The American Journal of Human Genetics, Volume *98*

Supplemental Data

Frequency and Complexity of De Novo Structural Mutation in Autism

William M. Brandler, Danny Antaki, Madhusudan Gujral, Amina Noor, Gabriel Rosanio, Timothy R. Chapman, Daniel J. Barrera, Guan Ning Lin, Dheeraj Malhotra, Amanda C. Watts, Lawrence C. Wong, Jasper A. Estabillo, Therese E. Gadomski, Oanh Hong, Karin V. Fuentes Fajardo, Abhishek Bhandari, Renius Owen, Michael Baughn, Jeffrey Yuan, Terry Solomon, Alexandra G. Moyzis, Michelle S. Maile, Stephan J. Sanders, Gail E. Reiner, Keith K. Vaux, Charles M. Strom, Kang Zhang, Alysson R. Muotri, Natacha Akshoomoff, Suzanne M. Leal, Karen Pierce, Eric Courchesne, Lilia M. Iakoucheva, Christina Corsello, and Jonathan Sebat

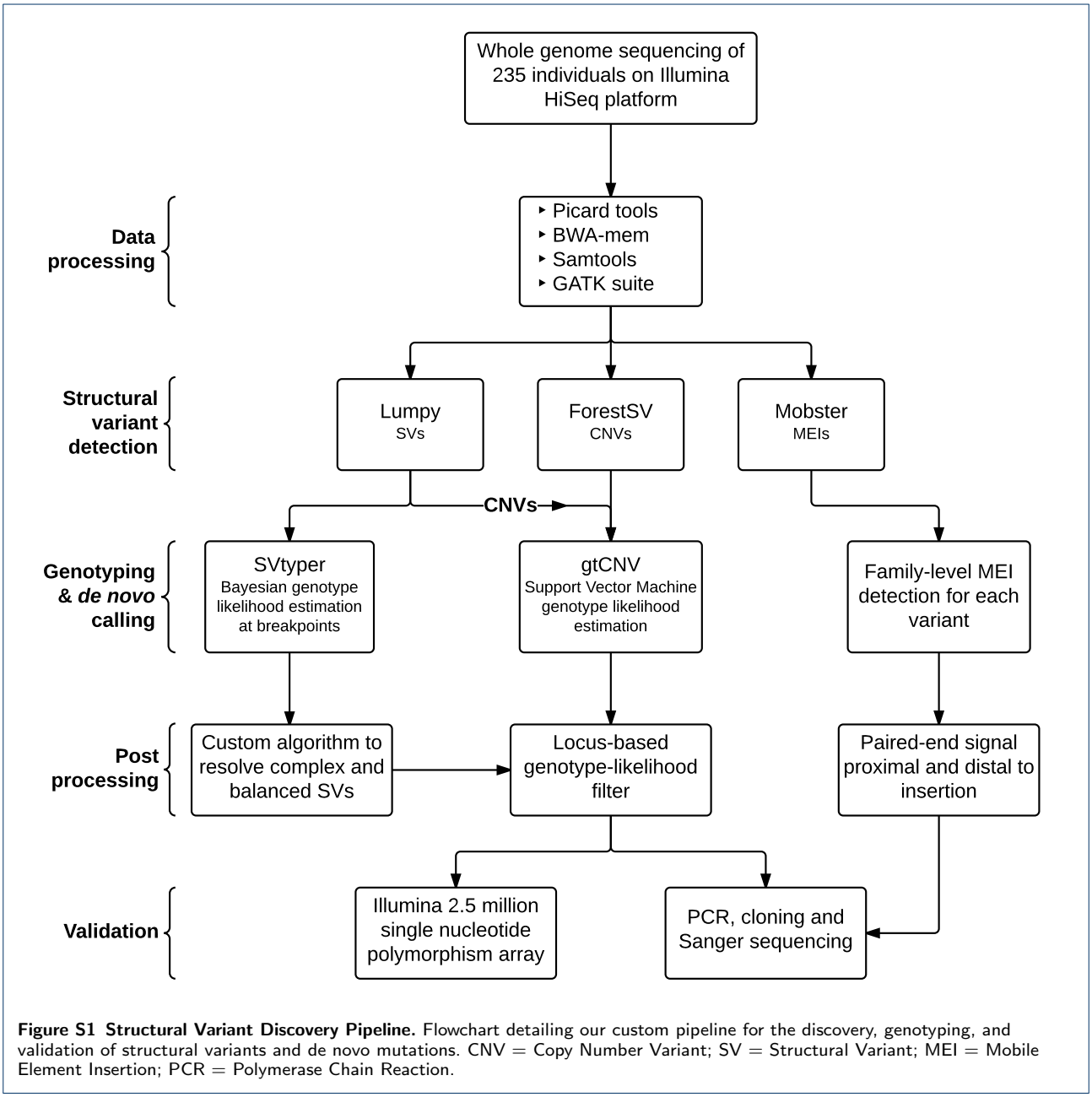

Figure S1 Structural Variant Discovery Pipeline. Flowchart detailing our custom pipeline for the discovery, genotyping, and validation of structural variants and de novo mutations. CNV = Copy Number Variant; SV = Structural Variant; MEI = Mobile Element Insertion; PCR = Polymerase Chain Reaction.

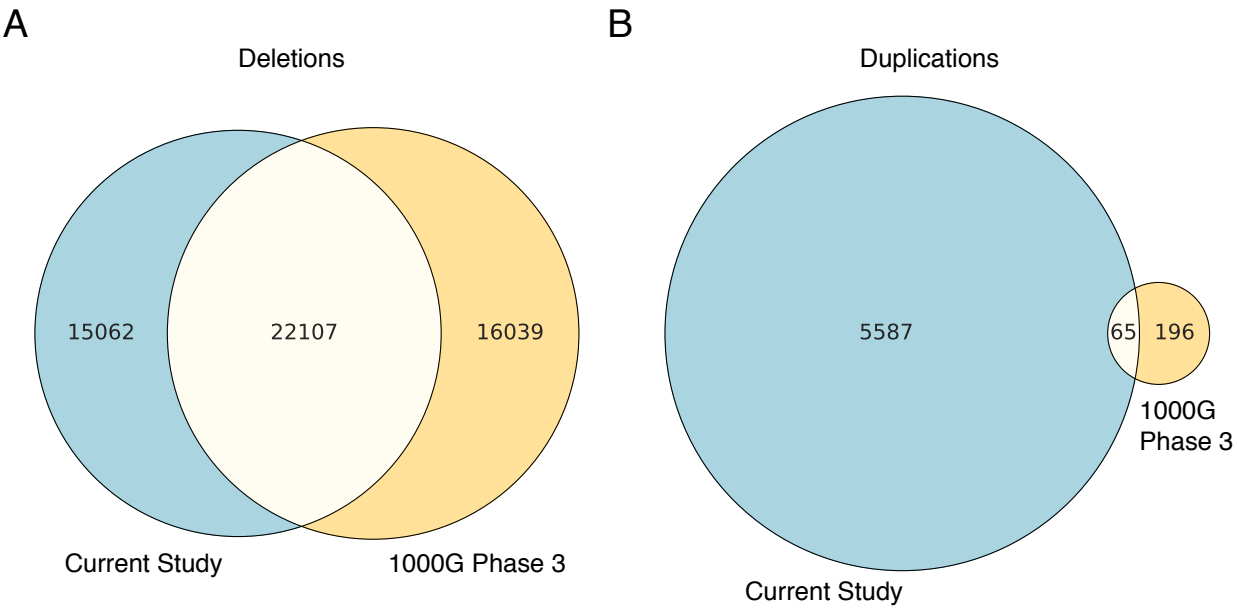

Figure S2 Overlap between SV calls made using our methods and 1000 genomes phase 3 methods on high-coverage genomes. Venn diagrams indicate the overlap of non-reference deletion and biallelic duplication calls made on 27 individuals sequenced at high coverage as part of the 1000G project.

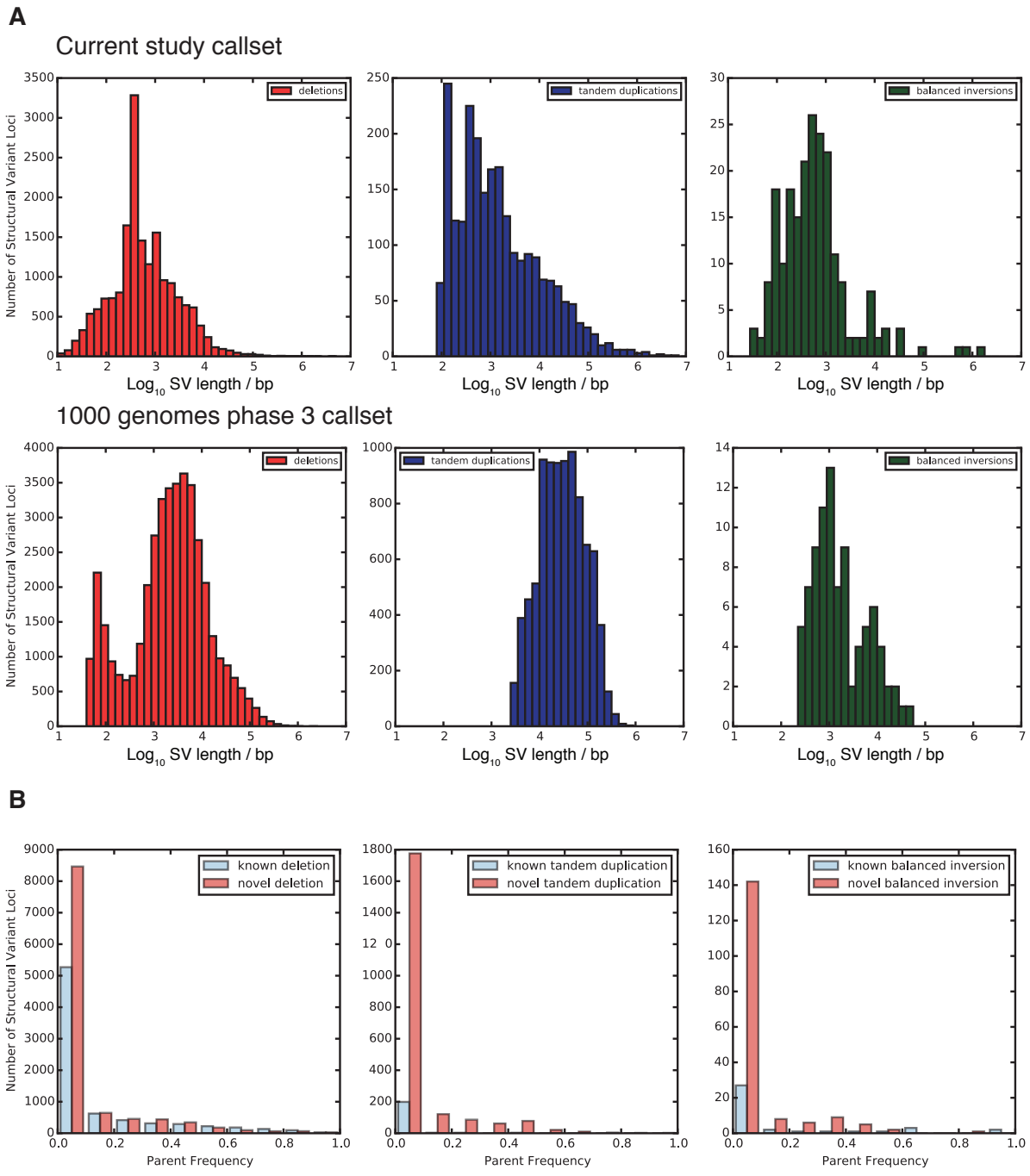

Figure S3 Comparison between current study call set and 1000 Genomes Phase 3. A) Histograms of the log10 structural variant (SV) size distributions for deletions, tandem duplications and balanced inversions in our study and 1000 genomes phase 3 (1000G) SV call set. B) Histograms showing the number of novel versus known SVs across a range of parent frequencies.

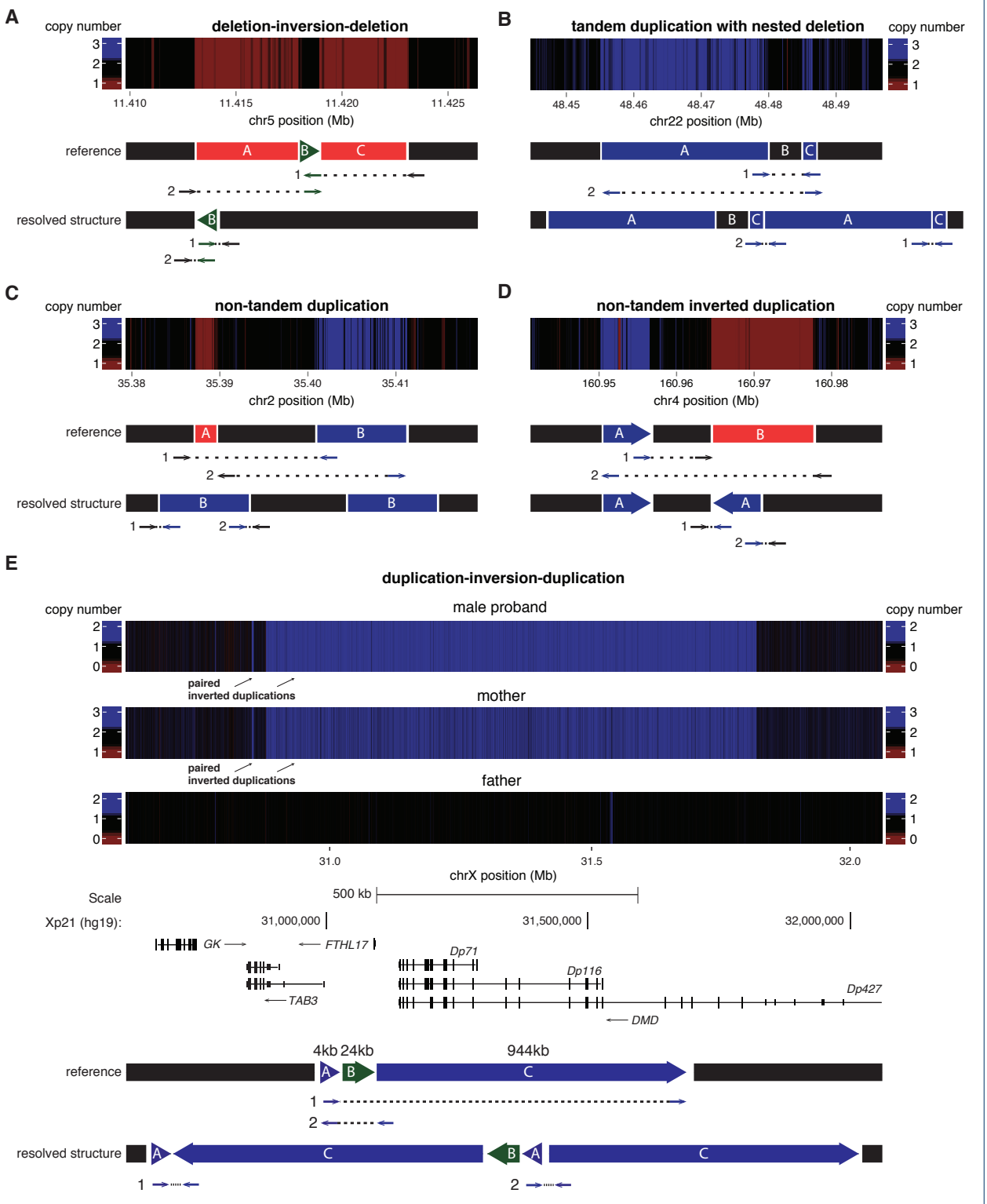

Figure S4 Complex Structural Variation Detected using Genome Sequencing. Examples of each class are taken from our call set of SVs. A) deletion-inversion-deletion, B) tandem duplication with nested deletion, C) non-tandem duplication, D) non-tandem inverted duplication, E) duplication-inversion-duplication (including genes in the vicinity of the SV event). Heat maps indicate changes in copy number observed from the depth of coverage at each locus, normalized to the chromosomal average. Lettered segments indicate the structure of the chromosome in the reference and the observed genome. Black segments are unchanged in the SV events, green segments are inverted, blue segments are duplicated, and red segments are deleted. Arrows indicate the discordant orientation and location of paired-end reads relative to the hg19 reference genome and the concordant pattern of paired end reads relative to the resolved structure. n.b. segments not shown to scale.

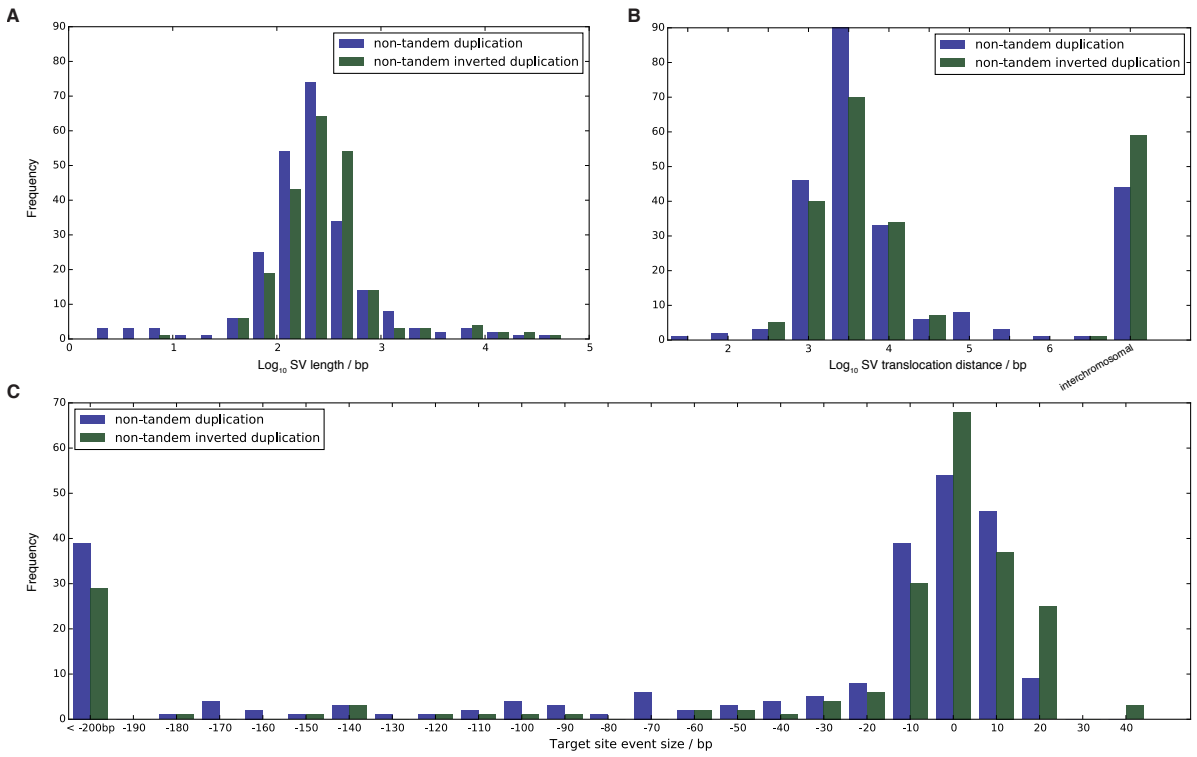

Figure S5 Distribution of Non-Tandem Structural Variants. a) Histogram of the lengths of non-tandem duplications (blue) and non-tandem inverted duplications (green). b) Histogram of translocation distances. c) Histogram of target site deletions or duplications at the non-tandem event's insertion point.

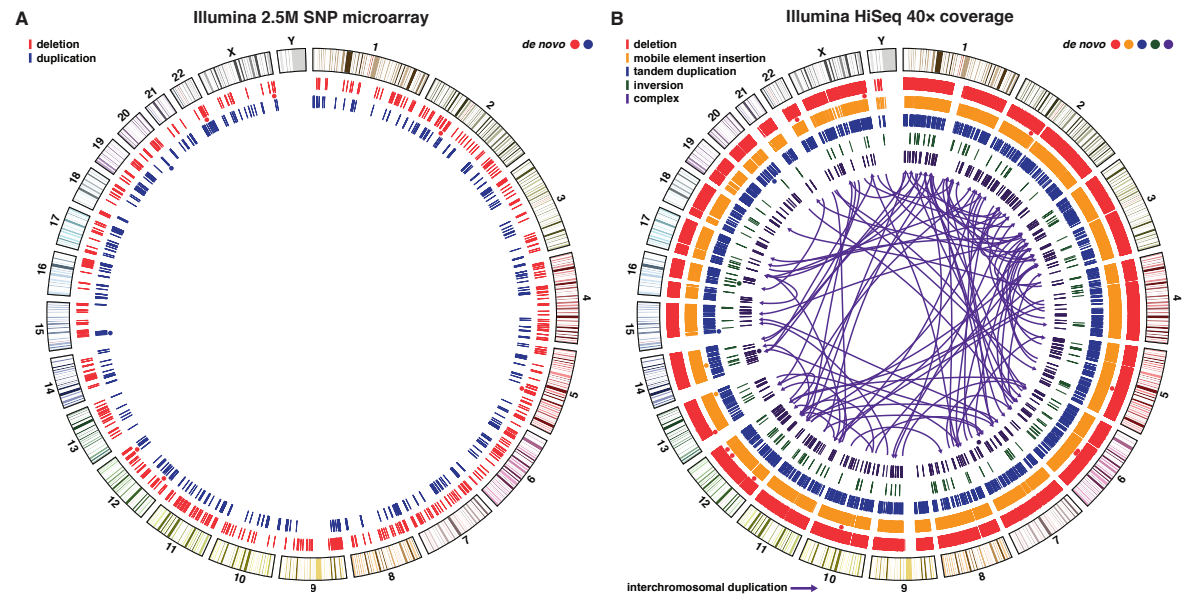

Figure S6 Comparison of Structural Variation Detection using Microarrays and Genome Sequencing. Circos plots comparing structural variant calls for 205 individuals in this study derived from a) Illumina 2.5 million single nucleotide polymorphism (SNP) microarray, and b) from WGS at 40× coverage on the Illumina HiSeq platform. Concentric circles represent from outermost to inner in panel: ideogram of the human genome with karyotype bands (hg19), deletions, mobile element insertions (four different classes), tandem duplications, balanced inversions, complex structural variants (four different classes). The circles indicate the location of de novo SVs, and their colors match the five SV types. Arrows represent interchromosomal duplications.

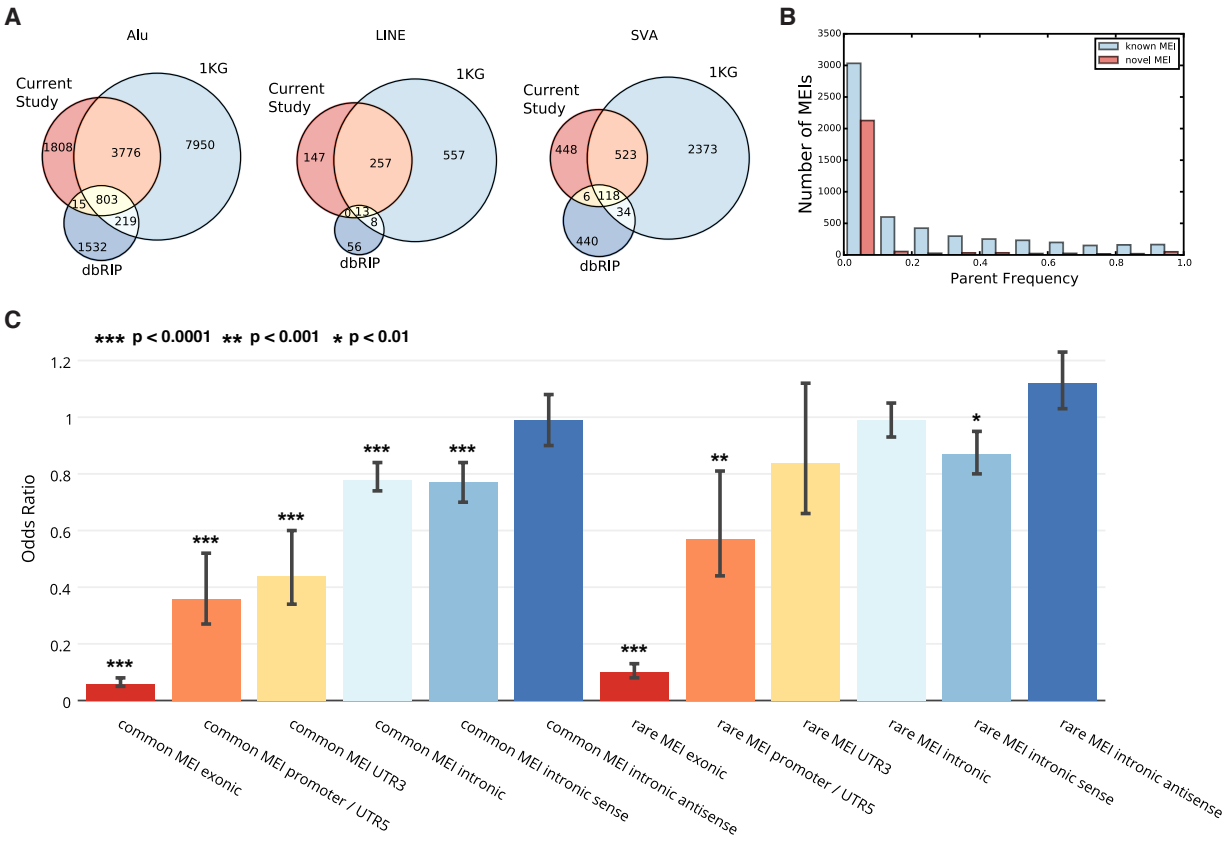

Figure S7 Mobile Element Insertion Overlap with Published Databases and Genomic Features. A) Venn diagrams showing the overlap of MEIs detected in our study with MEIs from the 1000 genomes project (1KG) phase 3 integrated SV call set and the database of retrotransposon insertion polymorphisms (dbRIP), calls were considered to overlap if they were within 100 base-pairs of each other. B) Histogram showing the number of novel versus known MEIs across a range of parent frequencies. C) Bar chart showing the odds ratio of the overlap of observed common (frequency $\geq$5%) and rare MEIs with genomic functional elements compared to expected overlap through permutation. Error bars represent the 95% confidence interval for odds ratio.