Online Methods

Unmapped fosmid end sequences were identified from nine individuals as previously described [11]. Sequence contigs were assembled using phrap (http://phrap.org) and have been deposited to GenBank under accessions GU266782 - GU269144. Two custom Agilent oligonucleotide arrays were designed to target these sequences, and a series of hybridizations were performed using genomic DNA from sample NA15510 as a reference. Artifactual and low-signal probes were removed based on an analysis of single channel fluorescent intensity and the correlation of probe responses across experiments. Additional contigs having high-identity BLAST hits against sequences from non-primate species were also removed. Copy-number polymorphism was assessed through both direct comparisons of array intensity values and a modal clustering method that attempts to fit array intensity data to distinct integer copy-number states. ArrayCGH data has been deposited in GEO under accession GSE20634. Additional details of array design, probe quality analysis, and polymorphism calling are described in the **Supplementary Note**.

Orphan clones were identified from the G248 (NA15510) clone library. Restriction profiles using four enzymes were obtained and used to link individual clones into contigs using the Contig Builder program [12].

Unmapped clone end sequences were determined relative to the build35 genome assembly, while completely sequenced clones were compared against the more recent build36 assembly (**Supplementary Table 9**). Two sequenced clones (AC234849 and AC226835) were tested and determined to be artifactual and were omitted from all analysis. Approximate breakpoint regions were identified using the program miropeats [27] and refined based on review of an alignment of sequences extracted from the corresponding breakpoint regions.

mRNA-seq data reported by Wang et al. [19] was downloaded from the short-read archive and mapped against the sequenced novel insertion using mrsFAST (http://mrfast.sourceforge.net). Only reads that did not map against the build36 genome sequence were considered. Splicing was ignored, and all mapped positions with up to two mismatches were recorded. We identified all segments with a depth of three or more reads.

Conservation analysis was based on the Ensembl Compara 51 alignments of nine mammalian genomes (http://www.ensembl.org). Segments corresponding to the human insertions were found by BLAST and conserved elements were identified using GERP version 2.1 (http://mendel.stanford.edu/sidowlab/downloads/gerp/index.html). All conservation analysis was limited to the eight non-human genome sequences for which alignments were available.

Diagnostic k-mers were identified to genotype each variant by searching overlapping 36-mers derived from sequenced breakpoints against a database of sequenced insertions and the build36 assembly. Only k-mers with a single hit (permitting one substitution) were retained. In order to be genotypeable, a variant must have at least one deletion k-mer and one insertion k-mer that met these criteria. Illumina sequence reads were then searched against this set of diagnostic k-mers (requiring a perfect match). A breakpoint search

score was computed for each variant based on the number of reads that match k-mers derived from the insertion or deletion alleles.