Supplemental Data

The Fine-Scale and Complex Architecture of Human Copy-Number Variation

George H. Perry, Amir Ben-Dor, Anya Tsalenko, Nick Sampas, Laia Rodriguez-Revenga, Charles W. Tran, Alicia Scheffer, Israel Steinfeld, Peter Tsang, N. Alice Yamada, Han Soo Park, Jong-Il Kim, Jeong-Sun Seo, Zohar Yakhini, Stephen Laderman, Laurakay Bruhn, and Charles Lee

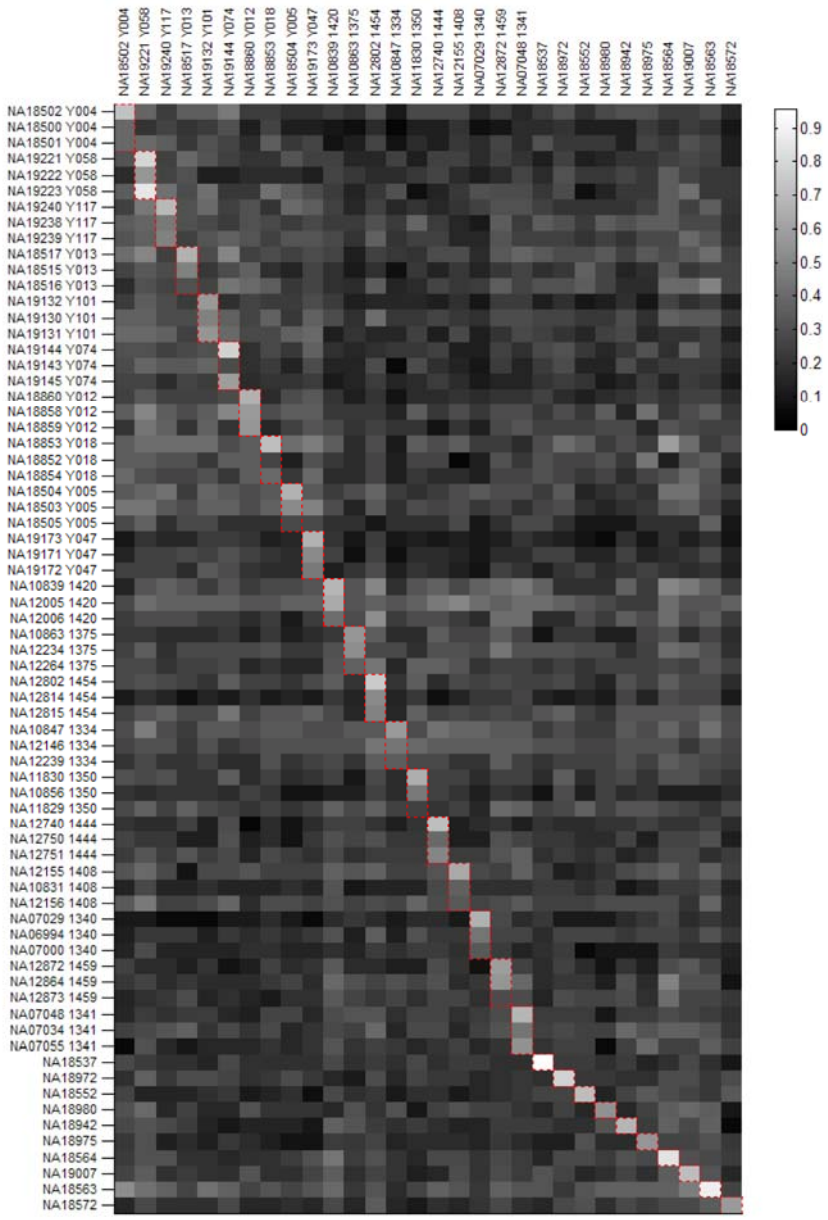

Figure S1. Identification of HapMap Samples, Based on CNV Call Concordance with a Previous Study

We assessed whether it was possible to identify our 30 samples from all 270 HapMap individuals from the Redon et al. study, on the basis of CNV call concordance. The correct sample was identified for 29 of 30 (97%) and 25 of 30 (83%) individuals on the basis of calls from the 500K EA and WGTP platforms, respectively. Of the five misidentified individuals from the WGTP analysis, four (80%) had differences of < 1% concordance for WGTP CNVs

between the true and misidentified individuals, including one individual who was most concordant with a first-degree relative. The only misidentified individual from the 500K EA platform was also most concordant with a first-degree relative. The Heatmap diagram depicted above indicates the proportion of calls from the Redon et al. study (on the basis of 500K EA data for this figure) that were replicated in this study for each sample. For each CEPH and YOR sample in this study, the corresponding family trio is shown. See Table S7 for data for all 270 HapMap individuals and both the WGTP and 500K EA platforms.

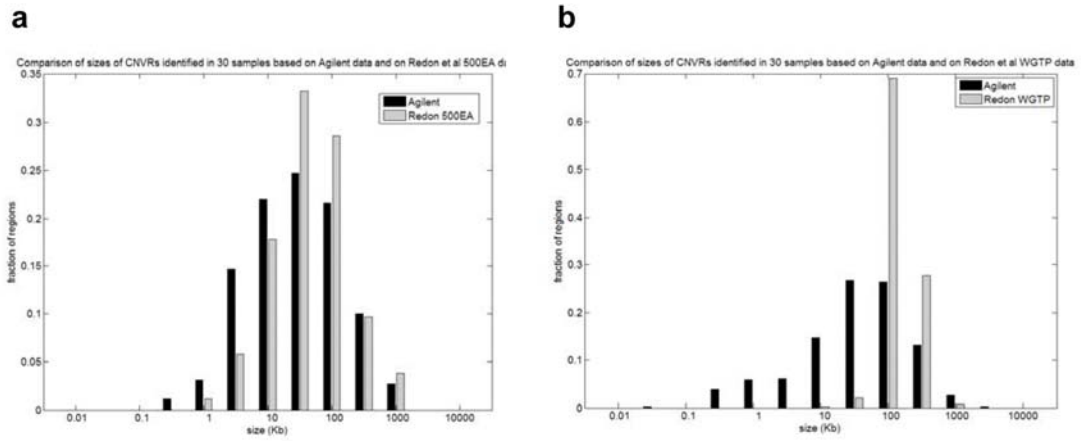

Figure S2. CNV Size Comparison between This Study and the Redon et al. Study for Calls Made in the Same Sample

Frequency distribution of estimated CNV sizes for calls made in the same individual and direction (i.e., gain or loss) for our study compared to the Redon et al. 500K EA platform (A) and the Redon et al. WGTP platform (B). Size distributions for these regions are shown in log scale, with 10-fold multiples of 1 and √10. The WGTP comparison depicts all CNVs from our study, regardless of size. However, we also removed CNV regions from our dataset with estimates < 20 kb in size. Of the 264 remaining regions that were identified both in our study and on the WGTP platform for the same individual, 213 (80%) were smaller in size based on our estimates, and 154 (58%) were smaller in size by more than 50%.

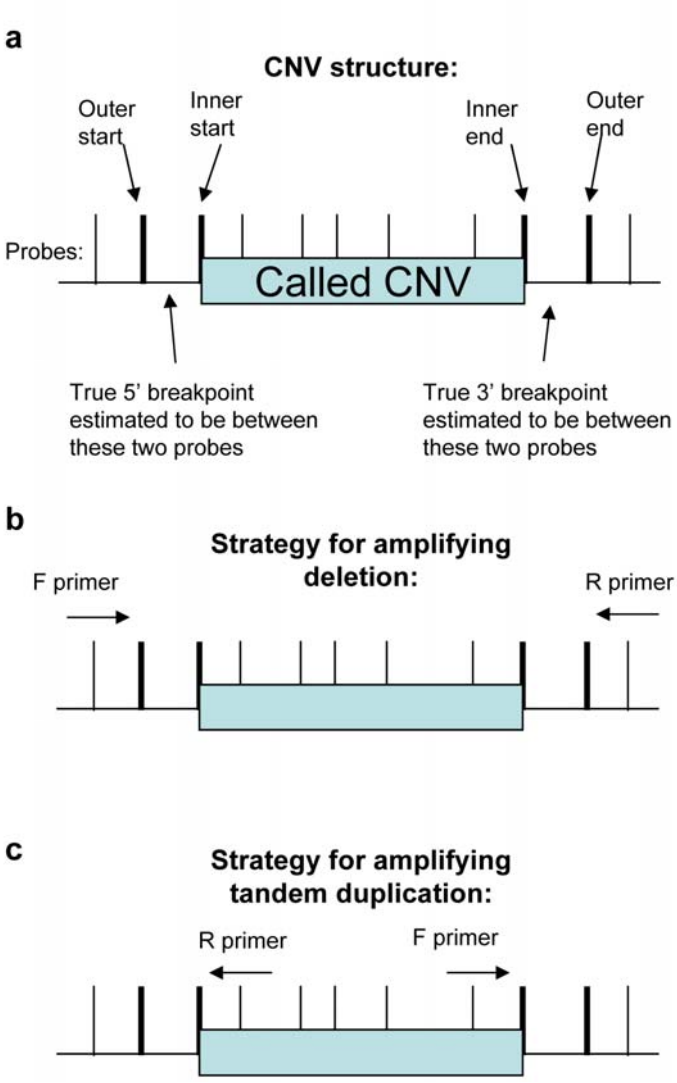

Figure S3. PCR Amplification and Sequencing Strategy for Resolution of CNV Breakpoints

For each CNV, we have the positions of the non-copy-number-variable probes flanking the called CNV (outer start and outer end) and the first and last probes considered copy-number variable (inner start and inner end) (A). Because we did not know whether the actual variants were deletions or duplications, we designed primers that would result in successful amplification in the presence of a deletion (B) or a tandemly-arranged duplication (C) if our CNV breakpoint estimates were accurate.

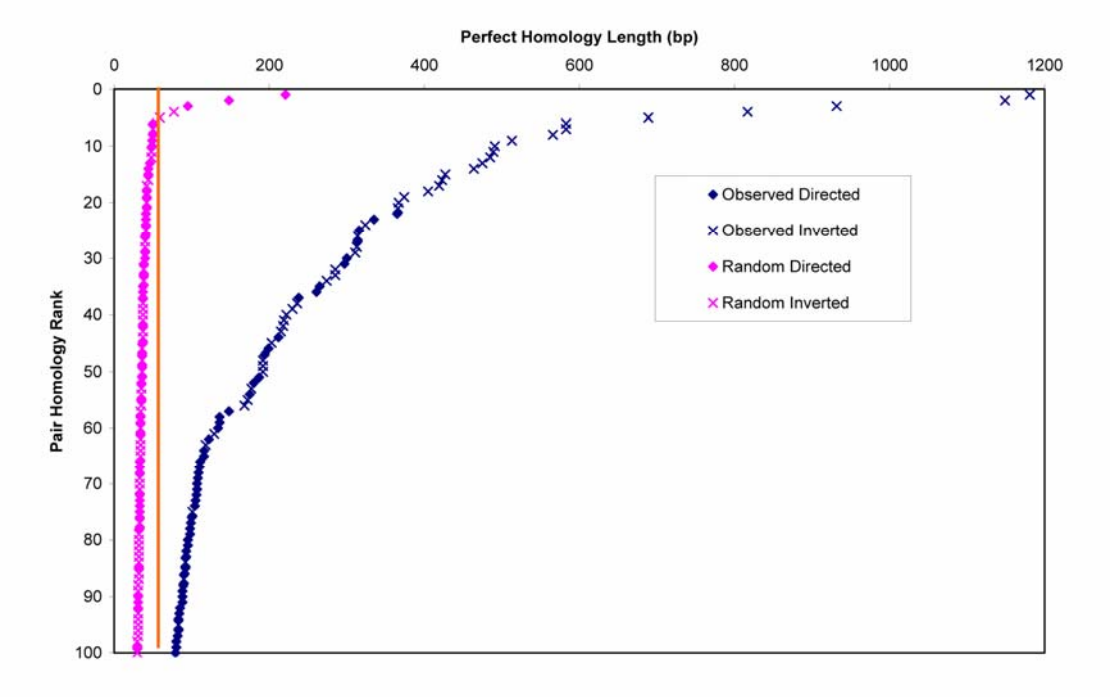

Figure S4. Enrichment for Direct- and Inverted-Orientation Homologous Sequences between CNV Breakpoint Pairs

This figure depicts the empirical cumulative distribution of the observed homology lengths between CNV breakpoints, λ(C), compared to the empirical cumulative distribution of λ(C) obtained from a random genomic shift of each of the CNV calls. The random sequences were selected such as to not alter the characteristics of the observed set of CNV calls, in terms of lengths and proximity of the end sequences. The graph reflects only the significant end of the distribution—the top 100 breakpoint sequence pairs. Direct (i.e., arranged in the same orientation) homologies are marked with diamonds, and inverted homologies are marked with Xs. The significantly higher abundance of long homology in the observed data, as compared to random, is readily seen. The orange vertical line represents homology of length 50 bp, and 145 observed CNV pass this threshold compared to only five sequences in the random distribution (hypergeometric $p < 10^{-36}$). Furthermore, the longer homologies tend to be in an inverted orientation. Under a null hypothesis of equal probability to both orientations, the high occurrence of inverted long homologies has a significant hypergeometric $p < 10^{-8}$ (HGT(N,B,n,b) computed for N = 5609 all homology pairs, B = 1980 inverted homologies only, n = 45 all homologies of length > 200 bp, and b = 35 the intersection—inverted homologies of length > 200 bp).

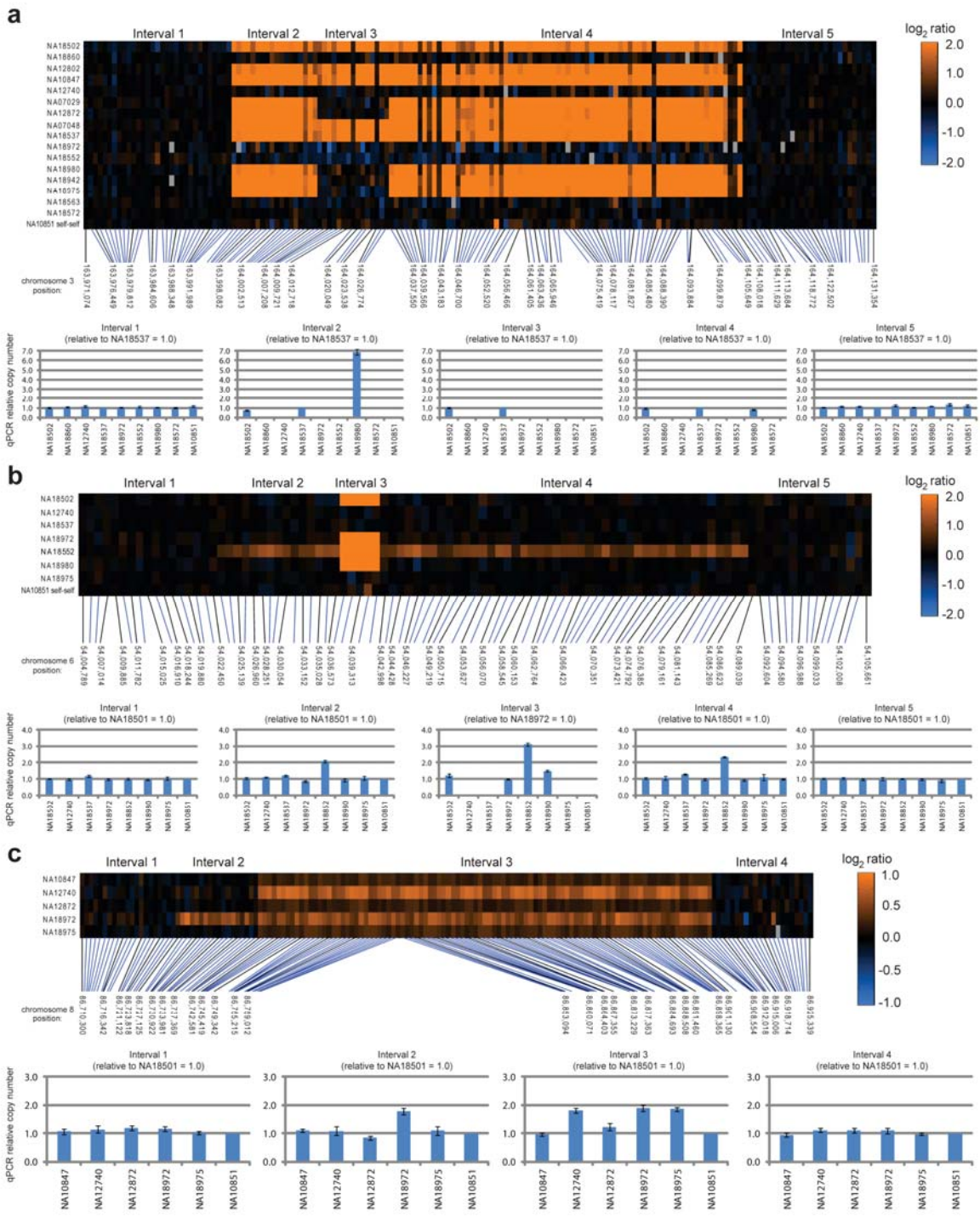

Figure S5. Validation of Architecturally Complex CNV Regions by Quantitative PCR

A series of qPCR probes (see Table S11) were positioned across putative complex CNV regions on chromosome 3 at 164.0 Mb (A), chromosome 6 at 54.0 Mb (B), and chromosome 8 at 88.7 Mb (C). For each region, probe-by-probe $\log_2$ ratios are depicted in heatmaps (see scale bars), with qPCR relative copy-number estimates for different intervals within each region shown below (error bars represent the SD). qPCR results are shown as relative to the individual that was used to construct the standard curve. For some experiments, as indicated, we used a different reference individual than NA10851 to construct the standard curve because NA10851 had a homozygous deletion for all or part of the CNV region.