the p_{CNV} metric showed the best performance with an AUC of 0.76, followed by the combination of CNV detection algorithms with an AUC of 0.59 (Figure S2).

The total overlap between the Illumina Omni2.5 calls and the 1000 Genome deletions was 7.7%, again small deletions and common deletions were the main causes of a failure to detect a validated CNV.

8.3. p_{CNV} and replicate Illumina microarray data

We next assessed the performance of the p_{CNV} metric between technical replicates of 90 samples from 30 SSC trio families run on two different Illumina microarrays: the Illumina 1Mv3 and the Illumina Omni2.5. Following quality control, data were available from 57 of the 60 independent parental samples. 2,604 CNVs were predicted on the 1Mv3 array compared to 5,447 on the higher resolution Omni2.5 array. A CNV on the 1Mv3 array was considered 'confirmed' if there was over 50% reciprocal overlap with a call of the same copy number on the Omni2.5 array; 1,026 CNVs met this criteria, while 1,577 did not. An ROC curve was used to compare the performance of p_{CNV} to other thresholds; to generate the ROC curve the following definitions were used:

- **True positive** 1Mv3 CNV overlapping with Omni2.5 CNV of same copy number with 50% reciprocal overlap

- **False positive** 1Mv3 CNV not overlapping with any Omni2.5 CNV

- **True negative** 1Mv3 CNV not overlapping with any Omni2.5 CNV and not meeting the detection threshold

- **False negative** 1Mv3 CNV overlapping with Omni2.5 CNV of same copy number with 50% reciprocal overlap and not meeting the detection threshold

The AUC is higher for the p_{CNV} metric (0.79) than for the combination of algorithms (AUC=0.76). The results are shown in Figure S2.

8.4. p_{CNV} and previously identified de novo *CNVs*

A similar approach was used to assess the performance for 144 CNVs that were identified as potentially *de novo* on the 1Mv3. 27 of these were confirmed on the Omni2.5 and with qPCR to represent true positives. To generate an ROC the following definitions were used:

- **True positive** 1Mv3 *de novo* CNV overlapping with Omni2.5 *de novo* CNV of same copy number with 50% reciprocal overlap
- **False positive** 1Mv3 *de novo* CNV not overlapping with any Omni2.5 *de novo* CNV
- **True negative** 1Mv3 *de novo* CNV not overlapping with any Omni2.5 *de novo* CNV and not meeting the detection threshold
- **False negative** 1Mv3 *de novo* CNV overlapping with Omni2.5 *de novo* CNV of same copy number with 50% reciprocal overlap and not meeting the detection threshold

The AUC is higher for the p_{CNV} metric (0.99) than for the combination of algorithms (AUC=0.82). Furthermore, at a p_{CNV} threshold of $\leq$1x10^{-9} the method achieves 100% specificity (on this small sample set) with 80% sensitivity. This threshold was therefore used for the main analysis. The results are shown in Figure S2. The *de novo* CNVs that were missed were generally small and sparsely covered on the 1Mv3 array.

9. Estimation of CNV frequency

Previously, CNV population frequency had been determined by comparison with the Database of Genomic Variants (DGV) (Macdonald et al., 2014) and defining a CNV as being rare if $\leq$90% of the CNV's length overlapped a list of regions at <1% population frequency in the DGV. The high-resolution Omni2.5 data is poorly represented in this database, therefore we elected to add in data from the parental samples to estimate the population frequency of CNVs. The CNV frequency was defined as the smallest number of parent samples with CNVs overlapping $\geq$50% of the CNV length. This metric was calculated both from parental samples with the same Illumina chip type and from all parental samples, with the higher of the two frequency estimates being used. Deletions and duplications were estimated separately. For each CNV this new approach was compared to the previous DGV-based approach and the higher of the two frequencies was used.

10. Identification of *de novo* CNVs

To obtain a list of high confidence rare *de novo* CNVs the following filters were applied:

1. Autosomal (not chrX, chrY, or chrXY)
2. Not a homozygous deletion
3. No evidence of a corresponding CNV in either parent
4. pCNV $\leq$1x10^{-9}
5. Estimated population frequency $\leq$0.1%

10.1. Confirmation of de novo *CNVs*

Following the application of these filters, 180 putative *de novo* CNVs were identified in 4,691 samples. All 180 were submitted for confirmation using SYBR green qPCR as described previously (Sanders et al., 2011). All qPCR analysis was performed blinded to affected status. A *de novo* confirmation rate of 97% was observed (175 out of 180), which is higher than the confirmation rate observed previously with a *de novo* detection pipeline that relied on visualization of the raw data (Sanders et al., 2011).

10.2. Identification of mosaic CNVs

The raw LRR and BAF data were visualized for all 175 validated *de novo* CNVs to identify mosaicism in the parents or child based on the appearance of the BAF and, to a lesser extent, the LRR. For CNVs with a mosaic appearance in one or more family members, the qPCR data and visual appearance were considered together to give a consensus result for the presence of mosaicism (Table S2).

11. Determination of parent-of-origin for *de novo* CNVs

To determine the parent-of-origin (PoO) informative SNPs within the CNV were identified by using the BAF to identify alleles that are present in only one parent. The interpretation of the informative allele depends on the type of *de novo* CNV. Deletions are comparatively simple, since only one copy is present and this copy does not have the *de novo* CNV, therefore if the informative allele is present in the child then the parent with this allele is not the PoO. For *de novo* duplications the number of copies of the informative allele in the child need to be determined. If two of the three copies include the informative allele then the parent with this allele is the PoO, however if only one of the three copies includes the informative allele then the parent with the allele is not the PoO. Using this logic the PoO of origin could be determined for 92% of *de novo* CNVs (Table S2). The parent of origin did not differ from the 50:50 expectation.

12. Estimate of mechanism of CNV generation

CNVs are hypothesized to arise through a variety of mechanisms; since these different mechanisms may affect the pathogenic potential of CNVs we attempted to classify CNVs by the mechanism of causation. Many CNVs associated with specific syndromes arise as a result of non-allelic homologous recombination (NAHR), for example 22q11.2 microdeletion leading to DiGeorge's Syndrome. NAHR occurs due to a high degree of homology between two segmental duplications with the intervening unique region of DNA being prone to deletion or duplication. To identify these events a database of paired segmental duplication regions was downloaded from the UCSC genome browser (Bailey et al., 2002). This list was filtered to include only segmental duplications on the same chromosome and reordered so that the first region was at the 5' end of the chromosome. CNVs were then annotated to identify reciprocal overlap with these regions and a CNV was considered to be mediated by NAHR if there was ≥50% reciprocal overlap. CNVs were also annotated against 1,367 previously identified gene-rich regions that are surrounded by homologous sequence (Girirajan et al., 2013). These regions are classified by the size of the total region and the size of the homolgous regions: NAHR hotspots (120 events, 50kbp-5Mbp, ≥10kbp homology), mini hotspots (253 events, 1-100kbp, 1-10kbp homology), micro hotspots (410 events, 1-50kbp, ≥100bp homology), and *Alu* hotspots (584 events, <50kbp, ≥15kbp *Alu* elements). A CNV was considered to match one of these regions if there was ≥50% reciprocal overlap. All other CNVs were marked as non-NAHR. Where conflicts arose between multiple mechanisms the following hierarchy was used:

1. NAHR hotspot
2. NAHR
3. Mini hotspot
4. Micro hotspot
5. *Alu* hotspots
6. Non-NAHR

As an additional check all *de novo* CNVs were visualized alongside the segmental duplication data in UCSC genome browser. As a result of this visualization step two *de novo* CNVs (0.8%) were reassigned as being non-NAHR instead of NAHR. Three non-NAHR CNVs were reassigned to new categories of 'unbalanced translocation' or 'subtelomeric'.

13. Burden of *de novo* CNVs

13.1. Filtering criteria

Starting with the complete list of qPCR validated *de novo* CNVs identified using the Illumina SNP genotyping arrays in the 2,100 matched probands and designated (.s1) siblings (Table S2) the following filters were applied:

- *De novo* with a p_{CNV} $\leq$1x10^{-9} in the child
- Deletion or duplication only (not HemiDel)
- Not germline or mosaic in the child or parent
- Autosomal (i.e. not chromosomes X, Y, or M)
- Rare (population frequency $\leq$0.1%)

All of these criteria are included as variables in Table S2 and the *de novo* CNVs that fulfill these criteria are labeled as 'SSC' in the column 'Main_analysis'. For burden analyses only samples from quartet families were included and these are labeled as 'Quartet' in the column 'Trio/Quartet'.

13.2. New vs. old samples

Our previous analysis of 1,124 families included 872 quartets and 252 trios (Sanders et al., 2011). Of the 1,124 families, 34 (25 quartets and 9 trios) are no longer included in the SSC or STC, often because an ASD diagnosis was made in the sibling during follow up; these families were removed from the analysis and are no longer included in the 'Sanders et al. 2011' or 'Combined cohort' analysis in the text and Figure 2 of the main manuscript. The sibling data for a further 4 quartet families were excluded through a more rigorous quality control and identity process in this manuscript, these families are analyzed as trios in this manuscript. In total 29 quartet families that were included in the prior burden analysis (Sanders et al., 2011) are no longer treated as quartet families in the present analysis.

The addition of genotyping data for 31 designated siblings allowed 31 families previously analyzed as trios in the (Sanders et al., 2011) analysis to be analyzed as quartets in the analyses presented in the main manuscript. In total 31 quartet families that were not included in the prior burden analysis (Sanders et al., 2011) are now treated as quartet families in the present analysis. The difference between these additions and exclusions (31 - 29 = 2) explains the description of 874 quartet families in the 'Sanders et al. 2011' in the text and Figure 2 of the main manuscript.

These differences affect 3.4% of the quartet samples previously analyzed (Sanders et al., 2011) and have no substantive effect on the results and estimates of burden previously presented. The families affected by these changes can be identified in Table S1.

13.3. Estimation of de novo *CNV burden*

For each of the 2,100 probands and 2,100 siblings the number of *de novo* CNVs was recorded. There were three probands (14091.p1, 11435.p1, 13036.p1) and one sibling (14473.s1) with more than one *de novo* CNV, therefore the 2.3% of probands and 3.0% of siblings have two *de novo* CNVs and the estimates of CNV burden (CNVs per sample) and sample burden (samples with at least one CNV) are only marginally different. The CNV burden was selected due to the comparability with other categories of mutation and variation. The mean burden was estimated by dividing the sum of *de novo* CNVs in each phenotype group by the 2,100 samples assessed, 95% confidence intervals were estimated based on the standard deviation and sample size, and p-values were estimated with a one-sided Sign test. A one-sided test was chosen since a clear hypothesis of increased burden in probands was being tested. A complete list of the analyses performed is shown in Table

S4, sheet 'Mutations'; due to the high degree of interdependence between the tests, and a clear hypothesis going into the analysis, correction for multiple comparisons was not performed for this, or other burden tests.

Proband:sibling ratios were estimated by dividing the number of *de novo* CNVs observed in 2,100 probands by the number of *de novo* CNVs observed in 2,100 siblings. The 95% confidence intervals associated with this metric were estimated using bootstrapping. A ratio was used in preference to an odds ratio to allow comparability across variant classes (including rare inherited LoF which are present in almost all samples).

13.4. Estimation of de novo *CNV gene burden*

For each of the 2,100 probands and 2,100 siblings the number of RefSeq genes with each *de novo* CNVs were recorded. The mean gene burden was estimated by dividing the sum of genes within *de novo* CNVs in each phenotype group by the 2,100 samples assessed, 95% confidence intervals were estimated based on the standard deviation and sample size, and p-values were estimated using the Wilcoxon Signed Rank Test (WRST). The WRST was selected for this analysis instead of the Sign Test since it considers the variability of gene counts between samples; this was not used in assessing CNV burden (section 13.3) due to the large number of ties (i.e. samples with the same rank). A one-sided test was chosen since a clear hypothesis of increased gene burden in probands was being tested. A complete list of the analyses performed is shown in Table S4, sheet 'Genes'.

Proband:sibling ratios were estimated by dividing the number of genes within *de novo* CNVs observed in 2,100 probands by the number of genes within *de novo* CNVs observed in 2,100 siblings. The 95% confidence intervals associated with this metric were estimated using bootstrapping.

13.5. Estimation of de novo *CNV gene burden by CNV size*

A linear model was used to explore the relationship between phenotype, CNV size, and the number of genes within CNVs for all filtered *de novo* CNVs from trio and quartet families (Table S2). Size and gene number were correlated (gene count $\sim$ size; $R^2 = 0.34$; $\beta = 4.7\text{x}10^{-6}$; $p < 2\text{x}10^{-16}$). Adding phenotype to the model showed 6.3 more genes per *de novo* CNV in probands compared to sibling (gene count $\sim$ size $+$ phenotype; $R^2 = 0.36$; $p = 0.04$ for phenotype); this is represented by the higher y-intercept for the line in probands than siblings in Figure S4. No interaction was observed between size and phenotype therefore the slope does not differ between probands and siblings (gene count $\sim$ size * phenotype; $R^2 = 0.36$; $p = 0.61$ for interaction term).

13.6. Estimation of de novo *CNV burden by sex*

Using a similar approach to the proband/sibling analysis (section 13.3), the number of CNVs per sample were estimated for the 275 female probands and 1,825 male probands from the 2,100 SSC quartet families. The mean burden was estimated by dividing the sum of *de novo* CNVs in each sex by the total number of samples of the same sex, 95% confidence intervals were estimated based on the standard deviation and sample size, and p-values were estimated using a Fisher Exact Test. A one-sided test was chosen since a clear hypothesis of increased burden in probands was being tested. A complete list of the analyses performed is shown in table S4, sheet 'MutationSex'.

Proband:sibling ratios were estimated by dividing the number of *de novo* CNVs observed in 275 female probands by the number of *de novo* CNVs observed in 1,825 male probands. The 95% confidence intervals associated with this metric were estimated using bootstrapping.

13.7. Estimation of de novo *CNV gene burden by sex*

Using a similar approach to the proband/sibling analysis (section 13.4), the number of genes within CNVs was estimated for the 275 female probands and 1,825 male probands from the 2,100 SSC quartet families. The mean burden was estimated by dividing the sum of genes in each sex by the total number of samples of the same sex, 95% confidence intervals were estimated based on the standard deviation and sample size, and p-values were estimated using the Wilcoxon Signed Rank Test (WRST). A one-sided test was chosen since a clear hypothesis of increased burden in probands was being tested. A complete list of the analyses performed is shown in Table S4, sheet 'GeneSex'.

Proband:sibling ratios were estimated by dividing the number of genes within *de novo* CNVs observed in 275 female probands by the number of genes within *de novo* CNVs observed in 1,825 male probands. The 95% confidence intervals associated with this metric were estimated using bootstrapping.

14. Burden of rare inherited CNVs

14.1. Filtering criteria

Starting with the complete list of predicted CNVs identified using the Illumina SNP genotyping arrays in the 2,100 matched probands and designated (.s1) siblings (Table S1), the following filters were applied:

- p_{CNV} $\leq$1x10^{-4}
- CNV observed in at least one parent
- Deletion or duplication only (not HemiDel)
- Autosomal (i.e. not chromosomes X, Y, or M)
- Rare (population frequency $\leq$0.1%)

All of these criteria are included as variables in Table S3. For burden analyses only samples from quartet families were included and these are labeled as 'Q' in the column 'Trio/Quartet'.

14.2. Estimation of rare inherited CNV burden

For each of the 2,100 probands and 2,100 siblings the number of rare inherited CNVs was recorded. The mean burden was estimated by dividing the sum of rare inherited CNVs in each phenotype group by the 2,100 samples assessed, 95% confidence intervals were estimated based on the standard deviation and sample size, and p-values were estimated with a one-sided Wilcoxon Signed Rank Test (WRST). A one-sided test was chosen since a clear hypothesis of increased burden in probands was being tested. A complete list of the analyses performed is shown in table S4, sheet 'Mutations'; due to the high degree of interdependence between the tests, and a clear hypothesis going into the analysis, correction for multiple comparisons was not performed for this, or other burden tests.

Proband:sibling ratios were estimated by dividing the number of rare inherited CNVs observed in 2,100 probands by the number of rare inherited CNVs observed in 2,100 siblings. The 95% confidence intervals associated with this metric were estimated using bootstrapping.

14.3. Estimation of rare inherited CNV gene burden

For each of the 2,100 probands and 2,100 siblings the number of RefSeq genes with each rare inherited CNVs were recorded. The mean gene burden was estimated by dividing the sum of genes within rare inherited CNVs in each phenotype group by the 2,100 samples assessed, 95% confidence intervals were estimated based on the standard deviation and sample size, and p-values were estimated using the Wilcoxon Signed Rank Test (WRST). A one-sided test was chosen since a clear hypothesis of increased gene burden in probands was being tested. A complete list of the analyses performed is shown in table S4, sheet 'Genes'.

Proband:sibling ratios were estimated by dividing the number of genes within rare inherited CNVs observed in 2,100 probands by the number of genes within rare inherited CNVs observed in 2,100 siblings. The 95% confidence intervals associated with this metric were estimated using bootstrapping.

14.4. Estimation of rare inherited CNV burden by sex

Using a similar approach to the proband/sibling analysis (section 14.2), the number of rare inherited CNVs per sample were estimated for the 275 female probands and 1,825 male probands from the 2,100 SSC quartet families. The mean burden was estimated by dividing the sum of *de novo* CNVs in each sex by the total number of samples of the same sex, 95% confidence intervals were estimated based on the standard deviation and sample size, and p-values were estimated using a Fisher Exact Test. A one-sided test was chosen since a clear hypothesis of increased burden in probands was being tested. A complete list of the analyses performed is shown in Table S4, sheet 'MutationSex'.

Proband:sibling ratios were estimated by dividing the number of *de novo* CNVs observed in 275 female probands by the number of *de novo* CNVs observed in 1,825 male probands. The 95% confidence intervals associated with this metric were estimated using bootstrapping.

14.5. Estimation of rare inherited CNV burden by inheritance

The number of rare CNVs inherited from the father and mother were estimated for the 2,100 probands (Table S4). A p-value was estimate using a binomial test.

14.6. Estimation of rare inherited parental burden

The number of rare CNVs inherited in the father and mother were estimated for the 2,100 probands (Table S4). A p-value was estimate using a binomial test. Applying the criteria from the (Desachy et al., 2015) did not alter this result. If the CNV list in Table S4 is filtered for size ≥30kbp, SNPs ≥25 for deletions and ≥35 for duplications then 438 riCNVs are observed in fathers and 478 riCNVs in mothers (p = 0.12, binomal test)

15. Recurrent *de novo* CNVs in the SSC

To determine a genome-wide threshold at which multiple *de novo* CNVs constitute a significant finding we applied the statistical approach we described in detail previously (Sanders et al., 2011). We used the distribution of *de novo* CNVs in siblings to assess the expectation of observing two or more overlapping *de novo* CNVs. Therefore, the sibling *de novo* CNVs were used to estimate the total number of regions in which a *de novo* CNV would be expected (C) using the unseen species method (Bunge and Fitzpatrick, 1993) in which $C = c/u + g2 * d * (1 - u)/u$ where: c = the total number of distinct *de novo* CNVs; $c1$ = the number of *de novo* CNVs with no overlap; d = total number of *de novo* CNVs observed; g = the coefficient of variation of the fractions of CNVs of each type, which is assumed to be 1 due to the small number of observations, and $u = 1 - c1/d$.

Excluding one of the 22q11.2 sibling deletions (section 1.1), a single example of overlap was observed in the siblings with a deletion and duplication at 16p13.11. Based on the sibling results, $c1$ = 34, c = 35, and d = 36, giving an estimate of C = 1,242 regions. No overlaps within sibling deletions or sibling duplications were observed so a conservative estimate of 621 (1,242 / 2) was used.

A permutation test was then performed to assess the proband data. For each iteration the number of *de novo* CNVs in probands (d = 132) were randomly assigned across the 1,242 regions (predicted from the sibling data) with replacement. If two *de novo* CNVs were observed in the same region then the iteration recorded a 'Yes' for 2-hit overlap. Triple overlaps, quadruple overlaps, etc. were recorded in the same manner. A total of 1,000,000,000 iterations were run and the number of iterations with 2, 3, 4, ... n overlapping *de novo* CNVs was divided by the number of iterations to yield a p-value. The total number of overlaps at a specific threshold across iterations was divided by the number of iterations to yield a q-value/false discovery rate (FDR). This process was performed for deletions (d = 69) and duplications (d = 63) separately based on 621 regions and these values were used if the region of overlap contained only deletions or duplications (Sanders et al., 2011). The results are summarized as p-values and q-values in Table 1 in the main manuscript and listed for every *de novo* CNV in Table S2.

15.1. Recurrent de novo *CNVs in the SSC and AGP*

The permutation test was repeated for the 273 *de novo* CNVs (153 deletions, 120 duplications) identified in the SSC ((Sanders et al., 2011), (Levy et al., 2011), and this analysis) or AGP cohorts ((Pinto et al., 2014)) listed in Table S2 using the same estimates of expected number of regions. The results are summarized as p-values and q-values in Table 2 in the main manuscript and listed for every *de novo* CNV in Table S2.

16. Estimate of total number of ASD risk loci for *de novo* CNVs

The unseen species problem was also applied to estimate the total number of loci at which *de novo* CNVs mediate ASD risk using the approached we described previously (Sanders et al., 2011). Based on the estimate of *de novo* CNV rate in siblings (0.016 per sample, Table S4), we estimate that we would expect 42 *de novo* CNV that do not mediate ASD risk to be present in 2,591 probands. Furthermore, these 42 non-ASD associated loci are likely to be the ones that do not overlap with other *de novo* CNVs in probands. Since 132 *de novo* CNVs were identified in SSC probands we would predict that 90 (132 - 42) mediate ASD risk. If we remove 42 non-overlapping (singleton) *de novo* CNVs from the proband list we obtain the following estimates: $c1$ = 25, c = 42, and d = 90, giving an estimate of C = 93 ASD risk loci. Repeating this for deletions, duplications, and with the inclusion of the AGP data gives the following results:

- SSC All: $c1$ = 25, c = 42, d = 90: Total risk loci (C) = 93
- SSC Deletions: $c1$ = 15, c = 26, d = 46: Total risk loci (C) = 61
- SSC Duplications: $c1$ = 7, c = 24, d = 45: Total risk loci (C) = 37
- SSC and AGP All: $c1$ = 66, c = 96, d = 193: Total risk loci (C) = 246
- SSC and AGP Deletions: $c1$ = 44, c = 64, d = 109: Total risk loci (C) = 181
- SSC and AGP Duplications: $c1$ = 40, c = 48, d = 84: Total risk loci (C) = 168

17. Phenotypic factors associated with *de novo* mutations

17.1. Non-verbal IQ and sex by de novo *status*

The 2,346 probands for whom CNV, exome, and NVIQ data were available were split by sex. Within each sex the probands were split into four groups: 1) No dnLoF or dnCNV; 2) At least one *de novo* deletion; 3) At least one *de novo* duplication; and 4) At least one dnLoF. Where a proband had a *de novo* mutation in more than one category they were included multiple times (this scenario included less than 40 probands). The difference in NVIQ between category 1 (no *de novo*) and each of the other categories was assessed using a generalized linear model as a predictor of *de novo* status. The results are shown in Figure 3A of the main manuscript.

17.2. Rate of de novo *mutations according to non-verbal IQ and sex*

The same 2,346 probands were split into six bins according to NVIQ. In each bin the percent of male and female probands with a dnLoF or dnCNV was calculated. A Fisher's exact test was used to compare the percentage of unaffected siblings with a dnLoF or dnCNV (10.7%) to the percentage of probands with a dnLoF or dnCNV. The analysis was repeated using on dnLoF and dnCNV in loci with a false discovery rate $\leq$0.1 (Table 2 and 4 in the main manuscript and Table S6). The results are shown in Figures 3B and 3C in the main manuscript.

17.3. Rate of de novo *mutations according to phenotype*

The rate of dnLoF and dnCNV was compared between groups of probands split by a variety of phenotypic factors. For each factor the groups of probands were compared using a Fisher's exact test. The head size z-score was taken from previously estimate the genetic deviation after correction for ethnicity, body size, and sex (Chaste et al., 2013). Factors with significant associations with *de novo* rate the results are shown in Figures 3D and 3E in the main manuscript. Other factors are shown in Figure S7.

18. Fraction of proband *de novo* mutations mediating ASD risk

To estimate the percentage of dnCNVs and dnLoF mutations that contribute risk we calculated the number of mutations per sample ($pMut$ in probands and $sMut$ in siblings) for each class of dnCNV (deletion/duplication) and dnLof (nonsense, splice-site, frameshift), see Table 3 and Table S4. Based on the assumption that none of the mutations observed in siblings mediate risk, the percentage of mutations mediating ASD risk in the probands ($pRisk$) was calculated by:

$$pRisk = \frac{(pMut - sMut)}{pMut}$$

The confidence interval was obtained using bootstrapping (Table 3).

19. Fraction of ASD cases with a *de novo* mutation mediating ASD risk

To estimate the percentage of probands in whom a dnCNV or dnLoF mutation contributes ASD risk we calculated the percentage of samples ($pSamp$ in probands and $sSamp$ in siblings) with at least one dnLoF or dnCNV, see Table 3 and Table S4. Based on the assumption that the mutations in siblings represent population variation that is not mediating ASD risk, the percentage of probands in whom a mutation is mediating ASD risk ($pRiskMut$) was calculated by:

$$pRiskMut = pSamp - sSamp$$

Of note, as the percentage of siblings with a mutation increases, the accuracy of this method to distinguish risk and non-risk mutations decreases. At an extreme example, if all siblings carried a *de novo* missense mutation then $sSamp$ would equal one so that $pRiskMut$ cannot be greater than 0. Furthermore, this highlights the risk of simply adding these estimates across categories which has the potential to greatly inflate the estimates derived. For example, if the deletion, duplication, nonsense, splice-site, and frameshift categories in Table three are added an estimate of 11.5% of probands carrying a risk mutation could be claimed, however making the estimate based on the combined category of all dnLoF and dnCNVs gives an estimate of 10.5%. The confidence interval in Table 3 was obtained using bootstrapping.

20. Burden of variants by size

Lists of samples and mutations were obtained for four categories of variants:

1. ***de novo* LoF** This list obtained by filtering Table S5 to mutations that had an LoF effect in samples in 1,911 quartet families in the SSC
2. ***de novo* CNVs** This list obtained by filtering Table S2 to CNVs included in the main analysis for 2,100 quartet families
3. **Rare inherited LoF** This list obtained by filtering LoF variants with a population frequency ≤0.1 for 1,911 SSC quartet families

4. **Rare inherited CNVs** This list obtained by filtering Table S3 to CNVs included in the main analysis for 2,100 quartet families. Only CNVs in which clear inheritance was observed from one of more parents were included. The p_{CNV} metric was estimated from the corresponding sibling at the same genomic co-ordinates to further refine the estimate of whether the CNV was inherited by one or both children. A p_{CNV} threshold was used since the insistence on evidence of inheritance increased the accuracy of the call.

Both sets of CNVs were divided into five bins based on the number of genes within the CNV. The ranges of the bins were selected to give approximately equal numbers of *de novo* CNVs in siblings (Figure 4).

20.1. Estimate of burden by variant size

For each *de novo* category the number of mutations in probands and siblings was calculated and this was divided by the number of families in the analysis to give an estimate of the number of mutations per sample. Confidence intervals were estimated using the standard deviation and sample size and p-values were estimated using a one-sided Sign test. The burden estimates are shown in Figure 4 and Table S4. For each category of mutation the rate of mutations in probands was divided by that in siblings to get the proband:sibling ratio. The confidence intervals of these estimates were obtained using bootstrapping.

For each rare inherited category the variants were split into three groups: those inherited in the proband only, those inherited in the sibling only, and those inherited by both children. Since ASD-association cannot be assessed through burden analysis of the variants inherited by both children these are excluded from the statistical analysis. The rate of variants was compared between the two remaining categories (proband only vs. sibling only) and the difference was assessed using a Wilcoxon Signed Rank Test (WRST). Confidence intervals were estimated using the standard deviation and sample size. These burden estimates are shown in Figure 4 and Table S4. For each category of inherited variants the rate of variants in probands was divided by that in siblings to get the proband:sibling ratio. The confidence intervals of these estimates were obtained using bootstrapping.

21. Intersection of *de novo* exome mutations and *de novo* CNV

A list of coding RefSeq genes from *de novo* CNVs in all 2,591 SSC probands (Table S2) was compared to a list of all 18,621 coding RefSeq genes to generate a list of genes within CNVs (1,610) and outside CNVs (17,011). These two gene lists were compared to exome mutations in all 2,508 SSC probands (Table S5). For each category (inside CNVs, outside CNVs) and exome mutation (LoF, missense, silent) the rate of mutations per gene was estimated. This analysis was repeated for four subsets of the *de novo* CNVs:

1. **Small *de novo* deletions** Deletion CNVs with $\leq$7 genes
2. **Small *de novo* duplications** Duplications CNVs with $\leq$7 genes
3. **Large *de novo* deletions** Deletion CNVs with $>$7 genes
4. **Large *de novo* duplications** Duplications CNVs with $>$7 genes

Since GC content and gene length increase the probability of a gene being targeted by both *de novo* CNVs and *de novo* exome mutations, the number of intersecting genes was corrected for the exome mutability (Table S6) and rounded to the nearest whole number. This correction had the effect of reducing the estimate of overlap in small *de novo* deletions by about a factor of two, but had minimal effect on the other categories (about 5% reduction in overlap). The difference between categories was assessed using the Fisher Exact Test. The results are shown in Figure S8.

The analysis was repeated using all the *de novo* CNVs from 4,687 SSC and AGP probands (Table S2) and the exome mutations from all 4,109 SSC and ASC probands (Table S5). The results are shown in Figure 5 in the main manuscript.

22. Enrichment of *de novo* exome mutations

The genes targeted by *de novo* LoF and missense variants from the SSC, ASC, and Deciphering Developmental Disorders (DDD) cohorts. (S5, (Deciphering Developmental Disorders, 2015)) were compared to genes targeted by *de novo* CNVs in the SSC and AGP (Table S2). Based on the number of genes in each list (CNV and exome) the expected number of overlapping genes was estimated. Enrichment was calculated as the number of observed matches over the number of expected matches. The significance of this finding was estimated using a hypergeometric test. The results are shown in Figure 5.

22.1. Variation of gene threshold

The enrichment analysis above (section 22) was repeated, but with the number of genes defining small and large varying between 1 and 10 (7 was used in the main manuscript since it is the median number of genes with dnCNVs from the SSC). The results show that the observed enrichment of dnLoF genes within small *de novo* deletions is robust across a range of CNV size thresholds.

22.2. Enrichment of ASD-associated genesets

To assess whether constrained genes (Petrovski et al., 2013) (Samocha et al., 2014), FMRP targets (Darnell et al., 2011), or CHD8 targets (Cotney et al., 2015) (Sugathan et al., 2014) the gene lists were considered as predictors of the presence of a small *de novo* deletion CNV including the same gene using logistic regression and a generalized linear model. To account for mutability as a co-variate the model of small *de novo* mutability built in section 25 was included, along with brain-expression (Sanders et al., 2012) (Kang et al., 2011).

23. Observed vs. Expected TADA values for CNVs

TADA values based on the combination of exome data from the SSC and ASC were used to identify the observed values in the genes within small and large *de novo* CNVs (Table S6). To estimate the expected TADA values of the equivalent number of genes a model of per gene mutation rate in CNVs was created by permuting CNVs based on their size and SNP content.

For each CNV a permuted CNV was generated by selecting a SNP at random from the Illumina Omni2.5 array and randomly selecting a direction (5' or 3' prime). The permuted CNV was extended from the initial SNP in the direction chosen until the same number of SNPs were selected as in the original sibling CNV. If the permuted CNV extended onto a different chromosome, or was 25% larger or 25% smaller in basepairs than the original CNV then it was rejected and the process was repeated until a suitable permuted CNV was generated.

Each iteration produced a permuted *de novo* CNV equivalent to the input CNVs. 100,000 permutations were run and the resulting CNVs were annotated against RefSeq. The number of permuted CNVs overlapping at least one exon was calculated for each gene and this number was divided by the total number of genes intersected by a CNV to get the per gene mutation rate.

For each gene observed in a proband CNV, one gene was selected at random, based on the mutation rate. For each gene the corresponding TADA FDR score was obtained and the TADA FDR scores for the equivalent number of observed genes were ranked from low TADA FDR to high TADA FDR. The entire permutation test was repeated 100 times.

The observed genes were also ranked from low TADA FDR to high TADA FDR. For the 1st ranked observed genes the median of the 100 permuted 1st ranked genes was estimated and this was used as the expected TADA score in Figure 6.

Repeating this method using the mutation rate calculated for small *de novo* deletions (section 25), or the mutation rate used as the input for TADA (Table S6) had minimal effect on the resulting plot.

24. Factors influencing mutation rate of small *de novo* deletions

To add the evidence of ASD association from small *de novo* deletions to TADA (He et al., 2013) it was necessary to estimate the per gene mutation rate based on the distribution in siblings. While many of the factors that influence the rate of single nucleotide variants (SNVs) are well recognized, including the length of the coding exons and GC content (Samocha et al., 2014), the factors influencing small *de novo* deletions are less well characterized.

We therefore assessed how eight factors were associated with the rate of small *de novo* deletions in siblings (Figure S10) and how these factors related to each other (Figure S11). Genes within small *de novo* deletions in siblings had a larger transcript size (including introns) and contained more SNPs on the Illumina Omni2.5; the larger transcript size remained significant if UTRs were excluded. Nominal association was observed with the gene size based on the length of the coding regions only. No enrichment was observed with the degree of overlap with segmental duplications or repetitive regions of DNA. The method of permuting CNVs based on SNPs, described in section 23, was also considered in the model using both sibling deletions and sibling deletions and duplications as the input for the permutations. While this method showed significant association with the genes within sibling deletions, the association was weaker than using gene length or number of SNPs alone. The measure was highly correlated with both of these factors (Figure S11).

25. Estimating the mutation rate of small *de novo* deletions

To create a model of the mutation rate of genes, the six significant factors out of the eight assessed were ranked by p-value (smallest to largest) followed by R^2 (largest to smallest) if the p-values were identical. The order of factors was therefore:

1. Length of gene transcript
2. Number of Illumina Omni2.5 SNPs within the gene transcript
3. Length of gene transcript excluding UTRs
4. Permuted sibling small *de novo* deletions
5. Permuted sibling small *de novo* deletions and duplications
6. Length of coding exons in the gene

The factors were added sequentially to a linear regression model until the additional factor was no longer significant. This occurred after adding the 'Length of gene transcript excluding UTRs' therefore only the first two factors (transcript length and Illumina SNPs) were selected. Given the high degree of correlation between transcript length with an without UTRs the same approach was tried, but excluding the 'Length of gene transcript excluding UTRs' factor; the addition of the next factor remained non-significant.

To further investigate the relationship between transcript length, the number of Illumina SNPs, and small *de novo* deletions the interaction between transcript length and number of Illumina SNPs was added to the model and found to be significant. Therefore, the model used for estimating per gene mutability was based on: transcript length, the number of Illumina SNPs within the transcript, and the interaction of these two values.

This model was applied to all genes in the genome. To convert this to a per gene mutability rate the minimum value estimated from the model was added to all genes so that every gene had a positive number. This number was then normalized based on the rate of genes within small *de novo* deletions in siblings.

25.1. Limitations of the mutation rate

Small *de novo* deletions are rare events in siblings. In the 2,100 siblings assessed within the SSC only 22 deletions were observed with 12 genes within them (11 of the 22 deletions do not overlap exons, 10 deletions overlpa exons of one gene, one deletions overlaps exons of two genes). Adding small *de novo* duplications allowed an additional 12 CNVs and 28 genes to be considered. Unlike silent *de novo* SNVs, there is no class of frequent neutral mutations to use to increase the power of this estimate calculation.

Following the logic that similar factors will govern the location of small *de novo* deletions in probands (despite the ascertainment bias of ASD association), the analysis was repeated with the addition of 116 small *de novo* proband deletions which overlap exons in 150 genes. The same six factors were found to be associated with genes within deletions (Figure S12), however the order was slightly different:

1. Number of Illumina Omni2.5 SNPs within the gene transcript
2. Length of gene transcript excluding UTRs
3. Length of gene transcript
4. Permuted sibling small *de novo* deletions
5. Permuted sibling small *de novo* deletions and duplications
6. Length of coding exons in the gene

A model of per gene mutability was built using the same methods as for the sibling deletions alone resulting in a model based on the top two factors (Number of SNPs and Gene size excluding UTRs) and their interaction. The resulting model was highly correlated with the first model ($R^2 = 0.86$), supporting the validity of the original model despite the small number of observations used as an input.

26. Modifying the TADA metric to include small *de novo* deletions

TADA was designed to analyze exome sequencing data, including single nucleotide variants (SNVs) and insertion/deletions (indels). The results are expressed as Bayes factors (BFs), one for each gene. To extend the TADA model, we developed a new method to include evidence of genetic association from genes within small *de novo* deletions. The resulting approach accounts for the presence of multiple genes within a CNV and expresses the results as a BF for inclusion in the original TADA model. In the new model, two BFs are estimated for each gene, one from the exome data and the other from small *de novo* deletion data. To obtain a single BF these two BFs are multiplied and the resulting BF is transformed into a false discovery rate (FDR/q-value) for each gene. Under this strategy, the analysis of CNV data is separate from SNV data so that the input data can come from different subjects (in a similar manner to the combination of *de novo*, transmitted, and case-control data in the original model.

26.1. Overview of the method

The input data for this addition to the TADA analysis will be small *de novo* deletion CNVs, spanning no more than 7 genes, from ASD cases. To obtain a BF for each gene we will perform the following steps:

1. Identify all CNVs in which this gene occurs.
2. For each identified CNV, assess the likelihood of contributing risk. This assessment uses the same logic as the original TADA model, effectively treating each CNV as a 'gene', and using the mutation rate to estimate a BF per CNV. A BF greater than 1 supports a role in ASD risk for the CNV, while a BF less than 1 is evidence against ASD risk.
3. To account for the presence of multiple genes within a CNV, the BF must be 'discounted' to account for the other genes present, before the BF is assigned to each gene in the CNV and used in the TADA model.
4. For each gene, the BF from the CNV component is multiplied by the BF from the exome data to obtain a final BF.

26.2. Computing the Bayes Factor for each CNV

The estimation of BF for each CNV uses the same logic as the original TADA model. This estimation requires two input metrics: 1) The mutation rate of the *de novo* CNV; and 2) The prior distribution of relative risk (RR), particularly the prior mean RR for small *de novo* deletions. For the per CNV mutation rate the per gene estimate of mutation rate within small *de novo* deletions (see section 25) was summed across all the genes within the CNV to estimate the per CNV mutation rate. We estimated the prior mean RR from the burden of small *de novo* deletions in the SSC dataset (Figure 2 and 4 in the main manuscript). For *de novo* CNVs, the burden (λ) is defined as the ratio of the number of CNVs in the 2,100 affected cases to the number of CNVs in the 2,100 unaffected siblings, estimated to be 3.0 for small *de novo* deletions in the SSC.

Based on an estimate that heterozygous disruption of 1,000 of the 18,665 genes in the genome contributes to ASD risk (Sanders et al., 2012) (He et al., 2013), we can estimate that 0.054 genes selected at random are ASD risk genes (1,000 / 18,665) and we will call this number π. The mean number of genes in a small *de novo* deletion is 2.7, therefore the expected number of ASD risk genes in a small *de novo* deletion under the null hypothesis is 0.14 (0.054 * 2.7) and we will call this number π_{CNV}. We can then estimate the average relative risk (prior mean RR: γ), using the following equation (see the Supplementary Methods of the original TADA paper (He et al., 2013) for derivation):

$$\pi_{CNV}(\gamma - 1) = \lambda - 1$$

Therefore:

$$\gamma = \frac{\lambda - 1}{\pi_{CNV}} + 1$$

Where π_{CNV} is the expected fraction of risk genes within the CNV, γ the prior mean RR, and λ the burden of small *de novo* deletions. So, in small *de novo* deletions in ASD:

$$\gamma = \frac{3.0 - 1}{0.14} + 1 = 15.3$$

This estimate of γ is used to estimate the Bayes Factor of the CNV, B_{CNV}, using the original TADA model (He et al., 2013), effectively treating each CNV as a 'gene'. Therefore B_{CNV} is estimated as function of x_{CNV}, which represents the number of observed small *de novo* deletions, the per CNV mutation rate, and π_{CNV}, which we estimate to be 0.14.

26.3. Computing the Bayes Factor of a gene within a CNV

Having estimated the BF for a CNV (B_{CNV}), our next goal is to estimate how much support this CNV provides to each gene contained within it. We assume that there is, at most, one risk gene in this CNV (an assumption based on the data presented in the main manuscript and the absence of any small *de novo* deletion CNVs with more than one gene with a low TADA FDR). Let H_0 be the hypothesis that none of the genes mediate risk, and H_k be the hypothesis that the k-th gene does mediate risk. The posterior probability of H_k is given by:

$$P(H_k|x_{CNV}) = \frac{P(x_{CNV}|H_k)P(H_k)}{P(x_{CNV}|H_0)P(H_0) + \sum_k P(x_{CNV}|H_k)P(H_k)}$$

Under our assumption of a single risk gene per CNV, the CNV mediates ASD risk if, and only if, one of its member genes mediates risk. Therefore, H_k is equivalent to the model where the CNV is causal, while H_0 is equivalent to the model where the CNV is non-causal. Thus the probability ratio of H_k vs. H_0 is BF of the CNV (B_{CNV}):

$$B_{CNV} = \frac{P(x_{CNV}|H_k)}{P(x_{CNV}|H_0)}$$

We also assume an equal prior probability for each gene (π):

$$P(H_k) = \pi$$

Based on an estimate of $\pi = 0.06$ in autism (see above), the equation above to estimate the posterior probability, $P(H_k|x_{CNV})$, simplifies to:

$$P(H_k|x_{CNV}) = \frac{\pi B_{CNV}}{1 - K\pi + \pi B_{CNV}}$$

Where K is the number of genes in this CNV. Next we obtain the BF of the k-th gene, B_k, that corresponds to this posterior probability ($P(H_k|x_{CNV})$). To do that, we solve this equation that expresses the posterior of the gene in two ways:

$$\frac{\pi B_k}{1 - \pi + \pi B_k} = P(H_k|x_{CNV}) = \frac{\pi B_{CNV}}{1 - K\pi + \pi B_{CNV}}$$

The result is a function of the CNV-level BF (B_{CNV}) that depends on K and π:

$$B_k = \frac{(1-\pi)B_{CNV}}{1 - K\pi + (K-1)\pi B_{CNV}}$$

We can better understand the effect of this function via some simple examples:

1. For a small *de novo* deletion that contains a single gene ($K = 1$), the per gene BF equals the per CNV BF ($B_k = B_{CNV}$), as expected:

$$B_k = \frac{(1-\pi)B_{CNV}}{1-\pi} = B_{CNV}$$

2. For a small *de novo* deletion with multiple genes observed many times in cases a high BF will be estimated for the CNV (e.g. $B_{CNV} = 20$). Since the BF is high compared with the estimate of π (0.06) we can simplify the equation by the approximation that $1 - \pi = 1$, $K\pi = 0$, $\pi B_{CNV} = 1$. This leads to a situation where the CNV-level BF (B_{CNV}) is evenly split between the K genes:

$$B_k \approx \frac{1 * B_{CNV}}{1 - 0 + (K-1) * 1} = \frac{B_{CNV}}{K}$$

3. For a small *de novo* deletion with multiple genes observed a few times in cases, the most common scenario in our data, the BF for the CNV will be more modest (e.g. $B_{CNV} = 2$). In this scenario the per gene BF (B_k) is similar to the per CNV BF (B_{CNV}) for small numbers of gene (K), for example, at $K = 2$, $B_k = 1.8$; at $K = 3$, $B_k = 1.81$:

$$B_k = \frac{(1-0.06)*2}{1 - K*0.06 + (K-1)*(0.06*2)} = \frac{1.88}{1 - 0.06K + 0.18(K-1)} = \frac{15.67}{K + 6.83}$$

26.4. Interpreting the TADA FDR

As the small *de novo* deletion data is added, the absence of the gene within any deletion will incur a small penalty, especially if the CNV mutation rate for that gene was high. Therefore, along with seven new genes meeting the FDRâĽď0.1 threshold (*NRXN1*, *SHANK2*, *SHANK3*, *SETD5*, *KAT2B*, *NLGN3*, *MBD5*), two previously identified genes now have FDR estimates in excess of 0.1: *RAPGEF4* with an FDR of 0.12 and *CACNA2D3* also with an FDR of 0.12. This is expected behavior for the TADA metric as it aims to create the best estimate from all available data for and against each gene. The FDR metric is a continuous measure so the application of a specific FDR threshold such as 0.1 is simply a convenient method of obtain a list of genes rather than an absolute threhold of 'truth'. Therefore, this result does not mean that *RAPGEF4* and *CACNA2D3* are no longer associated with ASD, rather that the evidence for their association has decreased incrementally.

27. Protein-protein interactions

The list if 28 genes with an TADA FDR $\leq$0.01 (Table S6) were submitted to DAPPLE 2.0 `www.broadinstitute.org/mpg/dapple/dappleTMP.php`); 25 genes were successfully mapped. Twenty of the submitted genes formed a single network shown in Figure 7A of the main manuscript. Two other genes (*WAC* and*TNRC6B*) formed a second network connected by a single indirect edge that is not shown. Using the full 65 genes with an FDR $\leq$0.1 resulted in one large network, shown in Figure 7B, and the same second network of *WAC* and*TNRC6B* with an indirect edge.

References

Adzhubei, I. A., Schmidt, S., Peshkin, L., Ramensky, V. E., Gerasimova, A., Bork, P., Kondrashov, A. S. and Sunyaev, S. R. (2010). A method and server for predicting damaging missense mutations. Nat Methods *7*, 248–9.

Bailey, J. A., Gu, Z., Clark, R. A., Reinert, K., Samonte, R. V., Schwartz, S., Adams, M. D., Myers, E. W., Li, P. W. and Eichler, E. E. (2002). Recent segmental duplications in the human genome. Science *297*, 1003–1007.

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

Chaste, P., Klei, L., Sanders, S. J., Murtha, M. T., Hus, V., Lowe, J. K., Willsey, A. J., Moreno-De-Luca, D., Yu, T. W., Fombonne, E., Geschwind, D., Grice, D. E., Ledbetter, D. H., Lord, C., Mane, S. M., Lese Martin, C., Martin, D. M., Morrow, E. M., Walsh, C. A., Sutcliffe, J. S., State, M. W., Devlin, B., Cook, E. H. J. and Kim, S.-J. (2013). Adjusting head circumference for covariates in autism: clinical correlates of a highly heritable continuous trait. Biol Psychiatry *74*, 576–584.

Colella, S., Yau, C., Taylor, J. M., Mirza, G., Butler, H., Clouston, P., Bassett, A. S., Seller, A., Holmes, C. C. and Ragoussis, J. (2007). QuantiSNP: an Objective Bayes Hidden-Markov Model to detect and accurately map copy number variation using SNP genotyping data. QuantiSNP: an Objective Bayes Hidden-Markov Model to detect and accurately map copy number variation using SNP genotyping data. *35*, 2013–2025.

Conrad, D. F., Pinto, D., Redon, R., Feuk, L., Gokcumen, O., Zhang, Y., Aerts, J., Andrews, T. D., Barnes, C., Campbell, P., Fitzgerald, T., Hu, M., Ihm, C. H., Kristiansson, K., Macarthur, D. G., Macdonald, J. R., Onyiah, I., Pang, A. W. C., Robson, S., Stirrups, K., Valsesia, A., Walter, K., Wei, J., Tyler-Smith, C., Carter, N. P., Lee, C., Scherer, S. W. and Hurles, M. E. (2010). Origins and functional impact of copy number variation in the human genome. Nature *464*, 704–712.

Constantino, J. N. (2002). Social Responsiveness Scale (SRS). Western Psychological Services 625 Alaska Avenue. Torrance, CA 90503-5124.

Cotney, J., Muhle, R. A., Sanders, S. J., Liu, L., Willsey, A. J., Niu, W., Liu, W., Klei, L., Lei, J., Yin, J., Reilly, S. K., Tebbenkamp, A. T., Bichsel, C., Pletikos, M., Sestan, N., Roeder, K., State, M. W., Devlin, B. and Noonan, J. P. (2015). The autism-associated chromatin modifier CHD8 regulates other autism risk genes during human neurodevelopment. Nat Commun *6*, 6404.

Darnell, J. C., Van Driesche, S. J., Zhang, C., Hung, K. Y. S., Mele, A., Fraser, C. E., Stone, E. F., Chen, C., Fak, J. J., Chi, S. W., Licatalosi, D. D., Richter, J. D. and Darnell, R. B. (2011). FMRP stalls ribosomal translocation on mRNAs linked to synaptic function and autism. Cell *146*, 247–261.

De Rubeis, S., He, X., Goldberg, A. P., Poultney, C. S., Samocha, K., Cicek, A. E., Kou, Y., Liu, L., Fromer, M., Walker, S., Singh, T., Klei, L., Kosmicki, J., Shih-Chen, F., Aleksic, B., Biscaldi, M., Bolton, P. F., Brownfeld, J. M., Cai, J., Campbell, N. G., Carracedo, A., Chahrour, M. H., Chiocchetti, A. G., Coon, H., Crawford, E. L., Curran, S. R., Dawson, G., Duketis, E., Fernandez, B. A., Gallagher, L., Geller, E., Guter, S. J., Hill, R. S., Ionita-Laza, J., Jimenz Gonzalez, P., Kilpinen, H., Klauck, S. M., Kolevzon, A., Lee, I., Lei, I., Lei, J., Lehtimaki, T., Lin, C.-F., Ma'ayan, A., Marshall, C. R., McInnes, A. L., Neale, B., Owen, M. J., Ozaki, N., Parellada, M., Parr, J. R., Purcell, S., Puura, K., Rajagopalan, D., Rehnstrom, K., Reichenberg, A., Sabo, A., Sachse, M., Sanders, S. J., Schafer, C., Schulte-Ruther, M., Skuse, D., Stevens, C., Szatmari, P., Tammimies, K., Valladares, O., Voran, A., Li-San, W., Weiss, L. A., Willsey, A. J., Yu, T. W., Yuen, R. K. C., Cook, E. H., Freitag, C. M., Gill, M., Hultman, C. M., Lehner, T., Palotie, A., Schellenberg, G. D., Sklar, P., State, M. W., Sutcliffe, J. S., Walsh, C. A., Scherer, S. W., Zwick, M. E., Barett, J. C., Cutler, D. J., Roeder, K., Devlin, B., Daly, M. J. and Buxbaum, J. D. (2014). Synaptic, transcriptional and chromatin genes disrupted in autism. Nature *515*, 209–215.

Deciphering Developmental Disorders, S. (2015). Large-scale discovery of novel genetic causes of developmental disorders. Nature *519*, 223–228.

Desachy, G., Croen, L. A., Torres, A. R., Kharrazi, M., Delorenze, G. N., Windham, G. C., Yoshida, C. K. and Weiss, L. A. (2015). Increased female autosomal burden of rare copy number variants in human populations and in autism families. Mol Psychiatry *20*, 170–175.

Dong, S., Walker, M. F., Carriero, N. J., DiCola, M., Willsey, A. J., Ye, A. Y., Waqar, Z., Gonzalez, L. E., Overton, J. D., Frahm, S., Keaney, J. F. r., Teran, N. A., Dea, J., Mandell, J. D., Hus Bal, V., Sullivan, C. A., DiLullo, N. M., Khalil, R. O., Gockley, J., Yuksel, Z., Sertel, S. M., Ercan-Sencicek, A. G., Gupta, A. R., Mane, S. M., Sheldon, M., Brooks, A. I., Roeder, K., Devlin, B., State, M. W., Wei, L. and Sanders, S. J. (2014). De novo insertions and deletions of predominantly paternal origin are associated with autism spectrum disorder. Cell Rep *9*, 16–23.

Fischbach, G. D. and Lord, C. (2010). The Simons Simplex Collection: a resource for identification of autism genetic risk factors. Neuron *68*, 192–5.

Genomes Project, C. (2010). A map of human genome variation from population-scale sequencing. Nature *467*, 1061–1073.

Girirajan, S., Dennis, M. Y., Baker, C., Malig, M., Coe, B. P., Campbell, C. D., Mark, K., Vu, T. H., Alkan, C., Cheng, Z., Biesecker, L. G., Bernier, R. and Eichler, E. E. (2013). Refinement and discovery of new hotspots of copy-number variation associated with autism spectrum disorder. Am J Hum Genet *92*, 221–237.

He, X., Sanders, S. J., Liu, L., De Rubeis, S., Lim, E. T., Sutcliffe, J. S., Schellenberg, G. D., Gibbs, R. A., Daly, M. J., Buxbaum, J. D., State, M. W., Devlin, B. and Roeder, K. (2013). Integrated model of de novo and inherited genetic variants yields greater power to identify risk genes. PLoS Genet *9*, e1003671.

Iossifov, I., O'Roak, B. J., Sanders, S. J., Ronemus, M., Krumm, N., Levy, D., Stessman, H. A., Witherspoon, K. T., Vives, L., Patterson, K. E., Smith, J. D., Paeper, B., Nickerson, D. A., Dea, J., Dong, S., Gonzalez, L. E., Mandell, J. D., Mane, S. M., Murtha, M. T., Sullivan, C. A., Walker, M. F., Waqar, Z., Wei, L., Willsey, A. J., Yamrom, B., Lee, Y.-h., Grabowska, E., Dalkic, E., Wang, Z., Marks, S., Andrews, P., Leotta, A., Kendall, J., Hakker, I., Rosenbaum, J., Ma, B., Rodgers, L., Troge, J., Narzisi, G., Yoon, S., Schatz, M. C., Ye, K., McCombie, W. R., Shendure, J., Eichler, E. E., State, M. W. and Wigler, M. (2014). The contribution of de novo coding mutations to autism spectrum disorder. Nature *515*, 216–221.

Iossifov, I., Ronemus, M., Levy, D., Wang, Z., Hakker, I., Rosenbaum, J., Yamrom, B., Lee, Y.-h., Narzisi, G., Leotta, A., Kendall, J., Grabowska, E., Ma, B., Marks, S., Rodgers, L., Stepansky, A., Troge, J., Andrews, P., Bekritsky, M., Pradhan, K., Ghiban, E., Kramer, M., Parla, J., Demeter, R., Fulton, L. L., Fulton, R. S., Magrini, V. J., Ye, K., Darnell, J. C., Darnell, R. B., Mardis, E. R., Wilson, R. K., Schatz, M. C., McCombie, W. R. and Wigler, M. (2012). De Novo Gene Disruptions in Children on the Autistic Spectrum. Neuron *74*, 285–299.

Itsara, A., Wu, H., Smith, J. D., Nickerson, D. A., Romieu, I., London, S. J. and Eichler, E. E. (2010). De novo rates and selection of large copy number variation. Genome Res *20*, 1469–81.

Kang, H. J., Kawasawa, Y. I., Cheng, F., Zhu, Y., Xu, X., Li, M., Sousa, A. M., Pletikos, M., Meyer, K. A., Sedmak, G., Guennel, T., Shin, Y., Johnson, M. B., Krsnik, Z., Mayer, S., Fertuzinhos, S., Umlauf, S., Lisgo, S. N., Vortmeyer, A., Weinberger, D. R., Mane, S., Hyde, T. M., Huttner, A., Reimers, M., Kleinman, J. E. and Sestan, N. (2011). Spatio-temporal transcriptome of the human brain. Nature *478*, 483–9.

Kent, W. J. (2002). BLAT–the BLAST-like alignment tool. Genome Res *12*, 656–664.

Levy, D., Ronemus, M., Yamrom, B., Lee, Y.-H., Leotta, A., Kendall, J., Marks, S., Lakshmi, B., Pai, D., Ye, K., Buja, A., Krieger, A., Yoon, S., Troge, J., Rodgers, L., Iossifov, I. and Wigler, M. (2011). Rare De Novo and Transmitted Copy-Number Variation in Autistic Spectrum Disorders. Neuron *70*, 886–897.

Liu, L., Sabo, A., Neale, B. M., Nagaswamy, U., Stevens, C., Lim, E., Bodea, C. A., Muzny, D., Reid, J. G., Banks, E., Coon, H., Depristo, M., Dinh, H., Fennel, T., Flannick, J., Gabriel, S., Garimella, K., Gross, S., Hawes, A., Lewis, L., Makarov, V., Maguire, J., Newsham, I., Poplin, R., Ripke, S., Shakir, K., Samocha, K. E., Wu, Y., Boerwinkle, E., Buxbaum, J. D., Cook, E. H., Devlin, B., Schellenberg, G. D., Sutcliffe, J. S., Daly, M. J., Gibbs, R. A. and Roeder, K. (2013). Analysis of Rare, Exonic Variation amongst Subjects with Autism Spectrum Disorders and Population Controls. PLoS Genet *9*, e1003443.

Lord, C., Risi, S., Lambrecht, L., Cook, E. H. J., Leventhal, B. L., DiLavore, P. C., Pickles, A. and Rutter, M. (2000). The autism diagnostic observation schedule-generic: a standard measure of social and communication deficits associated with the spectrum of autism. J Autism Dev Disord *30*, 205–223.

Lord, C., Rutter, M. and Le Couteur, A. (1994). Autism Diagnostic Interview-Revised: a revised version of a diagnostic interview for caregivers of individuals with possible pervasive developmental disorders. J Autism Dev Disord *24*, 659–85.

Macdonald, J. R., Ziman, R., Yuen, R. K. C., Feuk, L. and Scherer, S. W. (2014). The Database of Genomic Variants: a curated collection of structural variation in the human genome. Nucleic Acids Res *42*, D986–92.

Marshall, C. R., Noor, A., Vincent, J. B., Lionel, A. C., Feuk, L., Skaug, J., Shago, M., Moessner, R., Pinto, D., Ren, Y., Thiruvahindrapduram, B., Fiebig, A., Schreiber, S., Friedman, J., Ketelaars, C. E., Vos, Y. J., Ficicioglu, C., Kirkpatrick, S., Nicolson, R., Sloman, L., Summers, A., Gibbons, C. A., Teebi, A., Chitayat, D., Weksberg, R., Thompson, A., Vardy, C., Crosbie, V., Luscombe, S., Baatjes, R., Zwaigenbaum, L., Roberts, W., Fernandez, B., Szatmari, P. and Scherer, S. W. (2008). Structural variation of chromosomes in autism spectrum disorder. Am J Hum Genet *82*, 477–88.

O'Roak, B. J., Vives, L., Fu, W., Egertson, J. D., Stanaway, I. B., Phelps, I. G., Carvill, G., Kumar, A., Lee, C., Ankenman, K., Munson, J., Hiatt, J. B., Turner, E. H., Levy, R., O'Day, D. R., Krumm, N., Coe, B. P., Martin, B. K., Borenstein, E., Nickerson, D. A., Mefford, H. C., Doherty, D., Akey, J. M., Bernier, R., Eichler, E. E. and Shendure, J. (2012). Multiplex targeted sequencing identifies recurrently mutated genes in autism spectrum disorders. Science *338*, 1619–22.

Petrovski, S., Wang, Q., Heinzen, E. L., Allen, A. S. and Goldstein, D. B. (2013). Genic intolerance to functional variation and the interpretation of personal genomes. PLoS Genet *9*, e1003709.

Pinto, D., Delaby, E., Merico, D., Barbosa, M., Merikangas, A., Klei, L., Thiruvahindrapuram, B., Xu, X., Ziman, R., Wang, Z., Vorstman, J. A. S., Thompson, A., Regan, R., Pilorge, M., Pellecchia, G., Pagnamenta, A. T., Oliveira, B., Marshall, C. R., Magalhaes, T. R., Lowe, J. K., Howe, J. L., Griswold, A. J., Gilbert, J., Duketis, E., Dombroski, B. A., De Jonge, M. V., Cuccaro, M., Crawford, E. L., Correia, C. T., Conroy, J., Conceicao, I. C., Chiocchetti, A. G., Casey, J. P., Cai, G., Cabrol, C., Bolshakova, N., Bacchelli, E., Anney, R., Gallinger, S., Cotterchio, M., Casey, G., Zwaigenbaum, L., Wittemeyer, K., Wing, K., Wallace, S., van Engeland, H., Tryfon, A., Thomson, S., Soorya, L., Roge, B., Roberts, W., Poustka, F., Mouga, S., Minshew, N., McInnes, L. A., McGrew, S. G., Lord, C., Leboyer, M., Le Couteur, A. S., Kolevzon, A., Jimenez Gonzalez, P., Jacob, S., Holt, R., Guter, S., Green, J., Green, A., Gillberg, C., Fernandez, B. A., Duque, F., Delorme, R., Dawson, G., Chaste, P., Cafe, C., Brennan, S., Bourgeron, T., Bolton, P. F., Bolte, S., Bernier, R., Baird, G., Bailey, A. J., Anagnostou, E., Almeida, J., Wijsman, E. M., Vieland, V. J., Vicente, A. M., Schellenberg, G. D., Pericak-Vance, M., Paterson, A. D., Parr, J. R., Oliveira, G., Nurnberger, J. I., Monaco, A. P., Maestrini, E., Klauck, S. M., Hakonarson, H., Haines, J. L., Geschwind, D. H., Freitag, C. M., Folstein, S. E., Ennis, S., Coon, H., Battaglia, A., Szatmari, P., Sutcliffe, J. S., Hallmayer, J., Gill, M., Cook, E. H., Buxbaum, J. D., Devlin, B., Gallagher, L., Betancur, C. and Scherer, S. W. (2014). Convergence of genes and cellular pathways dysregulated in autism spectrum disorders. Am J Hum Genet *94*, 677–694.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., Maller, J., Sklar, P., de Bakker, P. I., Daly, M. J. and Sham, P. C. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. PLINK: a tool set for whole-genome association and population-based linkage analyses. *81*, 559–575.

Samocha, K. E., Robinson, E. B., Sanders, S. J., Stevens, C., Sabo, A., McGrath, L. M., Kosmicki, J. A., Rehnstrom, K., Mallick, S., Kirby, A., Wall, D. P., MacArthur, D. G., Gabriel, S. B., DePristo, M., Purcell, S. M., Palotie, A., Boerwinkle, E., Buxbaum, J. D., Cook, E. H. J., Gibbs, R. A., Schellenberg, G. D., Sutcliffe, J. S., Devlin, B., Roeder, K., Neale, B. M. and Daly, M. J. (2014). A framework for the interpretation of de novo mutation in human disease. Nat Genet *46*, 944–950.

Sanders, S. J., Ercan-Sencicek, A. G., Hus, V., Luo, R., Murtha, M. T., Moreno-De-Luca, D., Chu, S. H., Moreau, M. P., Gupta, A. R., Thomson, S. A., Mason, C. E., Bilguvar, K., Celestino-Soper, P. B. S., Choi, M., Crawford, E. L., Davis, L., Davis Wright, N. R., Dhodapkar, R. M., DiCola, M., DiLullo, N. M., Fernandez, T. V., Fielding-Singh, V., Fishman, D. O., Frahm, S., Garagaloyan, R., Goh, G. S., Kammela, S., Klei, L., Lowe, J. K., Lund, S. C., McGrew, A. D., Meyer, K. A., Moffat, W. J., Murdoch, J. D., O'Roak, B. J., Ober, G. T., Pottenger, R. S., Raubeson, M. J., Song, Y., Wang, Q., Yaspan, B. L., Yu, T. W., Yurkiewicz, I. R., Beaudet, A. L., Cantor, R. M., Curland, M., Grice, D. E., Gunel, M., Lifton, R. P., Mane, S. M., Martin, D. M., Shaw, C. A., Sheldon, M., Tischfield, J. A., Walsh, C. A., Morrow, E. M., Ledbetter, D. H., Fombonne, E., Lord, C., Martin, C. L., Brooks, A. I., Sutcliffe, J. S., Cook Jr, E. H., Geschwind, D., Roeder, K., Devlin, B. and State, M. W. (2011). Multiple Recurrent De Novo CNVs, Including Duplications of the 7q11.23 Williams Syndrome Region, Are Strongly Associated with Autism. Neuron *70*, 863–885.

Sanders, S. J., Murtha, M. T., Gupta, A. R., Murdoch, J. D., Raubeson, M. J., Willsey, A. J., Ercan-Sencicek, A. G., DiLullo, N. M., Parikshak, N. N., Stein, J. L., Walker, M. F., Ober, G. T., Teran, N. A., Song, Y., El-Fishawy, P., Murtha, R. C., Choi, M., Overton, J. D., Bjornson, R. D., Carriero, N. J., Meyer, K. A., Bilguvar, K., Mane, S. M., Sestan, N., Lifton, R. P., Günel, M., Roeder, K., Geschwind, D. H., Devlin, B. and State, M. W. (2012). De novo mutations revealed by whole-exome sequencing are strongly associated with autism. Nature *485*, 237–41.

Sebat, J., Lakshmi, B., Malhotra, D., Troge, J., Lese-Martin, C., Walsh, T., Yamrom, B., Yoon, S., Krasnitz, A., Kendall, J., Leotta, A., Pai, D., Zhang, R., Lee, Y. H., Hicks, J., Spence, S. J., Lee, A. T., Puura, K., Lehtimäki, T., Ledbetter, D., Gregersen, P. K., Bregman, J., Sutcliffe, J. S., Jobanputra, V., Chung, W., Warburton, D., King, M. C., Skuse, D., Geschwind, D. H., Gilliam, T. C., Ye, K. and Wigler, M. (2007). Strong association of de novo copy number mutations with autism. Science *316*, 445–9.

Sugathan, A., Biagioli, M., Golzio, C., Erdin, S., Blumenthal, I., Manavalan, P., Ragavendran, A., Brand, H., Lucente, D., Miles, J., Sheridan, S. D., Stortchevoi, A., Kellis, M., Haggarty, S. J., Katsanis, N., Gusella, J. F. and Talkowski, M. E. (2014). CHD8 regulates neurodevelopmental pathways associated with autism spectrum disorder in neural progenitors. Proc Natl Acad Sci U S A *111*, E4468–77.

Wang, K., Li, M., Hadley, D., Liu, R., Glessner, J., Grant, S. F., Hakonarson, H. and Bucan, M. (2007). PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *17*, 1665–1674.

Willsey, A. J., Sanders, S. J., Li, M., Dong, S., Tebbenkamp, A. T., Muhle, R. A., Reilly, S. K., Lin, L., Fertuzinhos, S., Miller, J. A., Murtha, M. T., Bichsel, C., Niu, W., Cotney, J., Ercan-Sencicek, A. G., Gockley, J., Gupta, A. R., Han, W., He, X., Hoffman, E. J., Klei, L., Lei, J., Liu, W., Liu, L., Lu, C., Xu, X., Zhu, Y., Mane, S. M., Lein, E. S., Wei, L., Noonan, J. P., Roeder, K., Devlin, B., Sestan, N. and State, M. W. (2013). Coexpression networks implicate human midfetal deep cortical projection neurons in the pathogenesis of autism. Cell *155*, 997–1007.