EXTENDED EXPERIMENTAL PROCEDURES

Additional Events without Resolved Breakpoints

We identified 81 loci with evidence for a nonreference structure for which we could not define variant breakpoints. This set includes 26 loci for which clones were not sequenced to a finished state because of the presence of difficult-to-assemble satellite or simple sequence repeats, or contained apparent collapses of arrays of tandem repeats. Sixty-three of the 81 loci are associated with segmental duplications, suggesting that 40-kb fosmid clones may not be sufficient to unravel the structures of such regions. Twenty-three of the sixty-three duplication-containing loci map near gaps in the genome assembly or to sequences that have been assigned to a chromosome but not fully integrated into the genome reference sequence (such as sequences labeled as "chr*_rand"). It is likely that many of these assembly gaps correspond to segments where distinct structural configurations are found among individuals. Ten of the 81 loci represent tandem duplications of sequences that are not duplicated in the assembly and have only been partially captured by a sequenced clone. The remainder correspond to variants in repeat- or duplication-rich segments for which clear breakpoint intervals could not be defined. We observe that 18 of these 81 unresolved loci involve genes, several of which have been associated with human phenotypic variation (Table S8A).

Tandem Insertions

The fosmid ESP-mapping approach can directly capture the entire sequence of insertions, relative to the genome reference, that are smaller than the clone fragment insert size (~40 kb). However, the approach also provides information about other types of insertions, including tandem duplications (Kidd et al., 2008; Marques-Bonet et al., 2009). The information gained about the structure of tandem duplications depends on the relative size and positioning of the tandem event and the fragment represented in the fosmid clone. When the entire duplicated array is spanned by the cloned fragment, sequencing will reveal the exact extent of the event as well as the features of the breakpoints (Figure S1A, blue lines). ESPs that map within the tandem duplication but span an internal junction (Figure S1A, green lines) will map onto the reference with a characteristic "everted" signature. This signature is diagnostic of tandem duplications involving large sequence blocks (Cooper et al., 2008). Clones that span one copy of the duplication unit and extend into the other tandem copy (Figure S1A, orange line) will appear as normal "insertion" clones when mapped onto the reference (that is, the end sequences will map closer together than expected). However, sequencing of such a fragment will only identify one of the insertion breakpoints.

Such events have been excluded from the breakpoint analysis since the variant extent and breakpoint structure could not be directly assessed. However, we identified ten events involving tandem duplications of sequence blocks represented as a single copy in the reference. Figure S1B illustrates one of these events, which involves duplication of the 3′ end of the *APBA1* gene.

Identifying Informative K-mers with a Support Vector Machine

We explored the possibility that specific sequence motifs might be associated with different classes of structural variants. We searched for specific sequence stretches that could discriminate among various classes of breakpoints with a Support Vector Machine (SVM) with a methodology previously employed to detect signals in DNA sequence associated with nucleosome positioning (Gupta et al., 2008; Peckham et al., 2007). We constructed a feature vector consisting of the frequencies of each k-mer sequence for k = 1 to 6 using the sequences in a window centered on each breakpoint. We then attempted to discriminate among different variant classes as well as random sequences sampled from the genome using these features. We used the area under receiver-operator characteristic (ROC) curve as a metric. A random classifier would have an ROC score of 0.5. Results are presented with both individual k-mer frequencies to discriminate event classes as well as with a support vector machine that utilized the entire collection of features. The SVM analysis was performed with Gist (Pavlidis et al., 2004).

To control for homologous breakpoints, we considered the sequences in a 200 bp window centered only one breakpoint from the "insertion allele" of each event. Based on a 10-fold cross validation strategy, we find one can distinguish variants with more than 10 bp of homologous breakpoint sequences from variants with less than 10 bp with k-mer frequencies (mean area under the ROC curve of 0.761). In contrast, events associated with microhomology were not distinguishable from randomly sampled genome fragments (mean ROC score 0.521) or from sequence variants showing no detected breakpoint homologies (mean ROC score 0.487; Table S8B). Direct ranking of k-mers based on their ability to discriminate between microhomology and extended homology sequences (Peckham et al., 2007) identifies GC-rich sequence motifs (such as CCG/CGG, CTCC/GAGG, and CGC/CCG) as being enriched among variants having extended homology between their breakpoints (Table S8C), a finding that suggests a preferential breakpoint composition for events that are consistent with the nonallelic homologous recombination mechanism.

SUPPLEMENTAL REFERENCES

Cooper, G.M., Zerr, T., Kidd, J.M., Eichler, E.E., and Nickerson, D.A. (2008). Systematic assessment of copy number variant detection via genome-wide SNP genotyping. Nat. Genet. *40*, 1199–1203.

Gupta, S., Dennis, J., Thurman, R.E., Kingston, R., Stamatoyannopoulos, J.A., and Noble, W.S. (2008). Predicting human nucleosome occupancy from primary sequence. PLoS Comput. Biol. *4*, e1000134.

Kidd, J.M., Cooper, G.M., Donahue, W.F., Hayden, H.S., Sampas, N., Graves, T., Hansen, N., Teague, B., Alkan, C., Antonacci, F., et al. (2008). Mapping and sequencing of structural variation from eight human genomes. Nature *453*, 56–64.

Marques-Bonet, T., Kidd, J.M., Ventura, M., Graves, T.A., Cheng, Z., Hillier, L.W., Jiang, Z., Baker, C., Malfavon-Borja, R., Fulton, L.A., et al. (2009). A burst of segmental duplications in the genome of the African great ape ancestor. Nature *457*, 877–881.

Pavlidis, P., Wapinski, I., and Noble, W.S. (2004). Support vector machine classification on the web. Bioinformatics *20*, 586–587.

Peckham, H.E., Thurman, R.E., Fu, Y., Stamatoyannopoulos, J.A., Noble, W.S., Struhl, K., and Weng, Z. (2007). Nucleosome positioning signals in genomic DNA. Genome Res. *17*, 1170–1177.

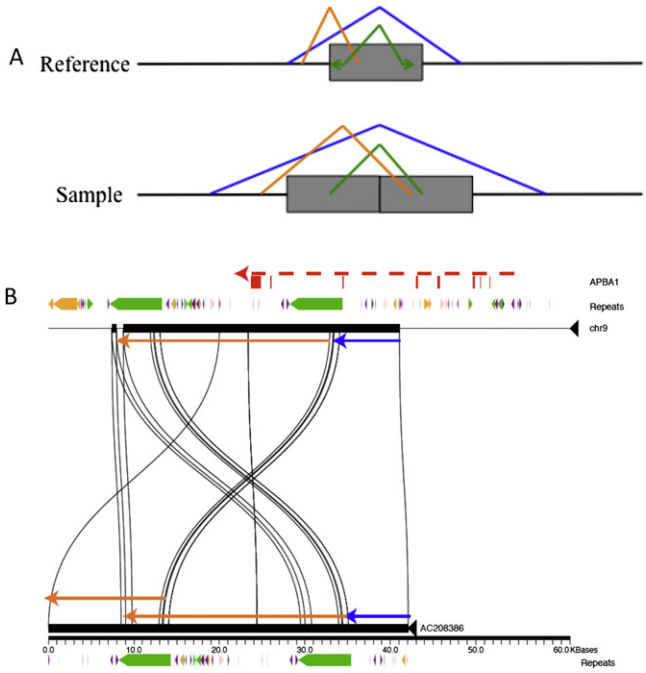


Figure S1. Detection Signatures for Tandem Duplications, Related to Figure 2

(A) Distinct clone signatures are observed when a sequence that is present in a single copy in the reference (gray box, top line) is tandemly duplicated in a sample. The blue lines represent the mapping positions of a fragment that entirely spans the event. Sequencing such a cloned fragment would reveal the complete content and breakpoint information for the tandem duplication. The green lines represent a clone that spans one of the tandem duplication junctions. Such clones map in an everted orientation when mapped to the reference. The orange lines represent a clone that extends into a tandem insertion. Sequencing such a clone would directly identify one of the breakpoints of the event.

(B) The sequence of fosmid AC208386 extends into a tandem duplication that includes the 3′ end of the *APBA1* gene on chr9. The segment represented by the orange arrow is duplicated in the clone. The extent of the insertion is not entirely captured, but it appears that both breakpoints would map to the pair of directly oriented LINEs (indicated in green).

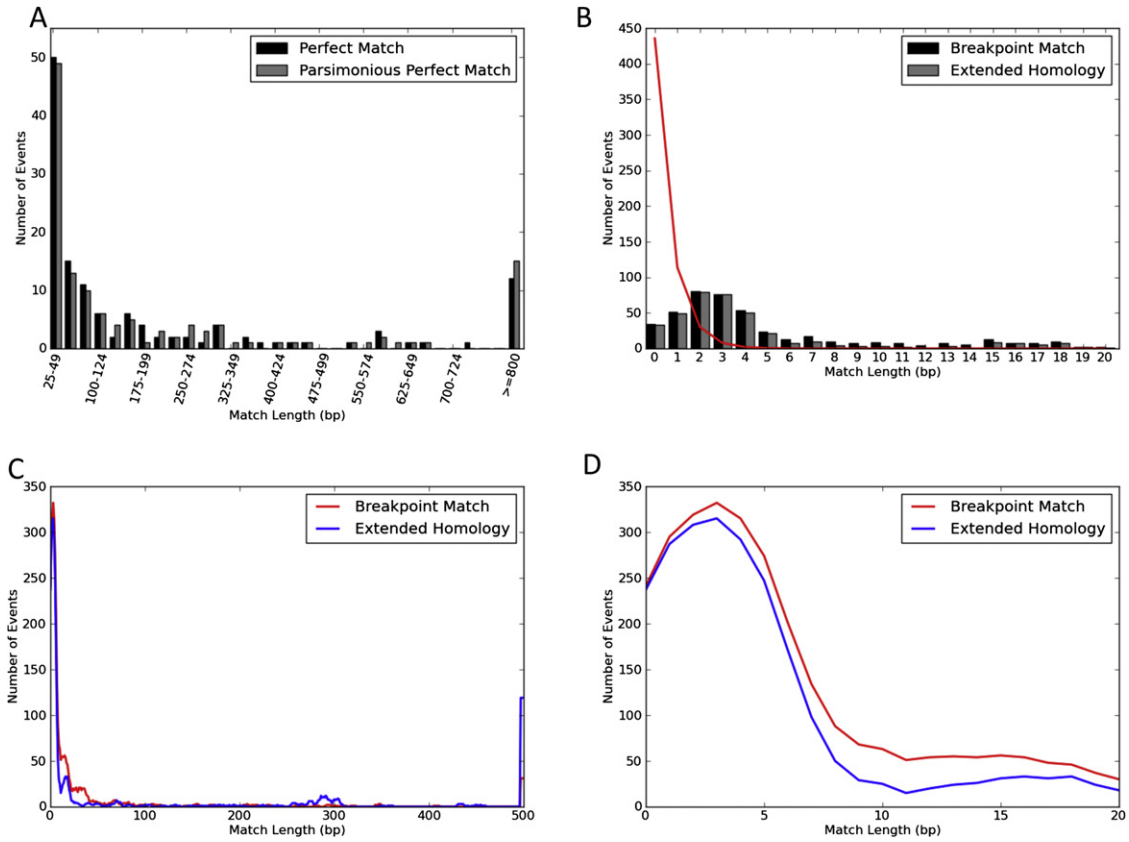


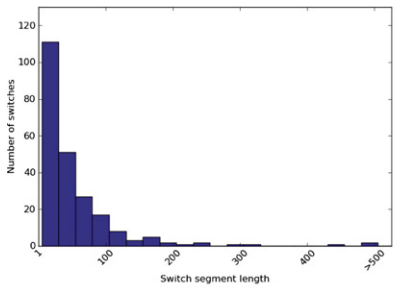
Figure S2. Characteristics of Class I Insertion and/or Deletion Variants, Related to Figure 4

(A) Distribution of perfect match length for putative NAHR events. The histogram shows perfect match length (100% identity among the 5′ breakpoint, the 3′ breakpoint, and the deletion haplotype) between the 5′ and 3′ deletion breakpoint sequences for 130 insertion and/or deletion events having at least 25 bp of perfect sequence identity. The gray bars represent the distribution of parsimonious perfect match length. In this definition, positions where the deletion allele matches neither the 5′ nor 3′ sequences (a configuration consistent with a new mutation that arose after variant formation) are discounted. Note that new mutations that occurred on the 5′ or 3′ breakpoint sequence after variant formation will still "break" the match length tract, resulting in a potential underestimate of the true extent of matching sequence present at the time of variant formation.

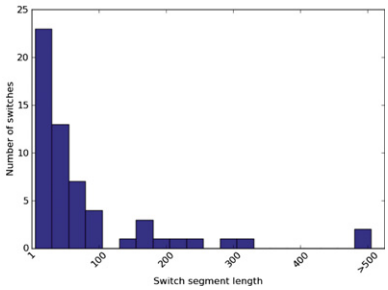
(B) Expected distribution of breakpoint match lengths. The red line corresponds to the expected distribution of breakpoint match lengths found from 100 random permutations.

(C and D) Breakpoint match lengths plotted as running averages in windows of size 6 with a step size of 1.

A



B



C

```
hg18.5'         GAGGGGAGGAAACTTAGAGGATGGGTCAACAGATGCAGCAAACCACCATGGCATACGTAT
AC226182.del    GCGGGGAGGAAACTTAGAGGATGGGTCAACAGATGCAGCAAACCACCATGGCATACGTAT
AC237150.del    GCGGGGAGGAAACTTAGAGGATGGGTCAACAGATGCAGCAAACCACCATGGCATACGTAT
AC237647.del    GCGGGGAGGAAACTTAGAGGATGGGTCAACAGATGCAGCAAACCACCATGGCATACGTAT
hg18.3'         GAGGGGAGGAAACTTAGAGGATGGGTCAATAGATGCAACAAACCACCATGGCATATGCAT
                * ************************** ******* ***************** * **

hg18.5'         ACCTATGTAACAAACCTTCACATTCTGCACATGTATCCCAGAACTTGTGCAGTAAAAATA
AC226182.del    ACCTATGTAACAAACCTGCACATTCTGCACATGTATCCCAGAACTTGTGCAGTAAAAACA
AC237150.del    ACCTATGTAACAAACCTGCACATTCTGCACATGTATCCCAGAACTTGTGCAGTAAAAACA
AC237647.del    ACCTATGTAACAAACCTGCACATTCTGCACATGTATCCCAGAACTTGTGCAGTAAAAACA
hg18.3'         ACCTATGTAACAAACCTGCACATTCTGCACACGTATCCCAGAACTTGTGCAGTAAAAACA
                ***************** ************* ************************** *

hg18.5'         AAGAAAAAGAAAGAAGAAAGAAAGAAAGAAAGAAAGAAAGAAAGAAAGAAAGAAAG
AC226182.del    AAG--------------AAGAAAGAAAGAAAGAAAGAAAGAAAGAAAGAAAGA---AAAG
AC237150.del    AAG------------------AAGAAAGAAAGAAAGAAAGAAAGAAAGAAAGA---AAAG
AC237647.del    AAG--------------AAGAAAGAAAGAAAGAAAGAAAGAAAGAAAGAAAGA---AAAG
hg18.3'         AAG--------------AAGAAAGAAAGAAAGAAAGAAACAAAGAAAGAAAGA---AAAG
                ***                 ****************** *************   ****

hg18.5'         AAAGAAAGGACCTAGCTCTATATTCTCAAGTGAACAAAAAATCTTTTTATATTATTGAGC
AC226182.del    AAAGA---GACCTAGCTCTATATTCTCAAGTGAACAAAATATATTTTTATATTATTTAGC
AC237150.del    AAAGA---GACCTAGCTCTATATTCTCAAGTGAACAAAATATATTTTTATATTATTTAGC
AC237647.del    AAAGA---GACCTAGCTCTATATTCTCAAGTGAACAAAATATATTTTTATATTATTTAGC
hg18.3'         AAGGA---GACCTAGCTCTATATTCTCAAGTGAACAAAATATATTTTTATATTATTTAGC
                ** **   ****************************** ** ************* ***

hg18.5'         AAGCTAATGTATTAATAAATTTAAAGTCAATGGTTATTGCTAGTCCTAATTTATGAAAAA
AC226182.del    AAGCTAATGTATTAATAAATTTAAAGTTAATGGTTACTGCTAGTCCTAATTTATGAAAAA
AC237150.del    AAGCTAATGTATTAATAAATTTAAAGTTAATGGTTACTGCTAGTCCTAATTTATGAAAAA
AC237647.del    AAGCTAATGTATTAATAAATTTAAAGTTAATGGTTACTGCTAGTCCTAATTTATGAAAAA
hg18.3'         AAGCTAATGTATTAATAAATTTAAAGTTAATGGTTACTGCTAGTCCTAATTTATGAAAAA
                ************************** ******** ***********************
```

Figure S3. Assessment of Variants Associated with Gene Conversion Related to Figure 5

Switch segment length distribution.

(A) Histogram of the sizes of switch segments inferred for 10 variants.

(B) Histogram of switch segment lengths excluding variant AC212911.

(C) Sequence from three different individuals shows the same pattern of alternating sequence matches for the AC226182 deletion. This event corresponds to a 109 kb deletion that removes the *UGT2B28* gene on chr4. Counting the deletion of "AAG" in the third row of the alignment as match to the 3′ breakpoint, this variant contains six distinct switches between the 5′ and 3′ breakpoint sequences.