The American Journal of Human Genetics

Supplemental Data

Genome Sequencing of Autism-Affected Families Reveals Disruption of Putative Noncoding Regulatory DNA

Tychele N. Turner, Fereydoun Hormozdiari, Michael H. Duyzend, Sarah A. McClymont, Paul W. Hook, Ivan Iossifov, Archana Raja, Carl Baker, Kendra Hoekzema, Holly A. Stessman, Michael C. Zody, Bradley J. Nelson, John Huddleston, Richard Sandstrom, Joshua D. Smith, David Hanna, James M. Swanson, Elaine M. Faustman, Michael J. Bamshad, John Stamatoyannopoulos, Deborah A. Nickerson, Andrew S. McCallion, Robert Darnell, and Evan E. Eichler

Supplemental Data

Supplemental Figures

Figure S1: Experimental approach.

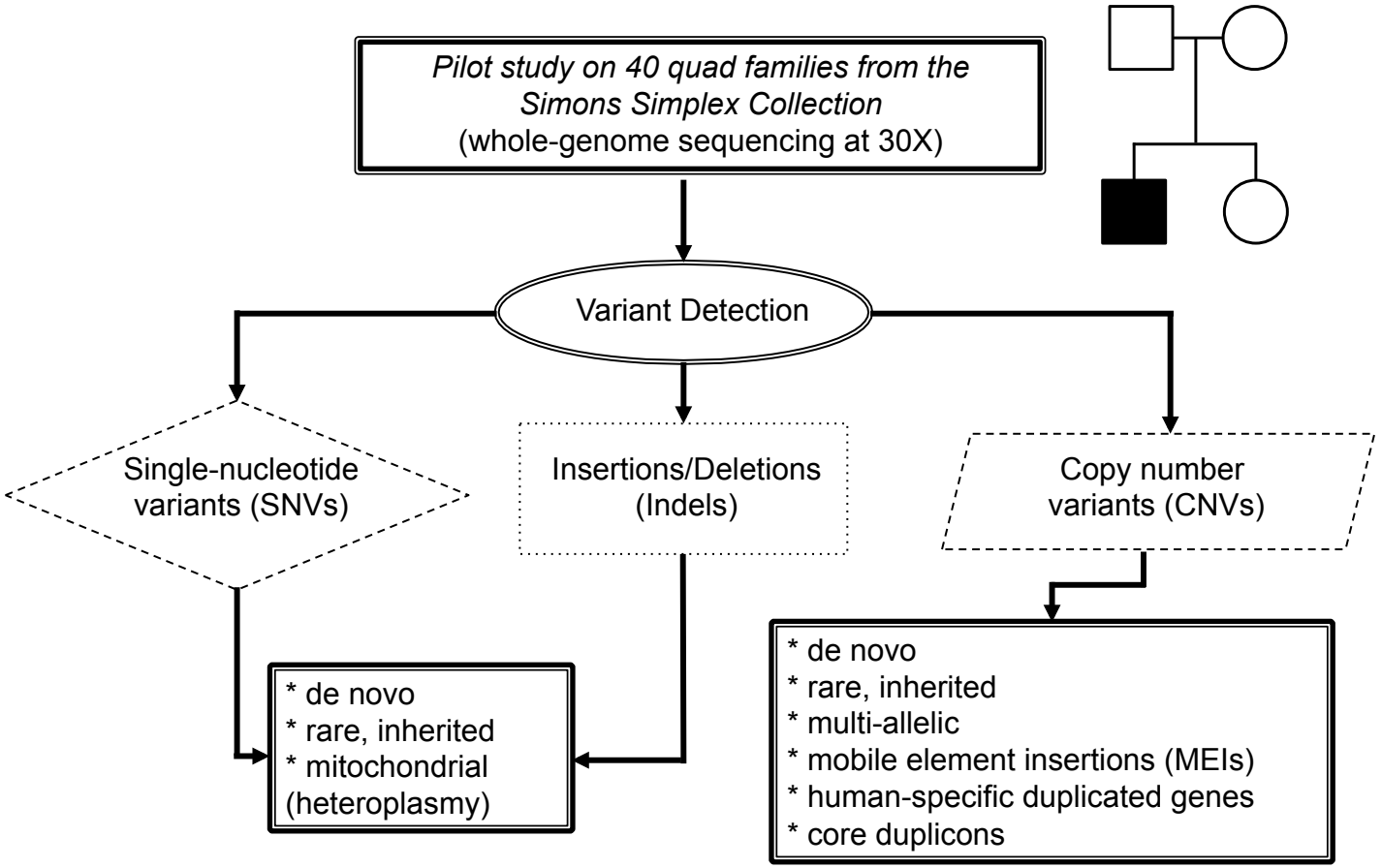

Figure S1: Flow chart detailing analyses performed on genome data.

Figure S2: Genome sequence properties.

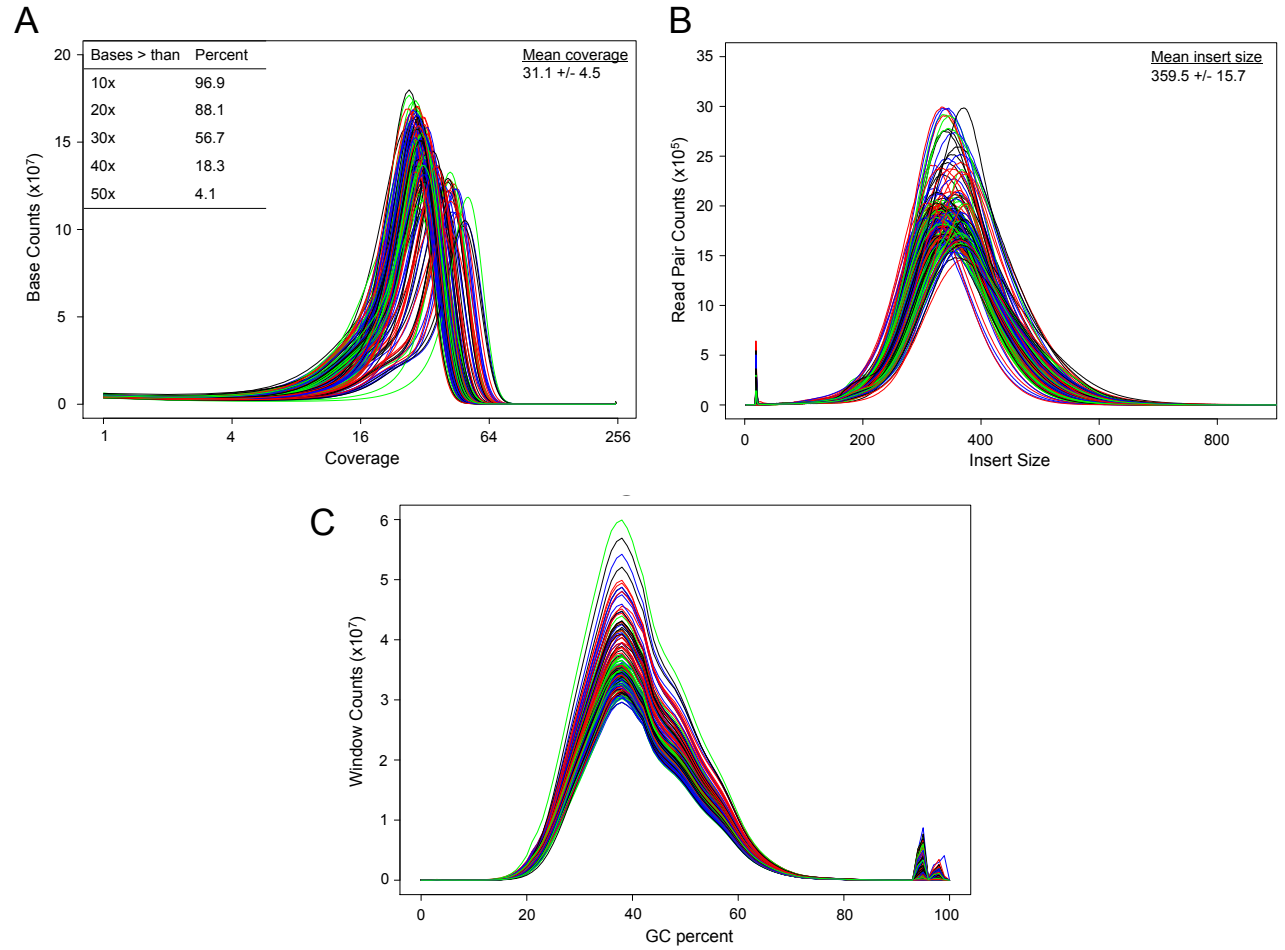

Figure S2: (A) Coverage of genome sequence data by sample. (B) Insert size metrics. (C) GC percent across genome data.

Figure S3: QC analysis.

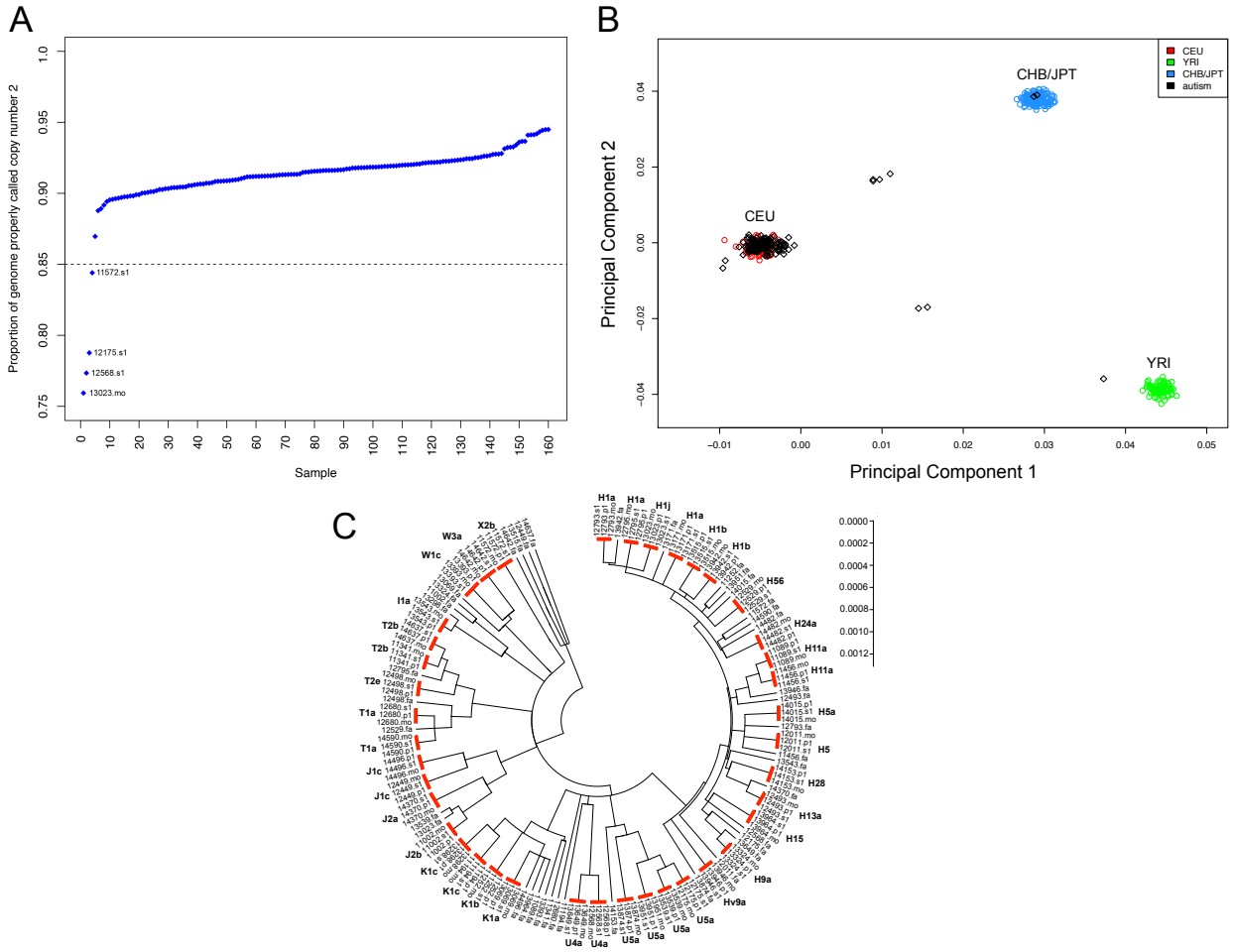

Figure S3: (A) Quality control based on read depth and proper calling of copy number 2 regions. Labeled are samples failing QC. (B) Principal component analysis of SNP data. Shown are three reference populations (CEU: European ancestry, CHB/JPT: Asian ancestry, YRI: African ancestry) and autism samples (shown in black). (C) Neighbor-joining tree of full mitochondrial genomes. Red bars indicate mother, proband, and sibling genomes from a family and haplogroups are labeled as well.

Figure S4: Exome versus genome: concordance.

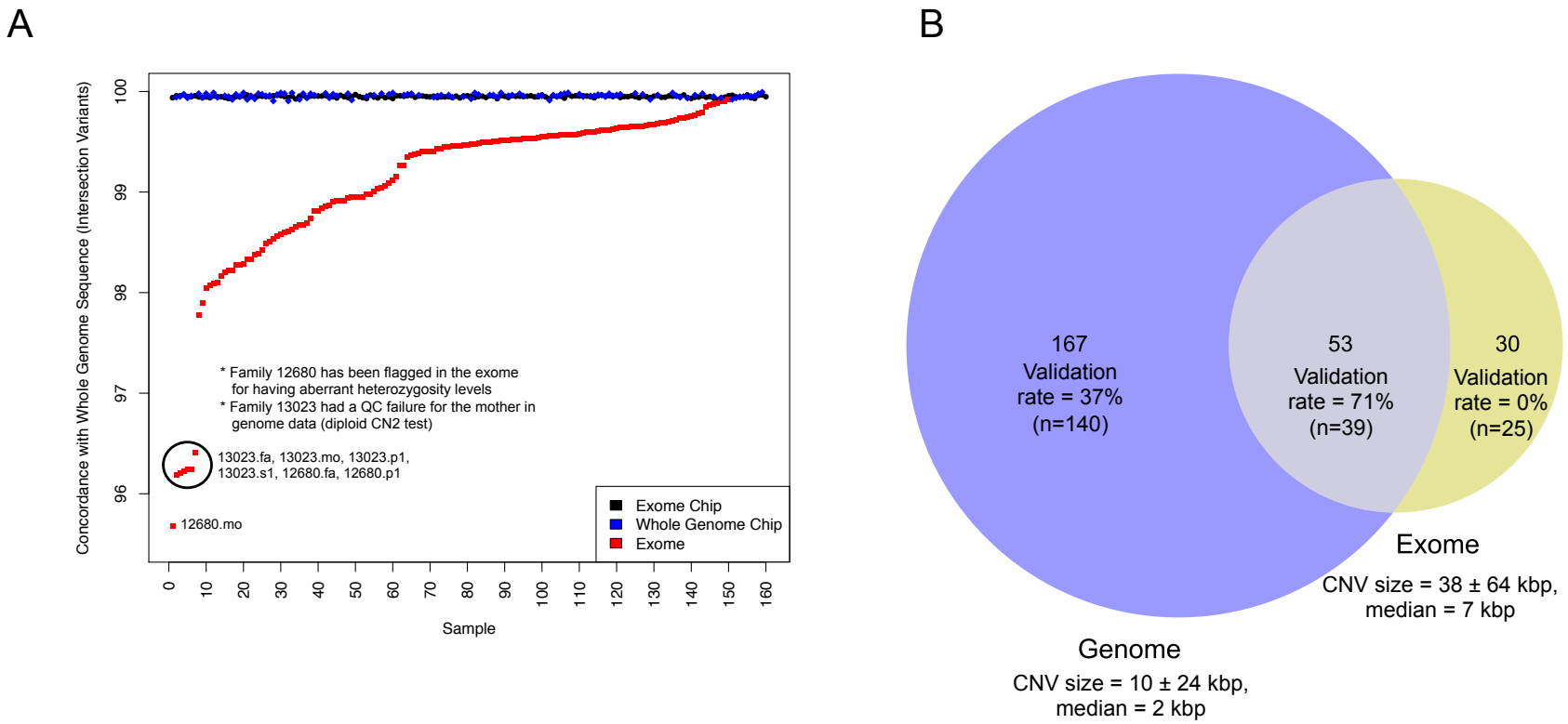

Figure S4: (A) Genotype concordance of genome sequence variants with exome chip, whole-genome chip, and exome data (B) Venn diagram of exome and genome CNV sites within the NimbleGen 36 Mbp exome capture region. Specifically, we compared CNV calls from this study generated by dCGH, GenomeSTRiP, and VariationHunter with calls from exome sequencing analysis [1] made by CoNIFER [2] and XHMM [3]. Calls made by the genome CNV tools are based on read depth, read pair, and split read information while calls made by exome CNV tools are based exclusively on read depth.

Figure S5: Permutation testing to assess clustering of de novo mutations in 100 kbp windows.

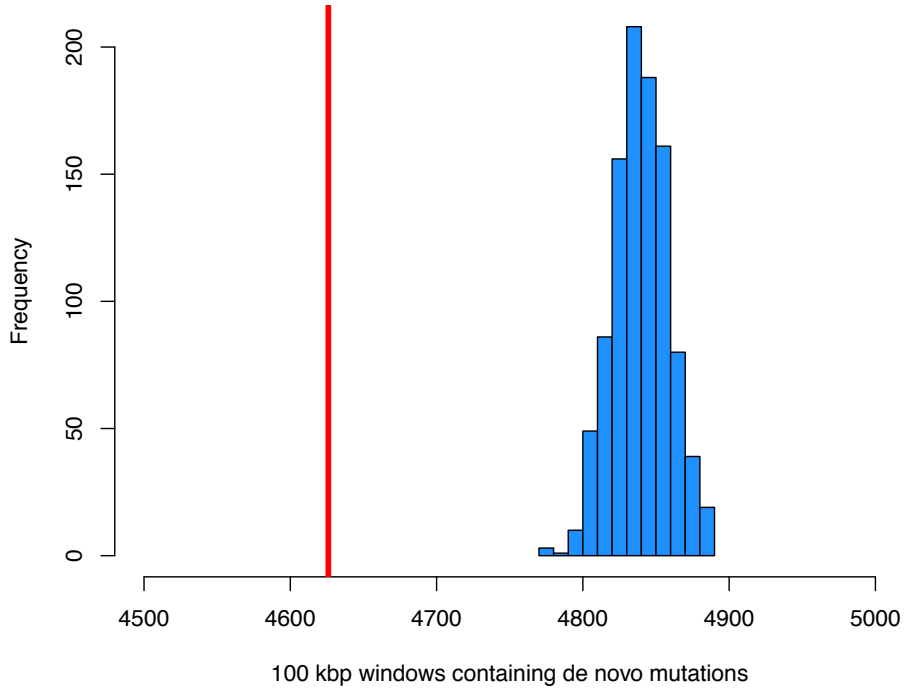

Figure S5: The red line shows the observed number of 100 kbp windows in which de novo mutations from this current study reside. The blue histogram represents the number of windows in which randomly placed de novo mutations reside (1000 permutations shown). The observed number of windows is far less ($p < 1x10^{-3}$) than what is expected if de novo mutations were randomly placed around the genome. The average distance between de novo events was 26.8 ± 31.2 Mbp and there were 189 new mutations where the next nearest de novo mutation within the sample mapped within 1 kbp.

Figure S6: Ranked order of feature importance for de novo variants based on random forest analysis.

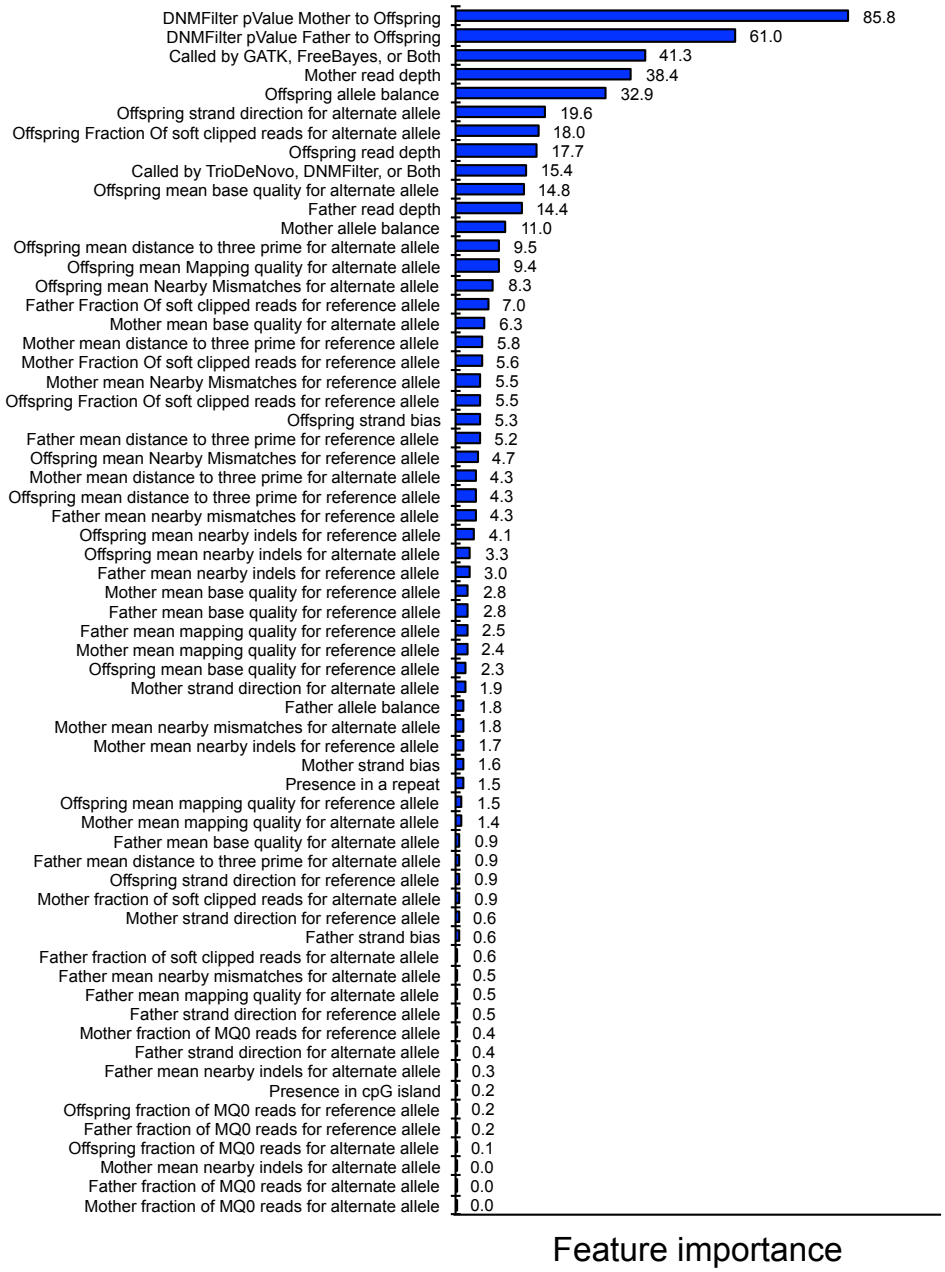

Figure S6: In our assessment of de novo variants we utilized two callers and attempted validation on all exomic and putative regulatory events. We also attempted validation by MIP sequencing on all events in five families. In total, we were able to assess validation status at **1,330** (986 validated and 344 not validated [false positive], validation rate = 74.1%) sites. To determine what may be happening in the ~25% of events that do not validate, we utilized the extract tool in DNMFilter to identify, from the original BAM files, the 59 features used by DNMFilter in its model [4]. In addition we considered four other features: (1) the initial caller (GATK, FreeBayes), (2) if the site maps to a CpG island, (3) if the site is in a repeat, and (4) the type of de novo caller (TrioDeNovo [5], DNMFilter [4]). In total, there were 63 features gathered for each valid / not valid (false positive) site. Using the random forest model, we derived a ranked list of the importance of features.

Figure S7: Distribution of Phred-scaled p-values in events that were validated as de novo (Valid) and those that were not (False positive).

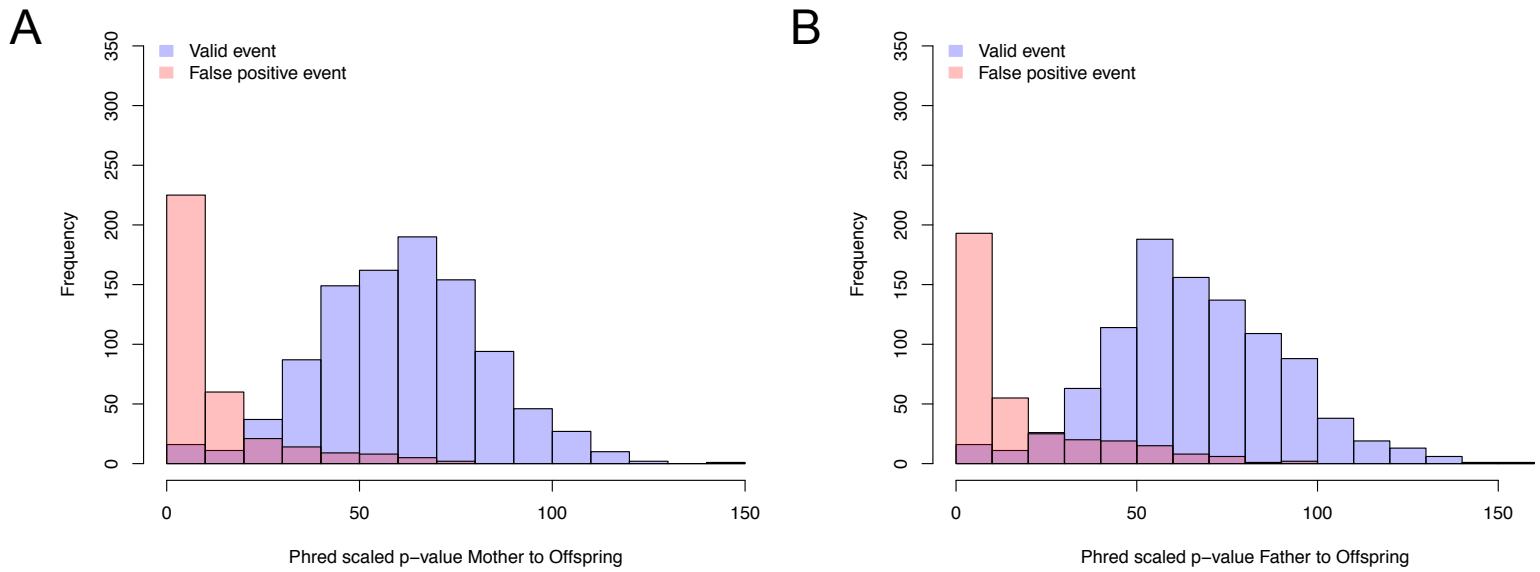

Figure S7: Validated events showed much higher Phred-scaled p-values. The two most important features based on the random forest model for invalidated variants was evidence of inheritance from the mother or father. This is represented as a p-value derived by DNMFilter and is defined in the DNMFilter paper as "the Phred-scaled P-value of a Fisher's exact test for father/mother and offspring, alt alleles versus ref alleles (two values)." Thus, the most common cause for invalidation was under-calling in one of the parents and a false classification as de novo. This could be easily remedied by considering relative number of alternate allele callers or by considering the DNMFilter p-values. These findings are consistent with the large number of false de novo calls that were found to be inherited during validation.

Figure S8: Counts of variants by initial caller in events that were validated as de novo (Valid) and those that were not (False positive).

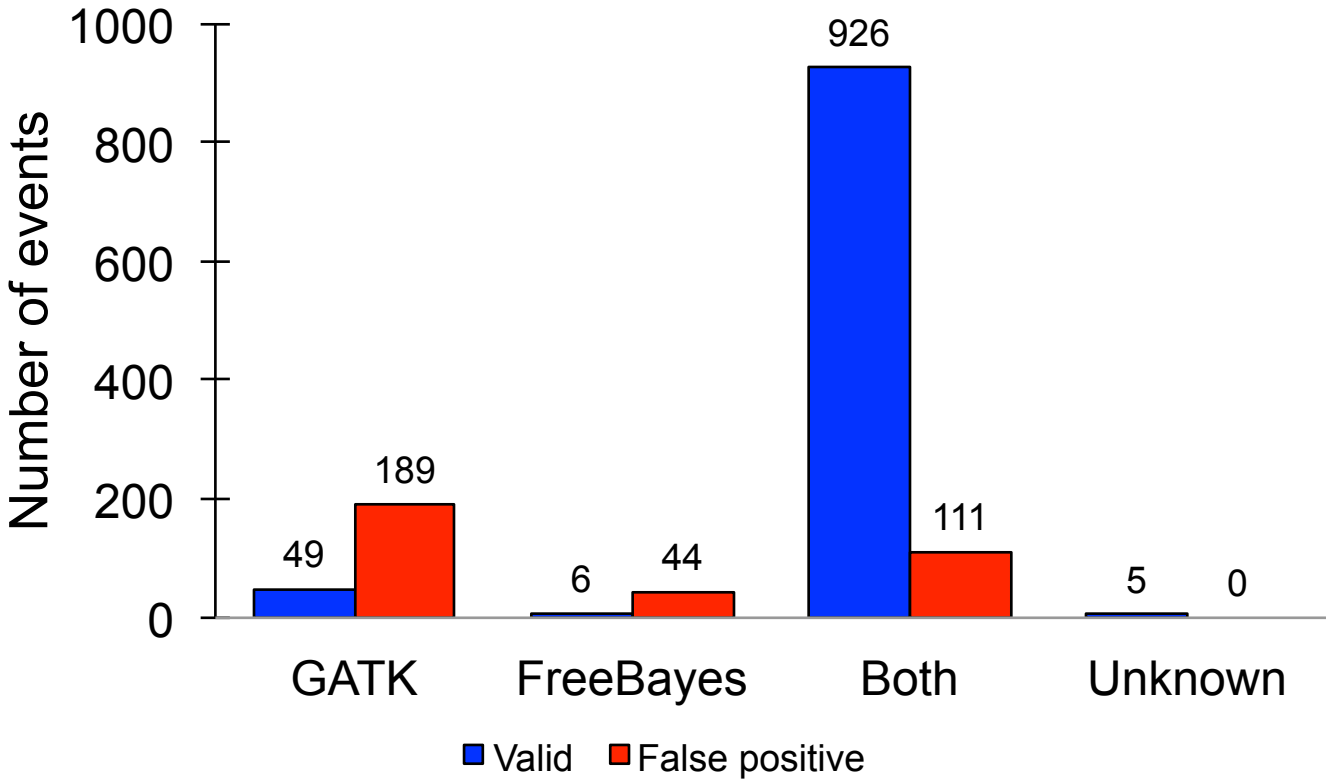

Figure S8: The utility of using both callers increased validation. The type of caller (either GATK or FreeBayes) was identified as another important feature of de novo variant validation. There is, however, higher false positive in either FreeBayes- or GATK-only call sets despite the fact that you recover additional de novo variants. Unknown refers to WES-specific sites.

Figure S9: Additional features important for events that were validated as de novo (Valid) and those that were not (False positive).

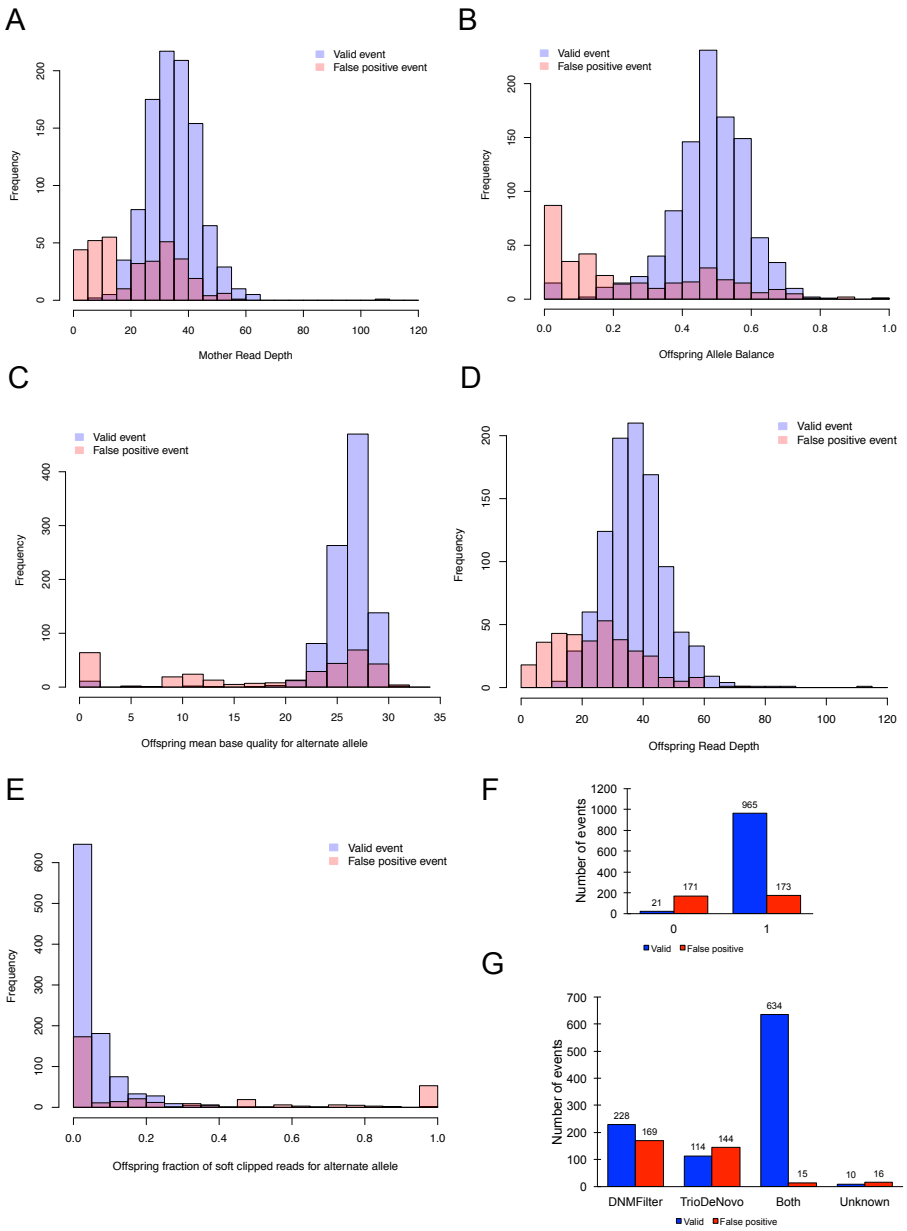

Figure S9: Included are (A) mother's read depth at the site; (B) offspring allele balance at the site; (C) offspring mean base quality for alternate allele; (D) offspring read depth; (E) offspring fraction of soft-clipped reads for alternate allele; (F) offspring strand direction at site where 0 indicates all reads are on the same strand and 1 shows presence on both strands; and (G) number of events based on the de novo caller (unknown refers to WES-specific sites).

Figure S10: Conditional inference tree for valid de novo and false positive de novo events.

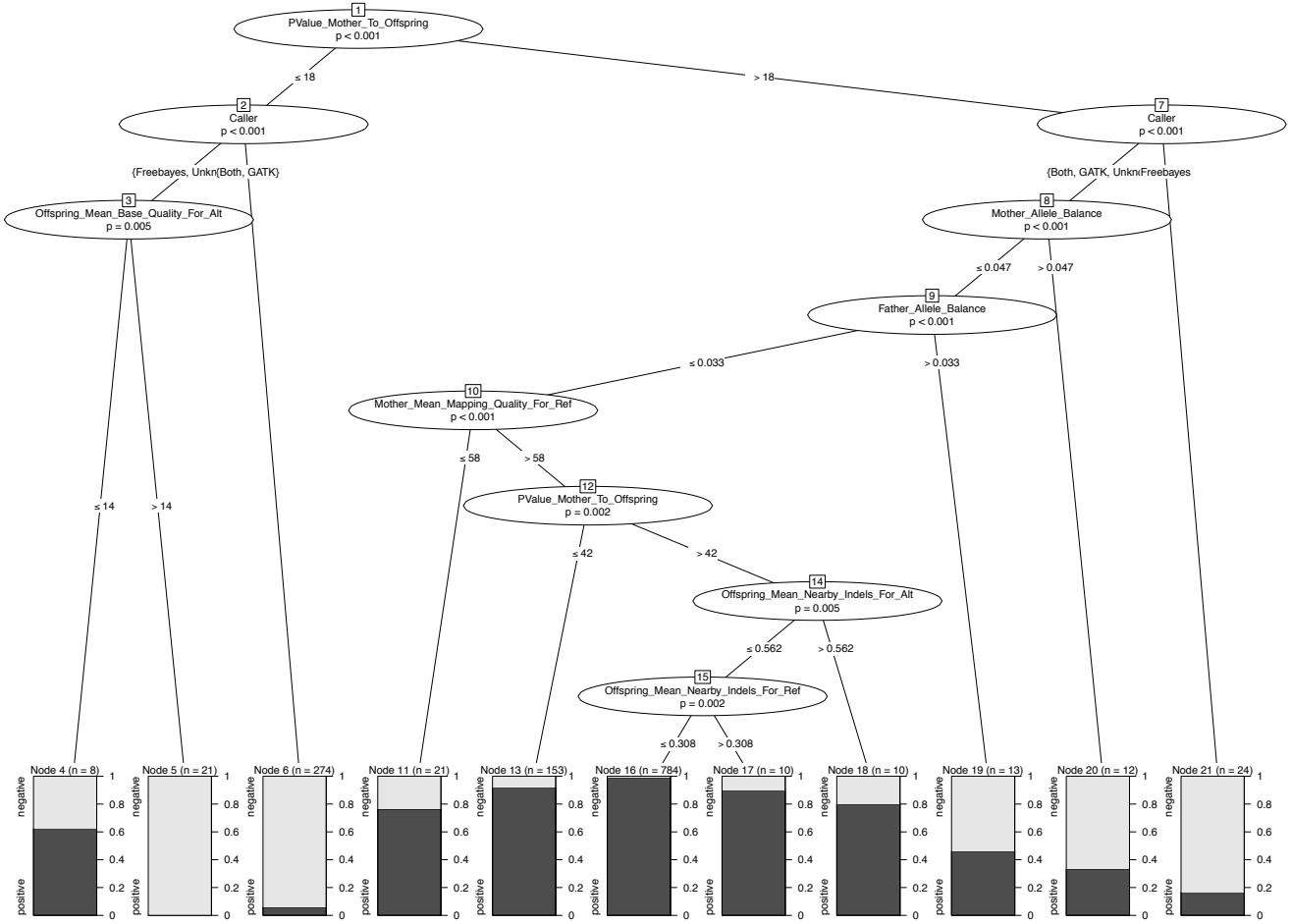

Figure S10: Shown is the best tree generated by the party R package. By following the paths in the tree the number of valid de novo (positive) and not valid de novo (negative) sites can be seen (at the bottom). We were able to generate a conditional inference tree (PARTY [6]) to guide researchers on the precise conditions to maximize discovery of future events. Similar to the random forest method, the tree indicated that the p-value test in the mother and the initial caller were each very critical features in the decision tree. Shown is the entire conditional inference tree and at the bottom is the proportion of validated events (positive) and invalidated events (negative) for each path of the tree. While individual researchers can make their own decision, our results indicate that read depth (Phred >18) and allele balance will discover the maximum set of true variants.

Figure S11: CNV validation.

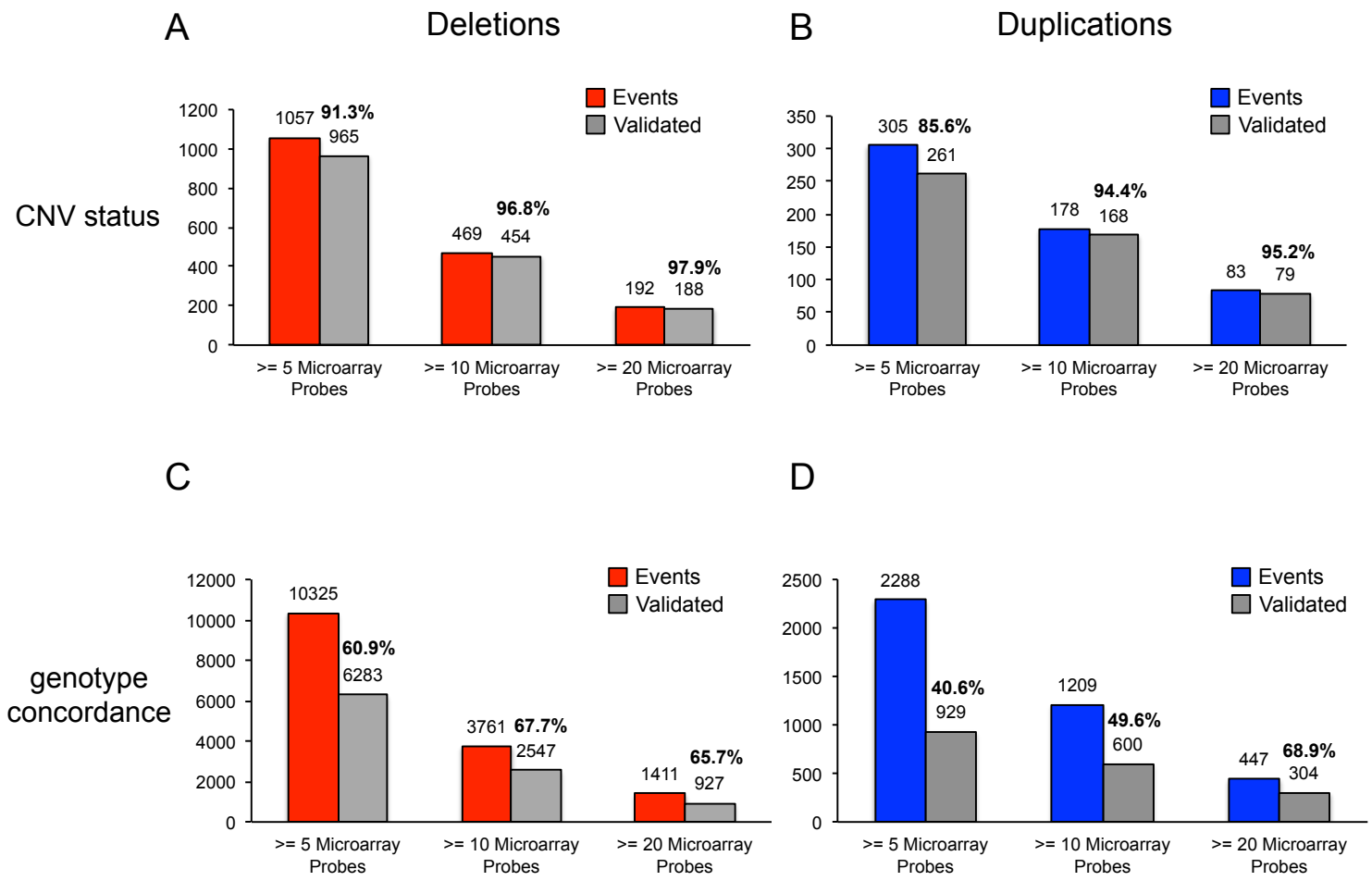

Figure S11: (A) CRLMM validation of deletion sites (B) CRLMM validation of duplication sites (C) Genotype concordance for deletions (D) Genotype concordance for duplications.

Figure S12: Exome versus genome: uniformity analysis.

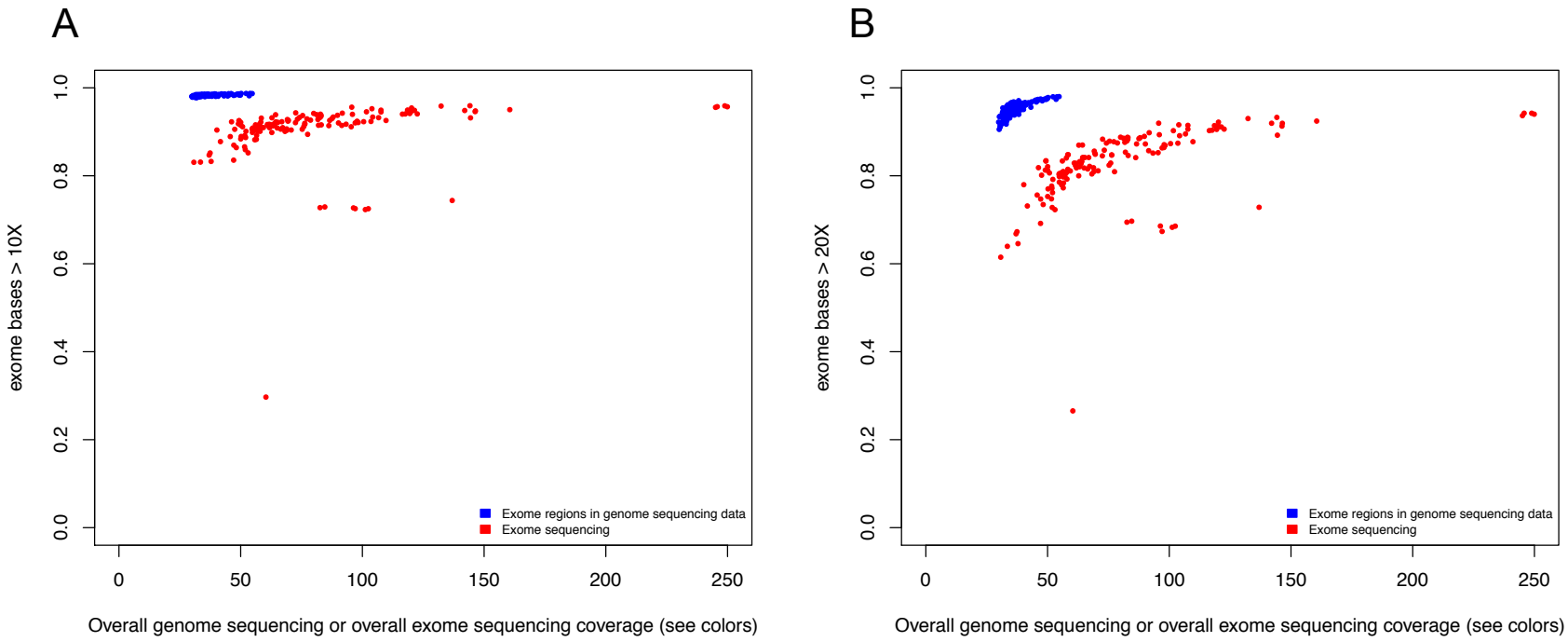

Figure S12: (A) Percent of bases **>10X** in exomic regions from genome sequencing and exome sequencing (B) Percent of bases **>20X** in exomic regions from genome sequencing and exome sequencing. Although WGS showed 36.6 ± 5.4-fold sequence coverage when compared to 81.2 ± 38.6-fold coverage depth by WES, the percent of basepairs with at least 10-fold coverage was greater for WGS (98.3 ± 0.2% vs. 90.4 ± 6.9% for WES) consistent with a more uniform coverage by WGS [7; 8]. As a result, we estimate that an additional 2,126 kbp of exome target was recovered by WGS compared to 42.3 kbp of the exome recovered only by WES.

Figure S13: Genome and exome sequencing identify unique SNV/indel events.

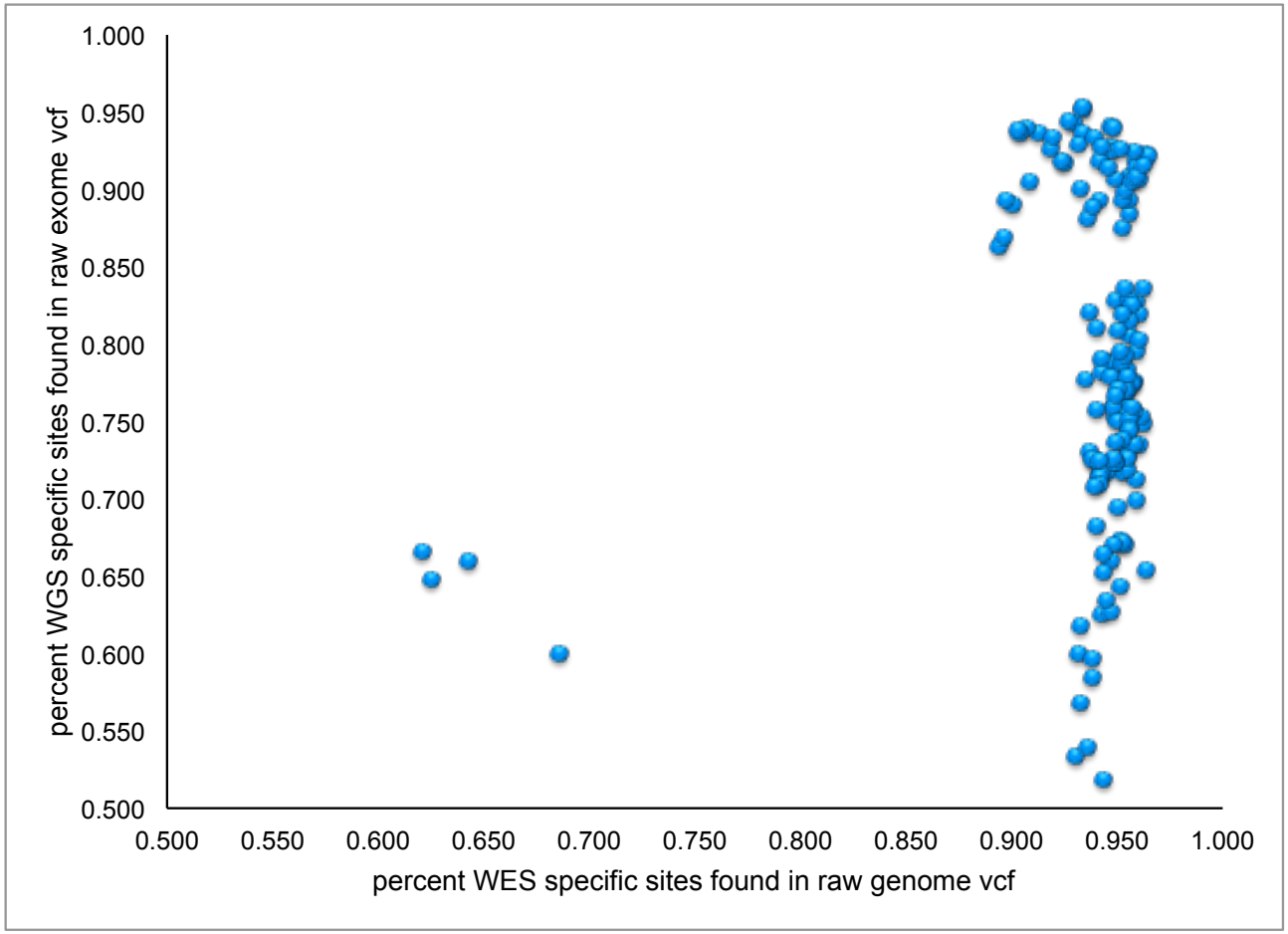

Figure S14: Genome and exome sequencing identify unique CNV events.

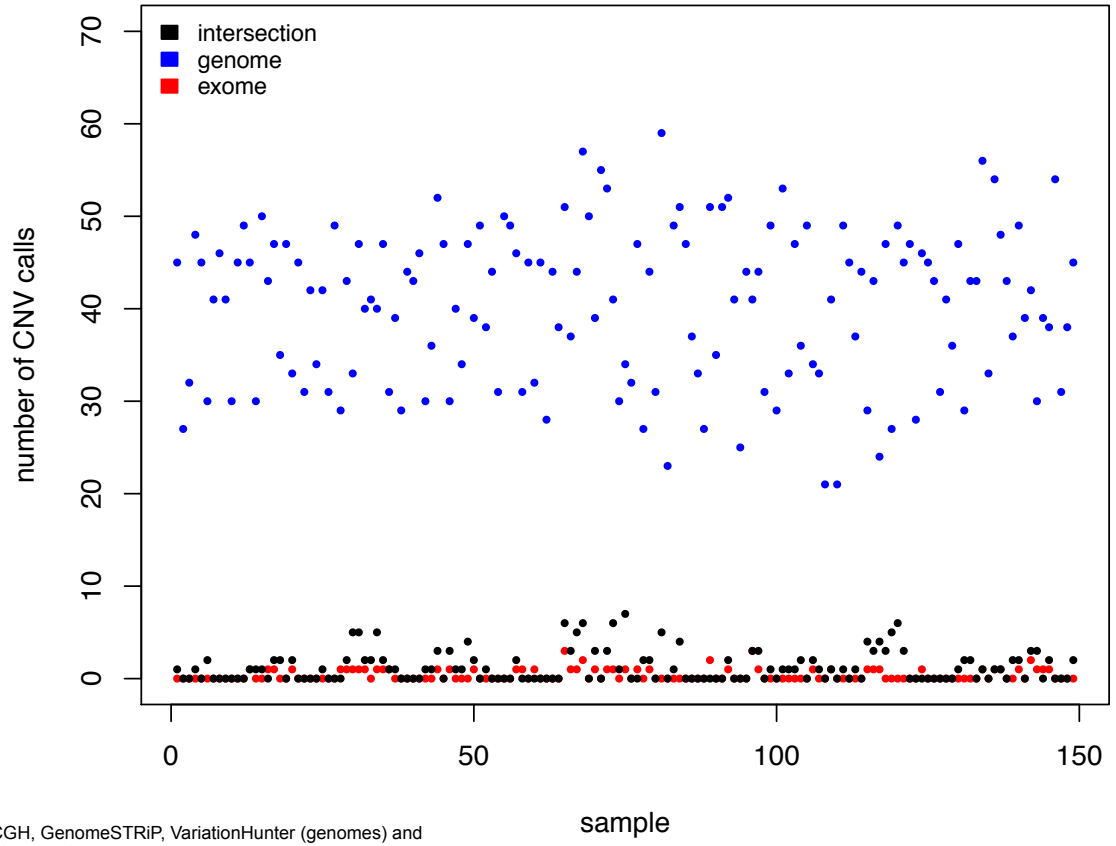

Figure S14: Shown are CNVs found in the exome in exome sequencing only, genome sequencing only, and in both exome and genome sequencing. Genome calls were made by dCGH, GenomeSTRiP, and VariationHunter and exome calls by CoNIFER and XHMM. As expected, genome datasets significantly enhance CNV detection. Overall, 167 CNV sites (67%) were called exclusively by WGS, 30 (12%) by WES only, and 53 (21%) by both (Figure S4b) with considerable variability by sampleWe used the SNP microarray validation approach to fairly assess validation rates in each of these sets and found that the intersection had highest validation (validation rate = 71%, n=39), followed by WGS-specific (validation rate = 37%, n=140) and lastly WES-specific events (validation rate = 0% (n=25).

Figure S15: CNV calls detected by WGS and by SNP microarray.

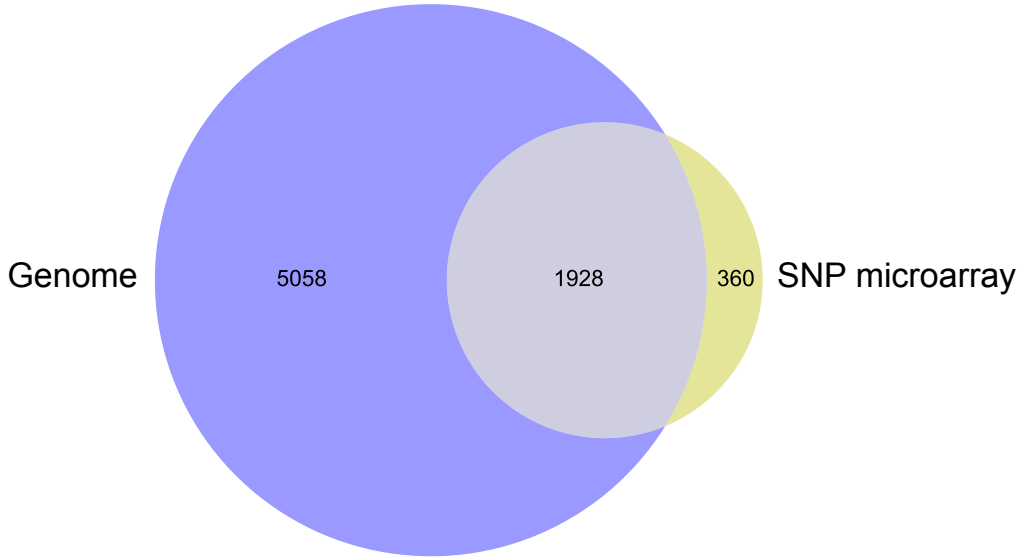

Figure S15: Shown is a Venn diagram of calls identified by one technology or the other and also those detected by both technologies.

Figure S16: Duplications in *SAE1* in autism patients from Lionel et al. 2011, Prasad et al. 2012, and the current study.

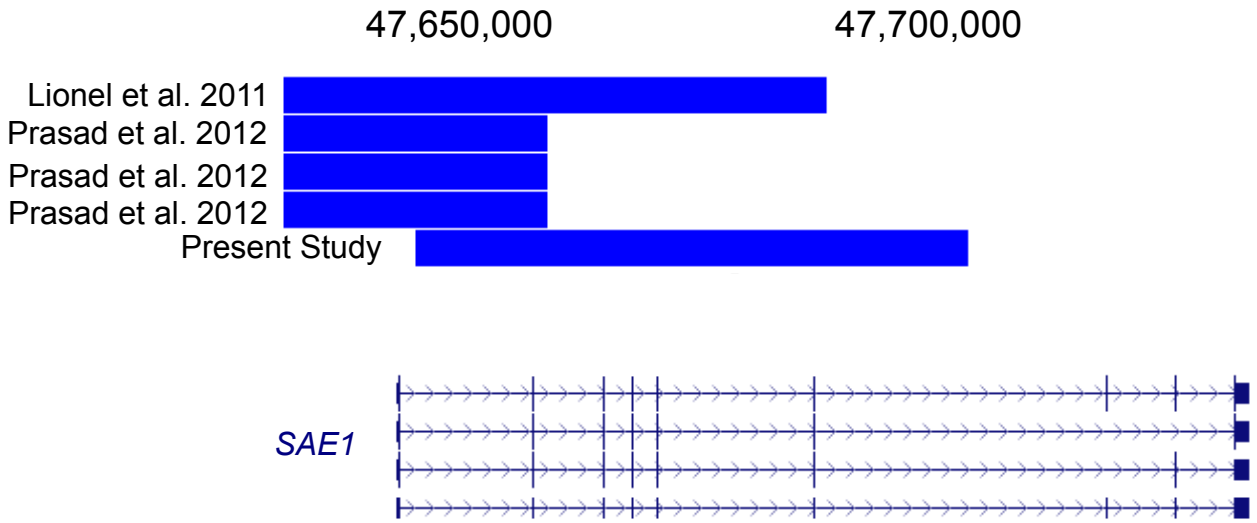

Figure S17: Supplemental images of all other constructs from the functional analysis of CNS DNase I hypersensitivity sites in DSCAM deletion.

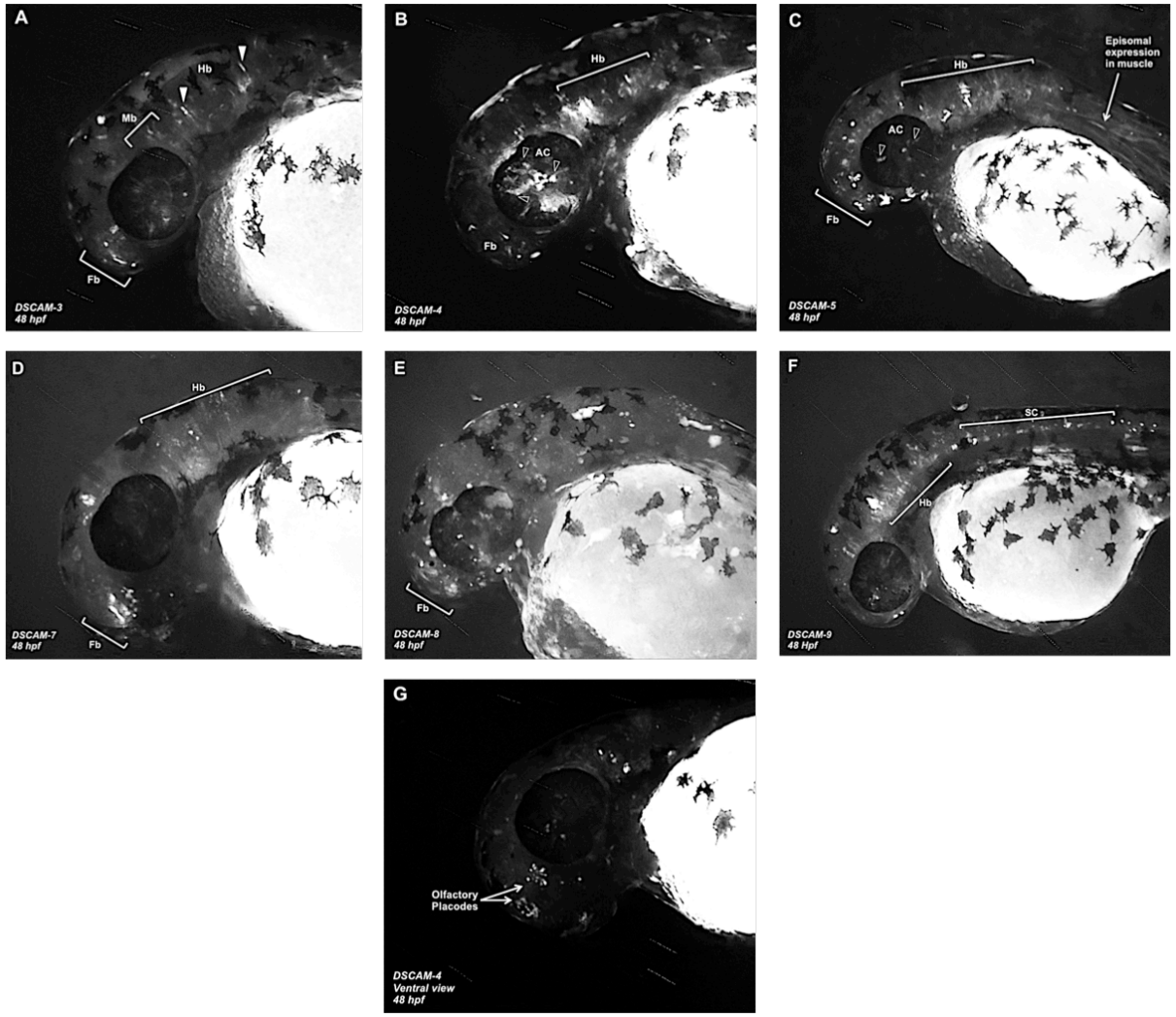

Figure S17: Fb – Forebrain, Mb – Midbrain, Hb – Hindbrain, Am – Amacrine cells, SC – Spinal cord, OP – Olfactory placode

Supplemental References

1. Krumm, N., Turner, T.N., Baker, C., Vives, L., Mohajeri, K., Witherspoon, K., Raja, A., Coe, B.P., Stessman, H.A., He, Z.X., et al. (2015). Excess of rare, inherited truncating mutations in autism. Nature genetics.
2. Krumm, N., Sudmant, P.H., Ko, A., O'Roak, B.J., Malig, M., Coe, B.P., Quinlan, A.R., Nickerson, D.A., and Eichler, E.E. (2012). Copy number variation detection and genotyping from exome sequence data. Genome Res 22, 1525-1532.
3. Fromer, M., Moran, J.L., Chambert, K., Banks, E., Bergen, S.E., Ruderfer, D.M., Handsaker, R.E., McCarroll, S.A., O'Donovan, M.C., Owen, M.J., et al. (2012). Discovery and statistical genotyping of copy-number variation from whole-exome sequencing depth. American journal of human genetics 91, 597-607.
4. Liu, Y., Li, B., Tan, R., Zhu, X., and Wang, Y. (2014). A gradient-boosting approach for filtering de novo mutations in parent-offspring trios. Bioinformatics (Oxford, England) 30, 1830-1836.
5. Wei, Q., Zhan, X., Zhong, X., Liu, Y., Han, Y., Chen, W., and Li, B. (2014). A Bayesian framework for de novo mutation calling in parents-offspring trios. Bioinformatics (Oxford, England).
6. Hothorn, T., Hornik, K., and Zeileis, A. (2006). Unbiased Recursive Partitioning: A Conditional Inference Framework. Journal of Computational and Graphical Statistics 15, 651-674.
7. Meynert, A.M., Ansari, M., FitzPatrick, D.R., and Taylor, M.S. (2014). Variant detection sensitivity and biases in whole genome and exome sequencing. BMC bioinformatics 15, 247.
8. Belkadi, A., Bolze, A., Itan, Y., Cobat, A., Vincent, Q.B., Antipenko, A., Shang, L., Boisson, B., Casanova, J.-L., and Abel, L. (2015). Whole-genome sequencing is more powerful than whole-exome sequencing for detecting exome variants. Proceedings of the National Academy of Sciences.