The American Journal of Human Genetics, *91*

Supplemental Data

Discovery and Statistical Genotyping of Copy-Number Variation from Whole-Exome Sequencing Depth

Menachem Fromer, Jennifer L. Moran, Kimberly Chambert, Eric Banks, Sarah E. Bergen, Douglas M. Ruderfer, Robert E. Handsaker, Steven A. McCarroll, Michael C. O'Donovan, Michael J. Owen, George Kirov, Patrick F. Sullivan, Christina M. Hultman, Pamela Sklar, and Shaun M. Purcell

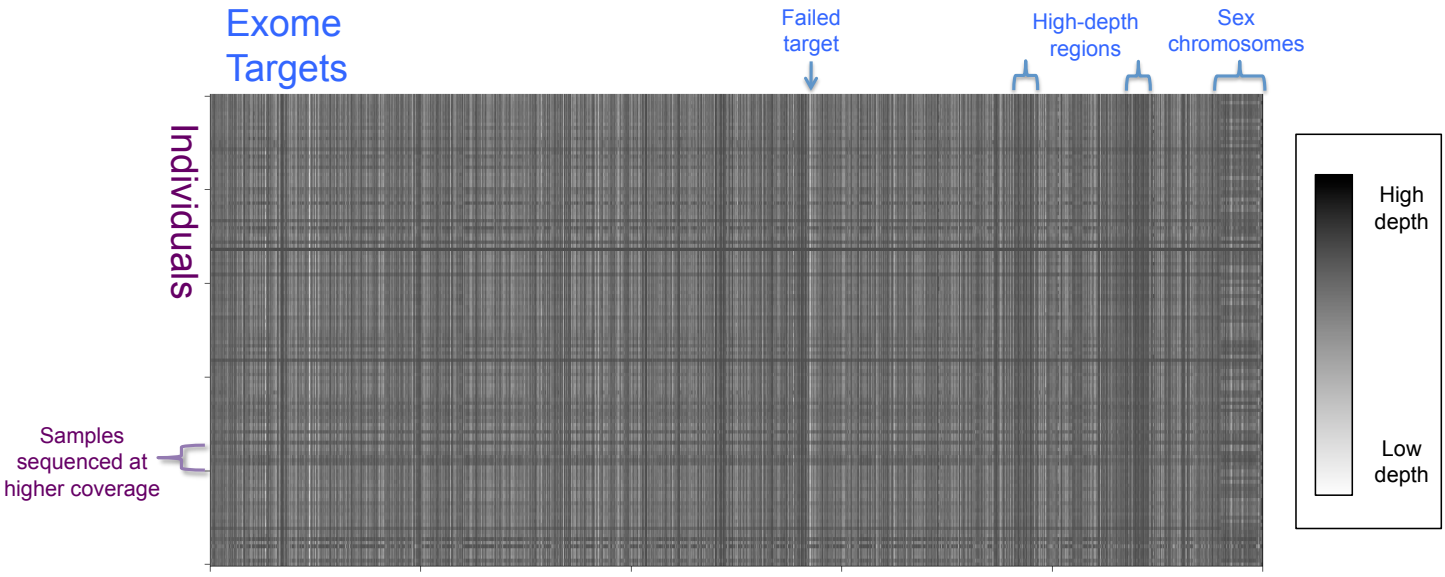

Figure S1. Depth of Coverage per Exon Target and per Sample

Mean per-target coverage is depicted for the subset of the first 130 case-control individuals (rows) by 160,000 exon targets (columns). The read depth shows clear systematic effects of target, region, chromosome, individual samples, and sample batching.

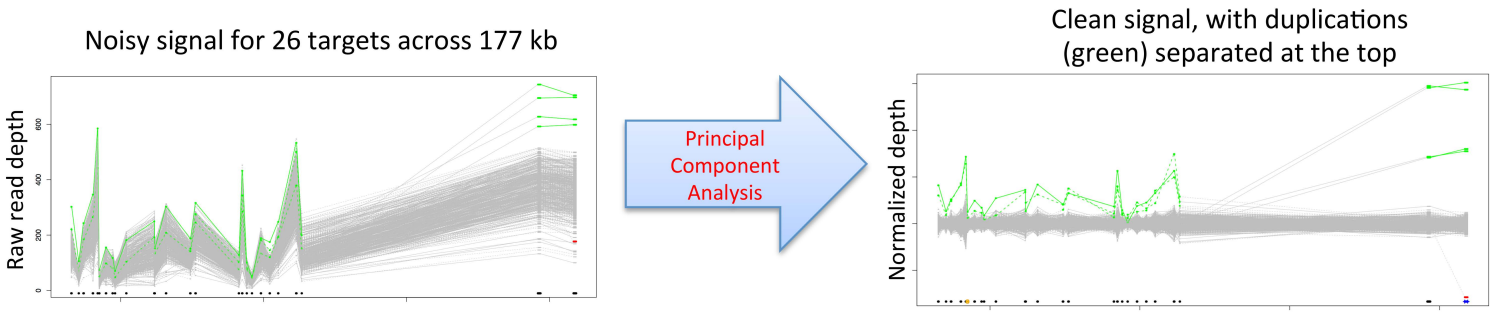

Figure S2. PCA Normalization Cleans up the Read Depth Data

The read depths for 1017 case-control samples in a region of 26 exon targets (marked by black dots on the bottom) across 177 kb are shown as directly calculated from the sequencing data (left) and after normalization and denoising using PCA - principal component analysis (right). In both panels, each line segment corresponds to the depths for a single individual, and depths corresponding to XHMM diploid calls are marked in gray and duplications in green. By taking read depths that vary by two orders of magnitude (left) and constraining most values to be uniform, XHMM's use of PCA permits the duplications to become apparent as having consistently high depth for a small number of samples (right).

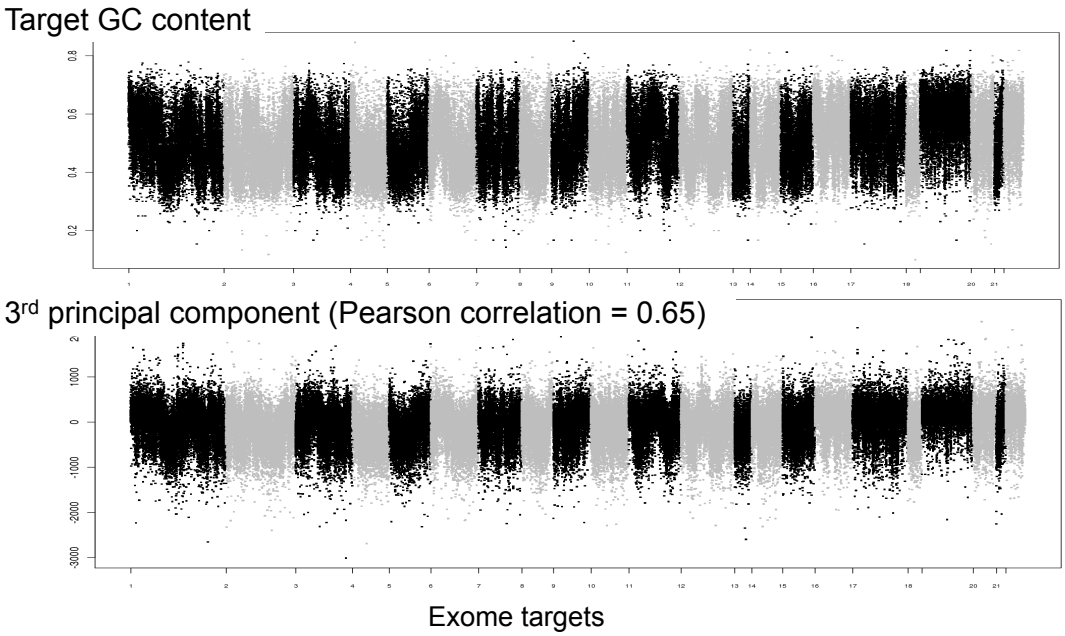

Figure S3. Unbiased PCA-Based Normalization Also Finds Factors Known to Affect Read Depth

Though the principal component analysis (PCA) does not require any pre-defined user input as to the sources of read depth variation and noise, it is still able to pick up expected confounders, such as GC content. In this case, the third principal component has a correlation of 0.65 with the per-target GC content, but note that GC content also loaded onto other components as well. Chromosome targets are colored in alternating black and gray.

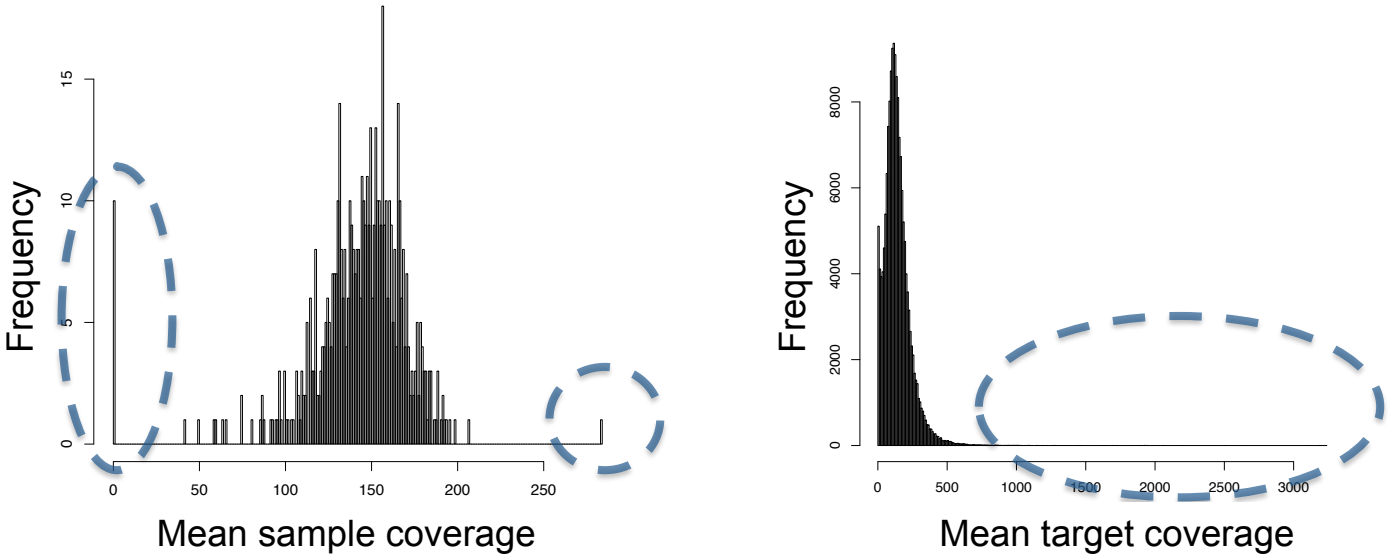

Figure S4. Distributions of Mean Read Depths

Left: Distribution of sample coverages, averaged over all targets. Right: Distribution of target coverages, averaged over all samples. Dotted lines mark samples and targets that were excluded based on consideration of atypically extreme read depths. The data presented here is for the first 130 samples of the Swedish case-control data.

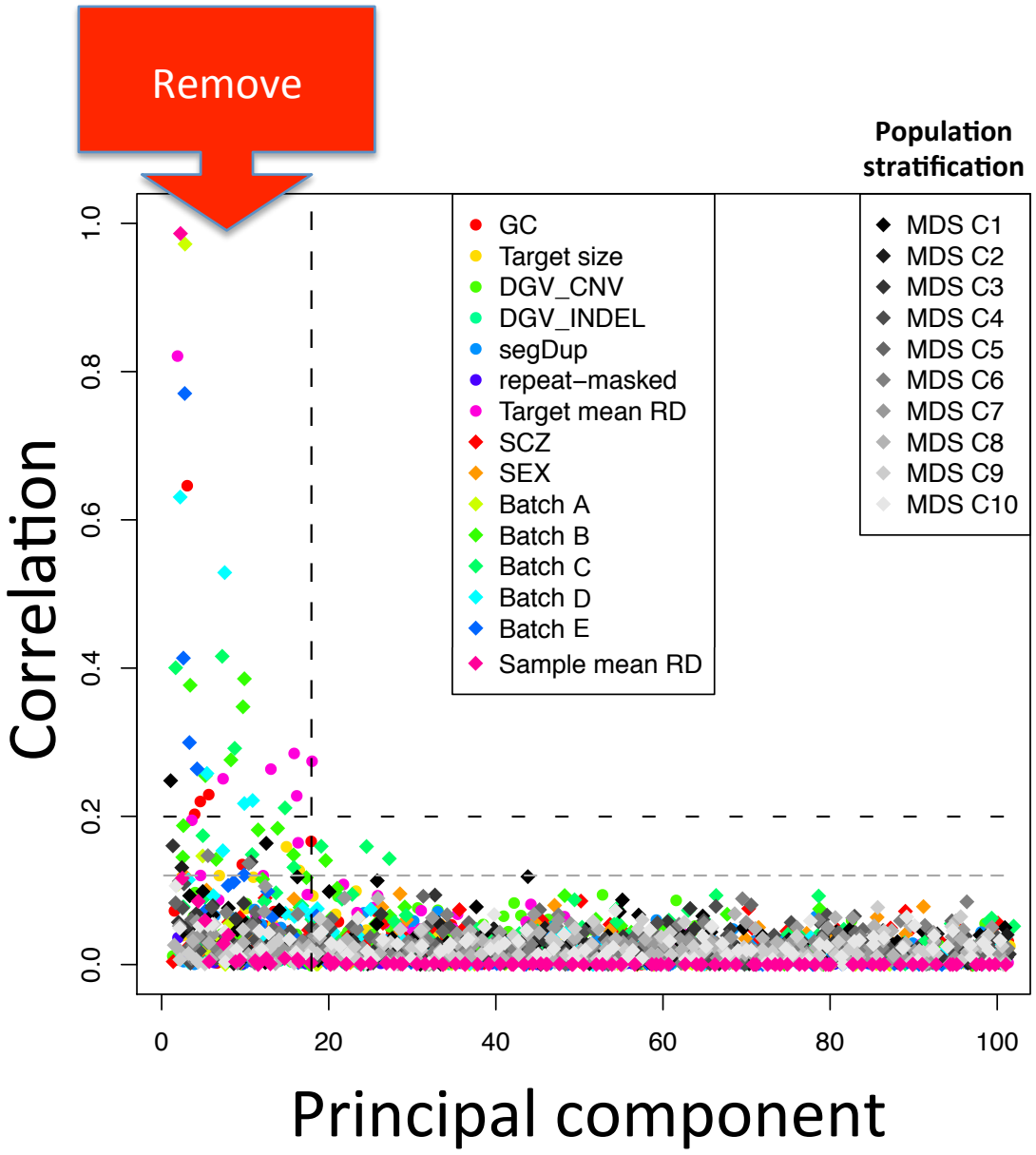

Figure S5. Comparison of the First 100 Principal Components for the Case-Control Read Depth Matrix and Predefined Phenomena Expected to Be Possibly Correlated

These factors include target-specific trends: GC content, target size, known CNV density (DGV_CNV), known indel density (DGV_INDEL), segmental duplication density, fraction of sequence that is repeat-masked, and mean target read depth; and sample-specific trends: case-control status (SCZ), gender (SEX), sample batch identity, mean read depth for the sample, and the 10 strongest population components (from PLINK's multi-dimensional scaling calculations, MDS, on GWAS SNP data). The first 18 principal components (demarcated by vertical line) show strong correlation with mean sample depth, batch effects, target GC content, and mean target depth, and these were removed in the automated normalization procedure.

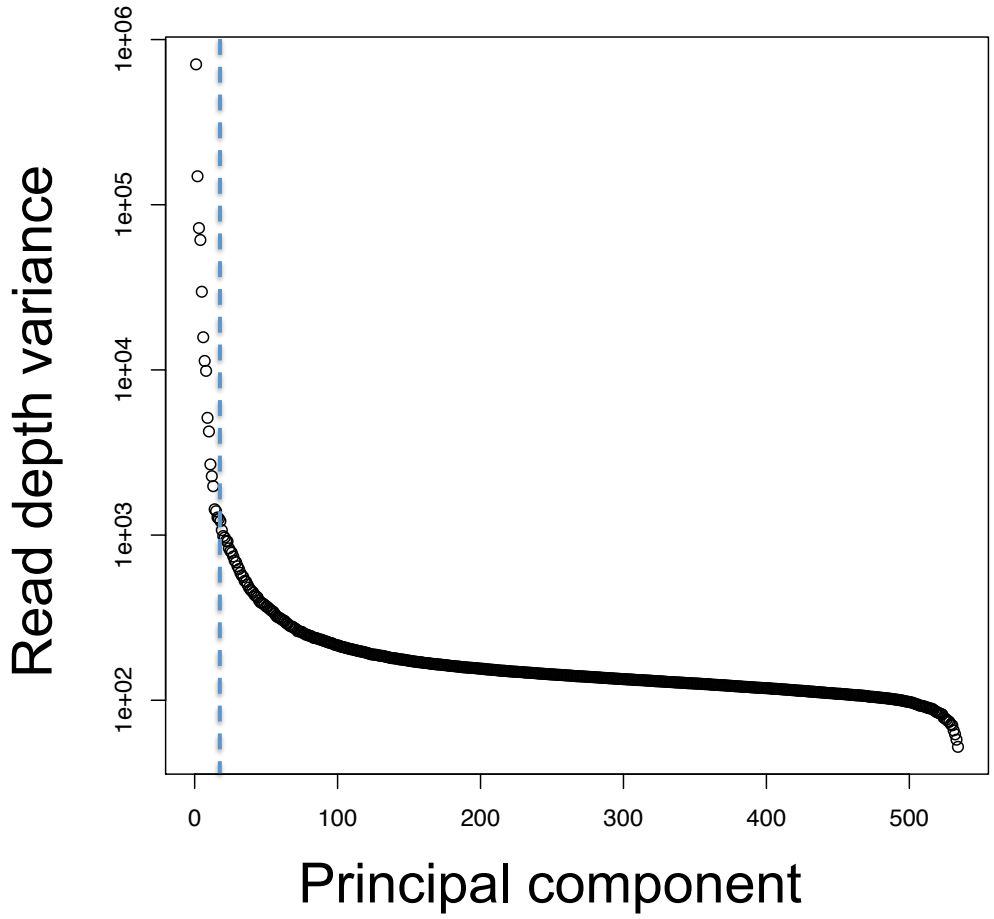

Figure S6. "Scree" Plot of Read Depth Variance Contributed by Each of the First 500 Principal Components for the Case-Control Data

The first 18 principal components (demarcated by vertical line) contribute exponentially more to the read depth variation than subsequent components and were removed in the automated normalization procedure. Note the log-scale of the y axis.

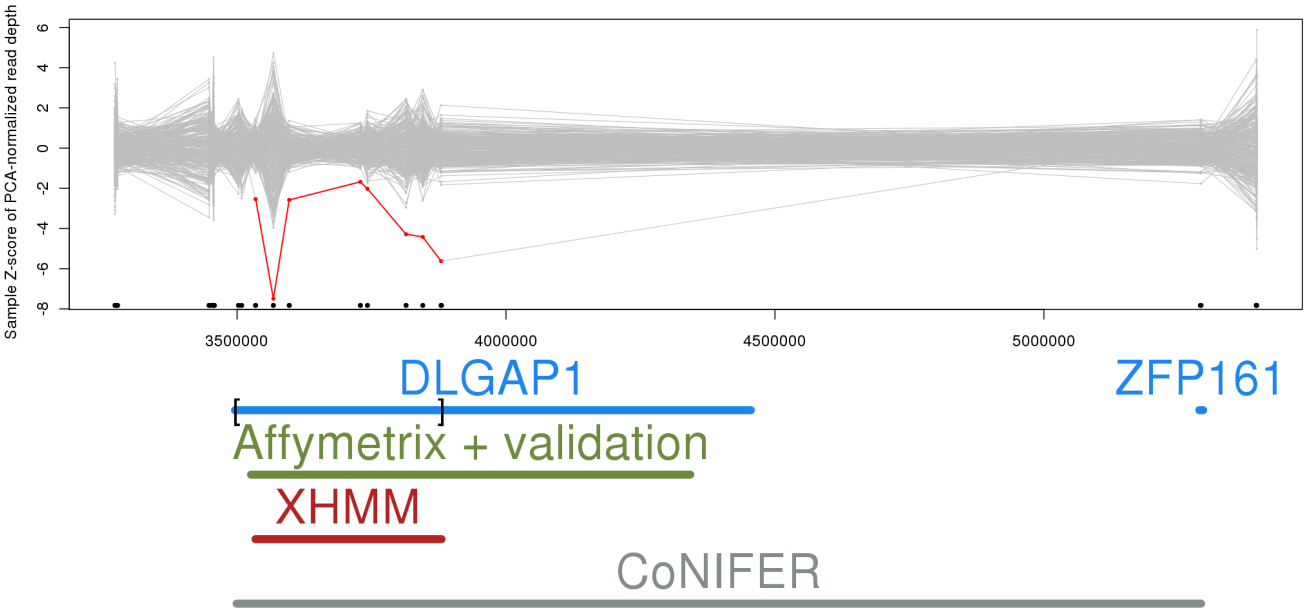

Figure S7. XHMM-Normalized Read Depths for De Novo Deletion in DLGAP1 (Discs, Large Homolog-Associated Protein 1)

Deletion is depicted as a red line, with depth well below the other samples (gray). The full extent of the genes and the validated Affymetrix-based *de novo* call are as marked, showing high overlap with the XHMM call. The exome-targeted region of each gene is delineated by brackets, and individual exome targets are marked by black dots on the bottom of the axis, indicating that the *DLGAP1* validated Affymetrix deletion overlaps 8 targets. The corresponding CoNIFER call is shown as a solid gray bar on bottom. XHMM exactly captures the exomic overlap of the validated Affymetrix-based call, whereas CoNIFER adds 3 targets on the 5′ end and 1 target on the 3′ end, yielding a call of 179 kb instead of the validated 82 kb. Also, note that the CoNIFER call now incorrectly overlaps a second gene (*ZFP161* [MIM 602126]). XHMM assigns a quality value of 57 to the 3′ breakpoint, indicating that it is highly likely that the breakpoint does not occur somewhere else.

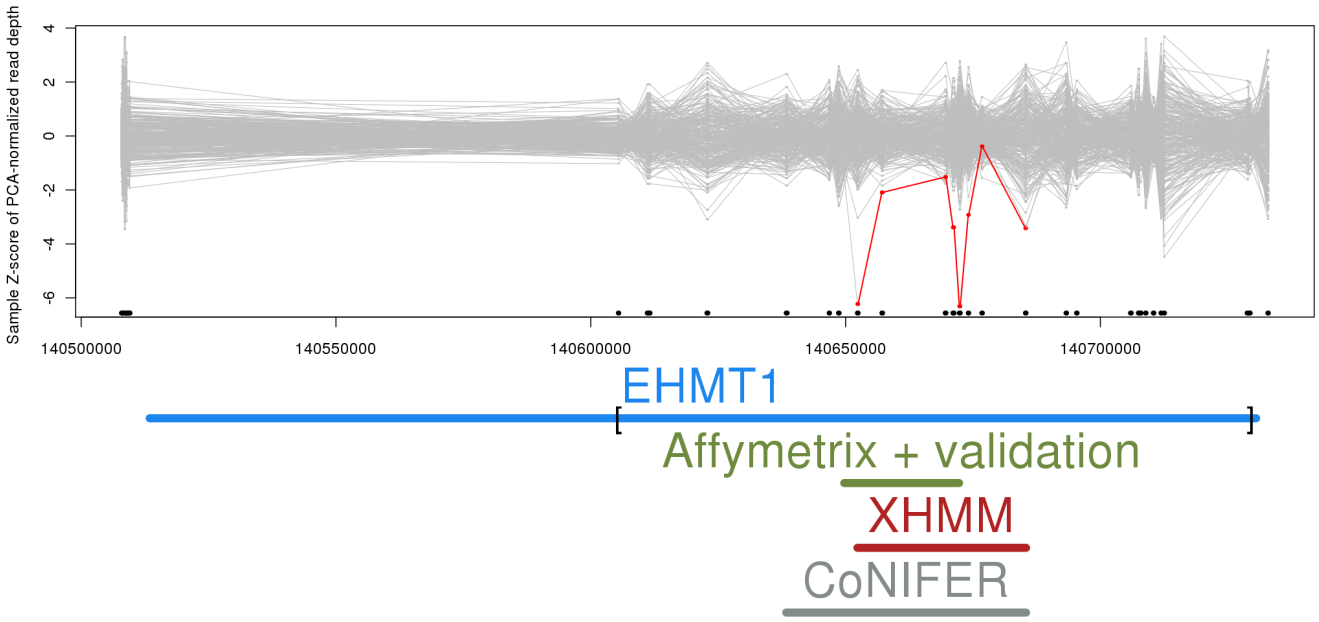

Figure S8. XHMM-Normalized Read Depths for De Novo Deletion in EHMT1 (Euchromatic Histone-Lysine N-methyltransferase 1)

Deletion is depicted as a red line, with depth well below the other samples (gray). The full extent of the gene and the validated Affymetrix-based *de novo* call are as marked, showing high overlap with the XHMM call. The exome-targeted region of the gene is delineated by brackets, and individual exome targets are marked by black dots on the bottom of the axis, indicating that the *EHMT1* validated Affymetrix deletion overlaps 4 targets. The corresponding CoNIFER call is shown as a solid gray bar on bottom. Note that, whereas XHMM adds only 47% in size to the validated Affymetrix-based call (adding 4 targets on the 3′ side), CoNIFER adds the same set of 3′ targets as well as another 3 targets on the 5′ side, more than doubling the length of the CNV as compared to the validated call.

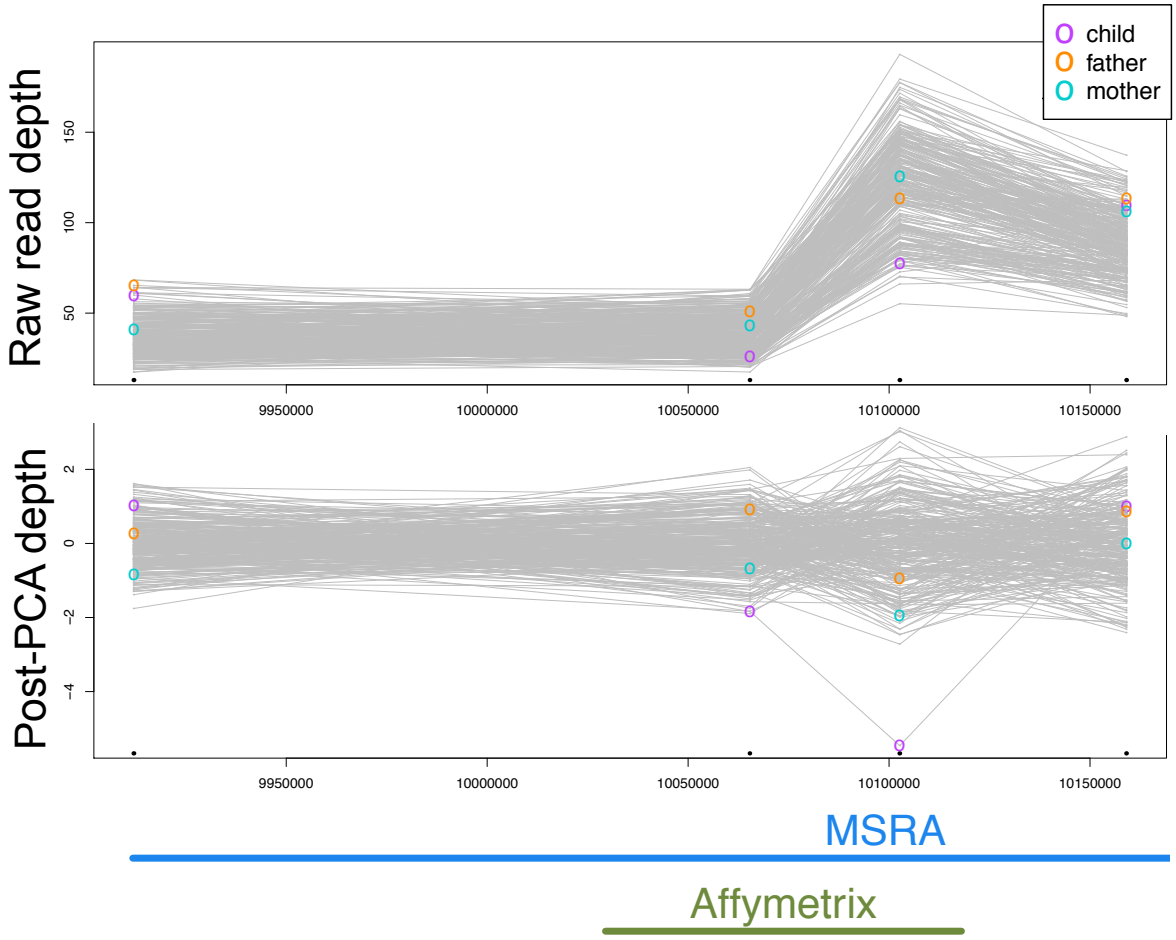

Figure S9. XHMM-Normalized Read Depths for Affymetrix-Based De Novo Deletion in MSRA (Methionine Sulfoxide Reductase A [MIM 601250])

Whereas no clear CNV would be apparent from the original read depths (top), a sub-threshold *de novo* deletion in a single target of *MSRA* is consistent with the validated Affymetrix-based deletion depicted. Child, father, and mother read depths are colored with open circles, highlighting the large gap between parental and child depth. Exome targets are marked by black dots on the bottom of the axis, indicating that the validated Affymetrix deletion overlaps only 2 targets, partly explaining why this event was not actually called by XHMM (see Table 3 and the *q* HMM parameter in Methods). Note also that, in our comparisons, CoNIFER did not make a call here either (though the validated call does not overlap the minimum of 3 targets that CoNIFER requires to make a call).

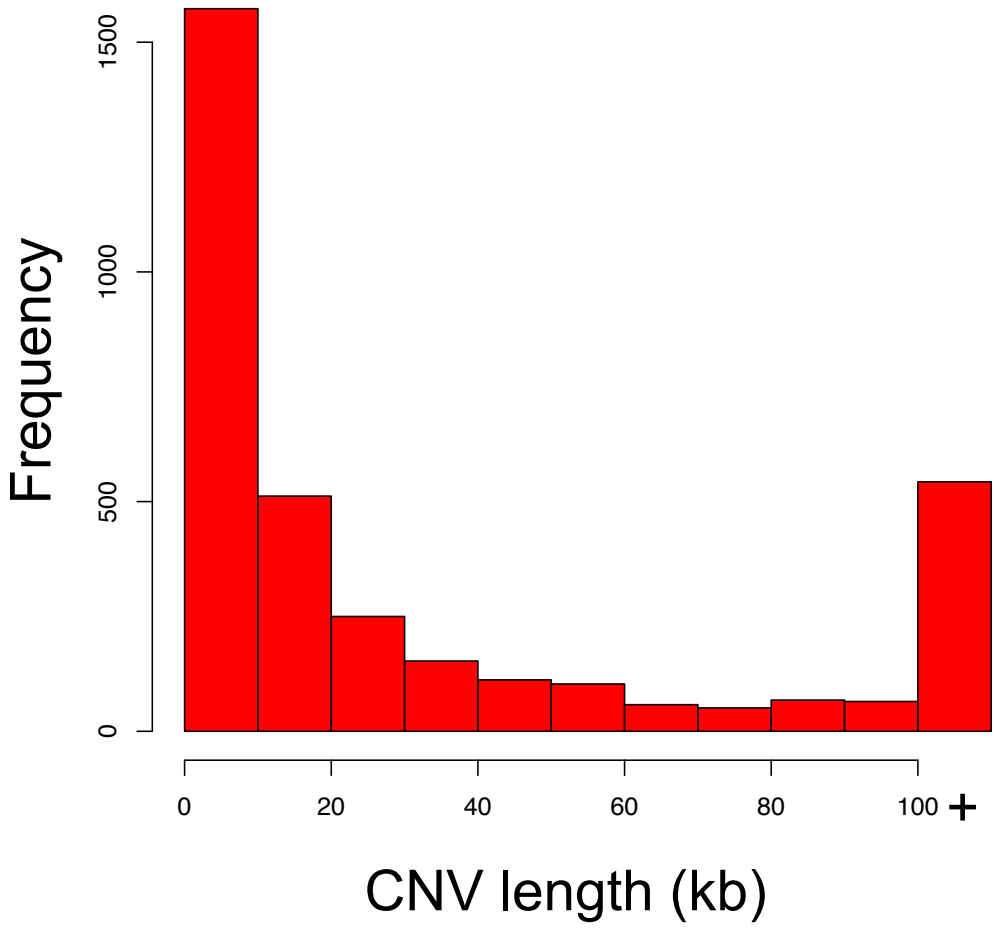

Figure S10. Distribution of Length for All Case-Control CNVs Spanning at Least Two Exome Targets

82% of CNV have size < 100 kb, and 94% are at least 1 kb in length. The median is 18 kb and the mean is 96 kb. The long tail of the distribution is truncated and lumped into the 100 kb bin.

Table S1. Breakdown of Rare (<1%), Reliable (>100 kb) CNV Calls for the 1,017 Swedish Schizophrenia Case-Control Samples, Discovered Using Birdsuite on Affymetrix Array Intensity Data

	Total	Deletions	Duplications
Affy calls	941	351	590
Affy, overlap exome	544	145	399

By necessity, we analyzed only those that overlap at least one exome target (bottom row).

Table S2. Breakdown of XHMM CNV Calls into Deletions and Duplications, and XHMM's Sensitivity toward Recovering All Affymetrix (Affy) CNVs, Deletions, or Duplications

	Total	Deletions	Duplications	Sensitivity to Affy	Sensitivity to Deletions	Sensitivity to Duplications
XHMM calls	5,441	2,478	2,963	394 (72%)	88 (61%)	306 (77%)
XHMM, MAF < 0.01	2,315	1,051	1,264	367 (67%)	70 (48%)	297 (74%)

The top row is for all high-quality XHMM calls, whereas the bottom is filtered at a sample frequency of 1%.