Mapping and Sequencing of Structural Variation from Eight Human Genomes
Supplementary Material

Table of Contents

1. Project Overview

The full project entails whole-genome shotgun sequencing of 62 unrelated human DNA samples and the establishment of clone-based resources with a special emphasis on resequencing structurally complex regions of the human genome[1]. The project is divided into two phases. Phase I targets an initial eight individuals and focuses on the construction of fosmid clone resources, generation of end-sequence pairs by Sanger-based sequencing, and full-insert sequencing of selected target regions. Phase II will focus on the remaining 54 individuals (40 fosmid & 14 BAC libraries). This two-phase plan was developed to allow the early production of sequenced structural variants, against which various genotyping platforms and new sequencing technologies could be benchmarked. It ensures that the initiative is yielding the expected information, justifying the continued use of Sanger-based sequencing capacity to generate additional data. The stated goals are to generate the first high-quality reference set of sequenced structural variants, provide insight into the molecular mechanisms underlying human genetic variation, and develop the necessary genotyping framework to assess phenotypic consequences in terms of human disease and disease susceptibility.

2. DNA Sample Selection

Eight individual DNA samples were obtained from EBV-transformed lymphoblast cell lines maintained by the Coriell Institute. All samples are part of the HapMap sample collection[2]. Wherever possible, individual samples were selected from pedigrees to assess the genetic transmission characteristics of variants. Five samples corresponded to samples analyzed as part of the ENCODE project[3]. Seven of the samples are female. ABC8, the only male, was sequenced to roughly twice the depth in order to provide sufficient representation of the Y chromosome from a single individual.

Table 1. DNA sample information

Sample ID	Library	Population	Sex	Family	Position in Pedigree	ENCODE
NA18517	ABC7	YRI	Female	Y013-2	mother	Yes
NA18507	ABC8	YRI	Male	Y009-3	father	Yes
NA18956	ABC9	JPT	Female			Yes
NA19240	ABC10	YRI	Female	Y117-1	child	No
NA18555	ABC11	CHB	Female			Yes
NA12878	ABC12	CEU	Female	1463-2	mother	No
NA19129	ABC13	YRI	Female	Y077-1	child	No
NA12156	ABC14	CEU	Female	1408-13	maternal grandmother	Yes

3. Library Production and End-Sequencing

Fosmid libraries were constructed individually from high-quality Coriell DNA using the pCC2FOS vector cloning system[4]. With the exception of the original published fosmid genomic library[5] (G248 a.k.a WIBR-2), all genomic libraries were constructed by Agencourt Biosystems. The power to detect smaller sites of structural variation depends on the standard deviation of library insert sizes, the depth of coverage of pairs, and the average length of the sequence read (longer length increases specificity of paired-end sequence placement). Of these parameters, the standard deviation of insert lengths has the largest effect. The standard deviation of libraries typically ranges from 2.5-3.5 kb, which

allows length variants > 8 kb to be readily distinguished when a three standard deviation threshold is applied. In an effort to detect smaller structural variants, we tested various experimental conditions in library construction to attempt to further reduce insert size variation. This entailed testing the *in silico* insert size distribution of small batches of sub-libraries (1,000-3,000 clones for paired-end sequence analysis) prepared from each sample. With a slight modification of the standard library production protocol, including the use of multiple buffer exchange columns and double pulsed-field gel purification of insert DNA, it became possible to routinely reduce the standard deviation of the insert size to 1.4-1.8 kb—a result that would theoretically allow insertions and deletions as small as ~6 kb to be systematically detected. The inserts of 800,000-1.4 million clones from each sample were then end-sequenced using Sanger based dideoxy sequencing.

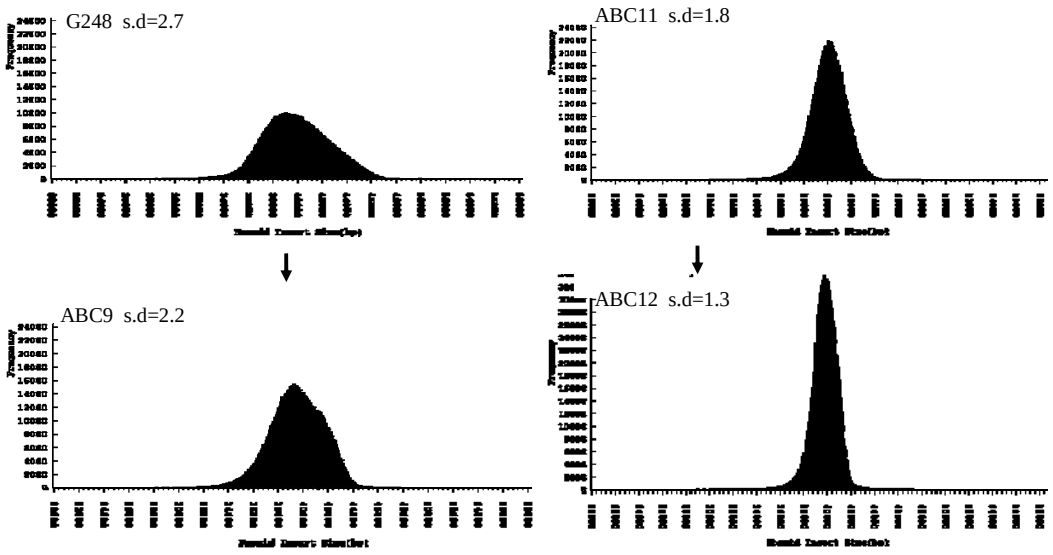


Figure 1. Improvements in fosmid library insert size distributions. The power to detect smaller variants increased as the variance of the libraries decreased. Since our length thresholds were set at 3 s.d. based on the mean insert size, a tighter distribution translated into a narrower length threshold allowing inserts and deletions of smaller size (previously 8 kbp, now 5 kbp) to be recovered. Since there is an inverse relationship between the number of structural variants and the size of an event, as the size threshold is reduced more structural variants are discovered.

4. End-Sequence Pair (ESP) Mapping

Detection of Structural Variants by ESP Mapping: All end-sequence pairs were mapped to the human genome assembly (hg17) using a previously described algorithm[5]. We selected sites where 2 or more clones within a given library showed evidence of discordancy by length (>3 s.d. beyond mean insert size) or orientation (Figure S3). In order to contribute to a discordant site, discordant clones had to pass more stringent mapping criteria: at least 150bp of non-repeatmasked bases included in the alignment (2% divergence threshold), a “best” map location (based on the previously described 13-

point scoring system), and a read length of at least 400 bp. We further required that ESPs map with a minimum sequence identity of 99.5% and include more than 30 basepairs of at least PHRED quality >Q30. As library insert sizes became more tightly distributed, we noted an apparent asymmetry in the *in silico* fosmid insert size distribution (Supplementary Material Figure 1). Consequently, in later libraries we also selected sites based on a threshold set at the top and bottom 0.5% of the distribution of *in silico* insert sizes between 40 and 60 kb. The employed cutoffs are listed below. We limited our analysis to insertion/deletion sites less than 1 Mbp and inversions less than 10 Mbp. Note: two independent libraries were constructed for ABC8, each of which was analyzed separately.

Table 2. Library detection thresholds

Sample ID	Library	Mean	StdDev	3 StdDev Lower Threshold	3 StdDev Upper Threshold	0.5% Lower Threshold	0.5% Upper Threshold
NA15510	G248	39892	2747	31651	48133		
NA18517	ABC7	37593	3877	25962	49224		
NA18507	ABC8_a	36704	3848	25160	48248		
NA18507	ABC8_b	36088	1913	30349	41827		
NA18956	ABC9	39512	2260	32732	46292		
NA19240	ABC10	41005	1837	35494	46516		
NA18555	ABC11	40033	1768	34729	45337	33386	44215
NA12878	ABC12	39752	1396	35564	43940		
NA19129	ABC13	39289	1775	33964	44614	32850	44356
NA12156	ABC14	39442	1727	34261	44623	33026	45332

Clones were assigned to different categories based on the nature of the discrepancy with respect to the human reference genome (see Table S1 for the number of clones assigned into different categories for each library). A random tiling path of clones was developed for each individual genome at a density of 1 clone/5kb. We estimate that a contiguous set of clones (and concomitant high-quality sequence) can be retrieved for each genome for over 98% of the euchromatin with 93% of the bases covered by four or more clones per individual (Figure S1). We focused on characterizing clones discordant by length (corresponding to insertions or deletions) or orientation (inversion breakpoints). A chromosome by chromosome view of the discordant fosmid ESP placements for each library is shown in Figure S3. Discordant clones are colored by library.

Table 3. Color code for discordant clone maps

Library	Population	Color
ABC7	YRI	green
ABC8	YRI	forestgreen
ABC10	YRI	blue
ABC13	YRI	cyan
G248	-	black
ABC9	JPT	purple
ABC11	CHB	red
ABC12	CEU	orange
ABC14	CEU	hotpink

The end-sequence placements are mapped in the context of gaps within the assembly (purple) and segmental duplications (grey bars). Sites defined by at least two clones from a single library are indicated by black bars, with validated sites indicated by a yellow bar. Deletions in the library source relative to hg17 are defined by clones whose apparent insert size is too large.

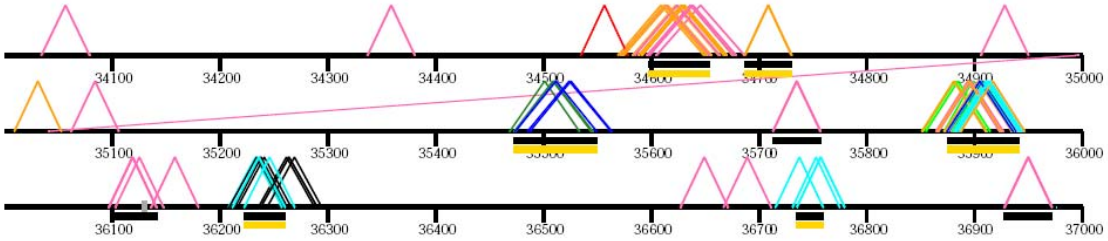


Figure 2. A 3 Mbp region on chr2 showing nine ESP predicted deletions. Six of the nine sites have been validated.

Clones whose apparent insert size is too small represent potential insertions. In addition to spanned insertions (smaller than 40 kbp), Figure S3 shows the position of one-end anchored clones (OEA clones, section 8 of Supplementary Material).

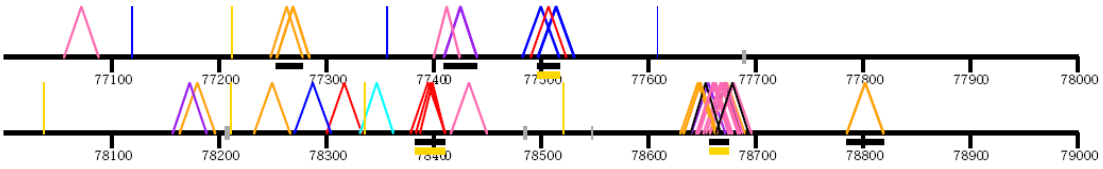


Figure 3. A 2 Mbp region on chr12 showing six ESP predicted insertions, three of which have been validated.

Clones whose ends map in the same direction (rather than mapping in an inward orientation) correspond to potential inversion breakpoints. The capture of a single inversion breakpoint is sufficient to identify a potential site of inversion. When both inversion breakpoints were captured they were merged together into a single inversion locus (Table S7).

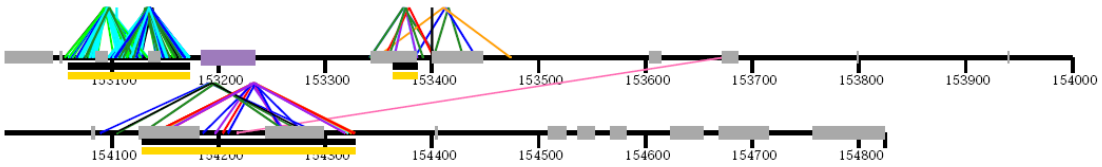


Figure 4. The last 2 Mbp of chrX contains three validated inversions. Both breakpoints for the first site were detected by clones from each of the four Yoruba libraries.

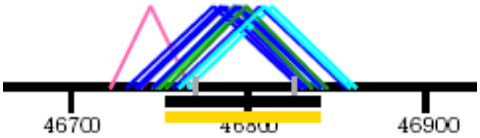


Figure 5. A deletion on chr3 only predicted in the 4 Yoruba libraries. Genotyping confirms that this variant is stratified (Table S5, deletion allele frequencies: YRI=0.537, CEU=0, and JPT+CHB=0.017; F_{ST}=0.49)

We also identified clones which are consistent with other types of rearrangements such as clones which appear to map to different chromosomes (termed trans-chromosomal clones) and clones where only one end maps against the reference assembly (termed one-end anchored [OEA] clones, Table S2).

5. Experimental Validation

We validated computationally predicted sites of structural variation using one or more of the following experimental methods: multiple complete digest (MCD) clone fingerprinting, microarray comparative genomic hybridization (arrayCGH), sequence analysis of clone inserts, or confirmation of a novel insertion of sequence. In addition, for larger inversion events, fluorescent *in situ* hybridization assays were developed. Combined, these methods confirmed and refined the location of 1,695 non-redundant sites of structural variation, including 747 deletion loci, 724 spanned insertion loci (< 40 kb in size), and 224 inversion loci. Supplementary Table 3 provides a complete list of insertion/deletion events predicted in each sample using the ESP approach, along with a summary of the validation results. Overlapping predictions were combined into a non-redundant list of predicted structurally variant loci given in Table S4. Similar information for inversions is given in Tables S6 and S7.

Table 4: Experimental validation of 1,695 structurally variant loci

Event Type	Total Validated Loci	Supported by MCD	Supported by array-CGH	Supported by Overlap with Novel Insertion Loci	Supported by FISH	Supported by Fosmid Insert Sequencing
Deletion	747	615	477	0	0	130
Insertion	724	567	142	194	0	100
Inversion	224	209	0	0	7	35

5.1 MCD Clone Fingerprint Analysis

"Multiple Complete Digest" (MCD) fingerprints were obtained for each fosmid from independent digestion with four restriction enzymes[6,7]. A total of 3,824 fosmid clones were analyzed using our semi-automated, high-throughput Large Insert Genome Analysis (LIGAN) pipeline, which confirms structural variants by comparing fosmid fingerprints and ESPs to the human genome reference sequence. This included 3,371 clones identified based on 3 standard deviation or 0.5% size thresholds, of which 1,331 were confirmed (39%). Additionally, we tested 453 predicted inversion clones of which 340 had fingerprint patterns consistent with inversion.

The 3371 clones represent a redundant subset of the total sites identified by the ESP approach throughout the course of this project. The selection process was driven by the need to exclude insertion/deletions sites where it appears that the same clone was end-sequenced twice, indicating that the identified site is not actually supported by two or more independent clones (termed "clonal sites"). Simply applying the described mapping criteria identifies 7184 potential non-redundant sites of insertion/deletion variation (2,777 deletions, 4,407 insertions, Table S4). However, defining clones from a single library which have an end maping within 30 bp of each other as "clonal" (as described in Tuzun et al.[5]) removes over half of these sites. Using this definition, there are a total of 2990 non-clonal insertion/deletion sites (1,345 deletions and 2,849 insertions). We have added this information (the "minspread" column, which gives the minimum distance between clones from an individual library supporting a site) to Table S4 as a measure of confidence for non-validated ESP predictions. Final clone selection was based on a manual review of the clone placement maps (Figure S3 a-c), and the manuscript focuses on a description of the sites which were validated using several different approaches. For completeness, all of the identified sites are given in Table S3 and S4 with the validated sites labeled as such.

The digests were electrophoresed in agarose gels with 48 sample lanes interspersed with five marker lanes. MCD fingerprint data was extracted computationally from gel lanes by Quantitative Gel Analysis Program (QGAP) (G. K-S. Wong, unpublished), which computes both the size and fragment multiplicity of each band. The accuracy of these measurements is critical for matching experimentally derived MCD fingerprints with the virtual fingerprints calculated from the reference sequence. The average enzyme-to-enzyme variation in estimates of clone size was typically <10%; greater discrepancies were primarily due to the presence of large, difficult-to-size fragments. Barcode labels, tracked in the Oracle database of the UWGC's laboratory-information management system (LIMS), maintained fidelity between the clones and their associated data sets.

Genomic Variation Analysis (GenVal) software [8] used ESPs and MCD fingerprints to tile fosmids against corresponding excerpts of the reference human genome. The reference sequence was used to anchor the ESPs and to create a virtual reference sequence-derived-restriction map (SDM), which was compared to the MCD fingerprints for fine-grained comparison of the fosmids and the reference sequence. If both end-sequences aligned in the correct orientation, a concordant or "full position" was formed; otherwise, one or more "half positions" were formed. Half positions occurred for a number of reasons, including inversions and duplicated sequences. Based on the tiling, GenVal computed the difference between clone length estimated from ESP alignment and average MCD insert size. These estimates, and the matching of MCD fingerprints to the SDM, were manually evaluated to identify structural variants. Clones with discrepancies confirmed by fingerprint mismatching of < -2.5 kb and > 2.5 kb were identified as spanning insertions or deletions, respectively. Clones spanning inversion breakpoints were validated: 1) by partial matching of MCD fingerprints to the SDM with each end sequence and 2) in the absence of a concordant position with one-to-one correspondence of the MCD fingerprints and SDM. (GenVal produced all end-sequence alignments within 5% of the

best alignment score; thus, clones could have multiple full and half positions. Alternates to the original placement of clone ESPs were selected only when they produced a concordant position with one-to-one correspondence of the MCD and SDM.

5.2 Array Comparative Genomic Hybridization (arrayCGH) Analysis
We designed two customized oligonucleotide arrays (NimbleGen and Agilent) targeted to regions identified as insertion or deletion, based on ESP mapping for the first seven analyzed fosmid libraries (G248, ABC7-ABC12). All hybridizations used sample NA15510 (G248) as the reference. Note that the array design covers some regions identified by less stringent size thresholds.

5.2.1 NimbleGen Oligonucleotide Microarray Design
We targeted 1,211 deletion intervals smaller than 250 kb in size (280,997 probes with an average density of one probe per 173 bp) and 11 deletions ranging from 250 kb to 1 MB in size (12,202 probes with an average density of 681 bp). Additionally, we targeted 272 insertion intervals which overlapped segmental duplications (30,330 probes with an average probe spacing of 173 bp), reasoning that these insertion intervals were most likely to harbor copy-number variation detectable by arrayCGH (gains and losses of duplicated sequences). As a control we included 4 kb of invariant flanking sequence on each side of each interval (66,127 probes, mean probe spacing of 1 probe/173 bp). In addition, we included 6 control regions not predicted to harbor structural variants (the regions around the CFTR, ALB, FOXP2, BRCA1, HoxA, and HoxD genes; 4,195 probes with an average probe spacing of 170 bp). Due to overlap among interval classes, these probe counts contain some redundancies.

Hybridizations and data normalizations were performed by the manufacturer using sample NA15510 as the reference. The results were analyzed using a simple hidden Markov Model, which was tuned using parameters from the control regions and a set of 283 manually reviewed intervals on chromosome 1. Results of the HMM analysis were post-processed with the requirement that called intervals span at least 10 probes and are reported in Table S3.

5.2.2 Agilent Oligonucleotide Microarray Design
We designed 2 x 244,000 probe-feature oligonucleotide microarrays comprised largely of sequences from the reference assembly hg17 and targeting intervals indicated by multiple discordant fosmids for both deletions and insertions. We selected 337,658 probes spanning 1157 putative deletion intervals with at least 10 probes over the interval and with a median probe spacing of 106 bp. Note that the same starting deletion intervals were used for the the NimbleGen design, but 54 failed Agilent design criteria. Unlike the NimbleGen design, we targeted all regions spanning putative insertions (n=1100), but selected probes at a lower density of approximately 1 probe per 300 bp, adding 67,000 probes. In both cases, high-density coverage spanned 1 kbp into flanking regions, and lower density coverage (1 probe/kbp) up to 5 kbp from end-points defined by the fosmid end sequence placements.

We designed a separate microarray for the new insertions based on one-end anchored sequence contigs (section 8). Probe selection methodology was quite similar: candidate probes were tiled across the non-repeat-masked portions[9]. These candidate probes were scored as described above, and final probes were selected by pair-wise filtering to achieve a targeted probe spacing of 1 probe/100 bp. This provided 1555 novel insertion sequences spanned by 14,053 probes with a median number of 9 probes per sequence (median spacing of 1 probe/95 bp). In addition, we selected 19,863 arrayCGH control probes selected from Agilent's Human Genome CGH Microarray 244A, including 18,692 autosomal probes and 1171 chrX probes. All chrX probes and 4100 autosomal probes were tested in replicate (present on both arrays). For background noise to signal characterization, 250 of the autosomal probes were printed in triplicate on each array. Finally, each microarray includes a QC grid of 5,045 features which also includes a probe set for normalization purposes.

Hybridization experiments were performed according to the manufacturer's instructions. Pairs of dye-reversed hybridizations were performed for each of the 8 samples relative to the reference (NA15510), as well as three self-self (NA15510 used as both 'sample' and 'reference') hybridizations. Other details on the sample labeling, hybridization, scanning, and normalization procedures can be found at http://www.chem.agilent.com/scripts/literaturePDF.asp?iWHID=39980 and http://opengenomics.com/pdf/pn_gs_feature_extraction.pdf.

Regions of statistically significant copy-number change were determined using two different methods. For intervals that mapped to the reference assembly, we used the ADM-2 (aberration detection module) algorithm on $\log_2$ ratios of fluorescent signals from the sample and reference. The ADM-2 algorithm uses an iterative procedure to identify all genomic regions where the average log2 ratios deviate from the expected value of 0 over a given interval; $\log_2$ ratios are weighted by $\log_2$ ratio error as calculated by Agilent Feature Extraction software. A statistical score is assigned to this deviation and the most significant score is reported at each iteration. We set the ADM-2 score threshold at 5 and calls with average log2 ratios less than 0.25 were excluded. With these settings, we demonstrated concordant genotyping results from 98% (158/162) of validated genotypes determined using a combination of quantitative PCR and custom Illumina GoldenGate assays for the same 8 samples.

For novel insertion sequence contigs (described below), we examined Cy3 and Cy5 signal levels from the arrays to identify sequences with significant signals above the baseline noise level. We did this by eliminating probes for which the maximal value of the mean ProcessedSignals (of the dye-flip paired measurements for each sample) did not exceed 100 counts for any of the samples, including the three self vs. self replicates. Note: 100 counts corresponds to approximately 10% of the median value across each array. This eliminated 11.5% of probes for all contigs. Intervals with fewer than 4 probes that passed the signal filter were removed from subsequent analyses, leaving 91% of the sequence contigs confirmed as human using this approach.

To identify intervals with copy-number variation, the mean and standard deviations of the $\log_2$ ratios of surviving non-outlier probes were calculated for each contig interval. Copy-number differences of contig intervals for each sample were considered significant relative to the self-self measurements of the reference by application of Student's 2-tailed t-test with a significance p-value cutoff of 0.005. This corresponds to an expected false discovery rate of approximately 5% (based on 1,152 significant calls made from a total of 10,064 tests). Since the reference individual (G248) is female, we restricted the copy-number analysis of the novel sequence contigs to those which mapped to the autosomes. Overall, 90% (1,299 of 1,435) of the OEA sequence contigs confirmed as human, with 32% showing evidence of copy-number variation at an expected false-discovery rate of approximately 5%.

5.3 Reference Individual Effect

As described above, automated calls were post processed with only calls spanning at least 10 probes (NimbleGen) or 5 probes and 1kb (Agilent) retained. Accurate genotyping by arrayCGH is dependent upon the (typically unknown) genotype of the reference sample. We attempted to correct for this by using our set of predicted deletion variants in NA15510. If NA15510 was predicted to harbor a deletion variant, arrayCGH "gain" calls over the interval were considered to support the prediction. This set of NA15510 predictions contains both false positives and false negatives and is dependent on a set size threshold, highlighting the potential utility of a well-defined, sequence-confirmed set of structural variants for use in future studies.

5.4 MCD Fingerprint Analysis vs. arrayCGH

The clone based discovery and validation procedures we employed involve comparisons against a reference genome sequence, whereas the arrayCGH validation experiment we undertook involved a comparison against a reference sample of uncertain CNV genotype. Since NA15510 was used as the reference sample in all hybridizations, we focused our comparison on variant loci which was supported by fingerprint analysis of a non-NA15510 clone. Given the different probe selection schemes used, we assessed deletion and insertion loci separately.

This study identified 747 validated deletion loci. These locus definitions are based on a merging of FES predicted sites from each of the nine libraries. 306 of the 747 loci encompass a sample-level prediction from a library other than G248 which is supported by MCD analysis and is covered by both arrayCGH platforms at a density of 200 bp per probe. 74% (227/306) of these loci were confirmed by a "loss" call in a predicted sample by at least one of the platforms. An additional 7 loci were supported by a "gain" call coupled with an ESP-predicted deletion in the G248 reference. This leaves 24% (72 of 306) of the loci which were not confirmed by arrayCGH. The median predicted deletion size (based on MCD analysis) of these 72 loci is 6.2 kb with 86% (62 of 72) smaller than the G248 ESP detection threshold of 8.25 kb.

Predicted insertion sites were targeted at different probe densities using different selection criteria (section 5.2 of this document). Therefore, for the insertion sites we restricted our analysis to regions covered at a density of at least 1 probe every 400 bp.

There are 30 insertion loci which meet the coverage criteria on both arrays and are supported by MCD fingerprint analysis as described above. 14 of these loci are supported by a "gain" call in at least one of the arrays for the predicted sample, leaving 16 insertion loci (53%) which were not confirmed (the median predicted insertion size for these 16 loci is 5.8 kb). The number of insertion loci meeting these criteria for both microarrays is low because of the different targeting methodologies employed in the two designs. The Agilent design (which targeted all ESP predicted insertions) confirmed 23 of 220 loci using the criteria described above, leaving 90% which were not confirmed (197 of 220, with a median size of 5.9 kb). For the NimbleGen design, which was restricted to those predicted insertions which overlapped segmental duplications, 43 of 113 sites were confirmed and 70 were missed (62%, median size of 7.9 kb).

The lower confirmation rate for insertion sites is not surprising and can be accounted for by several factors. First, our analysis indicates that ~25% of ESP identified spanned insertion loci may involve sequences not present in the reference assembly (Supplementary Material Table 3). The remaining insertions, which involve sequences represented at least one time in the assembly, may not actually be interrogated by the probes placed on the array since the probes were designed across the interval identified by ESP analysis rather than against the (potentially unknown) sequence which is inserted.

6. Genotyping

6.1 Reference Genotypes

Using a custom Illumina GoldenGate assay, we genotyped the 270 HapMap individuals for 34 polymorphic sites of sequenced structural variation[10]. Each of these sites was identified in the initial fosmid library and insertion-specific probes were designed based on complete fosmid sequence[5]. Additionally, we used quantitative PCR to genotype the HapMap panel for 23 of these 34 loci. The resulting set of genotypes (Table S12) were used as a benchmark for validating genotyping results using other platforms.

6.2 Genotyping Validated Sites using the Illumina Human1M Genotyping BeadChip

We developed a novel approach based on mixture likelihood clustering to genotype biallelic insertion/deletion variants using Illumina Infinium genotyping assays (Cooper, Zerr in preparation). For a given SNP within a deletion, the algorithm attempts to identify distinct clusters of fluorescence intensity values that, in principle, correspond to distinct copy-number states. We identify 'null' (homozygous deletion) samples as those samples with near-zero normalized intensity values and assume that remaining heterozygotes (as called by the Illumina BeadStudio software) are diploid. The remaining samples are subsequently analyzed using a two-component (hemizygote and diploid) mixture model, with parameters estimated using the EM algorithm maximizing the likelihood of the observed fluorescence data. We require multiple SNPs to support consistent copy-number genotypes within any given deletion event. Furthermore we require that samples known to harbor a deletion (the fosmid-end sequenced sample) were correctly classified as either null or hemizygous. This scoring framework has been implemented in the context of a greedy search algorithm to identify subsets of informative probes in and around an annotated deletion event, allowing for the exclusion

of noisy probes (which may result from probes within the deletion but overlapping duplicated/repetitive sequence) and/or breakpoint uncertainty.

We utilized publicly available SNP genotyping data from the Illumina Human1M genotyping platform for 125 HapMap DNAs of African, European, and Asian descent. These samples include 28 parent-child trios and 7 sample-level replicates. Since neither balanced events nor novel insertion sequences could be assessed using this platform, we focused solely on genotyping validated and breakpoint-refined deletion events. We used our algorithm to search for informative probes within 520 deletion sites (113 of these sites had sequence-defined breakpoints and 407 had breakpoints defined by arrayCGH) using the Illumina Human1M data. This number differs from the total number of validated deletion events reported in this study (see Table 1), as we focused on the subset of variants with good breakpoint estimates and excluded those deletions with only MCD validation.

Our initial search yielded putative copy-number genotypes for 144 non-overlapping sites (150 'sites' including redundant events). Many (255) of the original 520 sites could not be genotyped due to a lack of two or more probes within the deletion interval. An additional 115 sites spanned two or more probes but did not yield genotypes. Manual inspection of these data suggests multiple potential sources of genotyping failure. First, the sensitivity of the clustering approach is directly related to allele frequency, and thus rare alleles, in particular variants that are only present within one individual, cannot be reliably genotyped. Second, many probes within deletions showed obvious signals of cross-hybridization (as can happen for probes that overlap duplicated sequences), which disrupt accurate copy-number inference. Finally, some sites may actually harbor multiple distinct alleles (see Figure 4) that affect different subsets of probes within the interval and thus fail to yield consistent genotype calls. We are unable to quantify the relative influence of each of these potential sources of error on SNP genotype information alone.

To verify genotype accuuracy, we independently analyzed 27 deletions that had been previously genotyped using a combination of quantitative PCR and custom Illumina GoldenGate assays on the complete HapMap panel (Table S12), 13 of which provided good initial genotypes. Of these 13 intervals, 12 had concordancy rates of 97% or better and 1 exhibited a concordancy of 94%. Thus, we conclude strong overall reliability of our genotype inference method. However, manual review of the remaining sites did indicate that some sites provided questionable genotypes, and we conservatively eliminated these as being suspect. After eliminating these, we identified 130 deletions that were present at greater than ~1% frequency. Genotyping replicates (n=7) revealed >98% reproducibility with >90% of all sites exhibiting no genotyping inconsistencies. More than 98% of the children's genotypes were consistent with Mendelian transmission based on an analysis of 28 parent-child transmissions for each of these deletions (~3,600 trio genotypes). Allele frequencies of the 130 sites are listed in Table S5 and plotted as histograms in Figure S5.

7. Cross-Platform Comparisons

7.1 Comparisons Among Studies

We performed a comparison between the validated sites of structural variation identified in our analysis using fosmid-end sequence pair mapping (ESP) with a set of CNV annotations generated for the same 8 HapMap samples in other studies, which employed both Affymetrix SNP array genotype data and BAC arrayCGH experiments (McCarroll and Kuruvilla et al., unpublished and Redon et al.[11]). Affymetrix SNP 6.0 data was analyzed using a hidden Markov algorithm (*Birdseye*)[12] that makes joint use of data from SNP and copy-number probes to identify regions of gain and loss in the ESP samples. Using copy-number differences between chromosome X and the autosomes as a benchmark, it identifies regions that deviate in signal intensity and combines probe information (Viterbi algorithm) to generate a LOD score expressing the likelihood of copy-number difference. The algorithm is part of a software package, Birdsuite (developed by JM Korn). These CNVs were supplemented by an additional set of common CNVs by identifying regions of the genome in which a series of probes showed highly correlated patterns of intensity across the 270 HapMap samples (McCarroll and Kuruvilla et al., manuscript in preparation). All of these regions were required to be identified in two independent experimental runs of the samples. For our initial comparison, we used only those sites estimated to be greater than 5 kb, and determined for each study the number of annotations from the other studies that overlapped within each sample. In principle, this may be affected by random overlaps; however, we restricted our comparison to genotypes (i.e. comparison of events within a sample), and thus the amount of random overlap is expected to be small.

We find that the vast majority of events annotated within a sample are unique to one study. For example, of the 3302 annotations larger than 5 kb generated by ESP, only 415 (12.5%) overlap with a CNV annotation from Affymetrix 6.0 genotyping efforts, and only 186 overlap (5.6%) with an annotation from BAC arrayCGH of Affymetrix 500K arrays. There are ~70 (66-80) events within these 8 samples detected by all three analyses. ~75% of the annotations generated by Redon et al. and McCarroll et al. are unique to their respective studies, and ~85% of the annotations generated in this study are unique. These overlap statistics are summarized in the table below and in the context of a Venn diagram.

Table 5. Overlap counts for the CNV annotations generated for 8 HapMap samples using the Fosmid ESP approach (Kidd), Affymetrix 6.0 SNP Genotyping (McCarroll), and BAC-based ArrayCGH and Affymetrix 500K arrays (Redon et al.) We refer to these as Kidd, McCarroll and Redon sites.

	Redon	McCarroll	Kidd
Redon	810	254	178
McCarroll	218	2219	412
Kidd	186	415	3302

	3-way
Redon.McCarroll.Kidd	80
McCarroll.Kidd.Redon	67
Kidd.Redon.McCarroll	66

Note that the table is not symmetric, as the relationship between annotations and events is not necessarily one-to-one, and thus the overlap measure depends on the orientation of the analysis (i.e. the number of Redon et al. sites overlapping a Kidd et al. site is not identical to the converse). Several overlap counts from 3-way intersections are also shown, analyzed in the order indicated.

With respect to the variants identified by ESP, it is unlikely that the uniqueness is an artifact caused by a large number of false positives. There are many sites detectable by the ESP approach that would be missed by other platforms, such as insertions of sequence not represented in the reference assembly. Second, when we restrict our annotations to only those that have been directly validated (i.e. through MCD, CGH, or complete sequencing, rather than through overlap with a confirmed CNV locus), the proportion of overlapping vs. unique sites does not change (not shown). Finally, to completely eliminate the possibility of false positive contamination, we examined the 52 deletions in these 8 samples that have been validated and resolved to the base-pair level through complete fosmid sequencing. 11 of these are annotated by McCarroll et al, 3 are annotated by Redon et al, and only 1 is annotated by all three analyses; thus, 41 out of 52 (~80%) of even unambiguous deletions are unique to this study.

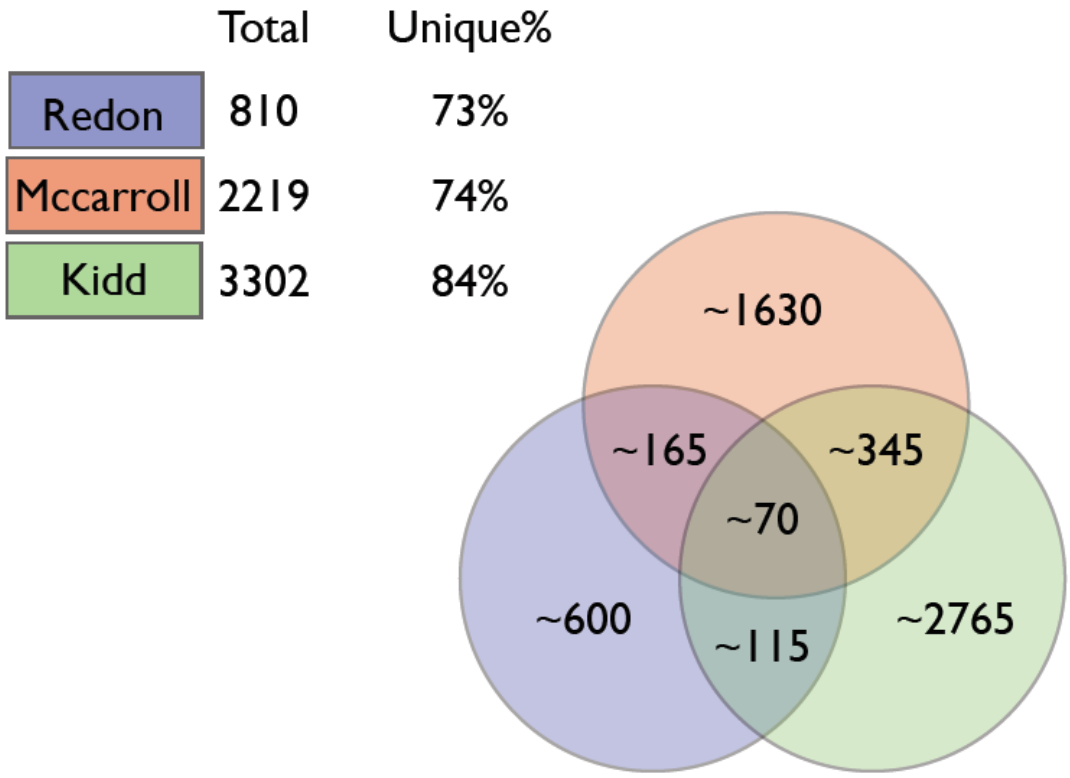


Figure 6. Visual representation of the data presented in the table above. Note that approximate values are indicated, resulting from the fact that the precise number of overlaps depends on the orientation of the overlap analysis (circles not drawn to scale).

We also conducted a broader analysis that included all sites predicted by ESP (including sites without physical validation) in these same 8 individuals, with the goal of determining the extent to which we could confirm additional variants annotated by Redon et al. and McCarroll et al. and also identify false negatives that may result from ESP mapping.

BAC arrayCGH and Affymetrix 500K arrays (Redon et al.): Of the 814 sample-level calls made on the 8 samples analyzed in common by Redon et al. (treating the WGTP genotypes separately from the 500k genotypes), 322 sites overlapped an ESP prediction in the same sample. Thus, 492 (~60%) sites do not overlap an ESP-annotated variant in our primary set of mutation predictions; for simplicity we refer to these CNV genotypes as 'missing' with respect to our default set of ESP-mapped sites. We investigated several possible explanations for the missing CNV annotations, including the possibility of small variants that are below the 3 standard deviation threshold for fosmid size detection, variants in genomic intervals that are rich in duplicated sequence, and variants in genomic regions for which we have poor fosmid clone coverage.

We find that ~20% of the 'missing' variants overlap an ESP prediction set with a reduced size threshold (2 standard deviations above or below the fosmid library mean), suggesting

there may be many smaller variants in this set. We find that ~10% of the missing annotations overlap regions that harbor multiple discordant fosmids that are not uniquely placed or align with inconsistent orientation, indicative of CNVs in duplication-rich regions that are difficult to precisely identify through ESP-mapping. Finally, ~33% of the missing CNVs overlap regions with poor clone coverage; a substantial fraction of these sites actually overlap with a single discordant fosmid clone, suggesting that deeper library sequencing would eventually identify many of them. However, non-random lack of clone coverage is also a problem, as we note that several of the libraries harbor HLA haplotypes that are sufficiently distinct from the reference assembly that they fail our alignment/mapping criteria.

Finally, we are still left with over 150 missing CNVs that cannot be accounted for by any of the above analyses. We conjecture that this may in part result from the experimental limitations resulting from utilizing a 'reference' genome. The reference sample used to generate the WGTP BAC arrayCGH data in Redon et al. is not the same DNA that is represented in the human genome assembly hg17, which we utilized as our 'experimental' reference. Given that many of the sites we identify are common in the human population, this change in reference may result in a distinct set of sites that are gained/lost in these respective analyses. We simulated the effect of changing references by considering how many sites would be 'missed' in each library assuming one of the other libraries was used as the reference assembly in place of hg17. We find that, on average, about 25% of sites would be expected to be missed (with other 'new' sites gained in distinct places) purely as a result of this change in reference. Thus, the final missing sites (~20% of all the calls) may potentially result from this experimental limitation.

Affymetrix 6.0 Comparison. We intersected our validated ESP CNV sites with predictions on the same eight individuals analyzed using the Affymetrix 6.0 platform With respect to these predictions, we find that only 858 out of 5,474 sites are captured by our analysis in these 8 samples, leaving thousands of 'missing' annotations. In this case, the vast majority of the missing CNVs are too small to be picked up by our ESP-mapping approach; the median length for these variants is ~2.5kb, well below the limit of detection for all of the libraries analyzed here. Accordingly, ~8% of the missing sites overlap an ESP annotation using a 2 standard deviation size threshold. We also find that ~8% of the missing sites reside in duplication-rich regions without unique fosmid placement, and an additional 14% in regions with poor overall clone coverage. After accounting for each of these factors, ~3000 variants were still not detected. However, these have a median size of only 1 kb, and thus are not detectable by fosmid ESP-mapping at any reasonable threshold.

8. Detection of Novel Sequences by Mapping OEA clones

One-end anchored (OEA) clones were defined as clones where only one end mapped against an hg17 chromosome (chr1-chrX, at least 150bp at 99.5% identity) and the other end did not. As part of the standard ESP analysis pipeline, each end-sequence is used as a query to search against hg17 using megaBLAST[13] version 2.2.9 with an e-value cutoff of 1e-40, an identity cutoff of 0.80, and a minimal hit score of 90 (options: -D 2 -v 7 -b 7 -e

1e-40 -p 80 -s 90 -W 12 -t 21 –F F). Only clones where the end mapped against chr1-chrX (no random or unk) were considered. This mapping must have been a "best" placement, with a similarity of at least 0.995 including at least 150 non-repeatmasked bases (repeatmasking performed at 2% divergence). For the unmapped end, we required more than 30 basepairs of PHRED quality Q30 and at least 200 basepairs of PHRED quality Q20. We focused our analysis on OEA clones from the first 7 libraries (G28, ABC7-ABC12). After excluding low quality sequence reads and common sources of contamination (Epstein-Barr Virus, bacteria, etc.), we identified 21,556 sequences from these libraries (Table S2).

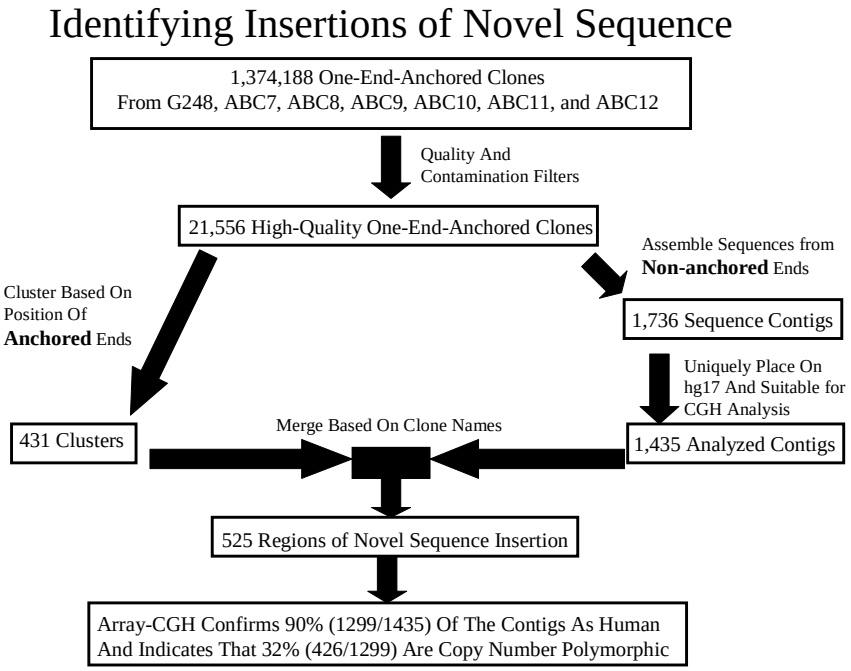


Figure 7. Flow chart summarizing analysis procedure of novel sequence insertions from the first 7 genomic libraries.

We analyzed these 21,556 OEA clones using two complimentary approaches. First, we focused on the anchored end of the OEA clones and identified 431 OEA clusters where each cluster contains at least 4 clones within 10 kb. The placement of these clones against hg17 was visualized using a coloring scheme where clones whose anchored end places in the forward orientation are colored gold and clones whose anchored end places in the reverse orientation are colored blue.

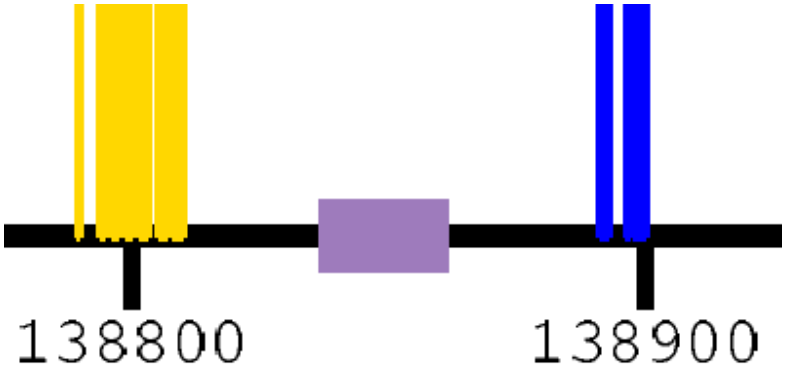


Figure 8. This example, from chr7:138.8-138.9 MB, shows two OEA clusters extending into a gap in hg17. Gold lines represents forward orientation OEA clones while blue lines represent reverse orientation OEA clones. The clones are pointing toward a gap (purple) in the reference assembly.

Second, we focused our analysis on the non-anchored end of the OEA clones. Using the TIGR assembler V2.0 we assembled the 21,556 non-anchored end sequences[14]. This procedure resulted in 1,736 sequence contigs (from 4,996 assembled end-sequences) and 17,063 unassembled singletons. The contigs had a median size of 983 basepairs, a GC content of 44%, and a repeat content of 32%. We additionally compared these 1,736 contigs against other assemblies of the human genome (hg18, and the Celera assembly[15]). Restricting the analysis to alignments at least 150 bp in length and 98% identity indicated that 48% (840 of 1,736) of the contigs showed some sequence similarity to an alternate human genome assembly. The common repeat content of these sequence contigs is summarized in the table below.

Table 6. Repeat content of 1,736 OEA sequence contigs (1,851,097 basepairs)

Repeat Class		Number of Repeats	Length Occupied (basepairs)	Percent of Total Sequence
SINEs		798	127438	6.88%
	ALUs	478	84762	4.57%
	MIRs	320	42676	2.30%
LINEs		631	218452	11.79%
	LINE1	425	173529	9.36%
	LINE2	188	41404	2.23%
	L3/CR1	18	3519	0.19%
LTR elements		376	111671	6.03%
	MaLRs	197	55295	2.98%
	ERVL	64	18020	0.97%
	ERV_classI	113	37833	2.04%
	ERV_classII	1	118	0.01%
DNA elements		158	33168	1.79%
	MER1_type	98	18117	0.98%
	MER2_type	36	11154	0.60%
Unclassified		0	0	0%
Total interspersed repeats			490729	26.48%
Small RNA		3	392	0.02%
Satellites		18	9128	0.49%
Simple repeats		429	72140	3.89%
Low complexity		355	29548	1.59%

We next sought to combine these two approaches. Of the 1,736 contigs, 1,435 had a consistent anchoring against hg17 (based on the placement of the anchored end of each OEA contributing to each contig) and sequence characteristics suitable for analysis using arrayCGH. Merging these 1,435 contigs with the 431 OEA clusters resulted in 525 distinct novel insertion loci (Figure S9).

9. Confirming Novel Insertions Via Optical Mapping

Optical Mapping is a high-throughput system for constructing whole-genome restriction maps from ensembles of single DNA molecules. Using directed fluid flow, high molecular-weight genomic DNA is deposited on a derivatized glass surface. The DNA is digested *in situ* with a restriction enzyme, then stained and imaged on an automated epifluorescence microscopy workstation. Custom machine-vision software deduces the mass of each DNA fragment from its integrated fluorescence intensity, creating an ordered restriction map from each DNA molecule. Using a process similar to sequence assembly, these individual-molecule restriction maps are assembled into a consensus map of contigs spanning up to 95% of the genome. These maps constitute a comprehensive, high-resolution representation of genome structure.

We applied the Optical Mapping platform to the GM15510 cell line from which the G248 clone library was constructed. We generated 2.2 million single-molecule restriction maps, then assembled them into 1140 consensus map contigs spanning 92.8% of the genome. We identified 11 clusters of OEA fosmids from the G248 library that did not map to gaps and examined the consensus map contigs aligning to these regions. In 8 of the 11 cases, the optical map identified a large insertion relative to the reference genome, with 6 of the identified insertions estimated to be larger than 40 kb (Table S8).

10. Sequence Analysis

We selected 405 fosmid clones predicted by fosmid ESP analysis to correspond to sites of inversion, insertion and deletion when compared to the reference genome. For each fosmid, we generated a clone shotgun sequence library and completely sequenced the insert of each clone . Sequences were assembled and viewed using phred/phrap/consed[16] software tools. A total of ~16.0 Mbp of finished or near-finished sequence was generated.

Table 7. Distribution of 405 sequenced fosmids

Library	Insertion	Deletion	Inversion	Total
G248	88	80	42	210
ABC7	9	16	6	31
ABC8	17	18	4	39
ABC9	11	10	0	21
ABC10	68	26	4	98
ABC12	4	2	0	6
Total	197	152	56	405

We compared fosmid insert sequences and corresponding human reference genome (hg17) sequences using BLAST and graphical visualization scripts (two-way_mirror.pl and miropeats[17]) to identify the extent of each rearrangement. We confirmed the ESP prediction for 278 clones, with 62 clones being ambiguous. Although experimental data confirmed 50/62 of these as harboring a structural variant, most of the breakpoints mapped to sites of large, complex regions of segmental duplications which complicated validation and, in some cases, prevented final sequence assembly (n=30).

The majority (53/64) of the clones that failed to confirm at the sequence level represented putative insertion events (Supplementary Material Table 8) as a result of a slight subcloning preference for “short inserts” as opposed to larger inserts. For example, instead of a clone carrying a 5 kb insertion and its end sequences mapping ~35 kbp apart, these clones simply carried an unusually short insert of 35 kbp. As insert size distributions became more tightly distributed, this effect became more pronounced. To eliminate this effect and still recover the additional structural variants afforded by a lower standard deviation, we performed a fingerprint analysis on every clone to eliminate those short inserts from subsequent analyses. Fingerprint (MCD) analysis indicates that 39 of the 53 clones would have been eliminated from sequencing as carrying inserts that were too short. Thus, with MCD analysis 93% (341/366) would have confirmed at the sequence-level.

Table 8. Clone Sequencing Summary

ESP Prediction	Clones Sequenced	Confirmed	Ambiguous	Not Confirmed
Insertion	197	103	41	53
Deletion	152	134	13	5
Inversion	56	41	9	6
Total	405	278	63	64

We sequenced 46 clones identified by ESP analysis which MCD analysis did not validate. 85% (39/46) corroborated the MCD analysis with 5 clones harboring a structural variant (2 were ambiguous). Each of these 5 MCD false-negatives corresponded to small insertion events (maximum size of 3.1 kb, mean of 2 kb), Based on these results, we estimate a false negative rate of 15%. Analysis of sequenced clones confirmed by MCD analysis indicates an overall false positive rate of 6% (3% for deletion predictions and 9% for insertion predictions).

We inferred the mechanism underlying structural variation through sequence comparison of the reference genome (hg17) and the complete fosmid insert sequence. This was performed in a series of steps. We initially compared each fosmid insert sequence against the human reference sequence using the program miropeats[17], which uses an indexing program (ICAass) to discover regions of homology and graphically displays an alignment of the reference and fosmid sequence as a postscript file. This successfully pinpointed large regions of homology beyond a specified threshold (s=400) and approximated the location of the four breakpoints between the reference and fosmid genome. Next, we annotated this alignment map of the reference and fosmid sequence by repeatmasking and annotating the positions of segmental duplications (Dupmasker, unpublished). This step frequently uncovered smaller repeat sequences (such as shorter duplicons or Alu repeats) that may have been missed by the miropeats analysis. Finally, once breakpoint regions were refined we created multiple sequence alignments (CLUSTALW[18]) to refine the location of the breakpoint. In order to classify an event as NAHR, we required at least 300 bp of paralogy at the breakpoint.

For each confirmed site of structural variation, we examined in detail the sequence content at the boundaries and assigned it a mechanism of origin. Four mechanisms were considered based on the following criteria: 1) Non-allelic homologous recombination (NAHR) based on the presence of direct repeats with high sequence identity (minimal length of perfect identity of 400 bp) at the boundaries of the rearrangement; 2) Non-homologous end-joining (NHEJ) based on the absence of homologous sequences at the boundaries, considering, although not requiring, the presence of additional (novel) sequences at the boundary of the rearrangement; 3) variable number of tandem repeats (VNTR) due to the expansion and contraction of a large number (>10) tandem repeat sequences (typical unit is less than 5 kbp in length); and 4) retrotransposition based on the absence or presence of a retrotransposon corresponding to the complete extent of the structural variant. Based on our thresholds of length detection, the latter was limited primarily to full-length LINE, HERV, and longer SVA retrotranspositions. We visually inspected and annotated each miropeats alignment using customized tracks corresponding

to human segmental duplication content (segmental duplication track [blue] and dupmasker [grey]), Refseq gene annotation (red) and common repeat content. An example of each mechanism follows. A complete set of all annotated alignments is provided organized according to assigned mechanism (see Figure S10).

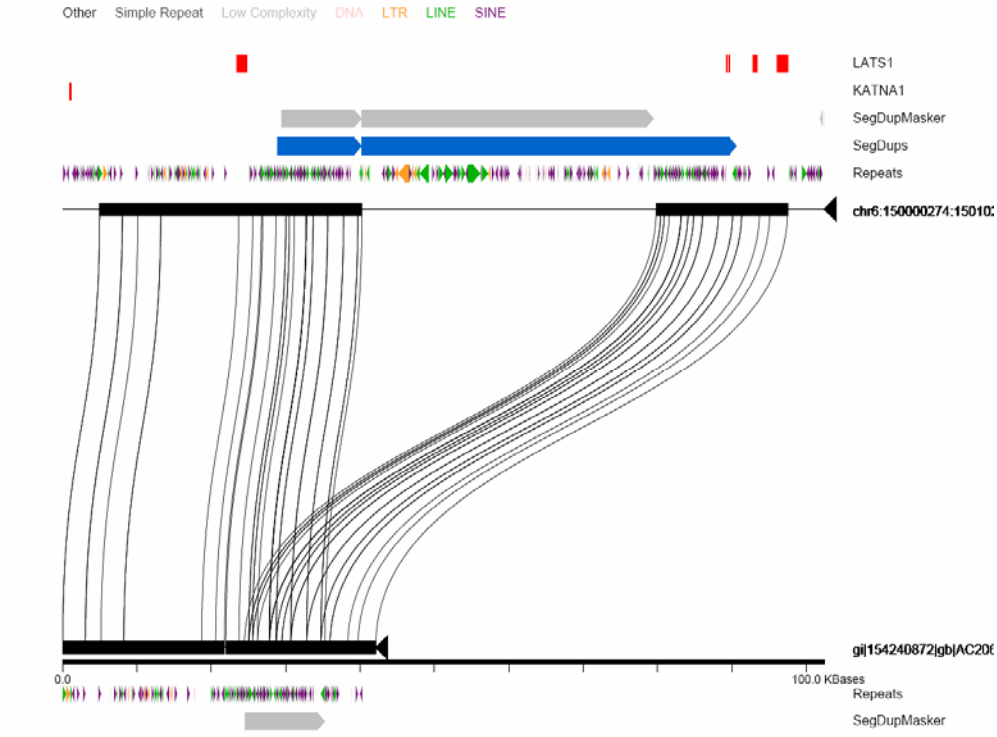


Figure 9a. Example of a deletion via non-allelic homologous recombination between segmental duplications within the intron of the LATS1 gene.

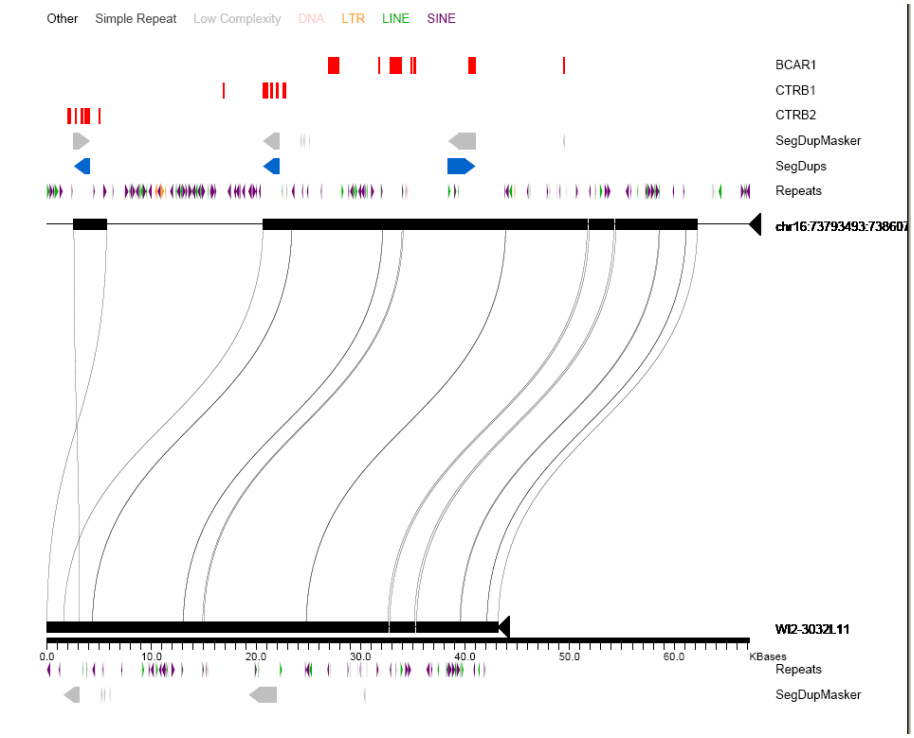


Figure 9b. Example of an inversion mediated by inverted segmental duplications (non-allelic homologous recombination) embedded within the CTRB1 and CTRB2 genes.

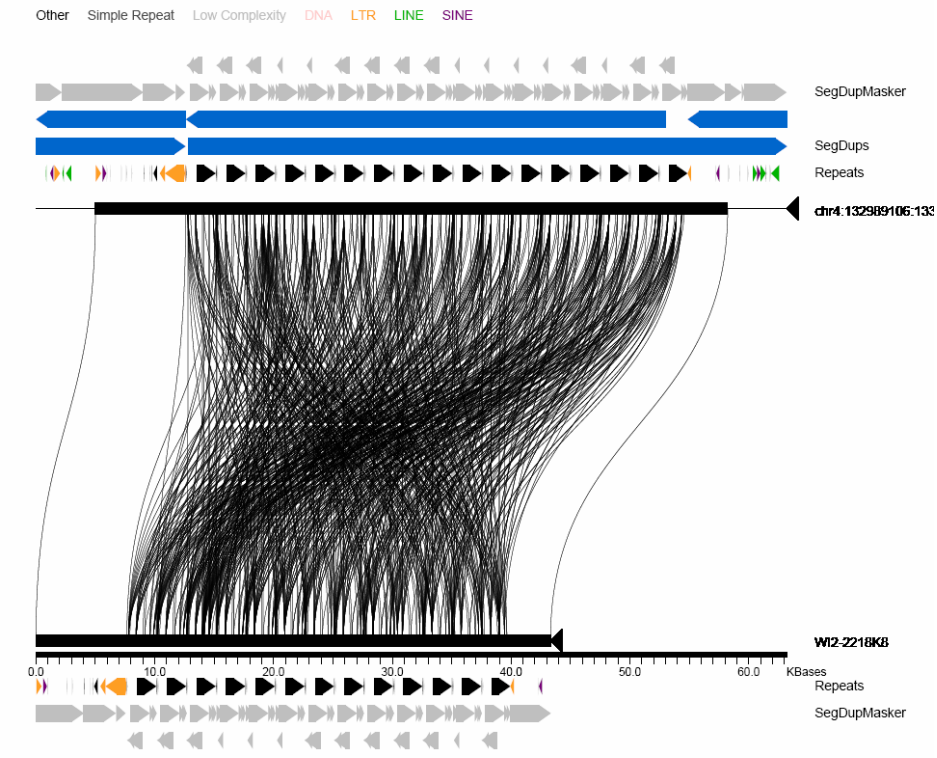


Figure 9c. Example of a contraction (deletion) mediated by variable number of tandem repeats.

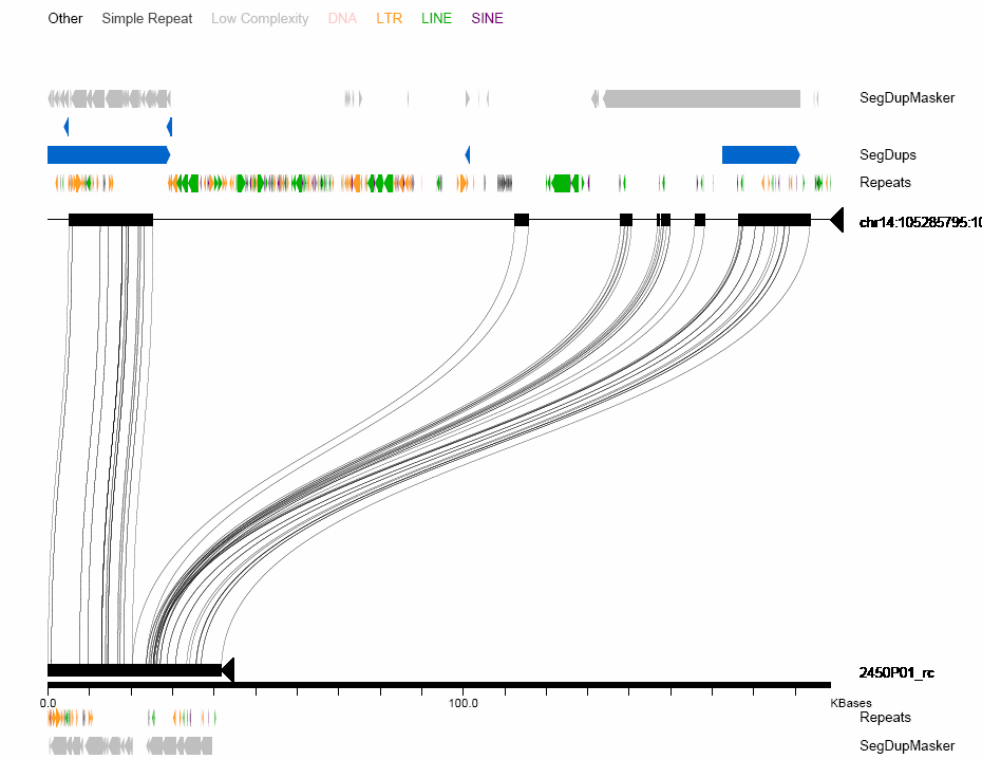


Figure 9d. Example of a complex deletion event involving 2 or more sequences likely by non-homologous end-joining.

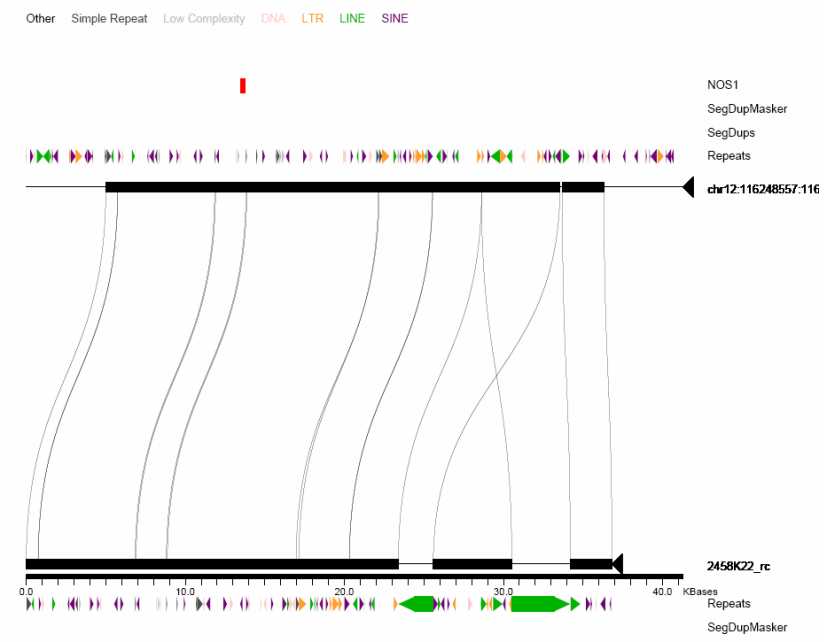


Figure 9e. An example of a possible L1 retrotransposition event followed by an inversion. Two or more events are required to convert the reference genome haplotype to the fosmid haplotype.

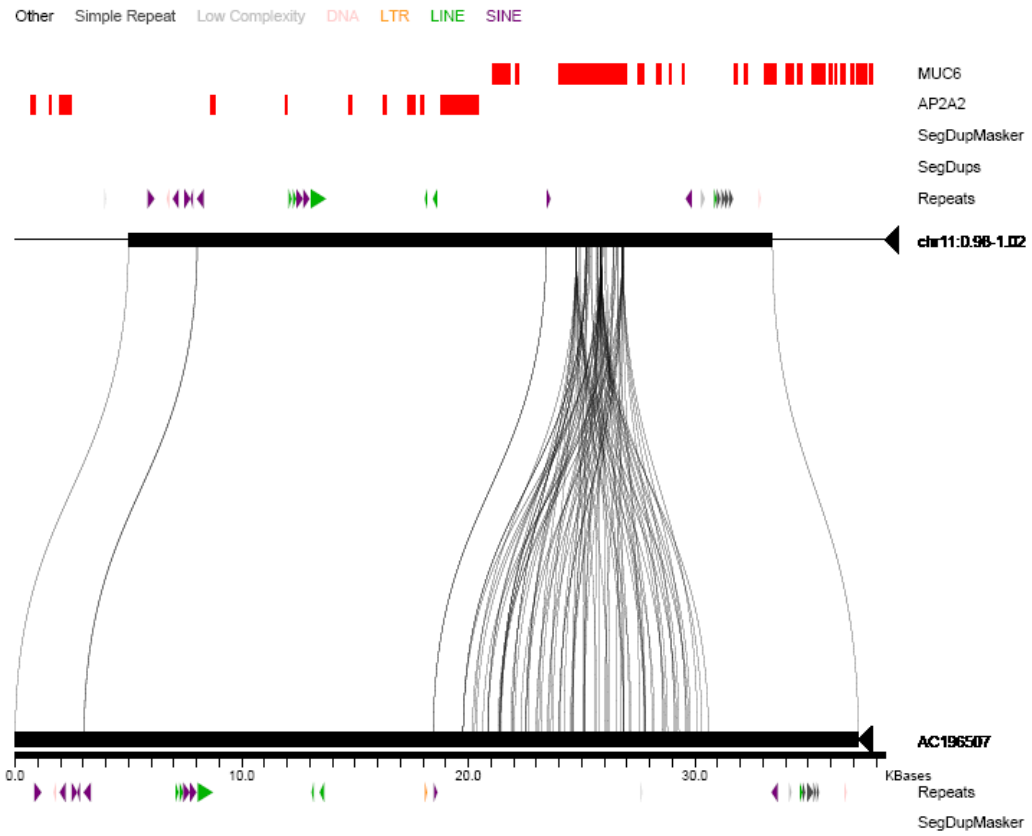


Figure 9f. An example of an VNTR insertion in a coding exon of the MUC6 gene.

11. SNP/Indel Analyses

We identified single nucleotide and small insertion deletion variants by comparing individual fosmid end sequences against hg17. Single nucleotide variants and deletion/insertion variants were identified using ssahaSNP[19] with cross_match (http://www.phrap.org) post-processing for more accurate deletion/insertion placements. For single nucleotide variation detection, NQS parameters were based on previous analyses which yielded a 5% false positive rate (Qsnp>=23 [minimum quality of variant base] Nnei=5 [number of neighbors in either side of a variant base] Qnei>=15 [minimum quality of neighbors] and Maxdiff=1 [maximum number of discrepancies in neighbors]). If the read aligns to more than one place in the genome, the best match was required to have 3-fold lower heterozygosity than the second match, or the read was discarded.

Validation*:* BCM and the Broad Institute have generated PCR directed sequence from 48 HapMap individuals across 10 ENCODE regions[2,3]. This data set includes 5 of the 8 individuals analyzed using fosmid ESPs. We used this data set as independent validation for our predictions based on fosmid ESPs. For the validation analysis, we only considered "perfect amplimers", i.e., those amplimers that included alignable forward and reverse PCR sequence reads for each of the 5 individuals. We determined the percentage of all fosmid-ESP SNPs/indels that were within the limits of these "perfect amplimers." Predictions within amplicons that are not unique ePCR hits were excluded. We used

polyphred[20] to detect SNPs from these amplimers, along with a pseudo trace generated from the reference sequence. If polyphred were a perfect detector of polymorphisms, then all true positive SNPs detected from the fosmids reads using ssahaSNP should be included in the polyphred output. In the case of variations detected from the fosmids and not detected in the same individual's PCR sequence, we inspected the PCR generated traces for any evidence of variation that was missed by polyphred.

Of 1988 total predicted genotypes, 1661 had a polyphred genotype consistent with the fosmid call (i.e., heterozygous including the fosmid allele, or homozygous for the fosmid allele, and were therefore counted as true positives). We randomly selected 100 of the remaining 327 predictions for visual examination. Of these 100, 39 had PCR reads that were of too low quality at the SNP position to determine genotype. We therefore removed 39% of the 327 (128) from the dataset total of 1988. An additional 38 predictions showed evidence of the correct genotype in the PCR reads, but were not called by polyphred. As a result, 38% of 327 (124) were added to the 1661 true positives total. 20 predictions showed no signs of the fosmid-derived SNP, so 20% of the 327 (65) are estimated to be false positive fosmid/ssahaSNP predictions. Finally, 3 predictions were classified as ambiguous. Thus, the overall SNP validation rate is 1-(65/1860)=96.5%. Due to the sparsity of fosmisd ESP coverage, a false negative rate could not be estimated.

Indel validation was done similarly to the SNP validation. We examined fosmid-read-derived DIP predictions within the "perfect amplimers" mentioned above, and determined by inspection whether each fosmid-based prediction was correct, a false positive, or undetermined due to low base quality or a microsatellite repeat (which are notoriously difficult to assess within the PCR-derived sequence). Microsatellites were defined as eight or more copies of a mono-, di-, tri-, etc. nucleotide. Predictions that were undetermined due to low base quality or the presence of microsatellites were not included in the calculation (in our aligned set there were 15 of the former and 105 of the latter). The false positive rate was computed as the number of false positives divided by the total of false positives plus the number of true positives verified by polyphred or manual inspection. We found 10 false positives, 40 predictions verified by polyphred, and 50 predictions verified by manual inspection, giving a false positive rate of 10%.

Single Nucleotide Variant Density Calculation. We calculated heterozygosity (number of observed single nucleotide variants divided by number of aligned bases) in sliding windows of 100 kbp with a 20 kbp step for each window. Each library was analyzed independently for both autosomes and the X chromosome using separate thresholds and regions which showed excess single nucleotide density (2 s.d. beyond the mean) were identified.

Data were also pooled across the libraries and analzed in 100 kbp non-overlapping windows to establish a heterozygosity distribution for genomic regions.

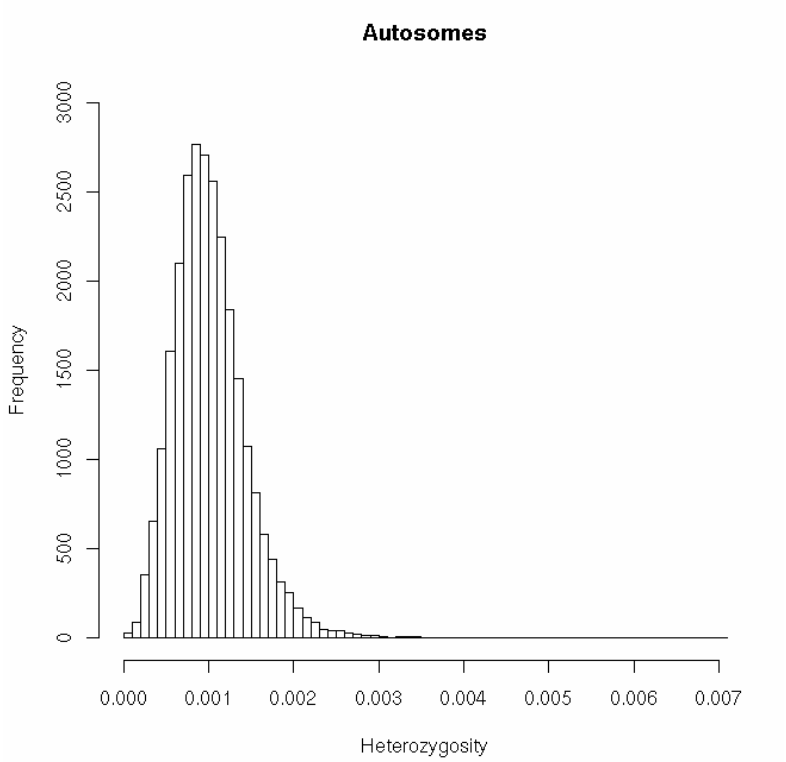


Figure 10. Histogram of single nucleotide heterozygosity in discrete 100 kbp intervals based on pooling all eight individuals.

Based on this distribution, we established a cutoff at the 99th percentile for regions showing “excess” heterozygosity.

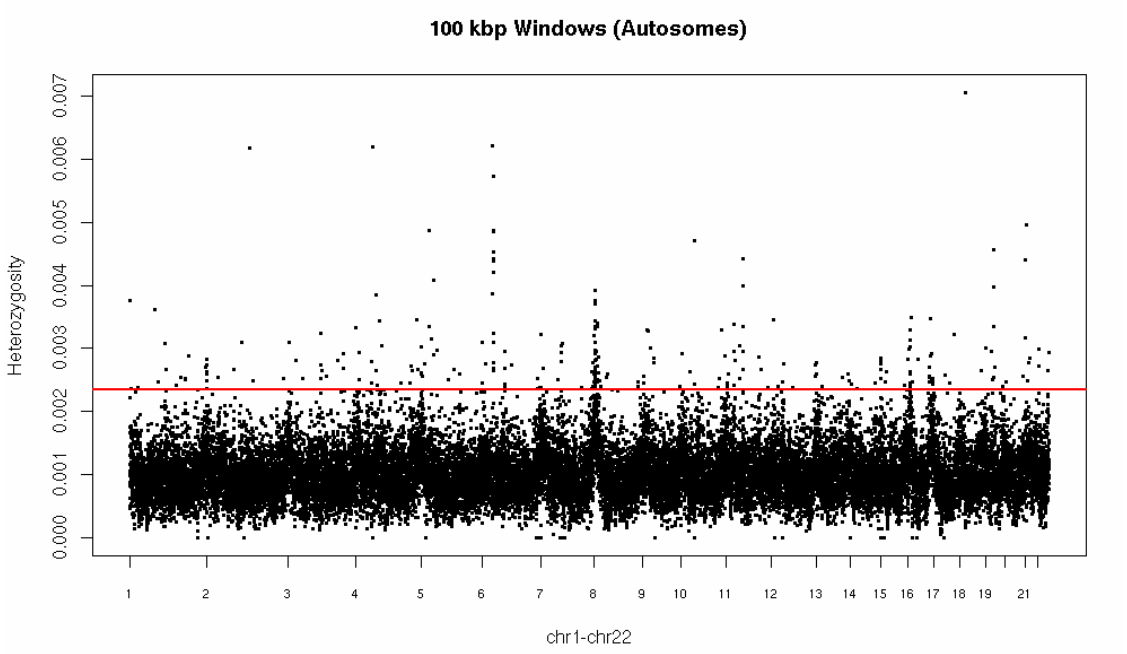


Figure 11. Heterzygosity as function of chromosomal location. Heterozygosity was computed based on the pooled reads from eight libriaries in discrete 100 kbp intervals for each chromosome. Red line denotes the 99th percentile.

The data showed clustering of regions of excess heterozygosity. There are 25 regions where there are 2 or more consecutive windows with heterozygsoity beyond the 99th percentile. These can be further collapsed into 14 locations in the human genome based on their proximity. These range in size from 200 kbp to a region ~10 Mbp in size located at chromosomal 8p23.

Table 9. Regions of Excess Single-nucleotide Heterozygosity

Chrm	Begin	End	Length
chr11	48,500,001	48,900,001	400,001
chr13	18,700,001	18,900,001	200,001
chr14	105,800,001	106,000,001	200,001
chr14	106,100,001	106,300,001	200,001
chr16	6,800,001	7,300,001	500,001
chr16	8,500,001	8,700,001	200,001
chr16	12,500,001	12,700,001	200,001
chr16	76,900,001	77,100,001	200,001
chr19	23,600,001	23,900,001	300,001
chr21	9,700,001	9,900,001	200,001
chr4	9,800,001	10,000,001	200,001
chr4	44,700,001	44,900,001	200,001
chr5	34,400,001	34,600,001	200,001
chr6	29,800,001	30,100,001	300,001
chr6	31,100,001	31,600,001	500,001
chr6	32,400,001	32,900,001	500,001
chr6	67,800,001	68,000,001	200,001
chr7	61,200,001	61,400,001	200,001
chr8	1,100,001	1,300,001	200,001
chr8	3,100,001	3,300,001	200,001
chr8	3,400,001	4,200,001	800,001
chr8	5,000,001	5,300,001	300,001
chr8	5,500,001	5,700,001	200,001
chr8	6,000,001	6,300,001	300,001
chr8	13,600,001	13,900,001	300,001

12. SNP Haplotype Analysis.

In order to provide preliminary data on whether the flanking SNP haplotypes are consistent with a single or recurrent rearrangement event, we performed a detailed analysis of SNP content of the structurally variant regions by focusing on those sites completely sequenced in fosmid clones (n=264). We unambiguously established the haplotype represented in the sequenced clone by mapping HapMap SNPs onto the finished clone sequence using BLAT[21] (we specifically examined the Phase 2 "phased" consensus SNP set [release 21]). (Note: 93% of the haplotypes were consistent between what was observed from the fosmid insert sequence and the inferred haplotype from Phase II data. This difference reflects a combination of potential sequencing and genotyping/phasing errors.) The resulting SNP haplotype was then compared with the HapMap SNP haplotypes determined for each additional sample predicted to carry the same structural variant. SNP allele frequency and ancestral state of each of the structural variants are important considerations in this analysis. Therefore, we limited our analysis to the subset of sites where the structural variant a) was seen at least twice but in less than five individual genomes and b) HapMap SNPs had a minor allele frequency > 20%. These criteria resulted in 72 informative sites. Next, we calculated the number of times that another sample carrying the structural variant also carried the same matching SNP haplotype.

Table 10: Structural Variants Mapping to the Same SNP Haplotype.

Event Type	Informative Clones	Each Additionally Predicted Sample Has A Matching Haplotype	Percent of Total
Deletion	39	25	64.1%
Insertion	28	15	53.6%
Inversion	5	3	60.0%
Total	72	43	59.7%

For 43 of these sites (59%), all of the additional samples predicted to carry the structural variant had the same haplotype as represented in the sequenced clone. In several of the remaining cases, the haplotype differences were limited to a few SNPs, indicating that the variant is present on closely related haplotypes. Therefore, we focused on sites where at least 3 SNPs differed among the inferred variant haplotypes. We identified 17 sites (23.6% of the total informative sites analyzed) that mapped to distinct haplotypes based on this definition. These are excellent candidates for recurrent duplication or deletion events, but this will not be conclusively resolved until such sites are sequenced from these individual libraries (similar to the SIRPB1). Data regarding these 17 sites are provided below.

Table 11: Variants Predicted on Differing Haplotypes.

Clone ID	Accession	Event Type	Chrm	Begin	End	Haplotypes
ABC10_45501700_B20	AC203593	Insertion	chr15	81322125	81356434	CAGCATG TAACGTG
ABC9_45366400_D20	AC206603	Deletion	chr6	19835459	19879746	TACTCG TATATA
G248P82007E10	AC192820	Deletion	chr22	37660970	37731665	ATAAAAATCGAA CCGGAAATCGAA
ABC9_43852100_D22	AC204974	Insertion	chr3	113338183	113368961	GCTC ATCT GCCC
ABC10_44477300_F21	AC203665	Insertion	chr16	26082704	26116844	TGTAGA TGTAGG GATACA
G248P80878F11	AC193146	Insertion	chr12	116253557	116284876	ATCC ATCT GCTC GTCC
ABC12_46795000_D12	AC206894	Deletion	chr4	108465511	108509774	GTGGG ACAAA
ABC8_684522_K19	AC213263	Deletion	chr4	173336129	173385132	TACAGAGTGA CCTAGGGTGG CCTAGAGTGA
G248P86370C8	AC153476	Deletion	chr19	40529245	40578845	ACGGGTCAGA ACGATATAGG
ABC10_45505500_C8	AC203630	Insertion	chr9	24959095	24993607	GTCGCACCT GCTCTACAG
ABC10_44088100_H17	AC203624	Insertion	chr18	49657533	49691438	CGACGGTGGTTACATGGTATATT CGACGATCCCTACACGGGATATT
ABC10_45521700_B16	AC203650	Insertion	chr12	125297297	125332491	GAAATAAATT GAAATAAACC GAAGCGTGTT
ABC9_43875600_C20	AC206438	Deletion	chr8	144736451	144787620	CCGGTCCTC CCGGTCTCC TAATCTCTC
ABC10_43667200_A22	AC203632	Deletion	chr18	50194581	50242381	AGTATATACTG AGTACGCGTCA GGTCTATACTG GAAATATACTG
G248P800782F5	AC158329	Deletion	chr1	34758633	34807266	CGGGCGCGGCTTGGTACAGATTGT CGGGCGCGGCTTGACTTGACCGAC
ABC10_44501400_N6	AC203655	Deletion	chr15	69460671	69508151	GATGTGTTGGGTAGGCG CACACCCCCTGTGAGCG CGCGCCCCGTGTGAGCG
ABC10_45534600_M1	AC203657	Insertion	chr18	46196360	46231673	CATAAACTACAGA CCTAGGTAGCAGA AACGAACTACAGA AACGAACTACGAC

The above table lists the name, sequence accession, event type, and genomic position for the 17 sites that mapped to distinct haplotypes. The last column of the table lists the inferred haplotypes for each of the individuals predicted to harbor the same variant. In each case, the top haplotype represents that found in the sequenced clone.

In summary, a preliminary analysis of the SNP content of sequenced sites suggests that approximately 24% of the variants predicted in multiple individuals may be found on different haplotype backgrounds. This could be the result of more ancient structural variation mutation events which are now found on different haplotypes or of recurrent mutation events. Sequence analyses of the corresponding sites from each of the individuals will be required to distinguish between these possibilities.

13. Table S3/S4 Description

Table S3 Description

Table S3 reports insertion/deletion sites identified by ESP analysis using length thresholds provided in Table 2 of this document. Sites which traverse gaps within the reference assembly (hg17) were removed. Sites with a span greater than 1 MB were also excluded (section 4). CGH breakpoint coordinates were arbitrarily assigned a position in the middle of the probe sequence. The columns of the table are as follows:

Library: Individual in which the site was identified
Chrm,start,end: Genomic interval of variation defined by overlapping discordant fosmids
Span: Size of this genomic interval
MinSpread: The minimum separation between placed clone ends supporting this site. An unusually small minspread suggests that the same clone may be been sequenced twice and that the site may not be supported by multiple independent clones.
Type: Type of event. Insertions are indicated by 'S' (clones appear to be small) and deletions by 'B' (clones appear to be large or big)
Avesize: Average size of discordant clones supporting this site
Overlap With Validated Locus: Whether this site overlaps with a validated locus
Site_id: Unique identifier for this site
Clone_IDs: IDs of discordant clones supporting the site.
Predicted_Variant_Size: Predicted size of the event. Identified by comparing the average size of discordant clones supporting the site with the average size of all clones from that library.
Refseq_Overlap: Intersection of the spanned interval with Refseq annotations.
MCD_size: Size of a clone from this site based on MCD fingerprint compared with size of the clone based on ESP placement (when multiple clones were tested, values are separted by ':')
Sequenced_Clone: Name of clone from this site selected for sequencing (multiple names separted by ':')
MCD_result: Result of fingerprint annotation, multiple clones are separated by ':'
G248Status: Overlap of spanned interval with predicted sites in G248. "Gain" indicates predicted G248 insertion and "del" indicates predicted G248 deletion.
Nim_probecount: Number of probes in this interval on the NimbleGen array
Nim_Deletion: Binary value indicating whether or not a deletion was called for this sample on the NimbleGen array
Nim_Gain: Binary value indicating whether or not a gain was called for this sample on the NimbleGen array
Nim_Hits: Total number of aberrations (gains+losses) called for this sample on the Nimblegen array
Nim_vs_Fos_Ratio: Ratio of size of the event based on CGH analysis compared to Predicted_Variant_Size
Nim_Types: Description of CGH aberration coordinates relative to the spanned interval
Agi_probecount: Number of probes in this interval on the Agilent array

Agi_Deletion: Binary value indicating whether or not a deletion was called for this sample on the Agilent array
Agi_Gain: Binary value indicating whether or not a gain was called for this sample on the Agilent array
Agi_Hits: Total number of aberrations (gains+losses) called for this sample on the Agilent array
Agi_vs_Fos_Ratio: Ratio of the size of the event based on CGH analysis compared to Predicted_Variant_Size
Agi_Types: Description of CGH aberration coordinates relative to the spanned interval
Supp_New_Seq: Name of new sequence insertion locus (if any) which intersects with the site
NimRefinedStart, NimRefinedEnd: Variant interval based on NimbleGen array
AgiRefinedStart, AgiRefinedEnd: Variant interval based on Agilent array
SequenceValidation: Summary annotation of variants observed in sequenced clone. D=deletion, I=insertion, V=inversion, U=Amiguous, N=no variant. Annotations of multiple clones are separated by ':'.
SeqBkPtStart SeqBkPtEnd: Breakpoints based on analysis of clone sequence.

Table S4 Description
We merged together the sites in Table S3 to create a non-redundant listing of variant region predictions.

Chrm, start, end: Coordinates of merged region
Locus_span: Size of merged region
Type: Type of event
Validation Status: Whether this locus is considered validated
MinSpread: The maximum minspread value from the individual sites (Table S3) merged into this locus.
Site_ids: Identifiers of sample level sites (Table S3) merged into this locus
Num_sites: Number of sites from Table S3 merged into this region
Num_lib: Number of different libraries contributing predicted sites to this region
G248_present—ABC14_present: Binary valued indicating whether named library contributed a site to this locus.
G248_fraction_spanned—ABC14_fraction_spanned: Fraction of the merged interval region spanned by a site predicted in each library
G248_predicted_size-ABC14_predicted_size: Predicted size of the events predicted in each library.
MCD_size: Difference between clone sizes based on MCD fingerprinting and ESP placement. Values from multiple clones are separated by ':'.
MCD_conclusion: Result of MCD analysis. "?" indicates not tested or no results
Sequenced_Clones: Names of sequenced clones from this region
Sequence_Status: Annotation summary
G248Status: Summary of G248 predictions over this interval
Min_Nim_probes: Minimum number of NimbleGen probes for all of the sample-level predictions merged into this region

ABC7_Nim_vs_Fos_Ratio—ABC14 Nim vs Fos Ratio: Ratio of predicted sizes based on CGH results compared to predicted sizes based on the sizes of discordant clones.
Called_Nimb_Loss,Called_Nimb_Gain: Binary value indicating whether or not a loss or gain call was made in any of the samples contributing to this region.
min_Agi_probes: Minimum number of Agilent probes for all of the sample-level predictions merged into this region
ABC7_Agi_vs_Fos_Ratio-ABC14 Agi vs Fos Ratio: Ratio of predicted sizes based on CGH results compared to predicted sizes based one sizes of discordant clones
Called_Agi_Loss,Called_Agi_Gain: Binary value indicating whether or not a loss or gain call was made in any of the samples contributing to this region.
Supp_NovelInsert: Indication of whether interval overlaps with a novel sequence insertion region
NimBrkpnts: Breakpoints from NimbleGen arrayCGH, breakpoints from multiple individuals are separted by ':'
NimBrkpntsAvg: Average value of breakpoints for merged region from NimbleGen array
AgiBrkpnts: Breakpoints from NimbleGen arrayCGH, breakpoints from multiple individuals are separted by ':'
AgiBrkpntsAvg: Average value of breakpoints for merged region from NimbleGen array
SeqBrkpnts: Breakpoints based on sequenced clones

14. Supplemental Figure Legends

Figure S1: Fosmid Genome Coverage. The fosmid library coverage is shown for each library as the fraction of nucleotides that are spanned by end-sequence pairs that map to a best location in the genome. The fraction of the genome with no best-placement spanning clones (n=0), 1 or more (>=1), two or more (>=2), and four or more (>=4) is indicated. Autosomes and the X chromosome are considered separately. ABC8 represented the sole male sample.

Figure S2: Fosmid Clone Tiling Paths. The browser snapshots (http://hgsv.washington.edu) show the clone ID and mapping location of concordant (black) and discordant (red) fosmid end-sequences mapped against the human genome (hg17) for each library. a) Clone tiling path across the leptin (*LEP*) locus is shown for two individual libraries, G248 and ABC7; all clones are concordant by length and orientation; b) A putative heterozygous deletion is shown for an individual library (ABC7) where both concordant and discordant clones are identified; note discordant clones from third library (ABC9) predict an insertion allele over the *CYP2D6* locus; c) The *GSTM1*shows an example of locus complexity. Analysis of the ABC9 library predicts a homozygous deletion (absence of concordant clones) and another genomic library (ABC10) predicts the presence of two non-overlapping insertions in the heterozygous state. Any region of interest in the human genome can be accessed using a UCSC browser interface (http://hgsv.washington.edu) and the corresponding fosmid clones, end-sequences, and alignments retrieved for further characterization.

Figure S3. End-sequence mapping of fosmids against the human genome. All discordant fosmids mapping to the human genome are displayed individually for each library using the following color scheme: ABC7=green, ABC8=forestgreen, ABC10=blue, ABC13=cyan, G248=black, ABC9=purple, ABC11=red, ABC12=orange, and ABC14=hotpink. The end-sequence placements are mapped in the context of gaps within the assembly (purple) and segmental duplications (grey bars). We required two or more discordant fosmids within an individual library in order to select a region for further characterization (predicted sites are represented by black bars, yellow bars indicate that a site was validated). Chromosome-wide maps are shown for putative a) deletions b) insertions and c) inversions. For the insertion map, we also show the location of one-end anchored clones as blue and gold vertical lines. The coordinates of all discordant sites as well as an interactive version of this map can be obtained via (http://hgsv.washington.edu).

Figure S4. Array Comparative Genomic Hybridization Validation. Array comparative genomic hybridization results of the same region shown in Figure 4. While a common deletion can be confirmed, breakpoint heterogeneity can not be discerned nor can genotype status be accurately predicted (see ABC9) unless the sequence structure of the reference genome (G248) is known.

Figure S5. Genotyping Structural Variation. Frequency spectrum for 130 deletion events identified through ESP, validated and breakpoint-refined using sequence or arrayCGH, and present in at least 1% of unrelated chromosomes. Each deletion event is indicated with tick marks along the X-axis, with the global allele frequency for that event indicated by the total height of the corresponding vertical bar (a). The allele proportions within the three distinct populations are indicated in blue (YRI), red (CEU), or green (CHB+JPT), respectively. b-d) Frequency of deletion event within each population plotted separately.

Figure S6. Inversion Identification and Validation. A schematic of clones used to identify 5 large inversions is shown (coloring as in Figure S3). Each of the depicted clones has an "inverted" orientation. Despite the presence of duplicated sequences (grey boxes), the clones have a single, best placement. For each locus, FISH validation is shown with the inverted allele indicated by the arrow. The 8q24 inversion was only observed in 3/20 nuclei and may be mosaic.

Figure S7 Size of Validated Sites. The size distribution was calculated for 1695 validated events based on the average size discordancy of supporting clones (for deletions or insertions) or the span of discordant clones (for inversions).

Figure S8. **Cross-Platform Comparisons.** Size comparisons for CNVs inferred by different platforms and methodologies on eight common HapMap samples. For each plot, the estimated size for a CNV annotated by fosmid ESP mapping is shown along the X-axis. Sizes for variants that overlap by any number of nucleotides within the same sample as annotated by Redon et al. (a) which used BAC CGH/Affymetrix 5.0 arrays, McCarroll et al. (b) which used Affymetrix 6.0 arrays, or Cooper et al. (c) which used IlluminaHuman1M BeadChips. The scale is in nucleotides on all three plots. Note that clustering of sizes can be seen, this is a result of the same CNV inferred within multiple samples.

Figure S9. Map of Novel Insertion Loci. The approximate locations of 525 putative new insertion loci based on positions of one-end anchored clones is shown for each human chromosome. Three categories of OEA clusters are distinguished: a) flanking a gap in the assembly (black) b) flanking a discordant fosmid predicting an insertion allele (spanned=blue) or c) neither (unspanned=red). The latter may correspond to larger insertion sequences.

Figure S10. Sequenced Structural Variation and Gene Structure. A graphical representation for sequenced sites (n=266) of structural variation (miropeats view) is provided. Each alignment compares the human reference genome (top) with the sequenced structure of the fosmid clone (threshold s=400). RefGene exons are shown as red bars above the human genome reference. Duplication and repeat content/orientation are shown using colored arrows (LINE=green, SINE=purple, transposon=orange, grey/blue=segmental duplications).

References

1. Eichler, E.E. et al. Completing the map of human genetic variation. *Nature* **447**, 161-5 (2007).
2. IHMC. A haplotype map of the human genome. *Nature* **437**, 1299-320 (2005).
3. ENOCDE Project Consortium. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447**, 799-816 (2007).
4. Donahue, W. & Eblingm, H. Fosmid libraries for genomic structural variation detection. *Current Protocols in Human Genetics* **5.20.1-5.20.18**(2007).
5. Tuzun, E. et al. Fine-scale structural variation of the human genome. *Nat Genet* **37**, 727-32 (2005).
6. Gillett, W. et al. Assembly of high-resolution restriction maps based on multiple complete digests of a redundant set of overlapping clones. *Genomics* **33**, 389-408 (1996).
7. Wong, G.K., Yu, J., Thayer, E.C. & Olson, M.V. Multiple-complete-digest restriction fragment mapping: generating sequence-ready maps for large-scale DNA sequencing. *Proc Natl Acad Sci U S A* **94**, 5225-30 (1997).
8. Bovee, D. et al. Closing gaps in the human genome with fosmid resources generated from multiple individuals. *Nat Genet* **40**, 96-101 (2008).
9. Smit, A., Hubley, R. & Green, P. RepeatMasker Open-3.0. (1996-2004).
10. Newman, T.L. et al. High-throughput genotyping of intermediate-size structural variation. *Hum Mol Genet* **15**, 1159-67 (2006).
11. Redon, R. et al. Global variation in copy number in the human genome. *Nature* **444**, 444-54 (2006).
12. Weiss, L.A. et al. Association between Microdeletion and Microduplication at 16p11.2 and Autism. *N Engl J Med* (2008).
13. Zhang, Z., Schwartz, S., Wagner, L. & Miller, W. A greedy algorithm for aligning DNA sequences. *J Comput Biol* **7**, 203-14 (2000).
14. Sutton, G.G., White, O., Adams, M.D. & Kerlavage, A. TIGR Assembler: a new tool for assembling large shotgun sequencing projects. *Genome Sci. Technol.* **1,** 9–19 (1995)
15. Venter, J.C. et al. The sequence of the human genome. *Science* **291**, 1304-51. (2001).
16. Gordon, D., Abajian, C. & Green, P. Consed: a graphical tool for sequence finishing. *Genome Res* **8**, 195-202 (1998).
17. Parsons, J. Miropeats: graphical DNA sequence comparisons. *Comput Appl Biosci* **11**, 615-619 (1995).
18. Thompson, J.D., Higgins, D.G. & Gibson, T.J. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* **22**, 4673-80 (1994).
19. Ning, Z., Cox, A.J. & Mullikin, J.C. SSAHA: a fast search method for large DNA databases. *Genome Res* **11**, 1725-9 (2001).

20. Nickerson, D., Tobe, V. & Taylor, S. PolyPhred: automating the detection and genotyping of single nucleotide substitutions using fluorescence-based resequencing. *Nucleic Acids Res.* **25**, 2745-2751 (1997).
21. Kent, W.J. BLAT--the BLAST-like alignment tool. *Genome Res* **12**, 656-64 (2002).

Supplementary Information Guide for "Fine-Scale Mapping and Sequencing of Structural Variation from Eight Human Genomes", manuscript ID 2007-11-11699

Figure S1: Fosmid Genome Coverage.
The fosmid library coverage is shown for each library as the fraction of nucleotides that are spanned by end-sequence pairs that map to a best location in the genome. (PDF; 16 kb)

Figure S2: Fosmid Clone Tiling Paths.
The browser snapshots (http://hgsv.washington.edu) show the clone ID and mapping location of concordant (black) and discordant (red) fosmid end-sequences mapped against the human genome (hg17) for each library. (PDF, 194 kb)

Figure S3. End-sequence mapping of fosmids against the human genome.
All discordant fosmids mapping to the human genome are displayed individually for each library using the following color scheme: ABC7=green, ABC8=forestgreen, ABC10=blue, ABC13=cyan, G248=black, ABC9=purple, ABC11=red, ABC12=orange, and ABC14=hotpink. The end-sequence placements are mapped in the context of gaps within the assembly (purple) and segmental duplications (grey bars). (S3a: PDF, 4.4 MB; S3b: PDF, 6.2 MB; S3c: PDF, 3.5 MB)

Figure S4. Array Comparative Genomic Hybridization Validation.
Array comparative genomic hybridization results of the same region shown in Figure 4. (PDF; 387 kb)

Figure S5. Genotyping Structural Variation.
Frequency spectrum for 130 deletion events identified through ESP, validated and breakpoint-refined using sequence or arrayCGH, and present in at least 1% of unrelated chromosomes. (PDF; 230 kb)

Figure S6. Inversion Identification and Validation.
A schematic of clones used to identify 5 large inversions is shown (coloring as in Figure S3). Each of the depicted clones has an "inverted" orientation. (PDF; 5.1 MB)

Figure S7 Size of Validated Sites.
The size distribution was calculated for 1695 validated events based on the average size discordancy of supporting clones (for deletions or insertions) or the span of discordant clones (for inversions). (PDF; 13 kb)

Figure S8. **Cross-Platform Comparisons.**
Size comparisons for CNVs inferred by different platforms and methodologies on eight common HapMap samples. (PDF; 870 kb)

Figure S9. Map of Novel Insertion Loci.
The approximate locations of 525 putative new insertion loci based on positions of one-end anchored clones is shown for each human chromosome. (Powerpoint; 200 kb)

Figure S10. Sequenced Structural Variation and Gene Structure.
A graphical representation for sequenced sites (n=266) of structural variation (miropeats view) is provided. Each alignment compares the human reference genome (top) with the sequenced structure of the fosmid clone. (PDF; 3.5 MB)

Table S1. Concordant vs. discordant clone placement summary statistics.
(Excel; 23 kb)

Table S2. One-end anchored (OEA) clone statistics.
(Excel; 15kb)

Table S3. All ESP predicted sites of insertions and deletions with associated experimental validation (See Supplementary Material Section 12 for description of column headers)
(Excel; 5 MB)

Table S4. ESP predicted sites of insertion and deletion loci (non-redundant) across the fosmid libraries (See Supplementary Material Section 12 for description of column headers)
(Excel; 4 MB)

Table S5. Genotyping results for a subset of ESP deletion variants based on analysis of genotypes from the llumina Human1M BeadChip
(Excel; 40 kb)

Table S6. ESP predicted inversion breakpoints
(Excel; 300 kb)

Table S7. Merged inversion loci (non-redundant)
(Excel; 64 kb)

Table S8. Large insertions of novel sequence confirmed by optical mapping
(Excel; 16kb)

Table S9. Genbank accession IDs of sequenced clones
(Excel; 73 kb)

Table S10: Sequenced structural variants that affect exons of genes
(Excel; 26 kb)

Table S11. Summary statistics of fosmid end sequences
(Excel; 17 kb)

Table S12. Genotypes based on custom GoldenGate Assay and qPCR
(Excel; 80 kb)

Figure S1: Fosmid Genome Coverage. The fosmid library coverage is shown for each library as the fraction of nucleotides that are spanned by end-sequence pairs that map to a best location in the genome. The fraction of the genome with no best-placement spanning clones (n=0), 1 or more (>=1), two or more (>=2), and four or more (>=4) is indicated. Autosomes and the X chromosome are considered separately. ABC8 represented the sole male sample.

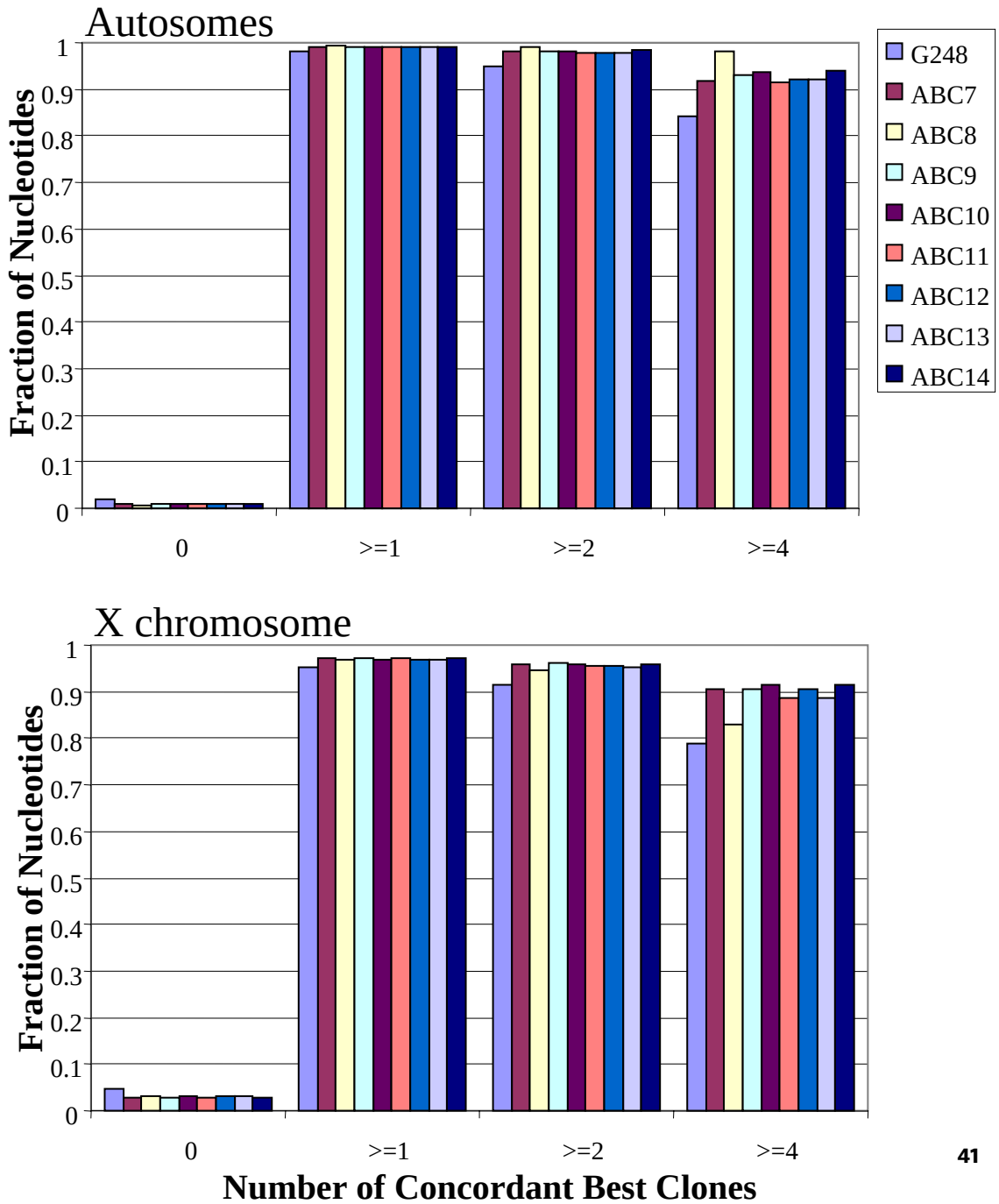


Figure S2: Fosmid Clone Tiling Paths. The browser snapshots (http://hgsv.washington.edu) show the clone ID and mapping location of concordant (black) and discordant (red) fosmid end-sequences mapped against the human genome (hg17) for each library. a) Clone tiling path across the leptin (*LEP*) locus is shown for two individual libraries, G248 and ABC7; all clones are concordant by length and orientation; b) A putative heterozygous deletion is shown for an individual library (ABC7) where both concordant and discordant clones are identified; note discordant clones from third library (ABC9) predict an insertion allele over the *CYP2D6* locus; c) The *GSTM1*shows an example of locus complexity. Analysis of the ABC9 library predicts a homozygous deletion (absence of concordant clones) and another genomic library (ABC10) predicts the presence of two non-overlapping insertions in the heterozygous state. Any region of interest in the human genome can be accessed using a UCSC browser interface (http://hgsv.washington.edu) and the corresponding fosmid clones, end-sequences, and alignments retrieved for further characterization.

Fig. S2 a) Clone Tiling path across LEP locus (1 clone/5 kb)

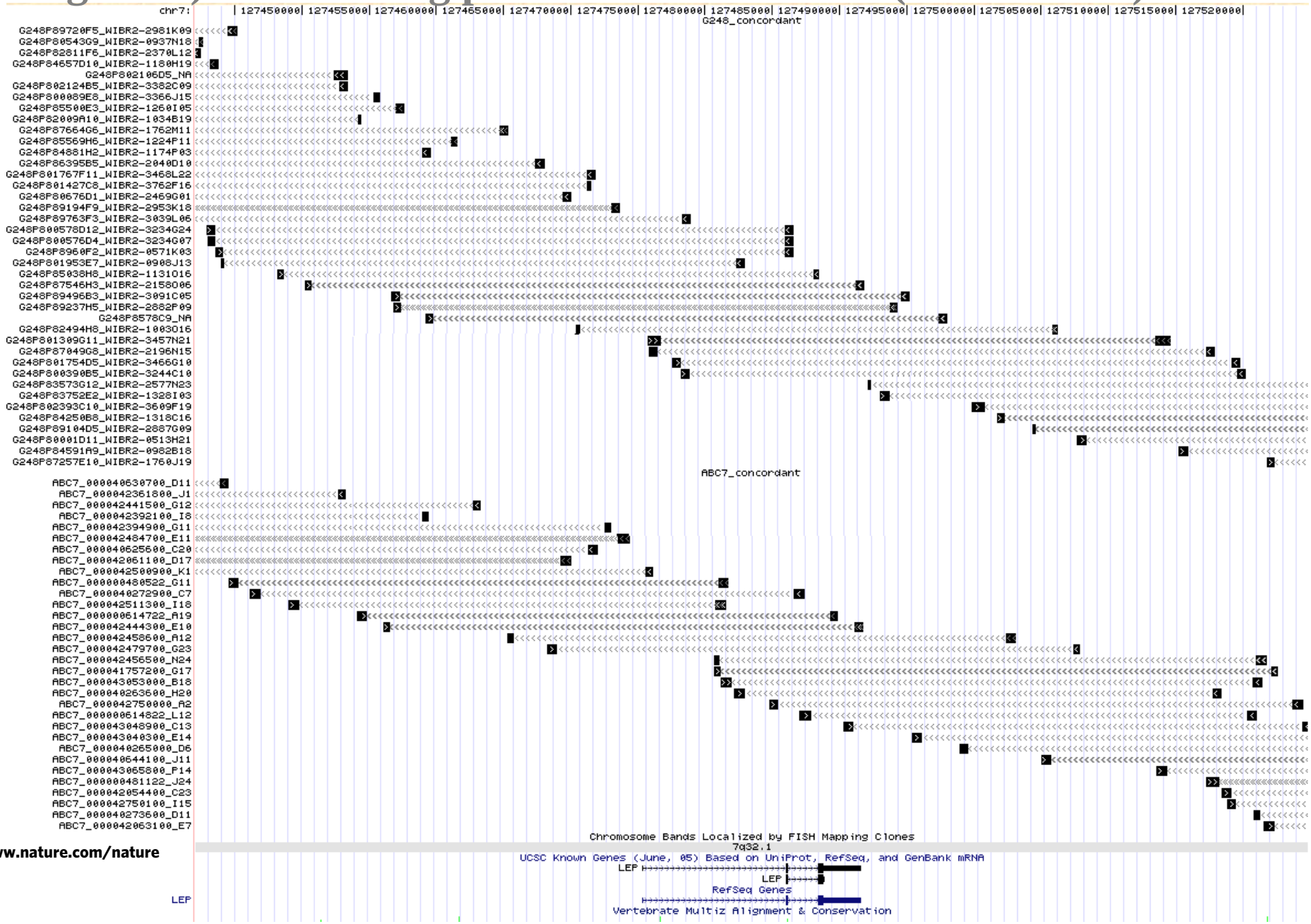


Fig. S2 b) CYP2D6 Deletion

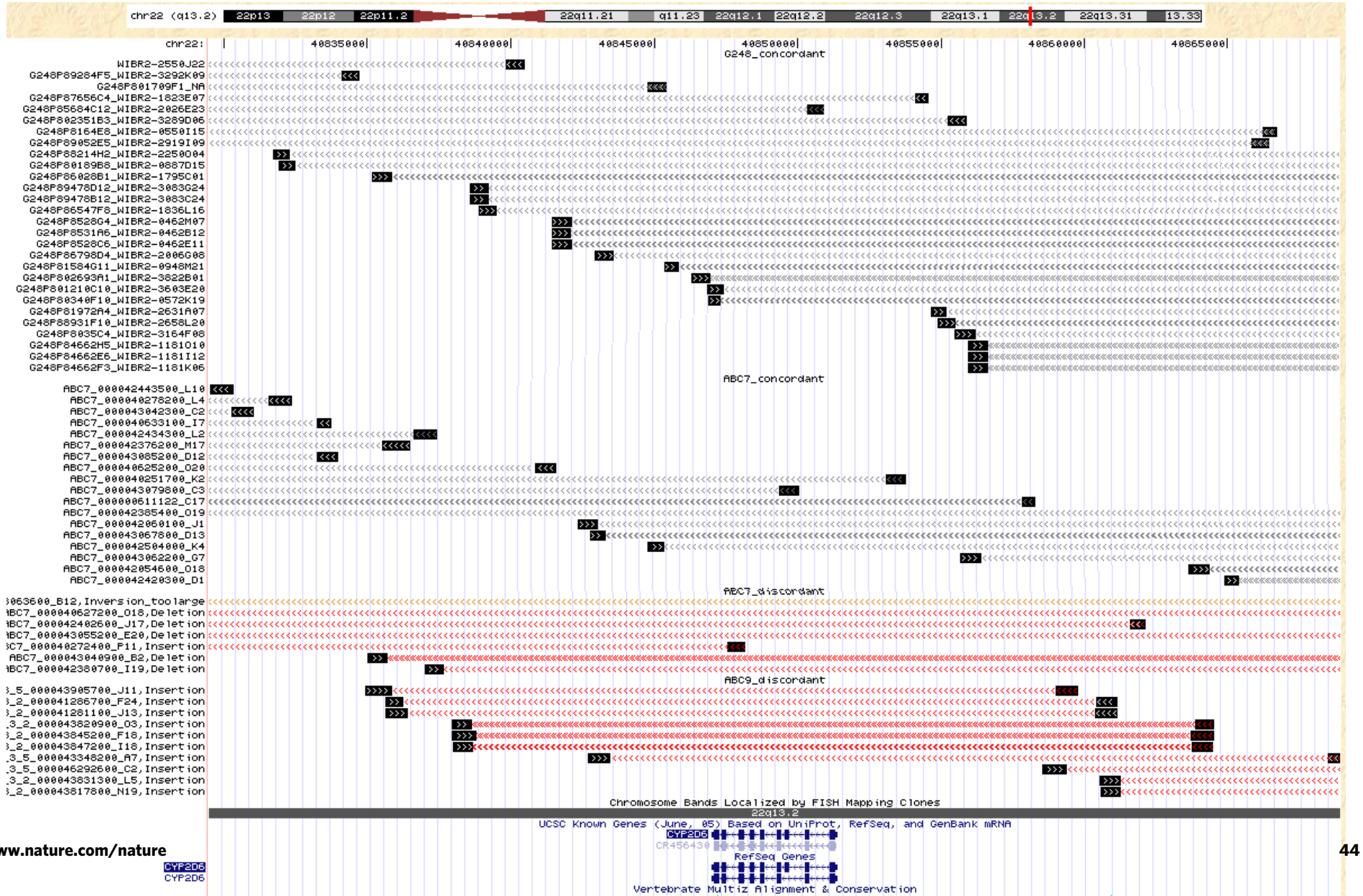
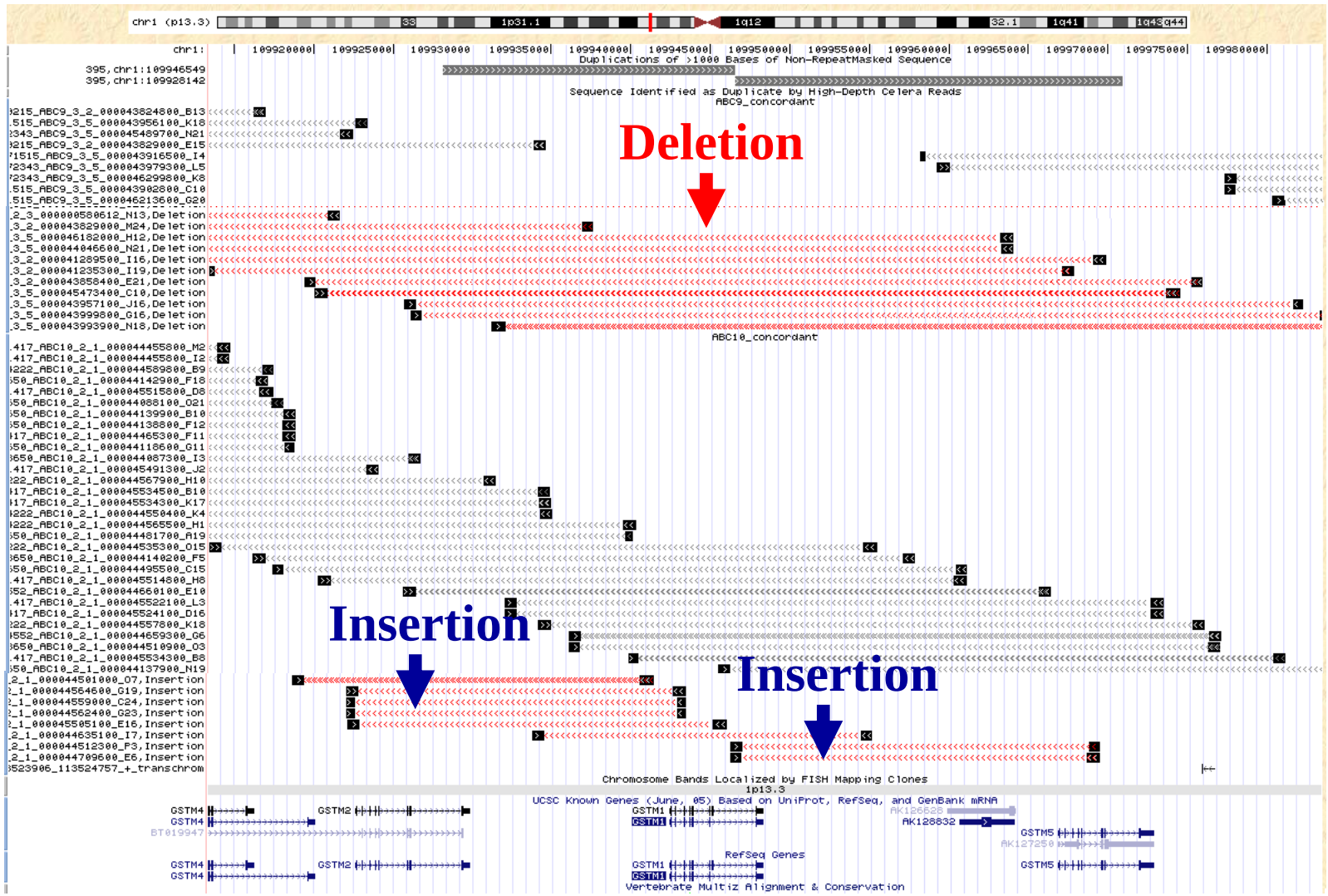


Fig. S2 c) GSTM1 Locus Complexity



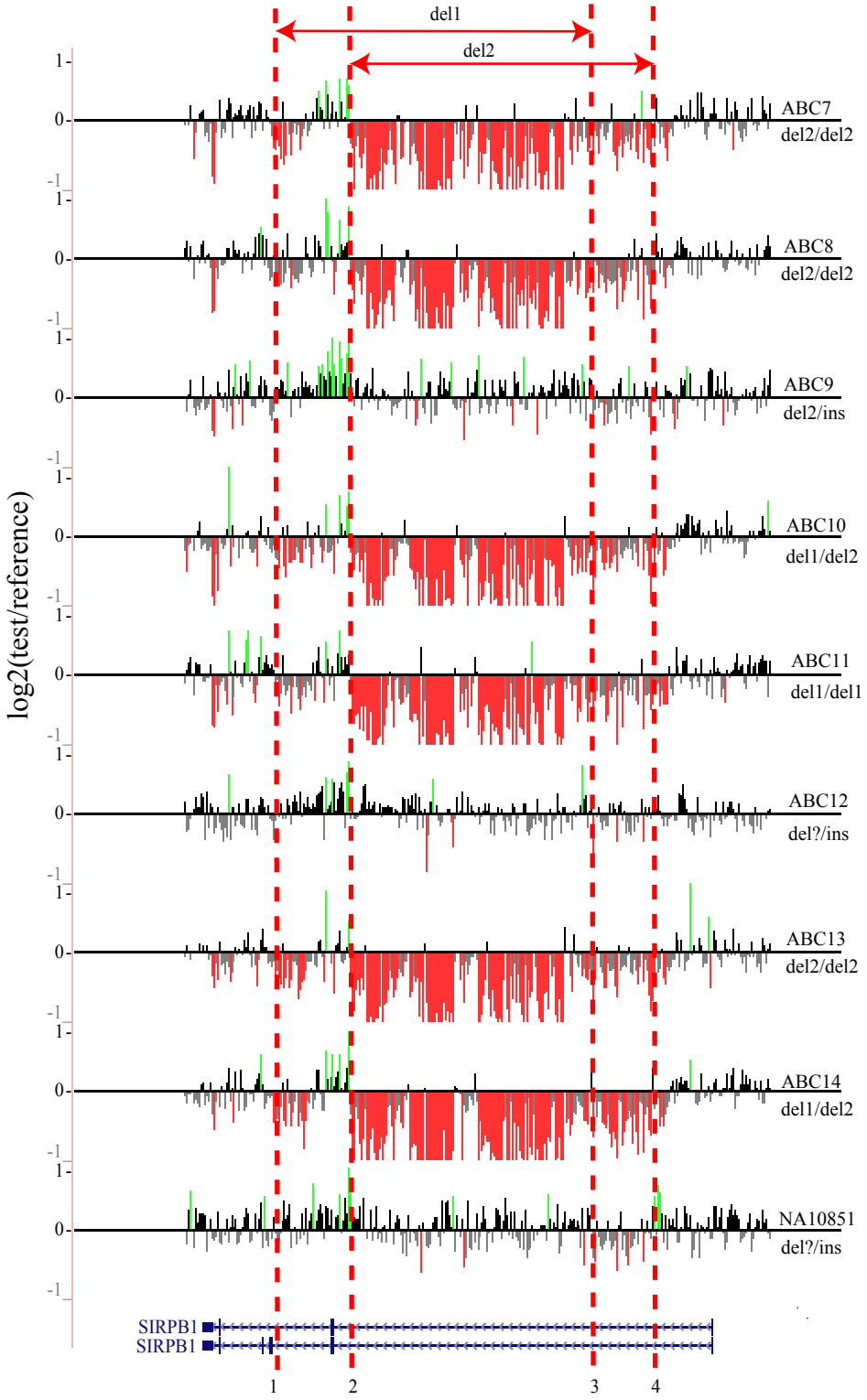
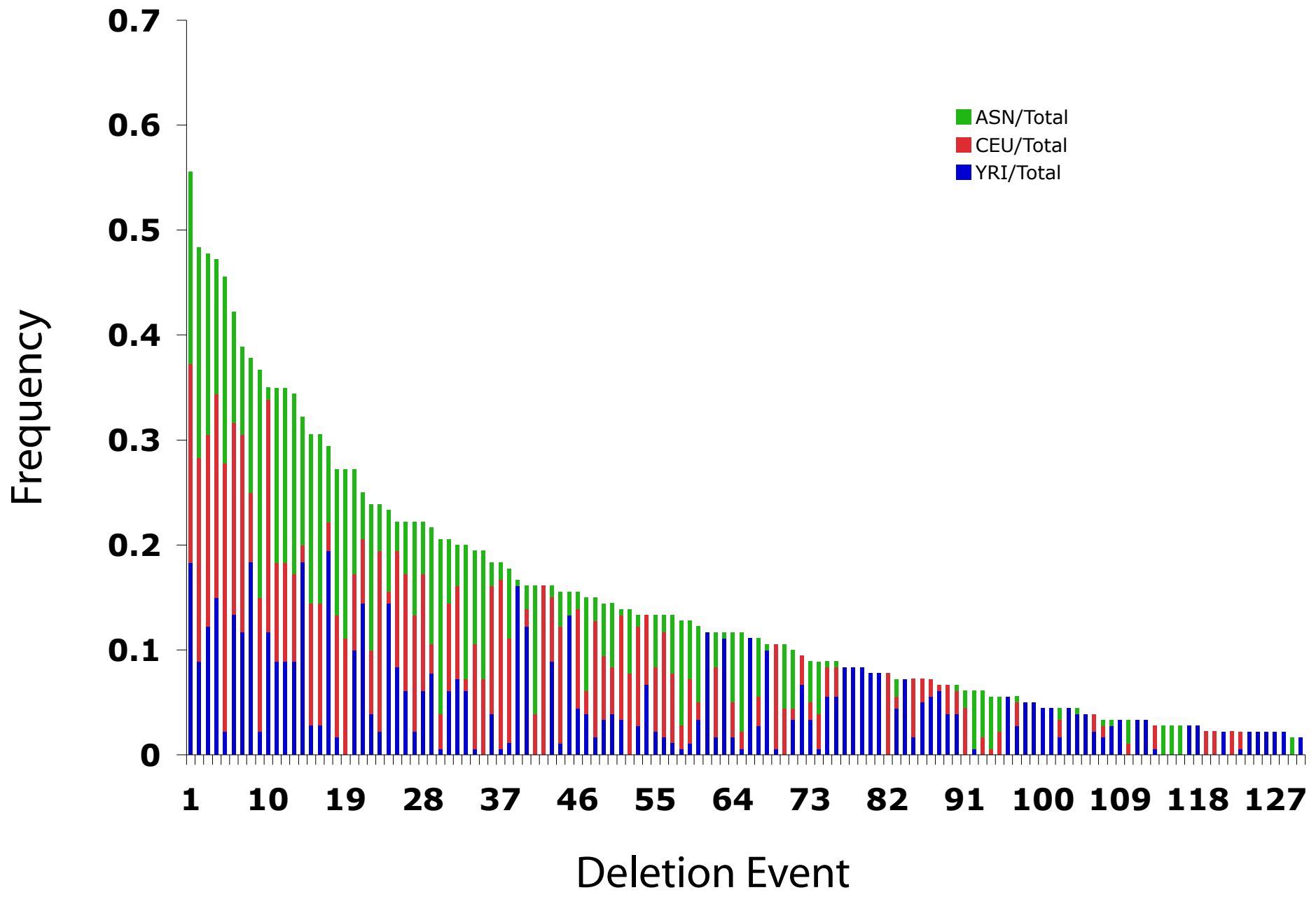


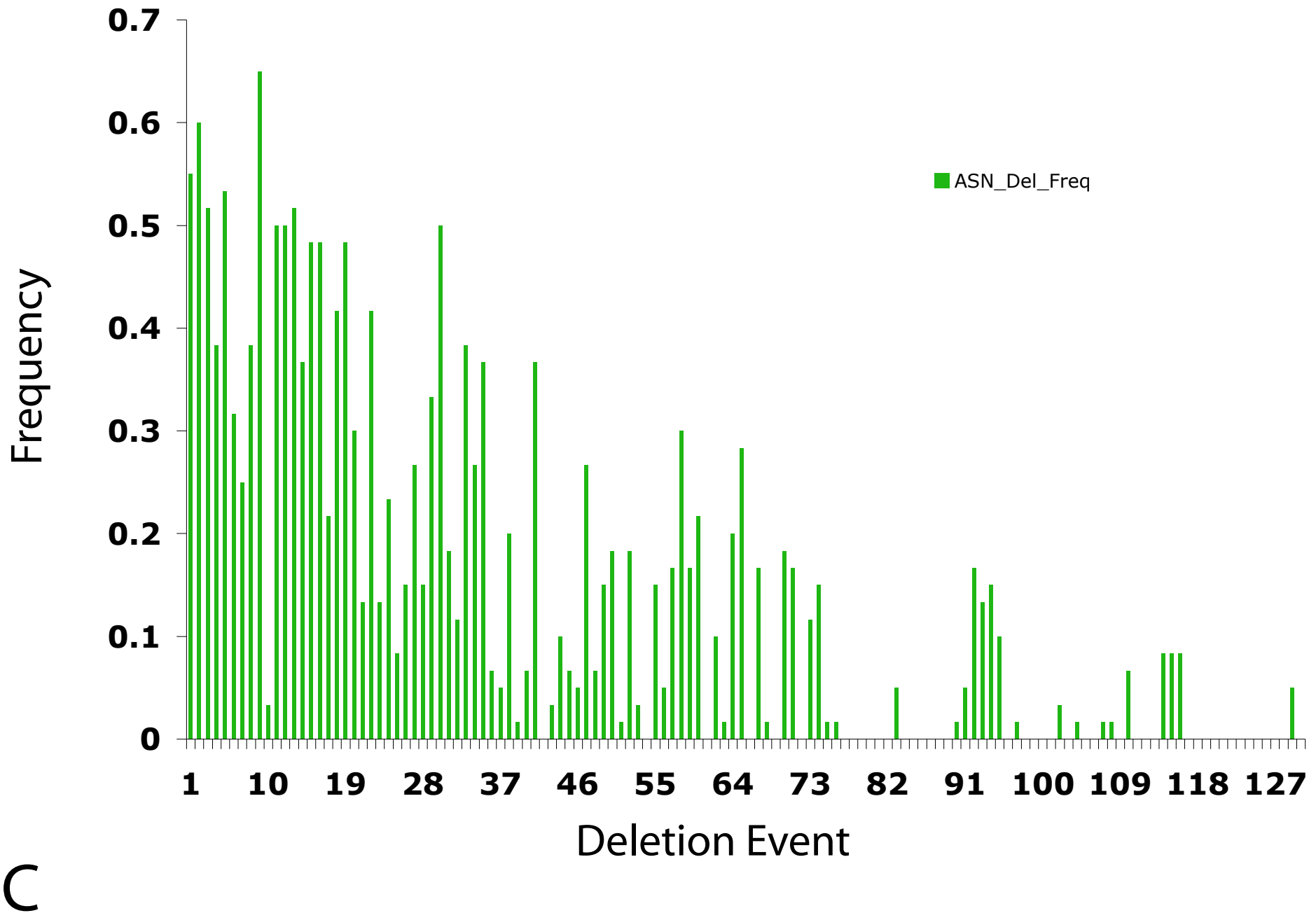
Figure S4. Array Comparative Genomic Hybridization Validation. Array comparative genomic hybridization results of the same region shown in Figure 4. While a common deletion can be confirmed, breakpoint heterogeneity can not be discerned nor can genotype status be accurately predicted (see ABC9) unless the sequence structure of the reference genome (G248) is known.

Figure S5. Genotyping Structural Variation. Frequency spectrum for 130 deletion events identified through ESP, validated and breakpoint-refined using sequence or arrayCGH, and present in at least 1% of unrelated chromosomes. Each deletion event is indicated with tick marks along the X-axis, with the global allele frequency for that event indicated by the total height of the corresponding vertical bar (a). The allele proportions within the three distinct populations are indicated in blue (YRI), red (CEU), or green (CHB+JPT), respectively. b-d) Frequency of deletion event within each population plotted separately.

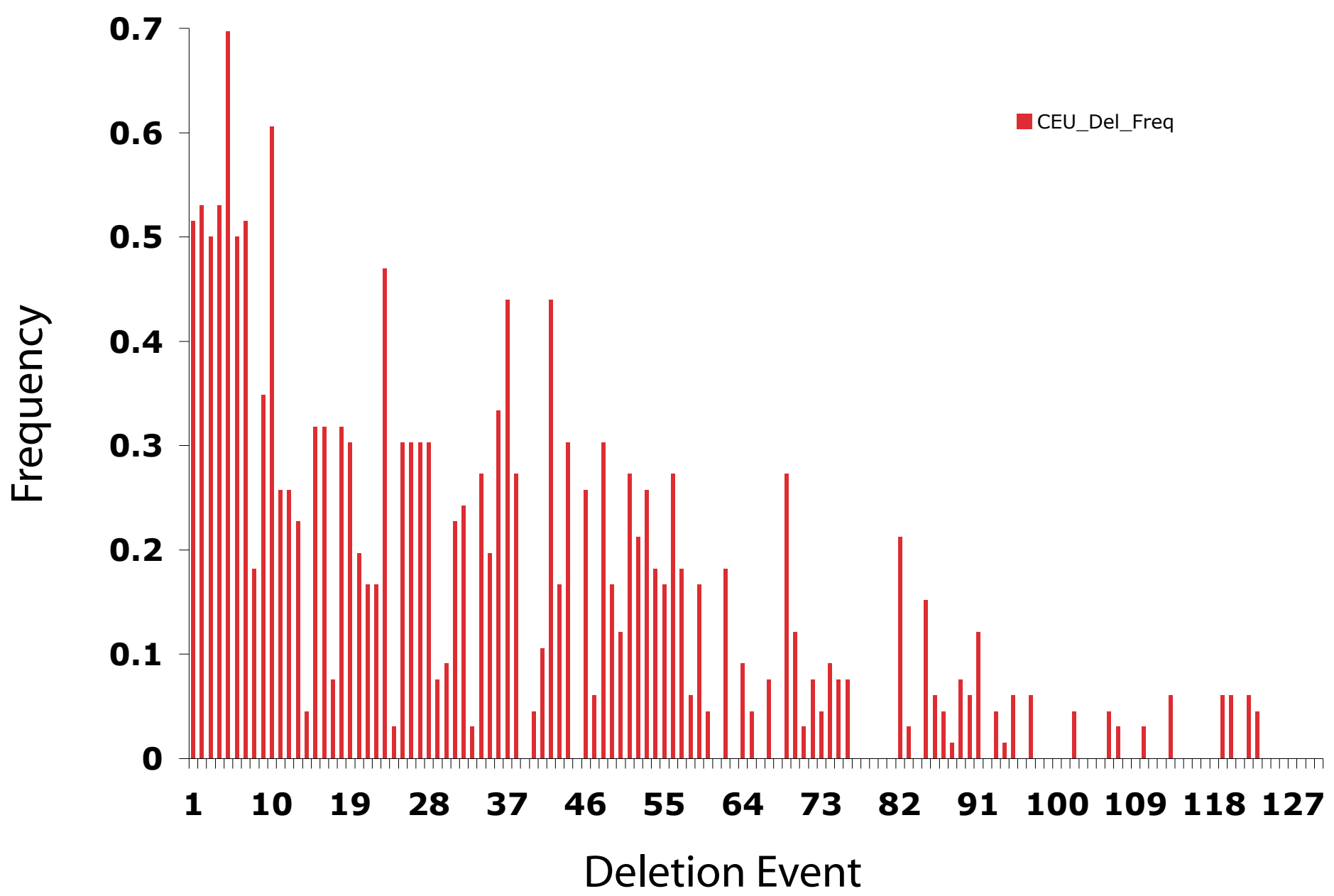
A


Frequency
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
ASN/Total
CEU/Total
YRI/Total
1 10 19 28 37 46 55 64 73 82 91 100 109 118 127
Deletion Event

B


Frequency
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
ASN_Del_Freq
1 10 19 28 37 46 55 64 73 82 91 100 109 118 127
Deletion Event

C


Frequency
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
CEU_Del_Freq
1 10 19 28 37 46 55 64 73 82 91 100 109 118 127
Deletion Event

D

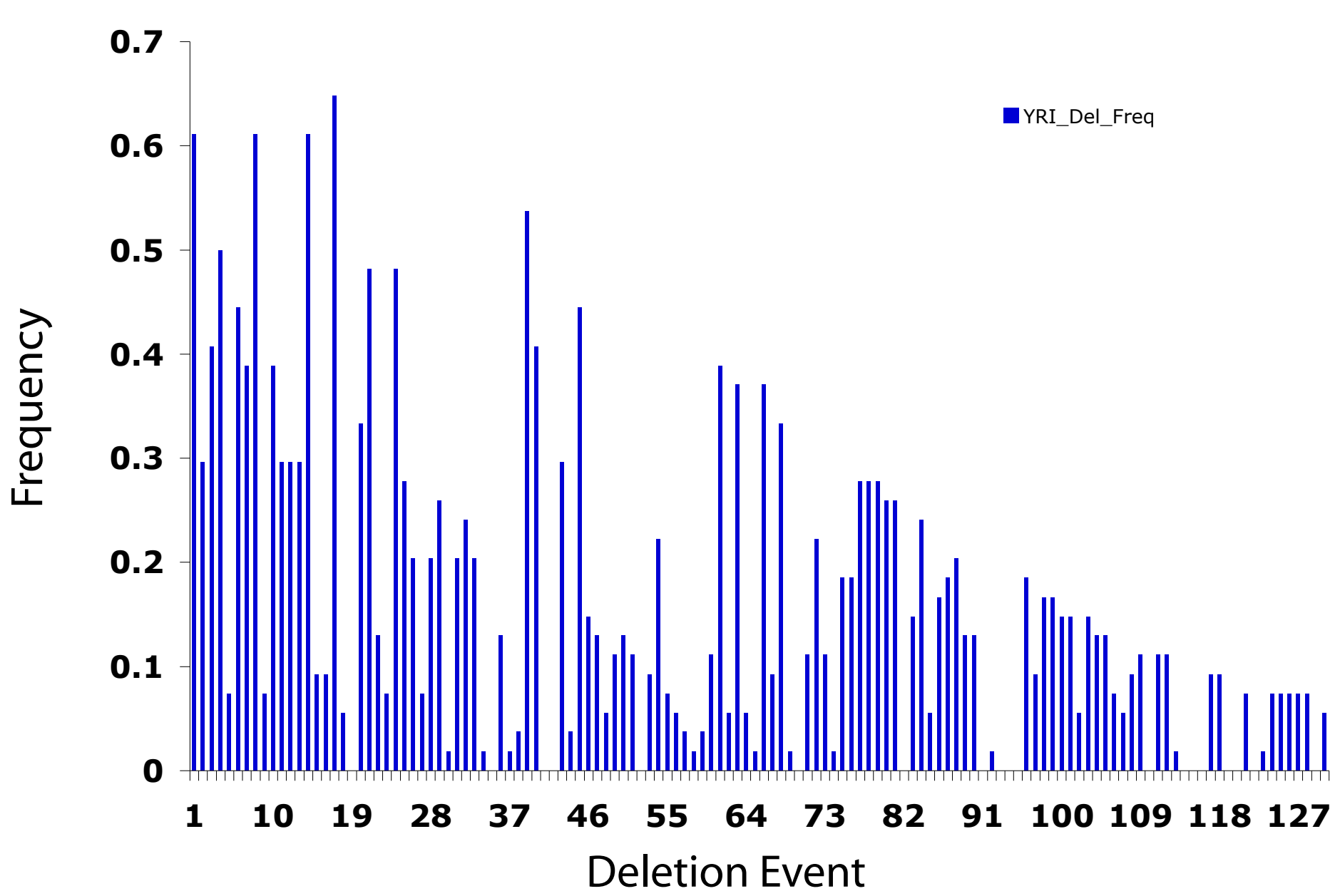
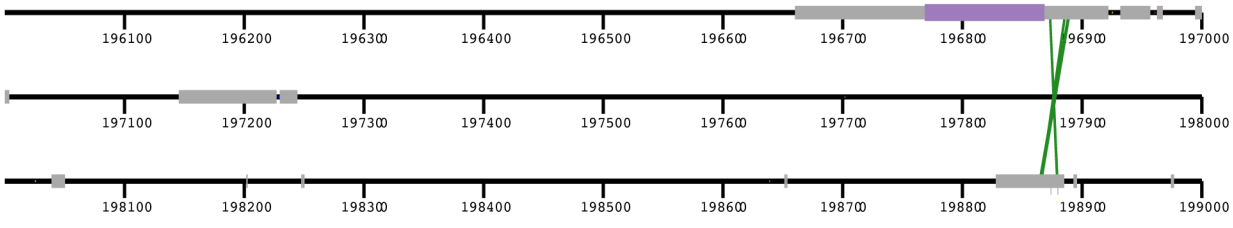
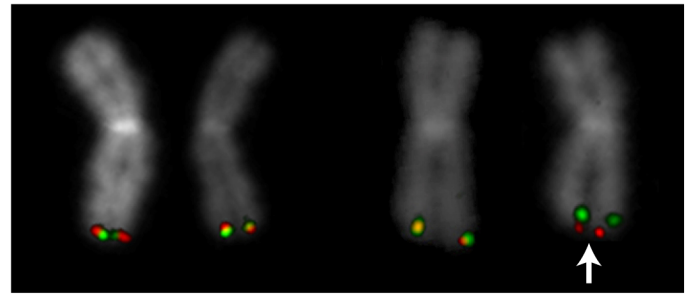

Frequency
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
YRI_Del_Freq
1 10 19 28 37 46 55 64 73 82 91 100 109 118 127
Deletion Event

Figure S6. Inversion Identification and Validation. A schematic of clones used to identify 5 large inversions is shown (coloring as in Figure S3). Each of the depicted clones has an "inverted" orientation. Despite the presence of duplicated sequences (grey boxes), the clones have a single, best placement. For each locus, FISH validation is shown with the inverted allele indicated by the arrow. The 8q24 inversion was only observed in 3/20 nuclei and may be mosaic.

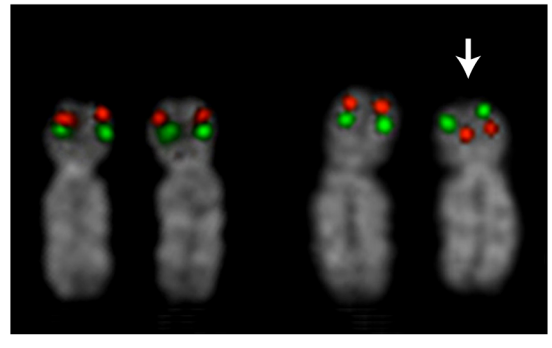
Figure S6

3q29

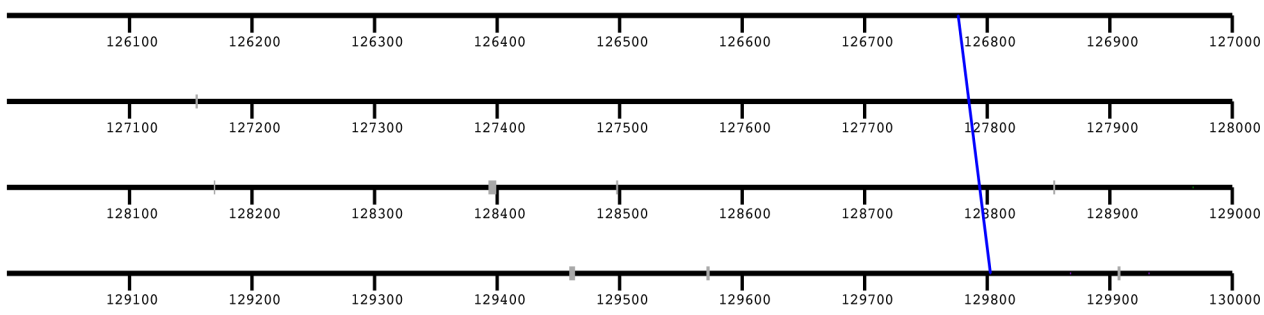


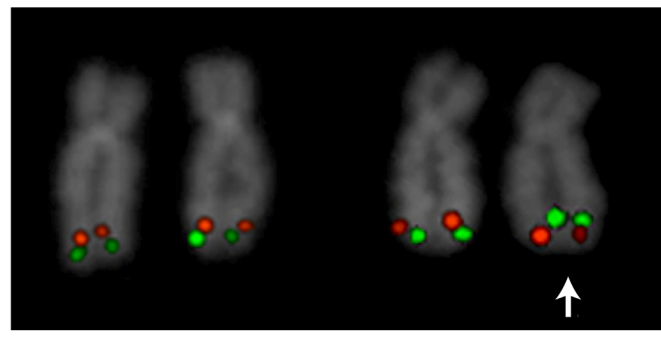


8p23

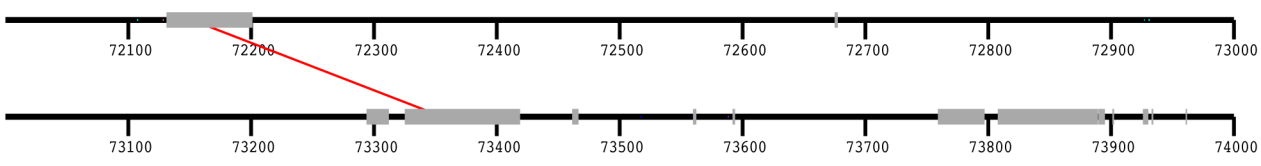


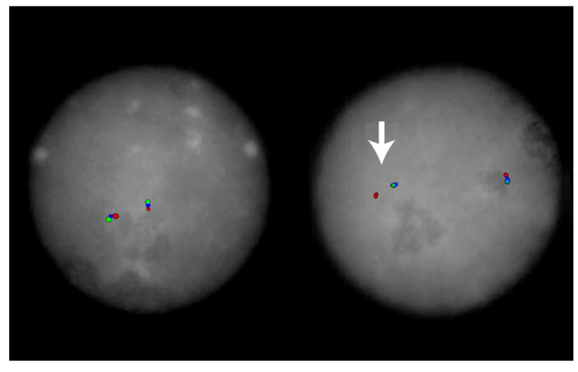
8q24



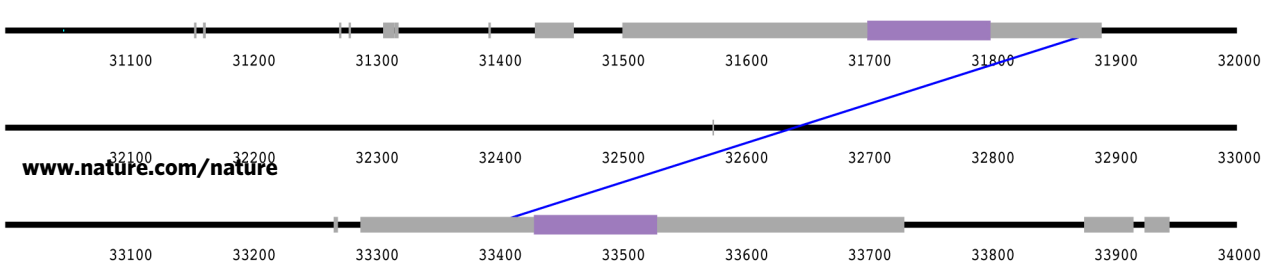


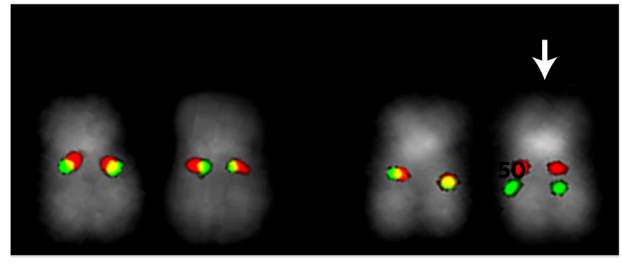
15q24





17q12





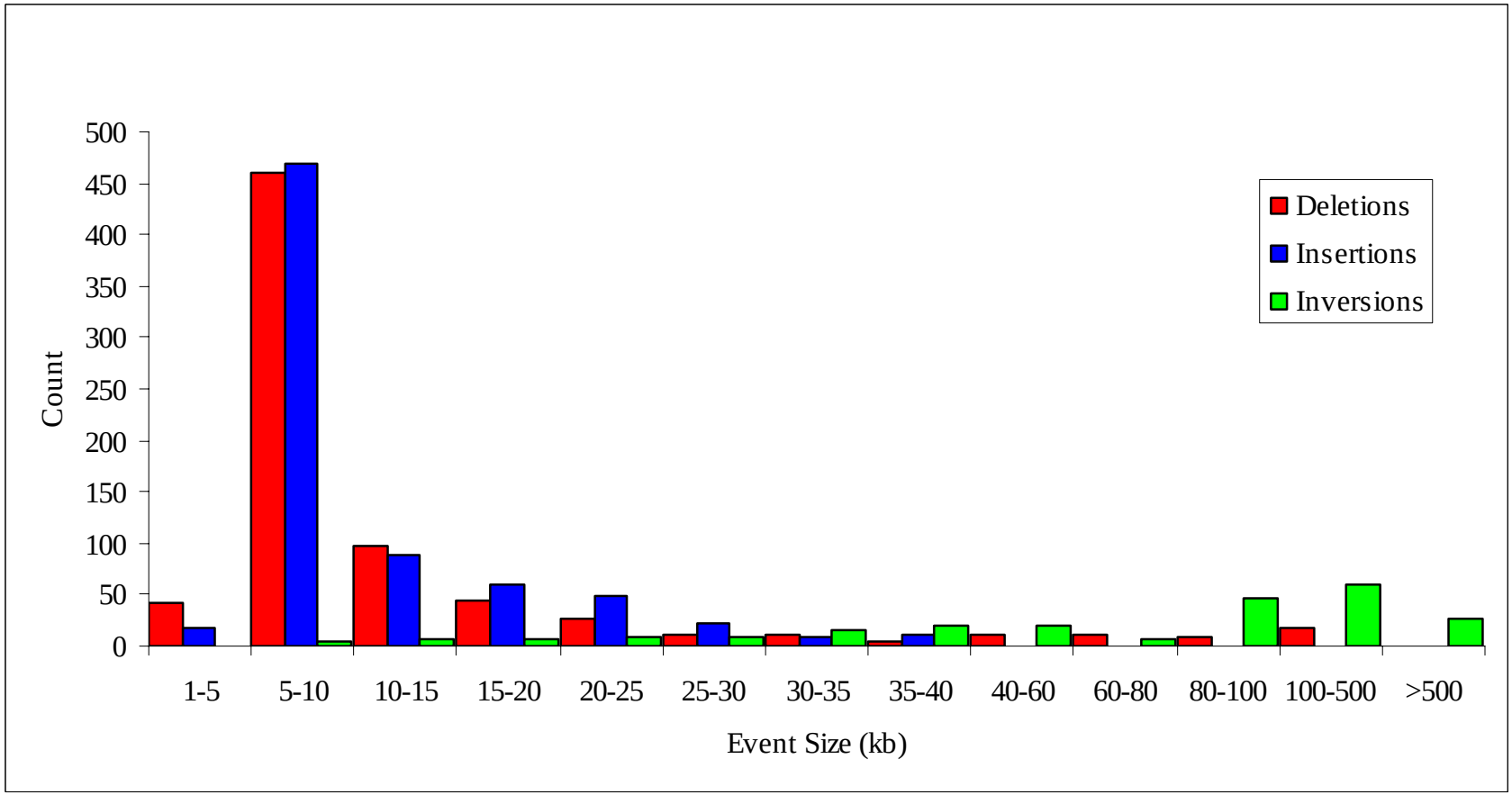


Figure S7 Size of Validated Sites. The size distribution was calculated for 1695 validated events based on the average size discordancy of supporting clones (for deletions or insertions) or the span of discordant clones (for inversions).

Figure S8. **Cross-Platform Comparisons.** Size comparisons for CNVs inferred by different platforms and methodologies on eight common HapMap samples. For each plot, the estimated size for a CNV annotated by fosmid ESP mapping is shown along the X-axis. Sizes for variants that overlap by any number of nucleotides within the same sample as annotated by Redon et al. (a) which used BAC CGH/Affymetrix 5.0 arrays, McCarroll et al. (b) which used Affymetrix 6.0 arrays, or Cooper et al. (c) which used IlluminaHuman1M BeadChips. The scale is in nucleotides on all three plots. Note that clustering of sizes can be seen, this is a result of the same CNV inferred within multiple samples.

Figure S8 a)

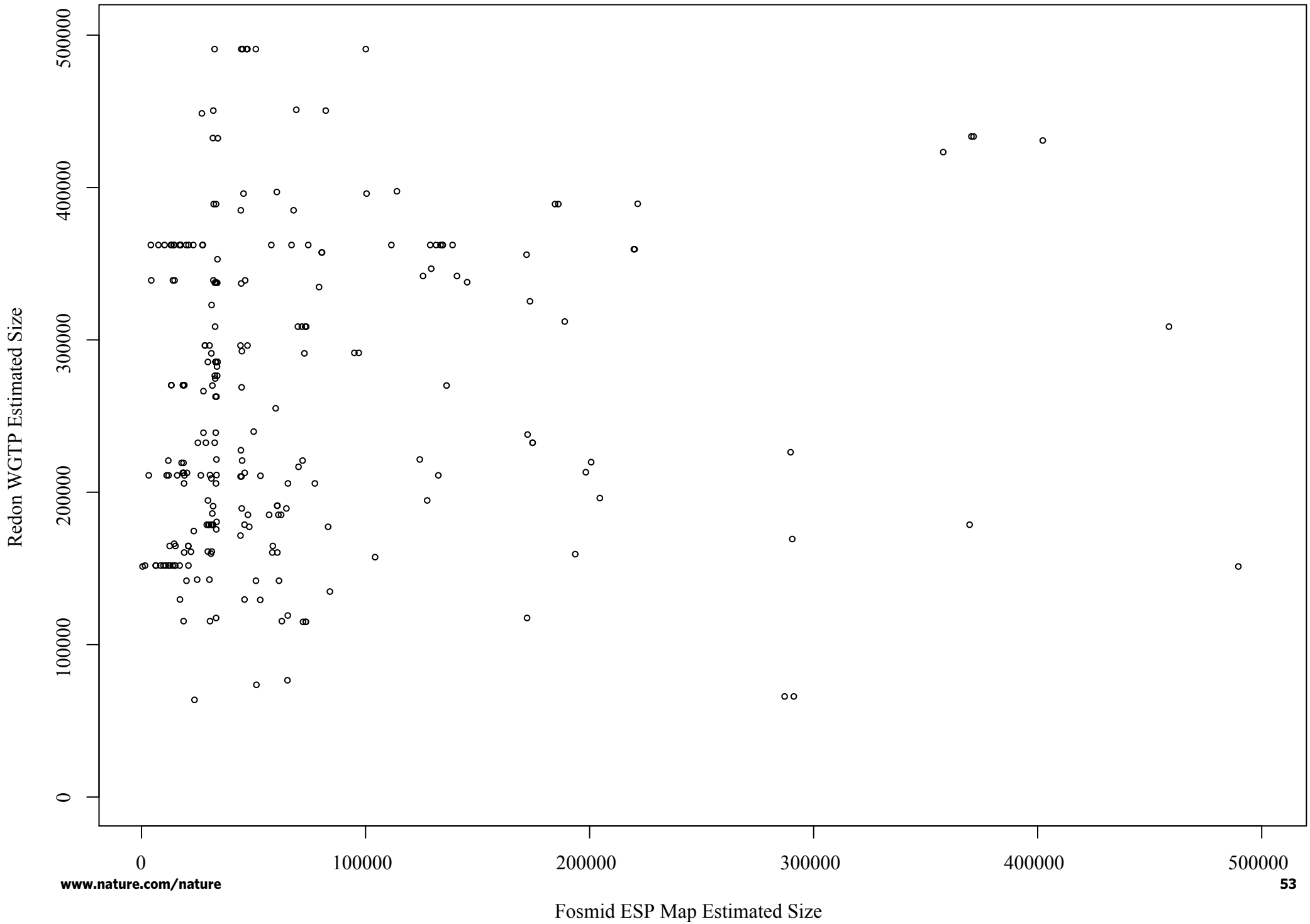

Redon WGTP Estimated Size
0
100000
200000
300000
400000
500000
0
100000
200000
300000
400000
500000
Fosmid ESP Map Estimated Size

Figure S8 b)

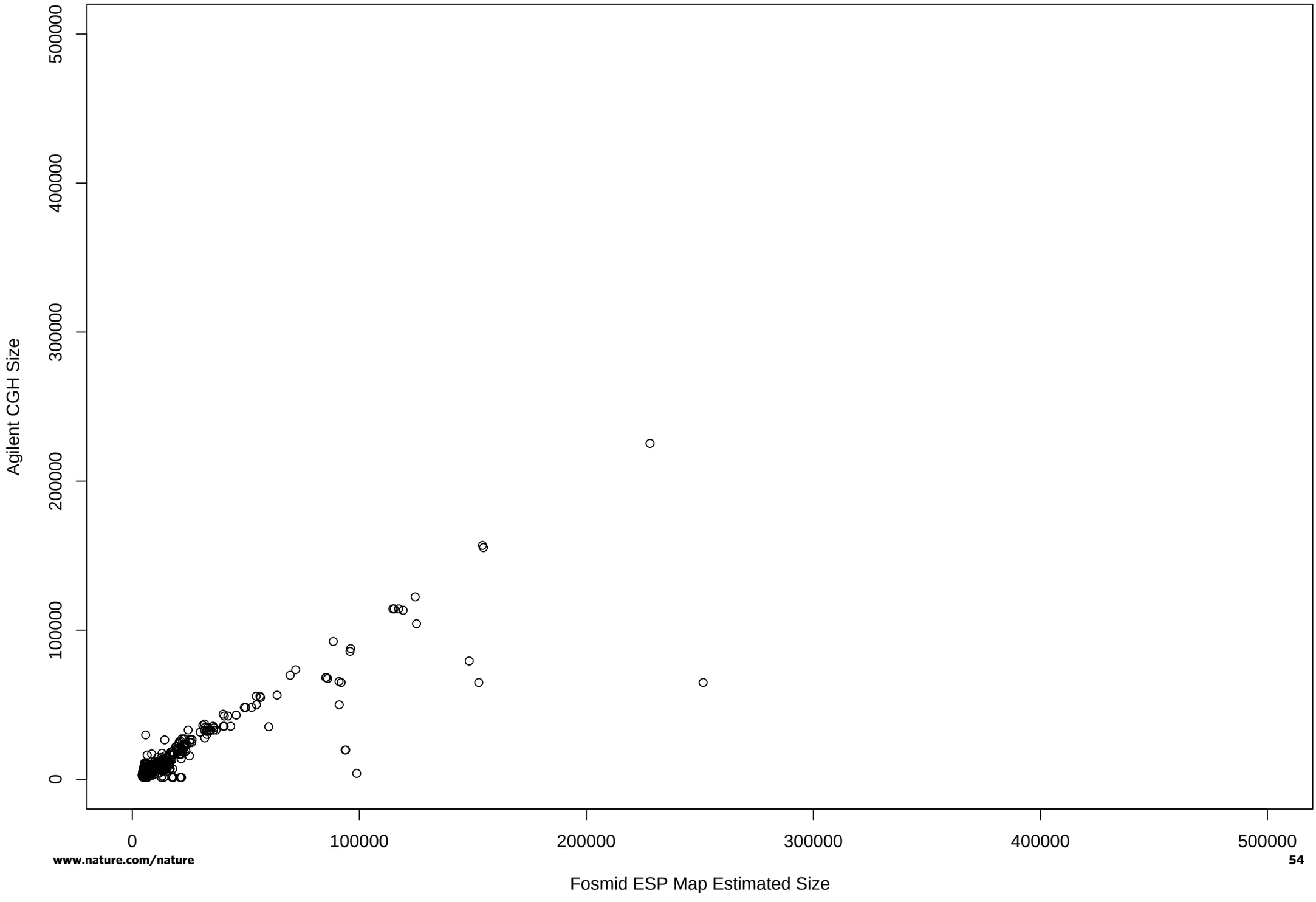

Agilent CGH Size
0
100000
200000
300000
400000
500000
0
100000
200000
300000
400000
500000
Fosmid ESP Map Estimated Size

Figure S8 c)

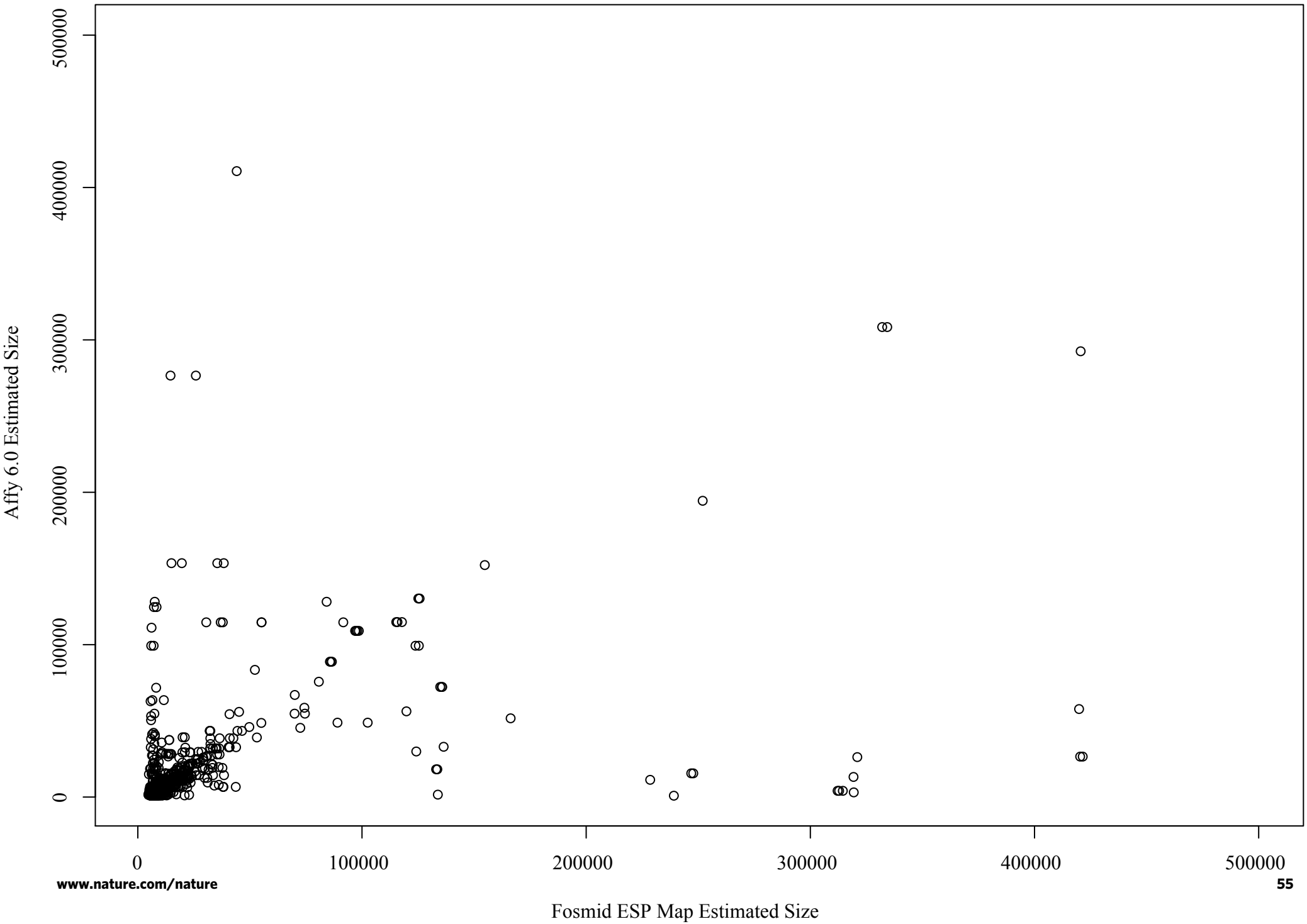
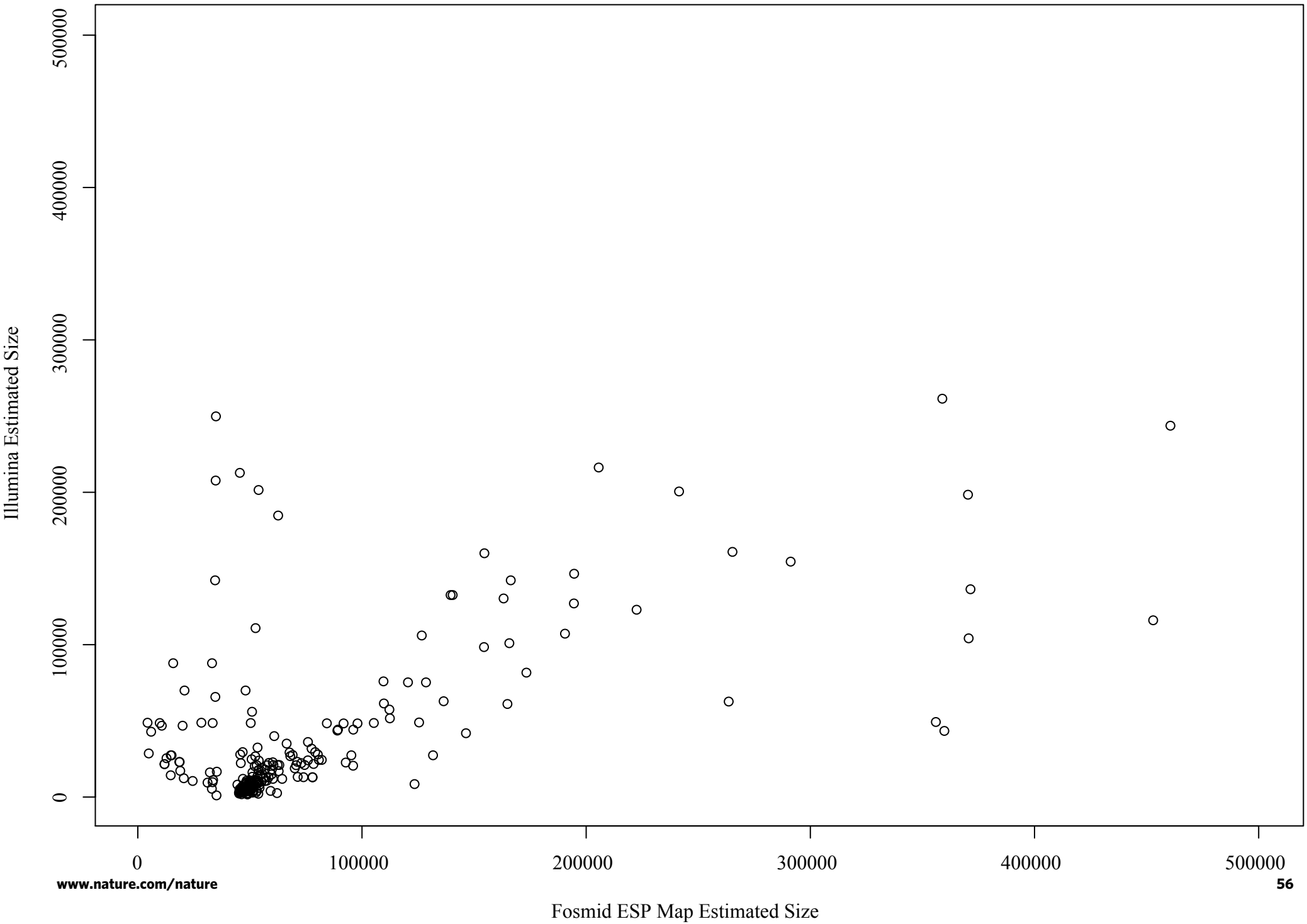

Affy 6.0 Estimated Size
0
100000
200000
300000
400000
500000
0
100000
200000
300000
400000
500000
Fosmid ESP Map Estimated Size

Figure S8 d)


Illumina Estimated Size
0
100000
200000
300000
400000
500000
0
100000
200000
300000
400000
500000
Fosmid ESP Map Estimated Size

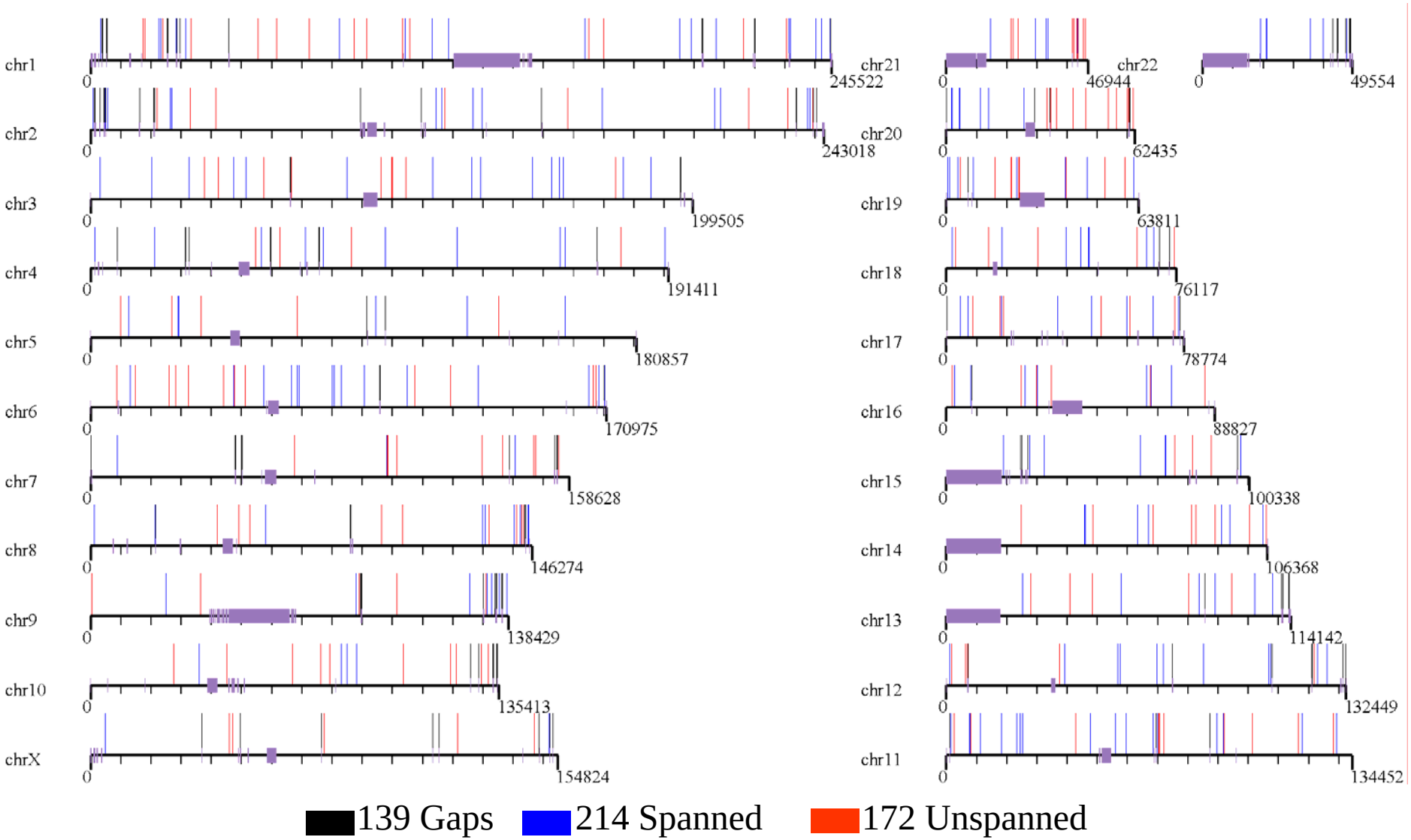


Figure S9. Map of Novel Insertion Loci. The approximate locations of 525 putative new insertion loci based on positions of one-end anchored clones is shown for each human chromosome. Three categories of OEA clusters are distinguished: a) flanking a gap in the assembly (black) b) flanking a discordant fosmid predicting an insertion allele (spanned=blue) or c) neither (unspanned=red). The latter may correspond to larger insertion sequences.