The American Journal of Human Genetics, Volume *93*

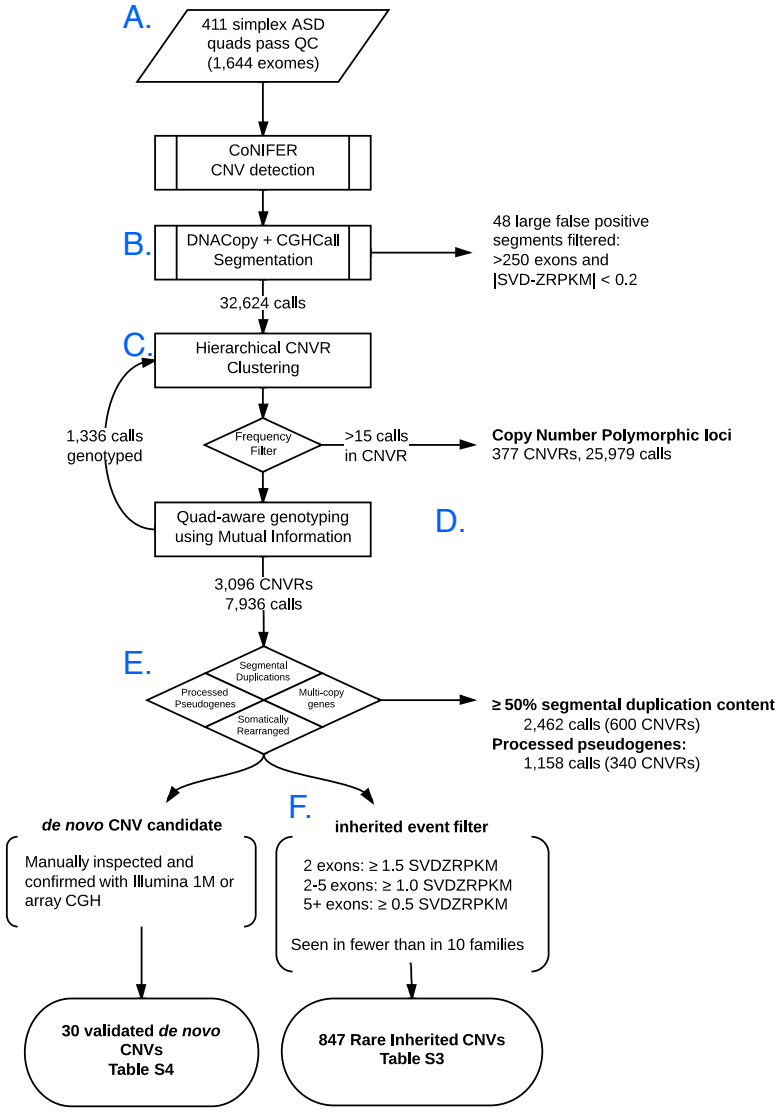
Supplemental Data

Transmission Disequilibrium

of Small CNVs in Simplex Autism

Niklas Krumm, Brian J. O'Roak, Emre Karakoc, Kiana Mohajeri, Ben Nelson, Laura Vives, Sebastien Jacquemont, Jeff Munson, Raphe Bernier, and Evan E. Eichler

Figure S1. Data Processing and CNV Calling Details



A.Previously generated FASTQ data from four exome sequencing studies (Iossifov et al., 2012; O'Roak et al., 2012; Sanders et al., 2012) was used in this study. In addition, we generated sequence for unaffected sibling in 20 published trios (O'Roak et al., 2011) for a complete set of 412 quads. Data was processed and analyzed using CoNIFER (Krumm, 2011, as previously described). SVD cutoff values was set to either 12 or 15 for each dataset. We excluded one family (12154) on the basis of significant contamination between members, resulting in 411 families QC'd families.

B. We used DNACopy and CGHCall to segment and assign deletion or duplication probabilities to SVD-ZRPKM values. Parameters for DNACopy were as follows: alpha = 0.01, using the undo.split="sdundo" option with undo.SD = 2.

C. Next, we grouped individual CNV calls into similar CNV Regions (CNVRs) using pairwise distances between all CNVs based on a modified reciprocal overlap (RO) heuristic that incorporates the size of the CNV as well as RO percentage.

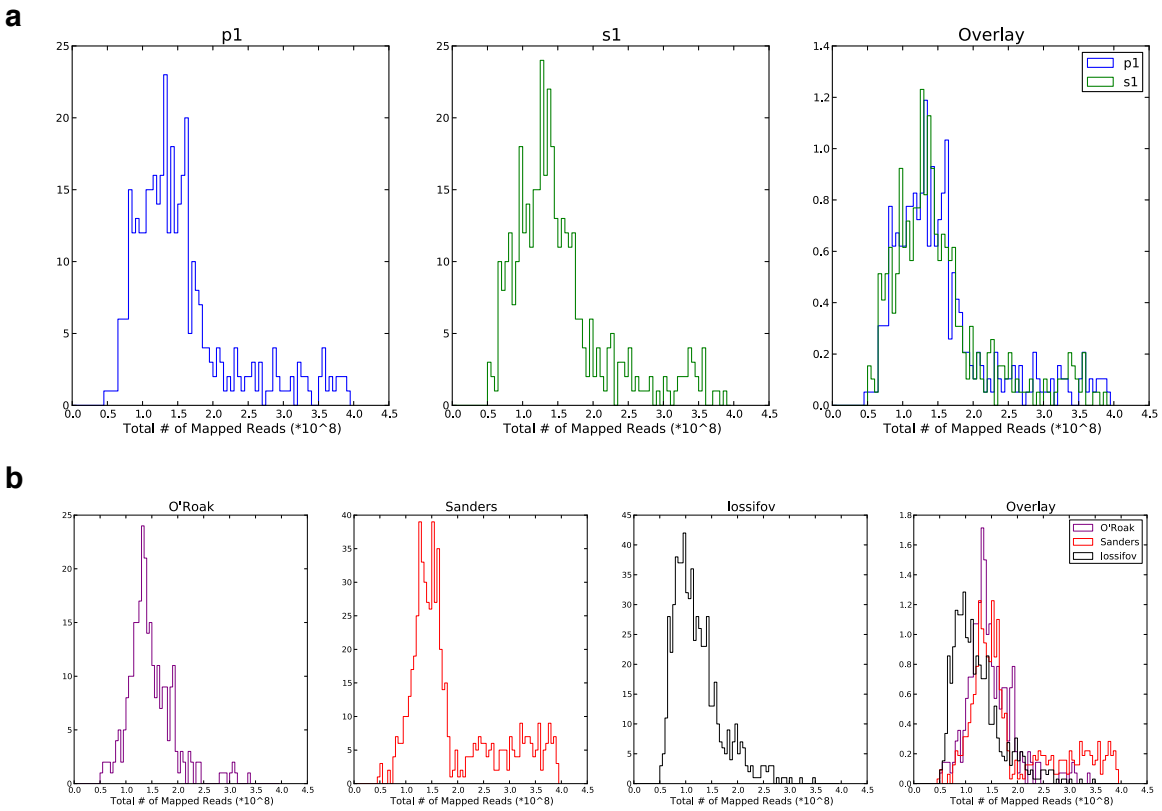
D.We reduced false-negative calls for inherited CNVs by applying a family-based genotyping method which uses a metric based on Mutual Information between the raw CoNIFER of each family member at a particular locus in order to determine missed calls.

E.CNVs were filtered based on overlap >50% with known polymorphic sites, processed pseudogenes, segmental duplications and other non-unique portions of the exome.

F. Our final set of calls was created by requiring an absolute median SVD-ZRPKM score (i.e., signal strength) of ≥ 0.5 for calls with 5 or more probes, ≥1.0 for calls 3-5 probes in length, and ≥ 1.0 for calls 2 probes in length. We excluded any calls on the X or Y chromosomes for all analyses in this work.

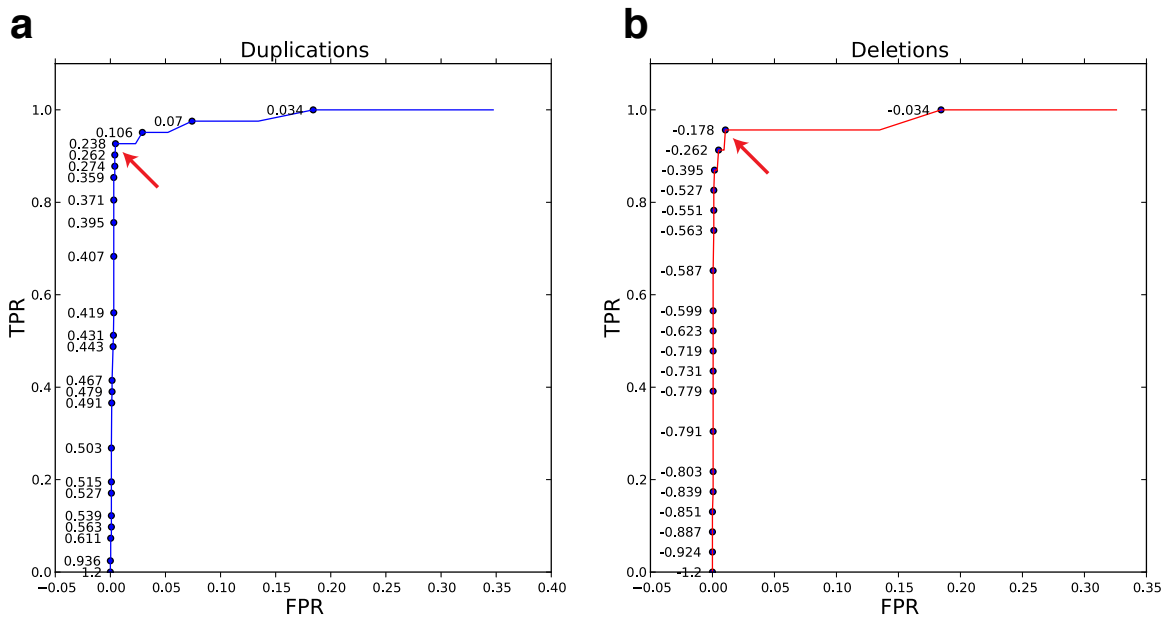
Details of these methods are available upon request.

Figure S2. Mapped Coverage between Probands/Siblings and by Data Source



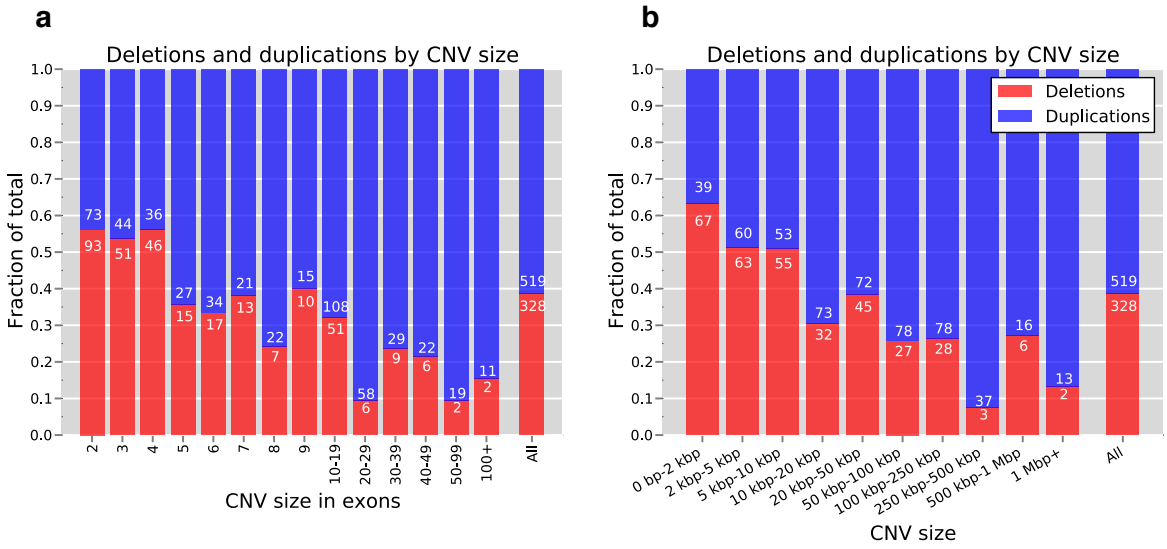
All panels: X-axis: total mapped 36mer reads (x10^8) by the mrsFAST alignment program to the human exome. **(a)** Histograms of Probands (left) and Siblings (center) and overlap (right) shows no significant difference in coverage levels (Paired t-test p= 0.09). **(b)**. Same as in **(a)**, but by dataset, revealing that the Iossifov dataset had lower coverage than the O'Roak or Sanders datasets.

Figure S3. Array-CGH Validation of CNVs



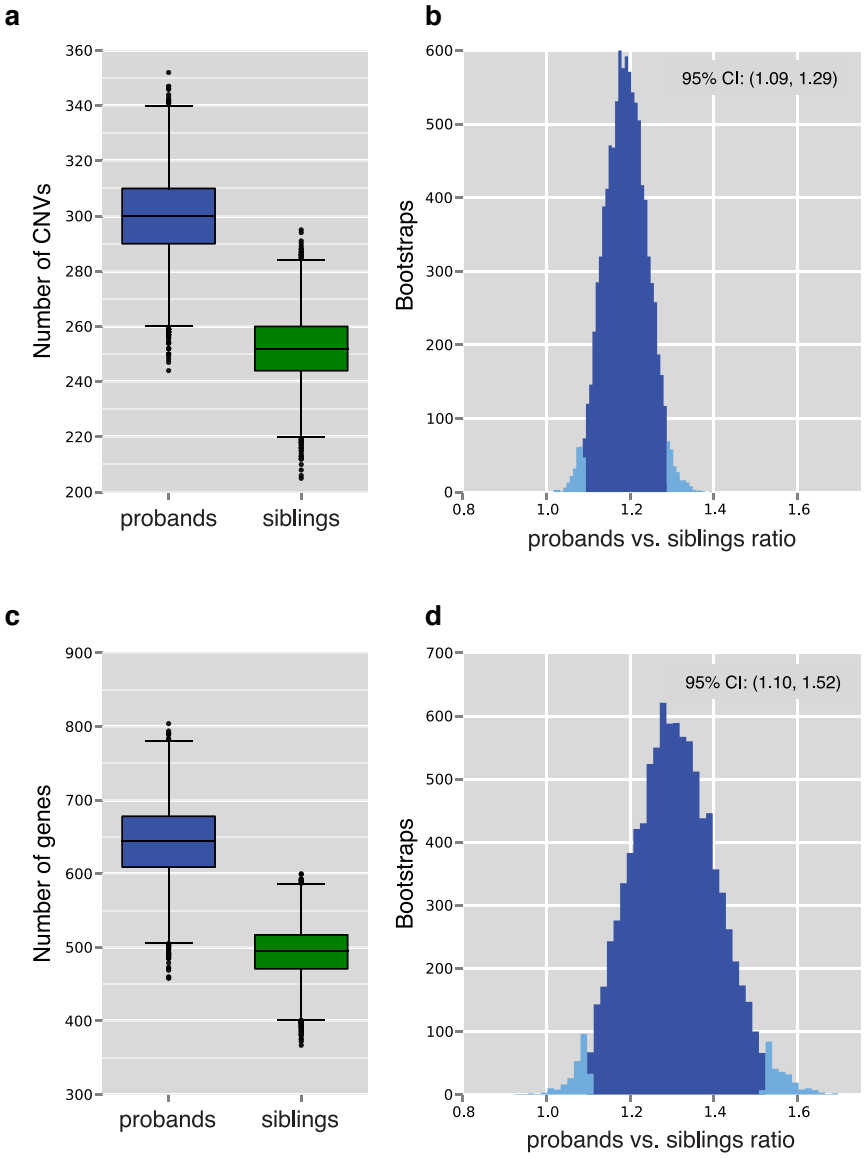
We designed a custom Agilent SurePrint G3 4x180k CGH microarray to confirm CNVs, using variable density spacing of probes, ranging from 125bp^{-1} for calls smaller than 10kbp to 5kbp^{-1} for large calls up to 500kbp, in order to insure at least 10 probes per call. (Note: Due to the high density of probes required for validation of small CNVs, some of the probes were of lower quality (as based on the manufacturer's quality score), and their performance was accordingly lower.) Test and reference DNA (we used DNA from HapMap sample NA18507) from each sample was labeled with Cy3 and Cy5 dye using a NimbleGen array labeling kit according to manufacturer's instructions. Five micrograms of labelled test and reference DNA was hybridized for 24 hours using Agilent reagents to the microarray slide and washed according to manufacturer's directions. Slides were scanned using an Agilent Microarray Scanner and analyzed using Agilent Feature Extract v10.5.1.1. Arrays with a per-sample standard deviation of LogR values > 0.5 were repeated. In order to reduce systematic and batch noise between probes and samples, we employed a similar normalization strategy to the CoNIFER pipeline and used SVD to remove the three strongest components of variance. We determined minimum logR thresholds for the validation arrays by leveraging the logR values across the 60 previously identified CNVs (from Sanders et al., 2011), each found in at least one of our validation samples. We calculated Receiver Operating Curves for (**a)** duplications (39 calls) and (**b)** deletions (21 calls), using the samples *without* the previously identified CNVs as the "true negatives". Next, we individually picked the optimal operating point (OOP) for deletions (median LogR OOP <= -0.178) and duplications (median LogR OOP >= 0.24), such that we maximally discerned our known true positives from true negatives. Both OOPs had a FPR of ~1%, and a recall rate >90%, indicating our array was highly specific and sensitive to true events. These logR cutoff values were used in assessing if novel CNVs were true positives or not: if the mean LogR across all probes in the call interval was greater than the duplication threshold (or lower than the deletion threshold), we considered the call validated. **Arrows** indicate chosen optimal operating point (OOP), which was used as the threshold for validation of unknown calls.

Figure S4. CNV Size and Copy Number



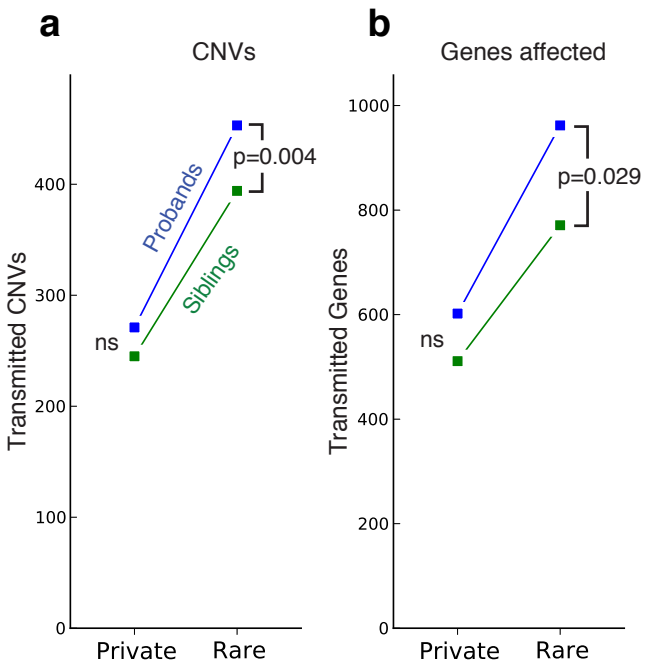
Inherited CNVs in probands and siblings, binned by size in exons **(a)** or estimated genomic size **(b)**. As expected, larger CNVs are more likely to be duplications, an effect we found true for both probands and siblings.

Figure S5. Bootstrap Results



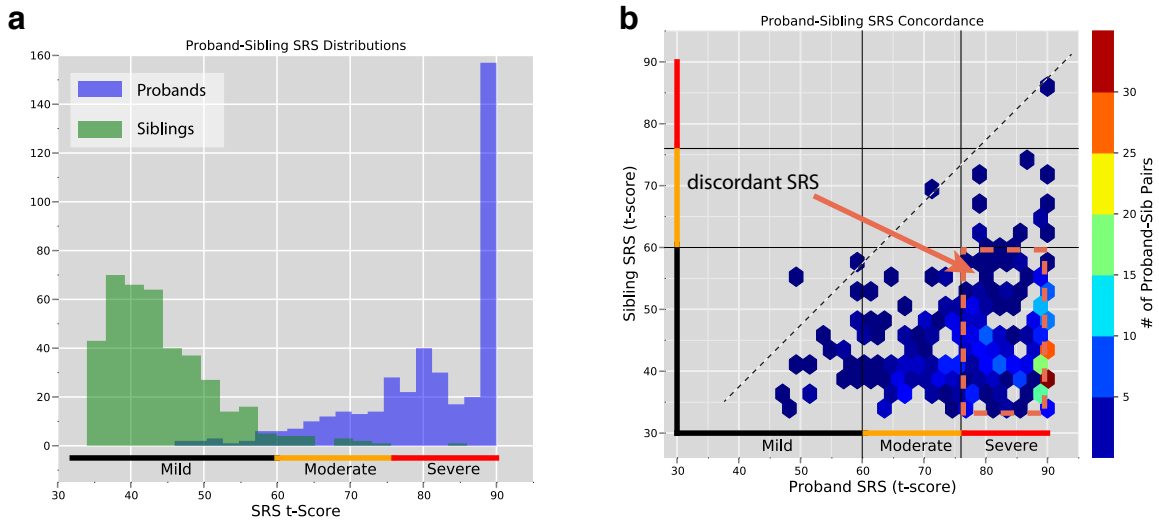
We tested the robustness of the overall effect of burden by a bootstrap method, in which we calculated the CNV burden ratio (for CNVs and genes) of 10,000 randomly sampled (with replacement) sets of families from the overall set of 411 quads. **(a)** Total CNV counts for probands (green) and siblings (blue) and CNV burden **(b)** between probands and siblings (dark blue: inner 95% of empirical distribution). In **(c)** and **(d)**, the results when counting total number of genes and genic burden.

Figure S6. Rare vs. Private Burden in 411 Quads



Rare vs. Private burden in 411 quads. There was no increased burden for CNVs **(a)** observed only once in 411 families, or for genes in those CNVs **(b)**.

Figures S7. Phenotypes (SRS and IQ) in Probands and Siblings

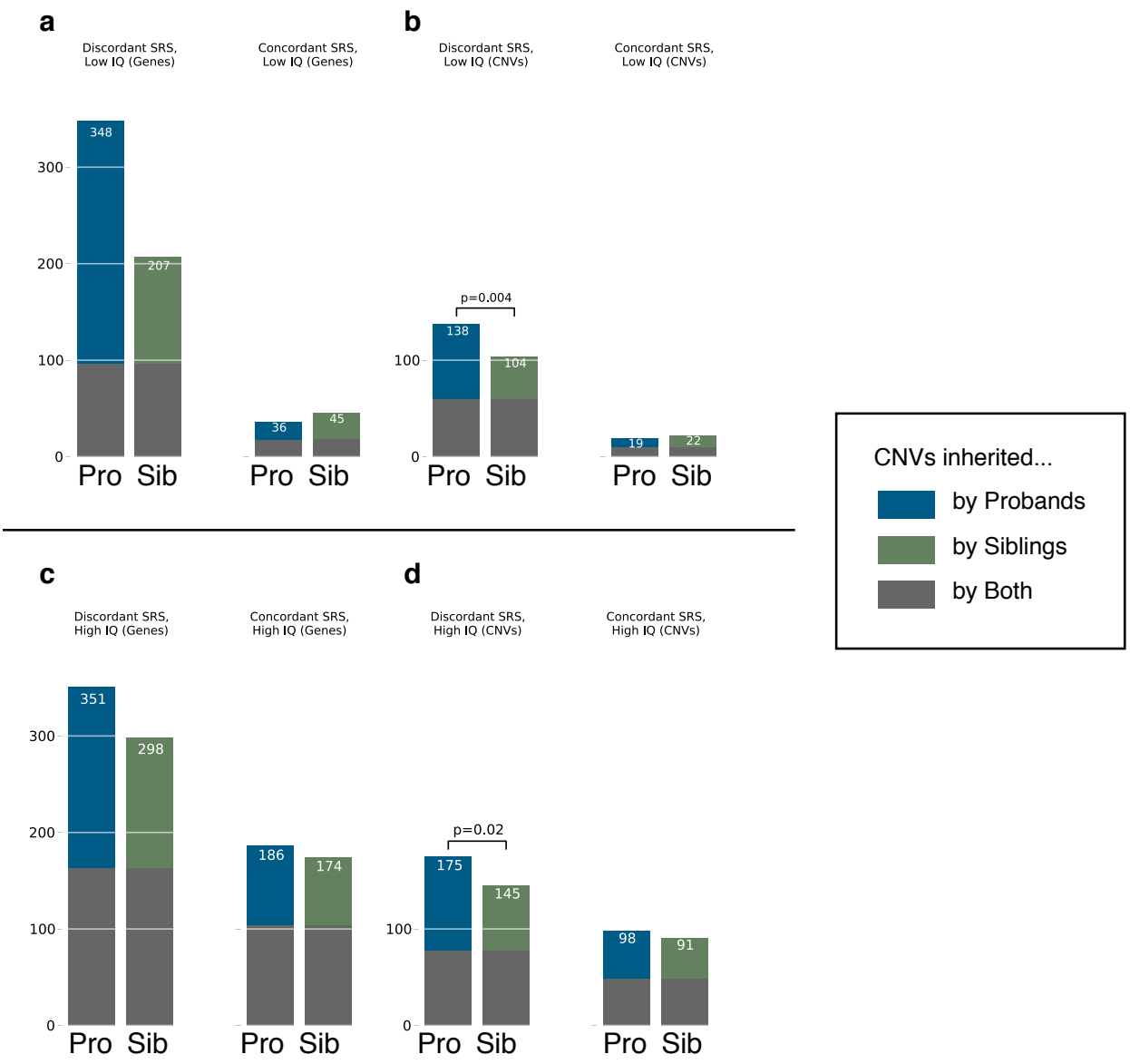


c

		Proband SRS		
		< 60 (“Unaffected”)	60-75 (“Moderate”)	76+ (“Severe”)
Sibling SRS	< 60	16 Concordant Quads	84 Concordant Quads	**276 Discordant Quads**
	60-75	—	1 Concordant Quad	13 Concordant Quads
	76+	—	—	1 Concordant Quads

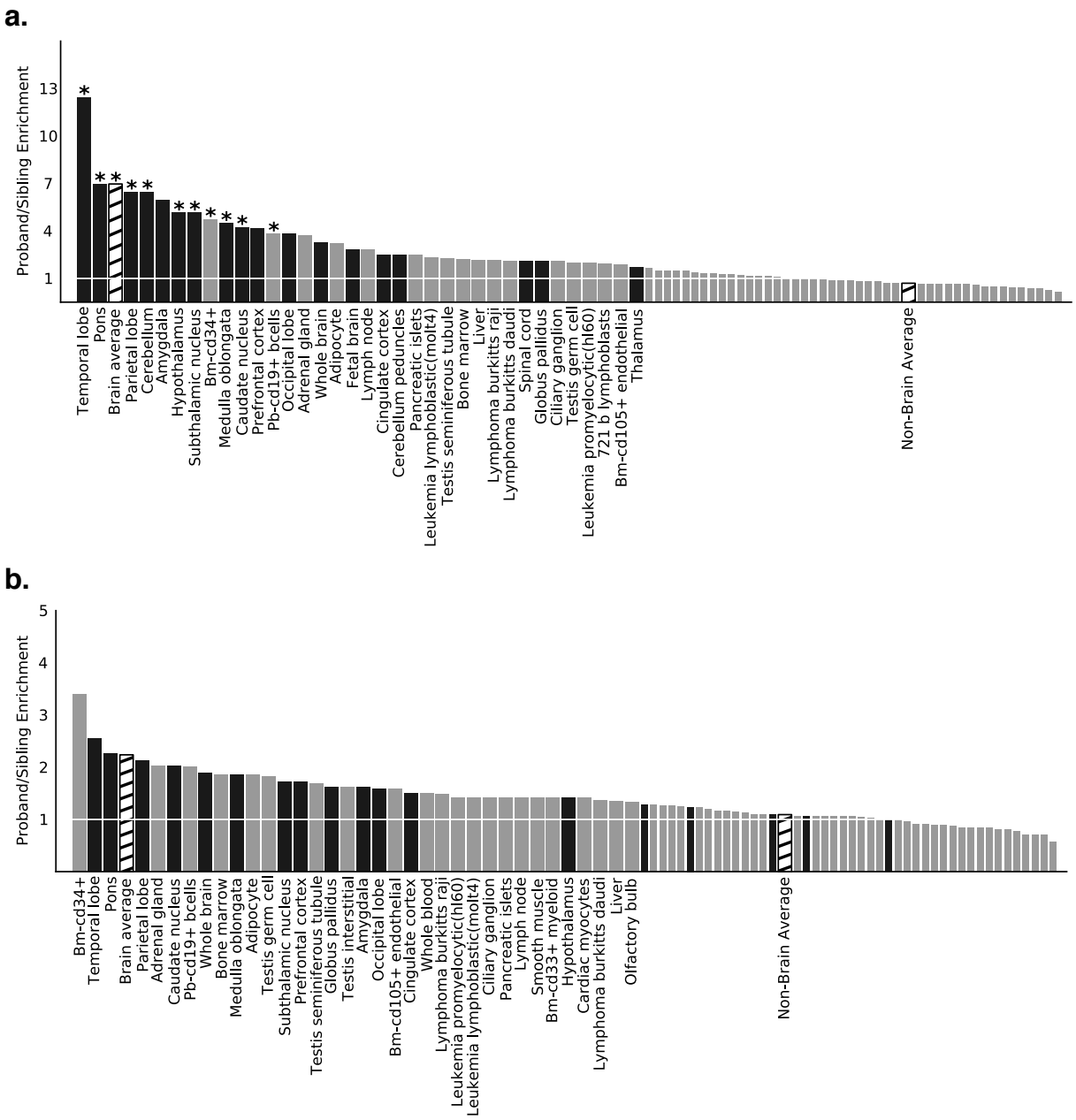
(a) Distribution of SRS t-scores in probands (blue) and siblings (green). Higher scores are more affected, and SRS t-scores greater than 75 are considered “severely affected”. **(b)** Heatmap plot of SRS values for probands (x-axis) and their siblings (y-axis). In almost all cases, the probands have higher SRS scores, but the difference in SRS score between probands and siblings varies widely among all pairs. We designated the pairs with the most extreme differences of SRS score between them as “Discordant SRS” pairs (indicated by arrow and dashed orange box, lower right). **(c)** Table clarifying discordant vs. concordant SRS quads.

Figure S8. Burden between SRS and IQ in Probands and Siblings



(a) Genic burden and (b) CNV burden for proband-sibling pairs where the proband has low IQ (< 70) for both discordant and concordant SRS quads. Burden for probands with high IQ (≥70) shown in (c) and (d). P-value bars drawn if two-tailed paired t-test p value is less than 0.0

Figure S9. Enrichment of Brain-Expressed Genes in SRS Discordant Quads (A) and All Quads (B)

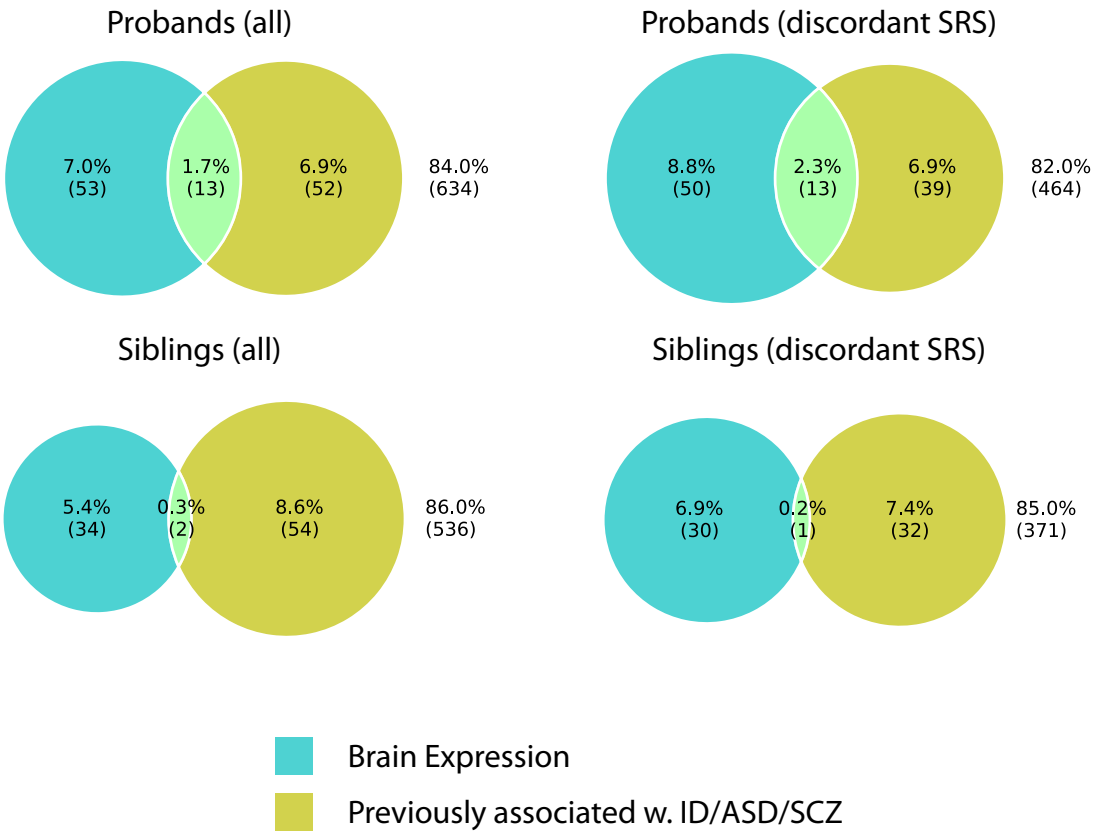


We used publicly available gene expression data from the Human U133A/GNF1H Gene Atlas (GEO: GSE1133), comprising 79 human tissues, including 18 nervous system tissues (Su, 2004). We associated the microarray probe IDs with HUGO gene names and average expression across multiple probes in the same gene. For each tissue, genes were sorted by expression and we considered the top 5% of each category to be “highly expressed”. To calculate enrichment, we took the unique sets of genes disrupted in probands and those disrupted in siblings and intersected each with the set of highly

expressed genes in each category. The ratio of counts between these two intersections constituted the fold enrichment for each category. Bars (y-axis) represent ratio of enrichment between proband and siblings for genes highly expressed in each tissue (defined as top 5%, see Methods). Black bars: tissue is part of brain or nervous system; white bars: non-brain or nervous system tissues; hatched bars: are computed averages. Asterix indicates significance using a FDR-based multiple testing correction q-value < 0.05. **(a)** probands from SRS discordant quads only show greater enrichment for brain-expressed genes than do all quads, **(b)**.

In order to correct for the 79 multiple comparisons, we employed a permutation and false discovery rate (FDR) strategy. First, we derived a null distribution of enrichment between probands and siblings by shuffling the proband-only and sibling-only sets of genes and recomputing the enrichment. Next, an empirical p-value was derived by scoring the actual enrichment value against the null distributions for each tissue. Using the FDR method described in (Storey & Tibshirani, 2003) and the R package qvalue, we calculated q values for each tissue and assessed statistical significance at $q < 0.05$. In order to calculate the brain and non-brain averages, we averaged gene expression across all 18 brain- and nervous system tissues and 61 non-brain tissues. These two categories were corrected for two comparisons each.

Figure S10. Intersection between Brain-Expressed Genes and Previously Associated Genes in Proband CNVs, but Not Sibling CNVs



We intersected the sets of genes found in probands (top row) and siblings (bottom row) that were either brain expressed (teal circles) or had previously been observed in ASD/Schizophrenia/ID (yellow circles). Probands—especially those in SRS discordant pairs— had a higher fraction of intersecting genes (13 genes, Table S11) than other groups or their siblings, suggesting that these genes may be top candidates for follow-up study in the pathogenesis of ASD.

To establish the list of genes previously associated with autism/ASD/intellectual disability/schizophrenia, we attempted to identify all genes that were associated with developmental delay, intellectual disabilities and schizophrenia. We conducted searches using the OMIM Gene Map feature with the following terms: “mental retardation” “intellectual disability”, “autism” and “schizophrenia”, and included all returned gene hits. We also included genes from the Simons SFARI autism candidate genes with” association scores” ranging from 1 to 4 (n=155 genes) (https://gene.sfari.org/autdb/submitsearch?selfld_0=GENES_GENE_SYMBOL&selfldv_0=&numOfFields=1&userAction=viewall&tableName=AUT_HG&submit2=View+All#GS). All genes from these searches were included in the list.

References

Iossifov, I., Ronemus, M., Levy, D., Wang, Z., Hakker, I., Rosenbaum, J., et al. (2012). De novo gene disruptions in children on the autistic spectrum. *Neuron*, *74*(2), 285–299. doi:10.1016/j.neuron.2012.04.009

Krumm, N., Sudmant, P. H., Ko, A., O'Roak, B. J., Malig, M., Coe, B. P., et al. (2012). Copy number variation detection and genotyping from exome sequence data. *Genome Research*. doi:10.1101/gr.138115.112

O'Roak, B. J., Deriziotis, P., Lee, C., Vives, L., Schwartz, J. J., Girirajan, S., et al. (2011). Exome sequencing in sporadic autism spectrum disorders identifies severe de novo mutations. *Nature Genetics*, *43*(6), 585–589. doi:10.1038/ng.835

O'Roak, B. J., Vives, L., Girirajan, S., Karakoc, E., Krumm, N., Coe, B. P., et al. (2012). Sporadic autism exomes reveal a highly interconnected protein network of de novo mutations. *Nature*. doi:10.1038/nature10989

Sanders, S. J., Murtha, M. T., Gupta, A. R., Murdoch, J. D., Raubeson, M. J., Willsey, A. J., et al. (2012). De novo mutations revealed by whole-exome sequencing are strongly associated with autism. *Nature*, *485*(7397), 237–241. doi:10.1038/nature10945

Storey, J. D., & Tibshirani, R. (2003). Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences*, *100*(16), 9440–9445. doi:10.1073/pnas.1530509100

Su, A. I. (2004). A gene atlas of the mouse and human protein-encoding transcriptomes. *Proceedings of the National Academy of Sciences*, *101*(16), 6062–6067. doi:10.1073/pnas.0400782101