

Supporting online material for

Diversity of Human Copy Number Variation and Multicopy Genes

Peter H. Sudmant*, Jacob O. Kitzman*, Francesca Antonacci, Can Alkan, Maika Malig, Anya Tsalenko, Nick Sampas, Laurakay Bruhn, Jay Shendure, 1000 Genomes Project†, Evan E. Eichler

*Both authors contributed equally to this work.

†A complete list of participants in the 1000 Genomes Project Consortium is available at the end of this document

TABLE OF CONTENTS

| | |
|---------------------------------------------------------------------------------|-----|
| MATERIALS & METHODS | 2 |
| SUPPORTING TEXT | 3 |
| 1. COPY NUMBER ESTIMATION | 3 |
| 1.1. Data and methods..... | 3 |
| 1.2. False discovery estimation..... | 4 |
| 1.3. Digital comparative genomic hybridization and the human CNV landscape..... | 5 |
| 2. EXPERIMENTAL VALIDATION | 7 |
| 2.1. SNP microarray validation..... | 7 |
| 2.2. FISH validation..... | 8 |
| 2.3. Quantitative PCR validation | 8 |
| 2.4. Array CGH validation..... | 9 |
| 2.5 Copy number aware array comparative genomic hybridization..... | 9 |
| 3. GENE ANALYSIS | 10 |
| 3.1. Missing human genes..... | 10 |
| 3.2. Most variable genes | 10 |
| 3.3. Highly population stratified genes..... | 11 |
| 3.4. Human-specific gene duplications..... | 11 |
| 4. SINGLY UNIQUE NUCLEOTIDE (SUN) ANALYSES..... | 11 |
| 4.1. SUN discovery and genotyping..... | 12 |
| 4.2. Validation of paralog-specific copy number estimation..... | 15 |
| 4.3. Paralogous regions of genomic variation..... | 16 |
| 4.4. Gene family analyses | 17 |
| 4.5. Validation of novel structural variants within segmental duplications..... | 20 |
| 4.6. Gene conversion analysis..... | 21 |
| SUPPORTING FIGURES | 24 |
| SUPPORTING TABLES..... | 82 |
| SUPPORTING REFERENCES..... | 100 |
| THE 1000 GENOMES CONSORTIUM | 102 |

MATERIALS & METHODS

Short read sequencing data were obtained from the NCBI Sequence Read Archive (<http://www.ncbi.nlm.nih.gov/sra> and Table S1) and mapped to a masked reference genome (Build36) masked with RepeatMasker and Tandem Repeat Finder. Mapping was performed with mrsFAST (*S1*) allowing for up to 2 mismatches. Reads exceeding 36 bp in length were truncated to 36 bp or divided into non-overlapping 36 bp constituents. Read depth profiles were then constructed independently for each sequencing library, corrected for G+C bias introduced during library construction (*S2*) using a smoothed correction factor, and combined. Copy number prediction was performed by regression against a standard curve fit to regions of known copy. Individual loci were genotyped by computing the median copy number across nonoverlapping windows of 1000 unmasked bases each. We constructed a map of 4.06×10^6 paralog-specific tags, or SUNs (Singly Unique Nucleotides) as follows: briefly, the human reference genome (Build36) was divided into its constituent k-mers ($k=30$) for both forward and reverse complement DNA. Kmers were screened for uniqueness by mapping to the reference with the mrFAST aligner (*S3*). We rejected k-mers with more than one perfect match in the genome as well as those with highly repetitive content. Within segmental duplications, we defined SUNs as the set of paralogous sequence variants within the underlying pairwise alignment between segmental duplications (Table S2).

We used fluorescence in situ hybridization (FISH), quantitative PCR (qPCR), and array CGH to validate our read depth-based copy number estimates.

FISH Analysis Methods: Metaphase spreads were obtained from 32 lymphoblast cell lines from Coriell Cell Repository, Camden, NJ, and one YH (Han Chinese). FISH experiments were performed using fosmid clones (Table S3) directly labeled by nick-translation with Cy3-dUTP (PerkinElmer) as previously described (*S4*), with minor modifications. Briefly, 300 ng of labeled probe were used for the FISH experiments; hybridization was performed at 37°C in 2xSSC, 50% (v/v) formamide, 10% (w/v) dextran sulphate, and 3µg sonicated salmon sperm DNA, in a volume of 10µL. Posthybridization washing was at 60°C in 0.1xSSC (three times, high stringency). Nuclei were simultaneously DAPI stained. Digital images were obtained using a Leica DMRXA2 epifluorescence microscope equipped with a cooled CCD camera (Princeton Instruments). DAPI and Cy3 fluorescence signals, detected with specific filters, were recorded separately as gray-scale images. Pseudocoloring and merging of images was performed using Adobe Photoshop software. A minimum of 50 interphase cells were scored for each probe.

Quantitative PCR assays: Quantitative PCR assays are listed in Table S4 using primer sequences listed in Table S5. Primer oligos were obtained from Operon (Huntsville, AL). Copy number estimates were obtained by the $\Delta\Delta C_T$ method and normalized using a primer set directed against a diploid control, the human albumin gene. Reactions were carried out in quadruplet using SYBR Green I Master Mix on a Roche LightCycler 480 thermocycler.

Array CGH: Copy number-aware CGH was performed by hybridizing two differentially Cy-labeled samples to a custom, high-density oligonucleotide 4x180K CGH microarray (Agilent, Santa Clara, CA) targeted at 17q21.31 with a density of 1 probe per 100bp. Labeling, hybridization, scanning, and data processing were performed as directed by the manufacturer. For the reference sample, an individual having intermediate copy in the target region was chosen (NA19240).

Additional validation was performed by comparison to SNP-chip based events from the same individuals called on the Affymetrix 6.0 and Illumina 1M DUO SNP platforms and CNV events and genotypes from the NimbleGen 41M probe and Agilent single-channel CGH platforms (S5-8). Paralog-specific copy number estimation was validated by comparison to fully sequenced clones corresponding to deletions within segmental duplications (S9), by paralog-specific quantitative PCR, and by Illumina short-read fosmid clone inset sequencing.

SUPPORTING TEXT

1. COPY NUMBER ESTIMATION

1.1. Data and methods

Copy number prediction was performed with read depth-based methods adapted from those we previously described (S3). A detailed description of the data and modifications to this method are described below.

Data: Sequence data were collected primarily from *Pilot 1* and *Pilot 2* of the 1000 Genomes Project, which includes two trios sequenced to high coverage and 179 individuals sequenced from 1-4X. A high coverage trio, including the published NA18507 genome and the genome of Jay Flatley, was also obtained from Illumina (<http://www.ncbi.nlm.nih.gov/sra/SRP000978>, <http://www.ncbi.nlm.nih.gov/sra/SRP001050>). In addition, 10 other published genomes of varying coverages were collected: NA10851 (S10), AK1 (S11), SJK (S12), KB1 ABT (S13), YH (S14), and five individuals from a Human Genome Diversity Panel (S15), as well as a collection of great apes including a gorilla, a chimpanzee (capillary sequence segmented into 36-bp fragments to mimic Illumina datasets as previously described, ref. (S3)) and a Bornean orangutan (Fig. S1). Additionally, we simulated the reference genome at 6X segmenting the assembly (Build36) into 36-bp fragments.

Read mapping: Short-read sequence data were mapped to the masked human reference genome (Build36) using the software *mrsFAST* (S1) with a maximum of two substitution mismatches allowing for no indels. *mrsFAST* returns all locations in the genome at which a specific read can map, given the specified parameters. All reads exceeding 36 base pairs (bp) in length were truncated to 36 bp, or divided into their constituent nonoverlapping 36-bp sequences to eliminate potential mapping biases between genomes sequenced at different read lengths. Using *mrsFAST*, we mapped 35.4 billion reads from this dataset to the human reference genome in one month using a 160 CPU Linux cluster farm. Masking was defined as all base pairs identified by RepeatMasker 3.2.7 and TandemRepeatFinder (S16) as defined in the UCSC genome browser (<http://hgdownload.cse.ucsc.edu/goldenPath/hg18/bigZips/>). Masking was extended 36 base pairs upstream and downstream of the original mask during construction of read depth maps to account for potential edge mapping effects.

G+C correction: Sequence from each genome library was mapped individually to the human reference genome (Build36) and each library was tested to account for potential biases that may have occurred during library preparation (Fig. S2). We applied a multiplicative G+C correction to every library to account for biases in construction and/or sequencing. This correction was determined by calculating the

average read depth across the control diploid regions (Build36 - [Database of Genomic Variants + gaps

$$k_{gc} = \frac{\mu_{total}}{\mu_{gc}}$$

+ genomic super dups]) binned by G+C content. The correction factor k_{gc} was defined as where μ_{total} is the total average read depth across all copy number 2 regions of the genome and μ_{gc} is the average read depth across regions of a specific G+C content. μ_{gc} was smoothed by LOESS regression, with piecewise linear regression at high and low G+C outliers. The corrected read depth at any base x in the genome $d'(x)_{gc}$ is then calculated as $d'(x)_{gc} = d(x)_{gc}k_{gc}$ where $d(x)_{gc}$ is the original depth at base x . A maximal G+C correction factor of two was enforced.

Quality control: To assess the quality of individual libraries we analyzed the correlation of the (G+C)-corrected read depth with the copy number of regions of experimentally validated fixed copy (Fig. S3).

As the effective coverage of a genome increases, the correlation between read depth and copy increases asymptotically as demonstrated by analyzing this metric on genomes generated from subsampling a single, high coverage genome (Fig. S4). We exploited this relationship as a quality control (QC) metric to judge the quality of each individual library (Fig. S5). Libraries that appeared to deviate from this asymptotic trend of increased correlation were discarded (i.e. all libraries that failed a correlation threshold of 0.85 were omitted). All libraries from a single genome passing QC were then combined by summing their (G+C)-corrected read depths at each base. 177/198 libraries from *Pilot 1* low coverage genomes passed our QC metric.

Copy Number Prediction: Copy number was predicted using a linear regression model based on our read depths versus copy number in regions of known fixed copy number (Fig. S3). We also developed a method to estimate copy number using a simple linear model maximum likelihood approach, as read depths were approximately normally distributed for each integer copy number state (Fig. S6). Using the predicted mean and variance of the Gaussian distributions underlying different copy numbers, a series of models (representing copy 0, 1, 2, 3 and so on) can then be generated to represent the likely distribution of read depths underlying a region of specific copy number. Given a region of interest, each of these models can then be tested to determine by maximum likelihood which model, and subsequently which copy number, best describes the underlying read depth distribution of that particular locus.

1.2. False discovery estimation

To determine an approximate upper bound on our false discovery rate (FDR) and how this changes as a function of sequence coverage and event size, we analyzed the copy number of diploid regions that are copy-number invariant in most humans. We defined these invariant diploid regions (cumulatively $\sim 1.2 \times 10^9$ non-RepeatMasked bp) by excluding from the whole genome known CNV regions (S5, 17) and all segmental duplications (WGAC+WSSD) (S18, 19). We limited our analysis to all contiguous autosomal sequence greater than 100 kb in length (i.e., regions of uninterrupted diploid copy). As a stringent measure of FDR, we analyzed all 1-kb windows of unmasked sequence in high coverage genomes ($>8X$) and found that 97-99% of individual windows in these regions were predicted at copy 2, as expected. However, a dramatic decrease in the fraction of predicted copy 2 regions was observed as coverage decreased for these small regions (Fig. S7), with the copy 2 fraction varying from 72-98%.

At 1-kbp resolution, we predict an FDR of ~10% for genomes of 4X sequence coverage or greater, however, the FDR rapidly decreases as larger regions are considered (Fig. S8). At 3-kb resolution the FDR is <5% in genomes with as little as 1.5X coverage. For regions >10 kb the FDR reduces to <2%. Similar results were observed for data simulated from subsampling a single, high coverage genome to lower coverage (Fig. S9), however, simulated genomes showed slightly better concordance to genomes at similar coverage.

Using these subsampled genomes, we explored factors contributing to false CNV discovery, again considering 1-kbp windows of diploid invariant sequence. False positives favored losses (predicted copy <2) as opposed to gains (predicted copy >2) (Fig. S10). These corresponded with reduced overall coverage in these regions arising from biases in library preparation and sequencing. Interestingly, at low G+C content, calls <2 contributed almost exclusively to the error, whereas at high G+C, there were errors in both directions: this explains the slight excess of deletion calls. These effects are constrained to a small fraction of the genome as a whole: ~73% of the genome lies between 35%-55% G+C content, where we observe high accuracy and little bias in the direction of error.

We estimated the copy number of all genes (discarding those <3kb in length) in individual NA18507 at full coverage (~43X) and in 15 simulated reduced-coverage genomes created by subsampling reads from NA18507 to ~1-25X coverage. Using the full-coverage genome as a gold standard, we computed deviation in copy number predictions for genes in different copy number classes (Fig. S11). As expected, deviation from the full-coverage estimates widened (i.e., accuracy was lower) as coverage decreased. Additionally, because the variance in read depth scales with copy number state, higher copy number states become increasingly difficult to predict within +/- 0.5 copies. Despite this trend, the magnitude of the errors remains small for genes of moderately elevated copy number, even with greatly reduced sequence data. For a genome at 3X coverage, ~96% (1702/ 1764) of genes >3kb in length with copies ranging from 3-10 remain concordant, changing by < +/- 0.5 copies of the full-coverage estimate (Fig. S12). For an 8X coverage genome, >98% of genes >1kb in length stay within this range.

We next evaluated the effect of coverage upon accuracy using two loci validated across a large panel of low and high coverage genomes. We compared read-depth and single channel array-based copy number predictions for the greatly copy-number expanded and highly variable gene *TBC1D3* (Fig. S13A,B). Consistent with the subsampling results, both low (<3X) and high coverage (>=3X) genomes showed strong correlation between sequencing- and array-based genotypes ($r=0.93$ and $r=0.97$, respectively). Another duplicated gene which we extensively validated was *CCL3L1*, a duplicated gene 1.9kb in length. We designed a quantitative PCR assay for this gene and genotyped 150 samples. This demonstrated a very strong correlation of $r=0.948$ across the entire range of variation (0-14 copies, Fig. S13C). Stratifying the individual genomes by sequencing coverage, we again observed strong correlation with only a modest reduction for low coverage genomes ($\leq 3X$ depth, $r=0.947$ with qPCR) versus those at higher coverage ($>3X$, $r=0.958$).

1.3. Digital comparative genomic hybridization and the human CNV landscape

In order to detect copy number polymorphism among our sequenced genomes we developed a comparative method similar to array comparative genomic hybridization (CGH), which we term digital CGH (dCGH). Windowed copy number estimations from one genome are subtracted from the windowed copy number estimations of a second genome, or from the mean value of a population of genomes. Similarly to array CGH, this provides a differential estimate to detect regions of gain or loss in

a particular genome compared to a reference genome, or a population of genomes (comparable to using a pooled reference experimentally). We utilized this method to detect both common and rare polymorphisms among our genomes.

To detect large copy number variants we performed dCGH on each of our samples versus a high coverage (28X) European individual, NA10852. This experimental procedure was analogous to that implemented with CGH in a recent publication (S5). The use of the same reference and test genomes facilitated direct comparison between these two datasets. Events were selected for which at least 15/20 contiguous 1-kb unmasked windows showed a differential of at least 0.5 (positive or negative, corresponding to at least a single copy gain or a loss, respectively). We spanned across gaps in the reference genome, and regions with a total length <20 kb or with fewer than 10 kb of unmasked base pairs were discarded. An average of 338 large CNVs were discovered per individual, however, the mean number of CNVs discovered in individuals with >10X coverage was 298 (n = 14) indicating an increase in the number of false positives in low coverage genomes because of their higher variability (Fig. S14). As this analysis is not optimized for boundary detection and may capture segmental duplication *mirror* effects (see below), it may inflate the number of copy number variable bases.

We merged these calls into a set of 1101 (~202,793 kb) copy number variable regions (CNVRs) and compared them to 1273 (~91,815 kb) calls matching the same size threshold criteria (>20 kb total, >10-kb unmasked sequence) reported by Conrad *et al.* (S5), discovered in 41 individuals. 81% of large Conrad CNVR base pairs were detected in our discovery set; after excluding private events from both call sets, this fraction increased to 85%. We focused on 952 large events >50 kb and noted that the majority (55%, 522/952) of these regions overlapped segmental duplications (>20% overlap). As expected, events of increasing size were at progressively lower frequencies (Fig. S15).

A substantial fraction of the observed polymorphism exists at <5% frequency within the population, with 53% (500/952) of events detected in fewer than eight individuals (Fig. S16). Interestingly, we noticed a significant change in the frequency spectrum after stratifying events by segmental duplication content (Fig. S17). Events containing no segmental duplication were overwhelmingly rare in frequency. For example, 71% (390/546) of such events were observed in less than three individuals, while 55% (302/546) without duplication were private events. In contrast, events containing segmental duplication were found at both high and low frequencies. For example, 16% (66/406) were observed in ≤8 individuals, while 42% (170/406) were highly polymorphic across most individuals.

Combining dCGH results with absolute copy number estimations from a population sample allows us to both discover and genotype variation. Additionally, combining paralog-specific copy number maps (see below, section 4) allows us to demarcate the boundaries of structural variants falling within segmental duplications. For example, we identify two distinct, large CNPs (210-kbp and 205-kbp) on chromosome 17q21.31 based on signals of gain from dCGH extending into the more highly duplicated region (Fig. S18). Paralog-specific copy number maps assist in determining the approximate breakpoint locations. We genotyped the resulting regions and observed absolute copy number differences between 2-6 as clearly distinguishable, discrete integer values, which we subsequently validated (Fig. S19 and see section 2.2).

We have formatted these genome-wide copy number maps as heatmaps (e.g., see main text Fig. 1, 2) for visualization across large loci and many individuals, as well as in a UCSC Genome Browser instance (S20) for fine-scale visualization. Both resources may be accessed at (<http://hgsv.washington.edu/>).

2. EXPERIMENTAL VALIDATION

We used four independent methods to validate the accuracy of our read depth copy number estimates: fluorescence *in situ* hybridization (FISH), SNP microarrays, quantitative PCR (qPCR) and array CGH. In addition, we demonstrate how our genome-wide assessment of copy number may be used to improve discovery and genotyping using array comparative genomic hybridization.

2.1. SNP microarray validation

CNV detection: To assess the power and accuracy of read depth CNV estimation against currently accepted experimental approaches for detecting copy number variation, we compared our absolute copy number estimates to CNV calls across 253 HapMap samples made on the Illumina 1M Duo SNP genotyping platform. Loci were discovered on the SNP microarray using a Hidden Markov Model or HMM as previously described (S7) and classified as a ‘gain,’ ‘loss,’ or ‘homozygous loss’ with respect to the population average (median size 39.6 kb; smallest event 1.6 kb). For each of these loci we took the median regression-based copy number call. We analyzed 152 samples (*Pilot 1*, *Pilot 2* and NA18507 trio) and determined the state of each individual as ‘loss’ or ‘gain’ with respect to the average copy number call across all individuals. 109 individuals were common to both sets of samples containing 2270 events with at least 1000 bp of unmasked sequence (503 gains, 1241 losses, and 526 homozygous deletions, respectively). Depending on the size and type of the event, we correctly predicted 94-100% of events using our read depth metric (Fig. S20). As expected, larger events are easier to detect, but the accuracy also increases as a function of effective read depth (Fig. S21).

CNP Prediction: SNP microarray genotypes and subsampling analysis

To assess our ability to accurately genotype copy number (as opposed to simple gains or losses), we compared our calls to those recently made by another group using Affymetrix 6.0 SNP microarray across 270 HapMap samples (S21), of which 114 overlapped with our samples. We selected 1015 of the 1320 regions assayed by McCarroll *et al.* containing at least 1000 bp of unmasked sequence for genotyping (S21). Our initial results indicated a 70% genotyping concordance with the SNP microarray copy number estimates (82,373/117,847). We identified 243 regions (n = 243) that showed little correspondence with our genotype calls (<20%). Of these, 94% (228/243) overlapped or were contained entirely within segmentally duplicated portions of the human genome, and encompassed 300 gene models.

After adjustment, we observed 86% concordance (101,631/117,847) between the SNP genotyping calls and our copy number estimates. Considering only genomes with high sequence coverage (seven genomes ranging from 13-29X effective coverage), this increased to 90% (8101/8980). Unadjusted concordance only increased to 73%, further supporting the misassignment of the population average copy number call among the SNP chip-based genotype calls. In contrast, analysis of unique regions of the genome (i.e. not containing segmental duplications) yielded 94% concordance (79,202/84,145) between call sets for unadjusted SNP-chip CN genotypes and 95% concordance (80,086/84,145) with adjusted SNP-chip CN genotypes (Fig. S22). The decreased concordance of genotype calls for segmental duplications stems in large part to the complex nature of these regions. In many cases though, the average copy number across the population for a region fluctuates, as detected in the genotypes from McCarroll *et al* (S21). The underlying structure of the region is, thus, much more complex, contributing to the observed discordance.

To further analyze the effect of variable genome sequence read depth on our ability to genotype copy number, we randomly subsampled reads from NA18507 and simulated genome coverage ranging from ~1-30X. We then analyzed the concordance between the McCarroll *et al.* genotype calls (n = 989) and calls adjusted according to the consensus difference observed in high coverage genomes (>13X, n = 6, NA18507 excluded) (Fig. S23a). We observe an asymptotic, rapid increase in the concordance of genotypes as coverage increases, leveling out at ~8X. To explicitly determine why higher coverage enables more accurate genotyping, we plotted the mean and standard deviation of coverage in 1-kb windows tiling a copy number invariant portion of the human genome across individuals of varied coverage (Fig. S23b). Though the variance of read depth increases linearly with increased coverage of a genome, the mean of the read depth increases at a more rapid rate facilitating more robust estimates of the underlying copy of a region.

2.2. FISH validation

We experimentally validated by fluorescence *in situ* hybridization (FISH) 25 predicted copy number differences of the genomes of 33 individuals using cell lines from the same individuals from which the computational predictions were generated (Fig. S24). There were three general categories of experiments: 1) highly copy number variable regions where individuals with the most extreme copy number difference were selected; 2) large, rare events seen in a single individual; and 3) duplications predicted to be duplicated in all individuals but represented as single copy in the human reference. FISH experiments were perfectly concordant with the computational predictions in 80% (47/59) of the cases, and in 93% (55/59) of the cases the FISH copy numbers differed from the prediction by one copy. In the rest, 7% (4/59) of the predicted copy numbers differed by two or three copies. Of these, three out of four predictions are in the *NPEPPS* region, for which we observed a very strong correlation between copy number predictions and those from the qPCR experiment ($r = 0.96$) and the single-channel array CGH validation ($r = 0.85$). This indicates a likely problem in the copy number estimation by FISH. In three highly copy number variable regions, FISH failed to accurately estimate copy number differences between individuals because the signals visualized on interphase nuclei were too numerous to provide an exact number of copies. In one case within the protocadherin gene cluster, FISH could not validate the computational predictions since the duplicated region was too small to be detected by FISH. The observed duplication structure has previously been confirmed in the literature (S22) confirming the failure of the FISH assay in this case (Table S3).

2.3. Quantitative PCR validation

We designed two primer pair sets across eight genic regions predicted to be both multi-allelic and highly polymorphic to validate our sequence-based copy number predictions. We additionally designed one paralog-specific qPCR assay. For several loci, we selected the most robust primer set among multiple ones tested. To determine the qPCR dynamic range we attempted to select individuals spanning the full copy number range. Copy numbers were estimated using the $\Delta\Delta C_T$ method normalized using a control directed against the diploid albumin gene. We observed very strong correlations between our copy number predictions and those from the qPCR experiment with R^2 values greater than 0.84 for seven of nine regions and, and only two regions failing all primer sets (Table S4, Figure S25).

As an example, we estimated the copy number of the gene *CCL3L1* in 150 individuals using qPCR to independently verify our sequence-based copy numbers. The *CCL3L1* gene is 1.9 kb long and known to be highly stratified between populations (S23-27). We observed a strong concordance between the sequencing-based copy number estimates and those made by qPCR, $r = 0.95$ (Fig. S26). Furthermore,

we demonstrate distinct population stratification of the *CCL3L1* region captured by our sequence-based copy number estimates and confirmed by qPCR.

2.4. Array CGH validation

One of the perceived limitations of array comparative genomic hybridization is its inability to distinguish differences for regions of high copy number. This is due to the individual variability of probe binding characteristics among loci as well as signal saturation within duplicated regions resulting in reduced sensitivity to detect smaller changes (e.g., 9 copies vs. 8 copies being harder to distinguish than 1 copy vs. 2 copies). Using our absolute read depth estimates for loci, we reassessed the accuracy of array-based methods by taking advantage of single-channel intensity data and leveraging the repeated hybridization of the reference genome over a set of experiments performed on multiple individuals.

We developed a CGH-based approach that leverages our absolute read depth copy number to perform more accurate genotyping of multi-copy CNVs. As a test dataset we used CGH data from (*S5*) in regions determined to be variable. These regions were run through a segmentation algorithm and merged across samples to obtain 7,498 distinct regions. Copy number was predicted from array data by first determining the median signal intensity for a region of diploid copy in the reference, S_{MR} . The reference sample copy number of a test region was then predicted by taking the median reference signal across all probes and arrays assayed, divided by S_{MR} . The copy number state of a sample was then predicted from the \log_2 ratio of the test versus the reference in a region multiplied by the inferred reference copy. This allows us to take advantage of the fact that the reference DNA sample used for array CGH is typically hybridized several hundred times, thus, copy number estimates of the median single-channel intensity of a region are likely to be robust.

Sequencing-based copy number predictions were compared to array-based copy number predictions by analyzing the correlation of copy number calls between samples. We noted in many regions that the underlying distribution of signal intensity on the array from the reference sample did not vary greatly with respect to the test samples, precluding this kind of analysis due to the lack of underlying variance in the copy number predictions (Fig. S27a), or inability of the array platform to detect variation at this locus. In order to filter by variance to select regions in which copy number variation was captured by the array, we developed a metric 'rCV'. $rCV = CV(\text{sample channel})/CV(\text{reference channel})$ where $CV = \text{std}(\text{sample signals})/\text{mean}(\text{sample signals})$. Higher numbers indicate a greater spread of signal intensities in the test samples than in the reference (Fig. S27b). For regions with >1 kb of unmasked sequence, we found that for rCV values greater than 2, 77% (501/648) regions have a correlation ≥ 0.7 . For regions with rCV greater than 3, 91% (248/272) have a correlation of ≥ 0.7 (Figs. S28 and S29).

Among regions with sufficient copy number response with respect to the reference as determined by the rCV metric (Fig. S30), we can calibrate array CGH copy number calls with read depth-based calls from a reference genome such as NA10851, which was used as a reference for CNV discovery in a large array CGH study (*S5*) and recently sequenced (*S12*).

2.5 Copy number aware array comparative genomic hybridization

Our analysis of 159 genomes provides a reference set of DNA and cell lines from which a specific reference DNA sample of known absolute copy may be used to maximize copy number sensitivity. We refer to this pre-selection of a particular reference of known absolute copy number as *copy number aware array CGH*. We demonstrate its utility and the accuracy of our predictions by reanalyzing the

17q21.31 region discussed above. We selected a DNA sample, NA19240, whose copy number across the mosaic region ranged from 1-5 copies (Fig. S31). This particular individual was an ideal reference sample because its copy number range approximated the median copy number for the populations suited. Comparative genomic hybridization using this reference sample was performed on five HapMap individuals with custom, high-density oligonucleotide 4x180K Agilent chips targeted to the 17q21.31 region with a density of 1 probe per 100 bp. Array CGH showed strong concordance with the computational predictions. We identified three distinct copy number polymorphisms of size 155, 205 and 135 kbp as well as numerous smaller common CNPs within this region. Additionally, using the copy numbers estimated in our reference genome, we were able to infer the true exact copy of the variable loci targeted in the test samples.

3. GENE ANALYSIS

We analyzed the genic portion of the human genome to assay for putative functional copy number variants. We performed copy number prediction on 25,832 nonredundant RefSeq gene models from the UCSC Build36 refFlat table using the median regressed copy number estimates for contiguous 1000-bp windows tiled across the genes. Gene models shorter than 1 kbp were excluded from this analysis (see Section 1.2). Genes were grouped into paralogous gene families using the relationships defined by NCBI Homologene and then further manually curated. As expected, the vast majority of genes were predicted as diploid, with 97% of non-duplicated genes having a median copy number of ~ 2 across all 159 individuals (median copy ≥ 1.5 and < 2.5 , fig. S32); 91% of these genes with > 10 -kb unmasked sequence were determined to be fixed in all 159 genomes analyzed. As expected, simulating copy number estimates for the reference assembly by splitting it into short read-sized fragments (36bp) and calculating copy number as before yielded almost exclusively estimates of diploid copy (fig. S32).

3.1. Missing human genes

We tested whether specific gene families were underrepresented in the human genome by searching for genes with median estimated copy among individuals greater than that in the reference genome Build36 (difference ≥ 5 copies between median copy in 159 individuals and reference genome) (Table S6; Fig. S33a). 167 gene models were identified corresponding to 44 gene families.

3.2. Most variable genes

We identified the most variable gene families among the 159 individuals analyzed (Table S7) selecting those with a high variance (> 3 copies²) across individuals. 260 genes were identified corresponding to 56 individual gene families (Fig. S33b). Enrichment for segmental duplication was calculated using these 56 gene families and a list of 17,601 nonredundant human genes, manually curated to include only one representative gene from each duplicated gene family (S3). Genes were considered to be in segmental duplications when $> 50\%$ of their length was covered by one or more segmental duplications. We found striking enrichments for segmental duplications in variable genes (OR 311.3, $P < 2.2 \times 10^{-16}$, Fisher's Exact). Notably, many of the genes found to be missing copies from the reference were also the most variable among humans.

We also observed 28 large regions showing extreme copy number variability with highly variable estimates within each region, suggesting complex variation in the underlying duplication architecture (available from <http://hgsv.washington.edu>). One such example spanning nearly one megabase and containing several brain and testis-expressed genes is shown in Fig. S34.

3.3. Highly population stratified genes

We identified the most population stratified gene families across individuals using the statistic V_{st} developed by Redon *et al.* (S28) (Table S8). Briefly, V_{st} was calculated as $(V_T - V_s)/V_T$, where V_T is the variance in copy number among all unrelated individuals and V_s is the population-specific variance. V_{st} values were calculated separately for each pair of populations. 176 genes were identified having V_{st} values greater than 0.2 clustering to 64 highly stratified gene families many falling within 22 large (>100 kbp) regions of increased population stratification (Table S9). Segmental duplication enrichment was calculated as above indicating a significant enrichment for segmental duplications in highly stratified genes (OR=311.3, $P < 2.2 \times 10^{-16}$). Moreover, genes exhibiting population stratification ($V_{st} > 0.05$) tended to have higher V_{st} values when they overlapped with segmental duplications ($P < 2.2 \times 10^{-16}$, two-sided KS test, fig. S35), indicating that segmental duplications harbor some of the most highly population-stratified sites in the genome. No correlation was observed between V_{st} and copy number indicating that increase of the V_{st} statistic is not an artifact of increased copy ($P = 0.2$; Fig. S36).

Among the most stratified gene families identified by our analysis, many have been previously described, including 3 among our top 12 candidates (*UGT2B*, *CCL3L1*, and *LILRA3*) (S23-27, 29-31) (See main text Fig. 3a; Fig. S37). Additionally, we identify highly duplicated gene families previously unassayable for copy number showing striking signatures of population stratification.

3.4. Human-specific gene duplications

We compared copy number from 159 human genomes to those of three outgroup ape species: the genome of a gorilla, a Bornean orangutan, and the “illuminized” chimpanzee genome. We identified 74 genes that were diploid among all nonhuman primates but showed an increase in the median copy in the 159 individuals analyzed. These genes cluster to 23 gene families that have been specifically duplicated in the human lineage (Table S10; see also Main Text Fig. 3b) and localized to 16 regions. We determined that 8 of these 23 gene families are largely fixed in the 159 samples we analyzed, having a variance < 0.15 in our population of 159 individuals.

To further develop our analysis of human lineage-specific gene duplications, we identified an additional 100 genes clustering to 30 gene families that have undergone human-specific expansions compared to all of the analyzed primate lineages (Table S11, Fig. S38) requiring the median human copy to show a differential of at least three from the nonhuman primates. Of interest, a number of genes involved specifically in human neuronal development and disease were identified among human-specific gene duplications and human expanded gene families including *HYDIN* for which we confirmed two rare deletion/duplication polymorphisms by FISH (Fig. S39), *GRPIN2* and *SRGAP2* (Fig. S40).

We compared the relative divergence of duplications underlying gene families specifically duplicated in the human lineage to those of all nonredundant segmentally duplicated genes. Both human-specific gene duplications (diploid in great apes) and genes increased in copy along the human lineage were found to have significantly higher sequence identity ($P < 7.0 \times 10^{-5}$ after multiple testing correction, Welch’s 1-tailed *t*) providing independent confirmation of the recent, human-specific origin of these gene duplications. The median percent identity of genes with >80% overlap in segmental duplications was 98.7, 97.0 and 95.1, respectively, for human-specific, human increased and all duplicated genes.

4. SINGLY UNIQUE NUCLEOTIDE (SUN) ANALYSES

Massively parallel short-read sequencing is becoming widely adopted for the discovery and typing of genetic variation. Most efforts to date have been confined to unique portions of the genome owing to an inability to uniquely map short reads originating from repeats or duplication tracts. We address this shortcoming by focusing on reads matching the small subset of divergent positions (<10%) within otherwise identical duplications. Using the underlying segmental duplication pairwise alignments from the human reference genome, we searched for paralogous sequence variants that could be used to uniquely distinguish and tag each paralog (Fig. S41). We termed these singly unique nucleotide (SUN) identifiers. We show even with short reads and modest coverage (≥ 30 bp, $< 5X$), it is possible to genotype both the individual content and copy number of paralogs within most of the highly identical segmental duplications in the human genome.

4.1. SUN discovery and genotyping

We created a database of singly unique nucleotides (SUNs) that uniquely tag a paralog within the length of short read (~30 bp). To construct this database, we first segmented the human genome (UCSC hg18/NCBI build 36.1) into its constituent 30mers (SUN kmers, or “SUNKs”) and their reverse complements ($n = 6.19 \times 10^9$). Next, we excluded all 30mers occurring more than once in the genome as these positions cannot uniquely tag a single genomic locus without additional information (e.g. longer reads or mate-pairing information). For each of the remaining 2.40×10^9 Watson-strand 30mers, we found all matches to the genome within edit distance 2 using the *mrFAST* aligner (S3) and discarded 1.40×10^8 30mers (5.83%) that, although unique, had a large number of close matches (> 500 within edit distance 1 or 2) or were comprised of highly repetitive 15mers (not shown). Such kmers are therefore at greater risk of spurious mappings from distant loci. After filtering, 2.256×10^9 SUNK positions remained throughout the genome (including 4.488×10^7 in segmental duplications).

Each cluster of SUNK positions within a duplicated tract is made unique by one or more underlying SUNs resulting from paralogous sequence variation. To catalog these, we analyzed all pairwise global sequence alignments between segmental duplications. We first consolidated the 44.8 million SUNKs within segmental duplications into 535,992 clusters spanning 68.3 Mb of the genome covered by one or more SUNKs. We then extracted intervals corresponding to each SUNK cluster from every pairwise alignment of segmental duplications in which that cluster overlapped. Within the resulting extracted fragments of pairwise alignments, we identified the underlying nucleotide differences and categorized them by class: substitutions (transitions and transversions), insertions, and deletions.

Because we sought to define per-position markers useful within the context of short reads, we counted SUNs by single-base physical coordinates within the reference genome. For example, a unique, 1-kbp insertion in one paralog relative to another would be counted as 1000 SUN positions in the former and only one in the latter, even though the underlying indel in this example may represent one discrete variant event. Nevertheless, for the purposes of short-read-based genotyping, the insertion-bearing paralog can be uniquely tagged by these 1000 inserted positions, while the cognate paralog is only tagged by the corresponding deletion junction position. This asymmetry was corrected by counting SUNs at each pairwise-aligned duplication (except in cases when one was on an unanchored chromosome, in which case only SUNs at the anchored copy were counted).

In total, we identified ~12.6 million SUN identifiers within segmental duplications. The transition/transversion (Ti/Tv) bias for these was 1.54, comparable to the overall ratio within segmental duplications as a whole. We excluded SUNs found in or near (± 36 bp) repetitive elements as identified

by RepeatMasker and TandemRepeatFinder. This step did not grossly alter the density nor makeup of the resulting SUN set. In total, there remained 4.08 million unique identifying base pairs, or 1-bp difference per 13.4 bp, which could be used to infer the specific identity of a paralogous sequence (Table S2).

We sought to quantify what fraction of the duplicated portion of the genome that could be ascertained for copy number using SUNs. As expected, the density of SUN positions diminished with increasing identity (Fig. S42). Indeed, the density of SUN positions might be expected to follow a linear relationship to the percent divergence or equivalently ($100 - \%ID$): the more highly identical a duplicate is, the fewer divergent bases and thus the fewer SUNs it contains. This relationship breaks down for some duplicon blocks, and the density of SUNs exceeds the minimum level of pairwise divergence. The following pathological example illustrates how this can arise: consider a 100-bp block that underwent nine duplications in close succession. Following these duplications, each new copy diverged at one different position. Relative to any one of these ten blocks, the nine duplicates are identical at 99/100 positions, but within all the blocks, there are by definition 10 SUNs differentiating that block from its nine duplicates. This leads to a SUN density of 10/100, which exceeds by the mean divergence by tenfold. Thus, pairwise divergence of duplicates is not always indicative of the density of markers suitable for paralog-specific genotyping.

To be useful for paralog-specific copy number genotyping, SUN markers must be in fixed association with a duplicate locus as defined by the reference assembly. Markers appearing to be SUNs may, in principle, instead represent SNPs within duplicated sequences, rare variants, or errors in the reference. In these alternative cases, a SUN's presence or absence among short reads does not necessarily reflect the copy number of its associated locus. Because we required a perfect read placement (edit distance 0) to consider a SUN present, several effects could cause a SUN to appear absent, including single-base variants affecting SUN itself or its flanks (\pm the read size, 36 bp), deletions or rearrangements eliminating the SUN or altering flanking sequence, and low sequencing coverage.

To determine the extent of these effects, we checked each SUN marker for presence or absence among reads from 12 unrelated human genomes sequenced to high coverage ($>10X$). We noted an overwhelming majority ($\sim 91\%$) of SUNs found in $\geq 11/12$ individuals (Table S12, Fig. S43), suggesting that these positions largely represent differences that can be used to tag specific paralogs. Turning to the full set of 159 individuals, we noted that $>81\%$ of SUNs were present in two-thirds or more of the genomes, with SUNs present at similar rates among each of the three populations sampled. At a more stringent threshold (presence among $>80\%$ of individual genomes), we noted a significant reduction in the SUNs' presence ($\sim 58\%$).

We noticed that SUNs dropped out of low-coverage genomes much more rapidly than in high coverage genomes (94.2% SUNs present among at least 10/12 = 83% of high coverage genomes compared with only 57.7% present among 128/159 = 81% of all genomes). Therefore, we reasoned stochastic "drop-out" due to low sequencing coverage was primarily responsible for the higher rate of SUN absence among the full set of genomes. Indeed, the fraction of SUNs present in each individual genome closely correlated with coverage — those genomes sequenced at lower coverage were overwhelmingly the ones contributing to the 'missing' SUNs (Fig. S43d). The rate of SUN drop-off appeared similar in all three populations rather than being skewed towards the African population which has relatively more rare

variation (S32), suggesting that there are relatively few allelic variants disguised as SUNs, or that their alternate alleles (i.e., not the ones represented in the reference) are quite rare.

We next modeled the effects responsible for missing SUNs by simulation. SUN markers were assumed to be polymorphic with allele frequencies drawn from a mixture with three components: (1) fraction p_{fix} of SUNs being present in all individuals (allele frequency=1), (2) fraction p_{most} of SUNs being present in most individuals (allele frequency= f) and (3) fraction $(1-p_{fix}-p_{most})$ drawn from a uniform distribution over allele frequencies $U[0,1]$. To determine the simulated genotype at each SUN, we divided each chromosome's SUN markers among these three components and for each SUN, drew two alleles (allele 0, SUN absent; allele 1, sun present) from the respective component's allele frequency distribution, except for chrX and chrY in males, for which only one allele was drawn for each SUN. This process was repeated for each sample to produce a simulated genotype of all SUNs for that sample. The depth of reads covering each SUN was then modeled separately within each sample, and was drawn from a Poisson distribution with parameter $0.5G_iC_j$, where $G_i \in \{0,1,2\}$ was the genotype of SUN i (respectively, homozygous absent, hemizygous, and homozygous present) and C_j was the mean coverage by edit-distance 0 reads in sample j . We then counted for each SUN the number of simulated samples in which that SUN was present (i.e., had simulated depth of mapping > 0), and repeated this process for all combinations: $p_{fix}=0,0.025,0.05,\dots,0.95,0.975,1.0$, $p_{most}=0,0.025,0.050,\dots,1-p_{fix}$ and $f=0,0.05,0.1,\dots,0.95,1.0$. The parameter set with lowest root mean squared deviation (RMSD) between the observed and simulated rates of autosomal SUN dropout was $(p_{fix},p_{most},f)=(0.425,0.275,0.800)$ meaning that 42.5% of SUNs were fixed, another 27.5% had allele frequency 0.8, and the remaining 30% had allele frequencies uniformly distributed between 0 and 1, for an overall mean SUN allele frequency of 0.795 (Fig. S44). Because of the multiple effects including flanking variation that can cause a SUN marker appear to be absent, this "allele frequency" actually refers to the allelic frequency at which (i). a SUN marker is present as annotated and (ii). there is no flanking variation. As such this is a lower bound on SUNs' actual allele frequencies – the alternate (non SUN) alleles may be much rarer and the extent of allelic polymorphism thus much lower. This further reinforces the notion that the primary reason for not observing a SUN is stochastic drop-out rather than polymorphism, and confirms the utility of these markers for genotyping across populations.

Relative to unique portions of the genome, segmental duplications are enriched for annotated SNPs (S33), many of which may actually be paralogous sequence variants (PSVs) in fixed association with one duplicate allele that are misannotated as polymorphisms at highly identical loci. Indeed, among the 3.28 million autosomal SUN positions found in all 12 high coverage genomes, which likely represent PSVs present at high frequency in the human lineage, we found 275,245 SUNs that were annotated as SNPs (dbSNP v130, quality filtered from UCSC Genome Browser). This was a highly significant enrichment for overlap between these fixed SUN positions and annotated SNPs, even beyond the overall enrichment for SNPs within segmental duplications. Some of these SUNs may indeed be SNPs with low minor allele frequency (MAF) – high-coverage sequencing of additional individuals will be required to resolve which of these are fixed SUNs devoid of common allelic variation and which are low-frequency SNPs.

Having characterized these SUN markers for presence or absence across 159 genomes, we next applied them in a quantitative fashion for paralog-specific copy number prediction. As for total copy number prediction, we estimated paralog-specific copy number (psCN) by the depth of short-read mapping, taking into consideration only reads hitting paralog-specific markers (specifically, reads starting at SUN

k-mer (SUNK) positions). Due to coverage effects and the scarcity of these markers within highly identical duplicates, we measured the average number of reads across all markers within each duplication interval. In order to obtain a psCN estimate, we performed linear regression using a model trained on regions of known copy (as described above for total copy number). For paralog-specific genotyping, we counted only reads mapping to SUNK positions with zero edits. This should exclude spurious mappings from paralogous copies at the expense of losing a minority of reads with one- or two-base sequencing errors or nearby SNPs. The paralog-specific copy number linear model was fit using only edit distance 0 placements over the same regions. One important difference between this and our total copy number estimation is that while for the latter we used depth-of-coverage (i.e., the number of reads overlapping each given position), for psCN estimation we considered instead the number of reads *starting* at each position. The reason for this difference is that reads starting at non-unique positions can “bleed in,” adding coverage to unique positions and skewing psCN estimates. Using reads restricted to the start positions of SUN kmers alleviates this problem.

4.2. Validation of paralog-specific copy number estimation

Comparison with high-confidence, fosmid-mapped deletions

As one means of validating our paralog-specific copy number (psCN) genotyping, we computed psCN estimates for 383 high-confidence deletion intervals previously identified by cloning and resequencing. These deletion intervals were initially identified within eight individuals (NA18517, NA12156, NA19129, NA18956, NA18555, NA12878, NA19240, and NA18507) by discordant fosmid clone end sequence pair mapping and were resolved to the deletion intervals at the single-base level by full capillary-based resequencing of the respective clone inserts (*S9*). GenBank accessions for the deletion breakpoint sequences are listed in Table S14.

Treating these deletion calls as a gold standard, we used short-read depth at SUNK markers to obtain a psCN genotype of each deletion interval in the same individual where it was discovered. As a control, we also computed the number of concordantly mapping fosmid clone end sequences within each deletion—true homozygous deletions tend to have zero clone ends mapping within the deletion interval. For a hemizygous deletion, there remains one copy of the deleted sequence in the genome; thus, except for very short deletions, the deletion interval tends to contain at least one concordantly mapped clone end originating from the non-deleted allele. We first examined the distribution of psCN genotypes within each deletion interval. We classified deletions with psCN less than 0.5 as homozygous, between 0.5 and 1.5 as hemizygous, and above 1.5 as diploid or greater.

We treated deletions in unique portions of the genome ($n = 310$) separately from those overlapping segmental duplications ($n = 73$). The vast majority of deletions within unique regions were correctly predicted (psCN < 1.5, 301/310, 97.1%, Fig. S45a). We called 81 (26.1%) as homozygous deleted. Of these, 65 (80.2%) and 78 (96.3%) were overlapped by 0 or ≤ 1 concordant clone end, respectively, confirming that we effectively discriminated between homozygous and hemizygous state. We called 220 (71.0%) deletions as hemizygous. Of these, a smaller proportion had 0 or ≤ 1 concordant clone ends as compared with the predicted homozygous deletions: 38 (17.2%) and 98 (44.5%), respectively. The remaining 9 deletion intervals (2.9%) were predicted as diploid or greater, reflecting errors in psCN genotyping (these 9 regions were on average shorter, with a median unmasked length of 605 bp compared to the 1567 bp for deletion intervals called with psCN < 1.5).

Among deletions overlapping segmental duplications, the majority of intervals were predicted as significantly less than diploid in paralog-specific copy number (psCN<1.5 in 59/73, 80.8%, Fig. S45b). The fact that such a large fraction was correctly predicted despite being extensively duplicated (mean of 7.6X per deletion interval) is strong confirmation of the specificity of psCN genotyping. We genotyped 22 (30.1%) as homozygous deleted and 37 (50.7%) as heterozygous deleted. Again, a majority of predicted homozygous deletions (psCN<0.5) had either 0 or <=1 concordantly mapped fosmid end sequences (15/22 = 68.2% and 18/22 = 81.8%, respectively). There remained a larger fraction (18.2%) of called homozygous deletions containing >1 fosmid clone end, which may represent clone end sequences from distant paralogs spuriously mapping in the deletion interval. We predicted diploid or greater paralog-specific copy number for 14 (19.2%) of the deletion intervals, of which some are embedded within flanking amplification events.

To further evaluate cases where we predicted a diploid (or greater) state for the deletion interval, we computed psCN at 10 kb flanking both sides of each deletion interval and compared that with the psCN estimate within the deletion interval (Fig. S46). If the deletion is accurately genotyped, we expected the estimate within the interval to be less than that of its flanks. We observed this to be so in the majority of cases: 308/310 (99.4%) for deletions outside of segmental duplications, and 68/73 (93.2%) for those inside of segmental duplications. Of the 14 segmental duplication-associated deletions with psCN>1.5, almost two-thirds (9/14, 64.3%) had lower predicted copy inside the deletion interval relative to their flanks, affirming the accuracy of these genotypes even when nested within highly duplicated regions such as upstream of the *NOTCH2NL* gene on 1q21. Such cases may arise when the reference genome is “missing” copies of a duplicated region (see above).

Experimental validation

Amylase

We compared predicted psCN genotypes at the amylase locus (Fig. S53) with single-channel array CGH copy number estimates (S6) and previously reported quantitative qPCR at this locus (S34). We summed our psCN measurements at *AMY1A*, *AMY1B*, and *AMY1C* and observed similarly strong correlations with estimates from the arrays ($r = 0.714$, Fig. S47) and from qPCR ($r = 0.713$, Fig. S48).

APOBEC

We compared our psCN genotyping with PCR-based genotypes at the *APOBEC3B* gene, which is commonly deleted among East Asian and Oceanic populations (S35). Treating these PCR-based calls as a gold standard, we achieved 86.8% genotyping accuracy this locus (Fig. S49).

CFHR

We validated our psCN genotyping at *CFHR3* using a quantitative PCR assay specific to this paralog. As with our other comparisons to qPCR data, we observed strong concordance ($r > 0.9$, Fig. S50). qPCR confirmed the population stratification predicted by read depth psCN analysis. This deletion shows a higher allele frequency among YRI.

4.3. Paralogous regions of genomic variation

Our SUN-based genotyping allows us to delineate specifically which copy of a polymorphic paralogous region is variable. Analyses of these regions uninformed by SUNs will spuriously detect polymorphisms that we term *mirror* copy number variants—those where the location of the true polymorphism cannot be distinguished due to homology (Fig S51). We re-genotyped 406 of the 952 large (>50 kbp) CNV

events discovered by dCGH that overlapped segmental duplications, estimating their paralog-specific copy number. We determined that 60% (245/406) of these regions showed no signature of variation using SUNs and hence represent mirror effects from cross-mapping between highly identical duplicated sequences.

To assess the extent of mirroring, we estimated total copy number and SUN-based psCN genotypes for 46,937 CNVs reported by a recent array CGH study (S5). We compared these genotypes among the 36 test individuals shared between that study and the current work and additionally took into account our estimates of total and paralog-specific copy number for the reference genome used in that study (NA10851). We focused on events within segmental duplications (>70% overlap) with a sufficient number of markers to allow paralog-specific genotyping (>400 SUNKs). Among these, we found that total copy number genotype concordance was 95% (4181/4412) between read depth estimates and array CGH. Strikingly, SUN-based paralog-specific genotyping concordance was only 62% (2734/4412). In contrast, concordance for regions outside of segmental duplications was essentially identical (read depth-based concordance 88%, 9235/10470; SUN concordance 91%, 9526/10470). Increasing the minimum number of required SUNKs to genotype a region to 2000 only increased the concordance to 61.5% (1820/2960), strongly suggesting that the decrease in concordance from read depth-based genotypes is not due to a lack of power to genotype paralog-specific copy number. We find that 35.1% (1547/4412) of calls on an individual level are confirmed copy number variable by read depth-based CNV genotyping. However, these events show no paralog-specific variation in copy and, hence, have been assigned incorrectly as result of sequence homology. To specifically tag the location of *mirrored* CNVs we manually curated 191 of the largest (>100 kb) CNVRs discovered using our dCGH approach. In 42 of these CNVRs we identified likely CNV *mirrors* within segmental duplications. For 28 of these *mirrors* we were able to resolve the specific true location of the CNV to an alternate paralogous region of the genome (Fig S51 and see Fig. 4B of Main Text). Overlapping regions showing likely mirror effects were consolidated and are listed in Table S13.

In addition to correctly defining loci of copy number variation, the method we developed will facilitate refinement of breakpoints within high-identity duplications. For example, as part of our analysis we observed a large deletion mapping between BP1 and BP2 of the Prader-Willi region on chromosome 15q11.2 in 2/159 individuals. This deletion is mediated by tandem clusters of segmental duplications separated by ~700 kbp of unique sequence including the genes *CYFIP1*, *NIPAI*, and *NIPAI2* (Fig. S70). Deletion of these genes is implicated in schizophrenia by GWAS (S36) and as a low-penetrance risk factor for developmental delay (S37, 38). These previous studies utilized SNP microarrays and array CGH thus only detecting the unique portion of the deletion. Based on SUN read depth, we determined that this deletion extends at least ~95 kbp beyond its previously assessed boundaries, encompassing at least three additional genes. We validated the unique portion of this deletion by FISH analysis and array CGH. We note that the SUN-based breakpoints showed Mendelian consistency between the two related individuals in which we observed the deletion. We also observed the reciprocal gain in this region in one individual (GM18555 CN = 3, validated by FISH). Within this background of amplification, we noticed a small diploid patch indicative of a nested deletion, which we validated by fosmid resequencing (see below).

4.4. Gene family analyses

In section 3, we used short-read-based absolute copy number estimation to discover gene families showing extreme variability, human-specific expansion, and population stratification in total copy.

Exploring the dynamics of the individual paralogs making up these duplicated families has been largely infeasible to date. To address this, we applied our ~4 million SUN markers to genotype specific paralogs among these highly duplicated genes. We focused on those identified by our total copy number analyses as showing extreme stratification or variability, in some cases allowing us to resolve the variability to specific paralogs (e.g., *NBPF1*) to the exclusion of others that showed near total fixation in the population (e.g., *NBPF7*).

We started with a global survey of variability among duplicated genes to define the complement of genes, which although segmentally duplicated, showed little variability in the 159 genomes sampled. We classified these invariant copy number genes as fixed within the human lineage. Genes within segmental duplications are often redundantly annotated—i.e., with one physical locus being annotated as containing numerous paralogous gene models. To avoid overcounting duplicated genes, we initially clustered gene models. We obtained from the UCSC Genome Browser 32,571 RefSeq gene models aligned to the hg18 reference assembly. We consolidated gene models sharing more than 50% reciprocal overlap into 22,038 clusters to reduce redundancy from different splice forms and multiple mappings of paralogous genes. Gene model clusters shorter than 50 kb were padded to 50 kb in order to utilize flanking paralog-specific tags for genotyping. Of these, 1240 overlapped with segmental duplications by at least half their length and, of those, we discarded clusters having <100 SUNK markers as well as those on chromosomes X and Y. We estimated the psCN for each of the remaining 990 gene model clusters across all 159 individuals sampled, taking each cluster as a single window and taking the mean signal across each cluster. For each gene we found the median psCN across all 159 individuals and, as a measure of variability, computed the fraction of individuals showing a deviation of >1 copy in psCN from the per-gene median. Approximately half (51%) of the duplicated genes showed common variation in psCN. For the purpose of this study, we defined common copy number variation where >5% of individuals deviated from the median copy number.

We compared this variability with that of 7834 clusters of gene models that do not overlap segmental duplications or other regions of defined CNV. As expected, the duplicated genes were overwhelmingly more copy number variable than their unique counterparts, which showed only slight deviation from median psCN (<0.1% of these genes by the above metric). We reasoned that the greater density of genotyping markers in the unique genes might account for this difference—i.e. most sequence within a unique gene will be informative. To model this effect, we randomly rejected 80% of the markers in unique genes to mimic the distribution of marker densities among duplicated genes. Even after controlling for marker density, the estimated variability of these unique genes remained low (<3.6% of genes showing deviation among >5% of individuals).

Next, we focused on the analysis of specific duplicated gene families (Fig. S52-S60). We were able to fully ascertain certain gene families (e.g., *LILR*, Fig. S52) as all paralogs are accounted for within the reference assembly and had sufficient density of SUN markers. For others, such as the highly expanded *NBPF* family (S39), several paralogs are not represented in the reference assembly and could not be accounted for by psCN analysis. We also excluded paralogs with spurious alignment to the reference (e.g., misaligned, crossing distinct duplication blocks and/or assembly gaps), along with those insufficient SUN markers (<200 SUN kmers) for reliable genotyping.

As one means of validation, we compared total copy number estimates for each of several gene families with the summed psCN estimates of the constituent paralogous. For example, as shown in the case of the

duplicates *ESPN* and *ESPNP* (Fig. S59), we observed strong correlation between the sum of individual paralog psCN estimates and the total copy number estimate for the corresponding gene family ($R > 0.8$ in all, $R > 0.9$ in 4/6). Additionally, these showed little bias in copy number (i.e., best-fit line had y-intercept ~ 0 , as shown for the duplicated genes *ESPN*) indicating that our psCN estimates at these loci are both accurate and precise.

CFHR

In addition to the two common deletion breakpoints at the *CFH* locus, we also noted two rarer, alternate copy number states in one individual each: single-copy gain of *CFHR1/4* and a deletion of *CFHR1/4* with novel breakpoints (Fig. S60). The former appears to be the reciprocal of the common *del2* deletion of *CFHR1/4*. The latter is shifted towards the centromere relative to the breakpoints of *del2* and, as with *del2*, appears to arise from homologous loci in two segmental duplications, suggesting that non-allelic homologous recombination (NAHR) mediates this event as it is presumed to for the more common deletions at this locus (S40).

LILR

The family of duplicated *LILR* family is found by total copy number genotyping to be highly population-stratified, with decreased copy number in Asians and vastly increased diversity in Africans (Fig. S52). Individual paralogs are functionally distinct, with diverse ligand binding characteristics and alternatively stimulatory or inhibitory effects (S29). We find the variability within this gene family to be specific to 5 of the 12 duplicated genes. *LILRA6* showed the highest variability, contributing to the bulk of the increased African diversity. We confirmed the deletion of *LILRA3* at high frequency among Asians (S30) to the near exclusion of other groups sampled. The remaining 8 *LILR* copies are diploid in nearly all individuals.

NBPF

The *NBPF* genes have undergone dramatic expansion in the human lineage. These genes exhibit neuron-specific expression and signatures of positive selection (S41). Analyzing the individual NBPF paralogs with sufficient SUN marker coverage (Fig S54), we found that of several paralogs (*NBPF3*, *NBPF4*, *NBPF7*, *NBPF15*, and *NBPF22P*) were diploid across most individuals with little variation. In contrast, *NBPF1* was both highly amplified and highly variable (median psCN=8.04, +/- median absolute deviation [MAD]=1.68 copies), with increased copy number and diversity among Africans (African median: 9.55 copies, non-African: 7.62; $P < 2.48 \times 10^{-12}$, Mann-Whitney U). The paralog *NBPF14* showed even stronger population stratification, with a common deletion strikingly enriched among Africans (Africans: 38/56=67.9% with psCN<2; non-Africans: 13/103=12.6%; OR=14.3, $P < 2.23 \times 10^{-12}$).

Amylase

Our analysis also reveals previously uncharacterized differences among well-studied gene families. For example, at the amylase gene cluster, thought to be important for human adaptation to starch-rich diets (S34), we find that the gene *AMY1C* is largely copy-invariant while the salivary amylase genes *AMY1A* and *AMY1B* are commonly subject to deletion. 23% of humans are diploid for both copies, while $\sim 20\%$ are homozygously deleted for either of these two genes (Fig. S53). Pancreatic amylase genes (*AMY2A* and *AMY2B*) by contrast show less variability; 7% of individuals carry 3-5 copies of *AMY2A*. *AMY2B* shows more variability in copy number, with 1 individual predicted to be homozygously deleted for this gene, a potential candidate for congenital pancreatic deficiency (S42). By comparison, the *AMYPI*

pseudogene shows the widest range in copy number with the highest copy number observed among Asians.

4.5. Validation of novel structural variants within segmental duplications

The specificity afforded by our SUN marker approach allowed us to discover novel structural variants within segmental duplications. We performed a genome-wide search for deletions within segmental duplications, selecting regions with psCN ≤ 1.5 in at least 2 of 159 individuals for ≥ 20 of 30 consecutive windows of 25-200 SUNs (maximum physical window size of 1000 bp). We identified 6712 putative deletions, and selected 31 of them for further validation.

Large-scale validation of CNVs within duplicated regions is challenging because available platforms (FISH, microarrays, qPCR, etc.) are hampered by poor genotyping specificity within such regions. We instead developed a clone-based validation approach. We leveraged fosmid clone libraries from eight individuals, which we had previously applied to discover structural variants (S9, 43). Based on our psCN predictions, we selected 144 clones spanning 31 deletion intervals from fosmid libraries derived from the same individuals in which each event was called (Table S15). We re-sequenced the selected clones by pooled, shotgun short-read sequencing on the Illumina platform and mapped the resulting reads back to the same SUN positions as originally used to infer the deletion. We considered these clone-based data to support our deletion calls when we observed hits to SUN positions flanking a region (as defined by its capillary end sequence mapped location) with little or no coverage corresponding to the predicted deletion interval. In some cases, we observed SUN hits only from one end of the insert; these nevertheless supported the deletion call when the observed coverage from the fosmid clone was significantly shorter than its physical size (i.e., 32 kbp or less).

Short-read sequence data from the selected clone inserts yielded interpretable patterns of SUN hits at 18 loci (Figs. S61-S70). These hits validate the specificity of our approach within these highly complex regions and confirm that these events predicted using whole-genome sequencing data arose at the indicated loci rather than as signals spuriously mirroring other duplicates. In addition, this approach confirmed that most of the SUN markers indeed specifically tag one paralog, as predicted. For instance, we confirmed a ~32-kbp deletion on 4p16.1 encompassing the 3' end of the gene *DEFB131* (homozygous in 4/8 individuals, hemizygous in 1/8) despite the presence of highly identical duplicates at >50 other loci (Fig. S61). In another example, we validated a hemizygous deletion in the individual NA18956 at the heavily duplicated *PSG* gene cluster, which encodes glycoproteins with known immunomodulatory roles during pregnancy (S44) (Fig. S62). Some of selected loci (e.g., *SMN*) were too sparsely populated with SUN markers to unambiguously interpret. In a minority of cases, we observed SUNs mapping outside of the clone's predicted location; these may represent errors in resequencing, polymorphisms relative to the reference, errors in the reference, or in a number of cases discussed below, interlocus gene conversion. A summary of the results of the clone insert sequencing is provided in Table S15.

In several cases, clone sequencing revealed additional complexities or potential artifacts. Among some psCN deletion loci, for example, the corresponding clones were initially classified by end-sequence mapping as inversions. We hypothesized that such inversions might be 'masking' an underlying deletion or a more complex rearrangement. For most of these, the SUN pattern confirmed the original inversion call but was uninformative regarding the putative underlying deletion (not shown). These deletions may be masked by the surrounding inversion, or they may represent false positives among our calls in which

polymorphisms prevented the corresponding reads from mapping to SUN positions with edit distance 0, creating patches without coverage and the false impression of a deletion. We noticed several other cases in which the majority of the SUNK hits mapped between the clone ends but were accompanied by patches of hits to paralogous sequences mapping outside the clone insert. We interpreted these either as errors in the reference (misassignment of sequence between highly identical paralogs) or as polymorphisms acting to convert unique markers from one paralog to those from another. A subset of these may represent interlocus gene conversion relative to the reference sequence, which we discuss further below.

Nextera Illumina sequencing of fosmid clones

We identified by their end sequence mapping locations (<http://hgsv.washington.edu/>) a total of 144 fosmid clones of shotgun genomic DNA from the following eight individuals: NA18517, NA18507, NA18956, NA19240, NA18555, NA12878, NA19129, and NA12156. Colonies were picked from freezer stocks, grown overnight at LB broth at 37°C, harvested, and clone DNA isolated by alkaline lysis as previously described (*S45*) with minor modifications for use in a 96-well format. After purification, clone DNA was arrayed on a single 96-well plate, in some cases combining two clones from unrelated loci. A sequencing library was created separately from each well using the Nextera Illumina-compatible DNA Sample Prep Kit (Epicentre, Madison, WI) as directed by the manufacturer, with ~100 ng of fosmid clone DNA per well as starting material. The subsequent PCR step was used to add a barcode tag specific to each well. Finally, the 96 resulting barcode-tagged libraries were pooled and sequenced with two lanes of paired-end, 76-bp reads on an Illumina Genome Analyzer IIx with an additional 9-bp index read to recover the barcode sequence. The resulting sequences were partitioned by well of origin corresponding to a single clone or two-clone pool on the basis of the barcode sequence, allowing for up to 2 edits to the nearest matching barcode. Reads from each clone or clone pair were then mapped to the genome using *mrsFAST* as described above. Lastly, we filtered the resulting read placements to perfect hits at SUNK positions to indicate the paralog-specific markers captured by each resequenced clone insert.

4.6. Gene conversion analysis

We observed patterns consistent with interlocus gene conversion among the paralog-specific read depths at several loci throughout the genome. Gene conversion arises from the non-reciprocal genetic exchange of information between two loci or two alleles. Conversion within the human genome is known to act preferentially upon segmental duplications as its activity is strongly dependent on the proximity and homology of acceptor and donor loci [reviewed in (*S46*)]. Previously characterized pathological gene conversion events in humans tend to be relatively short (<3 kbp per event), and we reasoned that we could detect evidence for these events by searching for small patches where the dropout of paralog-specific SUN signatures was accompanied by cognate patches of elevated copy at the corresponding paralogous position.

We noted several distinct patches following this pattern particularly among tandem segmental duplications (Fig. S71). One such 6-kbp patch appeared recurrently over the second exon of each of the duplicated genes *RHD* and *RHCE*—a known site of clinically relevant recurrent gene conversion (*S47-49*). Within our genome-wide psCN maps, this manifested as a localized loss in copy over exon 2 of *RHCE* accompanied by a gain over exon 2 of *RHD*. We interpret this as a signature of conversion acting to replace the paralog-specific markers at *RHCE* exon 2 with those from the paralogous exon and flanking intron of *RHD*.

To experimentally confirm this interpretation, we sequenced fosmid clone inserts spanning the conversion patch from one individual predicted to be homozygous for $RHCE \rightarrow RHD$ exon 2 conversion (NA18555) and one predicted to lack this conversion haplotype (NA19240). Because these clones are physically constrained to ~40 kbp and most (2/3) mapped unambiguously by capillary end-sequencing to $RHCE$, we considered reads mapping to SUN markers within the reference RHD sequence as evidence for conversion where RHD serves as a donor and $RHCE$ as an acceptor. Mapping sequence reads from these clone inserts back to SUN locations, we observed that the majority of SUNs corresponded to the $RHCE$ locus, confirming that all three clones mapped to $RHCE$. This finding was in agreement with the clone end sequence locations in two of three clone inserts and corrected the map location of the third, (ABC11_48195000_H6) for which the forward end capillary sequence had mapped one end sequence into the putative acceptor patch, spuriously creating an end sequence signature of an inversion. In both clones derived from individuals predicted to have an $RHCE \rightarrow RHD$ conversion, we observed a lack of SUNs precisely at the conversion acceptor patch accompanied by an isolated ‘island’ of SUN hits at the conversion donor patch. This pattern was not observed among orthologous clones from an individual where no gene conversion was detected. These findings confirm the gene conversion event and demonstrate that it is a polymorphism segregating within the population along with other paralog-specific variants such as the European-enriched RHD deletion, which we detect and confirm (*S50*).

We noted several other duplicated genes showing similar signatures, although these were often difficult to discriminate from background noise due to the low coverage of the genomes and the length of the interval. To discover other possible instances of recent inter-locus gene conversion, as well as to determine the genome-wide extent of its effect among segmental duplications, we developed a statistic to quantify this signature across all 159 genomes sampled. We evaluated this statistic individually for each annotated segmental duplication within the genome. We first split segmental duplications into windows of length <1 kb, discarding windows with <50 SUN kmers (SUNKs) and limiting each window to ≤ 400 SUNKs. We identified each window’s cognate, paralogous window by projecting it through the Needleman-Wunsch pairwise global alignment for the corresponding segmental duplication record. Next, we estimated the psCN of each cognate pair of windows as before, except that for this analysis we discarded SUN markers that did not align within a cognate pair of windows (e.g., in the case of an insertion in one duplication relative to its paralog). We quantified the relative shift in psCN of each window relative to the surrounding duplication by calculating its difference from the duplication-wide average. This was quantified for each window j of each duplication i in each sample s as:

$$a'_{sij} = \frac{a_{sij} - a_{si\cdot}}{\frac{1}{N_i} \sum_k (a_{sik} - \overline{a_{\cdot ij}})^2} \text{ where } \overline{a_{\cdot ij}} \text{ was the mean psCN across all samples for that particular window,}$$

window j belonging to duplication i ; $\overline{a_{si\cdot}}$ was the mean psCN of all windows making up duplication i in sample s ; and N_i was the number of windows in duplication i . These were calculated analogously for the other aligned duplication as b'_{sij} . We discarded short duplications as well as very highly identical ones with insufficient SUN markers (specifically, those with <2 windows meeting the SUNK criteria above). Having computed relative shifts versus a per-sample baseline across each segmental duplication, we measured whether there was a negative correlation versus random expectation. In other words, we tested whether the increase in SUN depth for one duplicated locus was significantly associated with reduction

in SUN read depth at the cognate positions in its paralog for the same sample. We quantified the degree of such correlation in the human population by counting the number of genomes with a positive shift at that window accompanied by a negative shift at the cognate window, and vice versa. For each window, we computed c_{ij} = the fraction of samples s with $a_{sij} > 0$ and $b_{sij} < 0$ or $b_{sij} > 0$ and $a_{sij} < 0$. To obtain a background distribution for c_{ij} , we shuffled the psCN values among the windows within each duplicate locus and individual (i.e., not shuffling data across individuals). We computed the degree of excess observed in this statistic at each window versus 1000 shuffled replicates. We standardized our empirically determined c_{ij} using the mean and standard deviation from the 1000 shuffled replicates, clipping this score to zero to remove negative values. Finally, we ranked each segmental duplication by taking the mean across all windows making up that duplication of the square per-window standardized c_{ij} values.

The analysis yielded 7862 segmental duplications ranked by the degree of enrichment of this signature relative to controls where the psCN values at each locus had been scrambled within each sample. The segmental duplication encompassing the Rh blood group locus scored very highly by this measure (score = 7.14, rank = 78/7864). We examined other top scoring loci by this measure and identified several that appear to show a similar signal (Table S16). Many of these mapped to duplicated genes as well as other sites of interlocus gene conversion suggesting potentially widespread functional impact of this signature.

We sought to determine whether this signature was widespread among segmental duplications, or if it instead reflected only a small number of cases. To do this, we recomputed the per-duplication scores as before, after having randomly permuted the sample labels (but keeping the data within each segmental duplication intact). We observed a highly significant excess of conversion signature in the top half of duplications by this measure relative to the signature among the same regions with randomly permuted sample labels ($P < 2.2 \times 10^{-16}$; when instead comparing signature of real vs. controls for top half of regions among permuted control, $P = 0.60$; one-sided Kolmogorov-Smirnov; Fig. S72).

Finally, we focused on different classes of segmental duplications to determine if the extent of this conversion signature was uniform throughout the duplicated portion of the genome. We observed a very striking enrichment for this signature among tandem (<1 Mbp, center-to-center) relative to distant or interchromosomal duplications, in accord with most known examples of gene conversion both in humans (*S51*) and mice (*S52*) (Fig. S73). Dividing segmental duplications by their level of sequence identity, we found this signature was nearly exclusive to duplications of >95% identity, in close agreement with known examples (*S51*) and that the gene conversion signals became more significant at higher levels of identity. Although beyond the scope of this study, the method we developed and the sites we identified should be valuable for future studies of gene conversion as more individual genomes (particularly from trios) become sequenced to high coverage.

SUPPORTING FIGURES

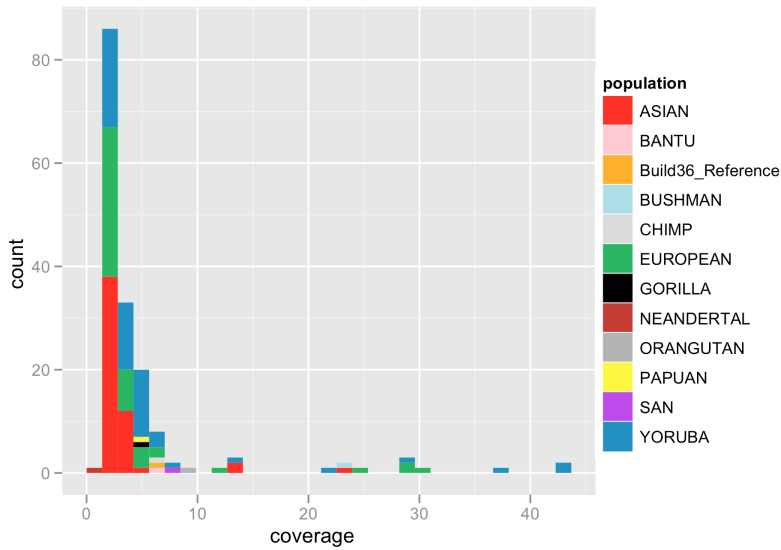


Figure S1. The number of genomes from each population as a function of sequence coverage.

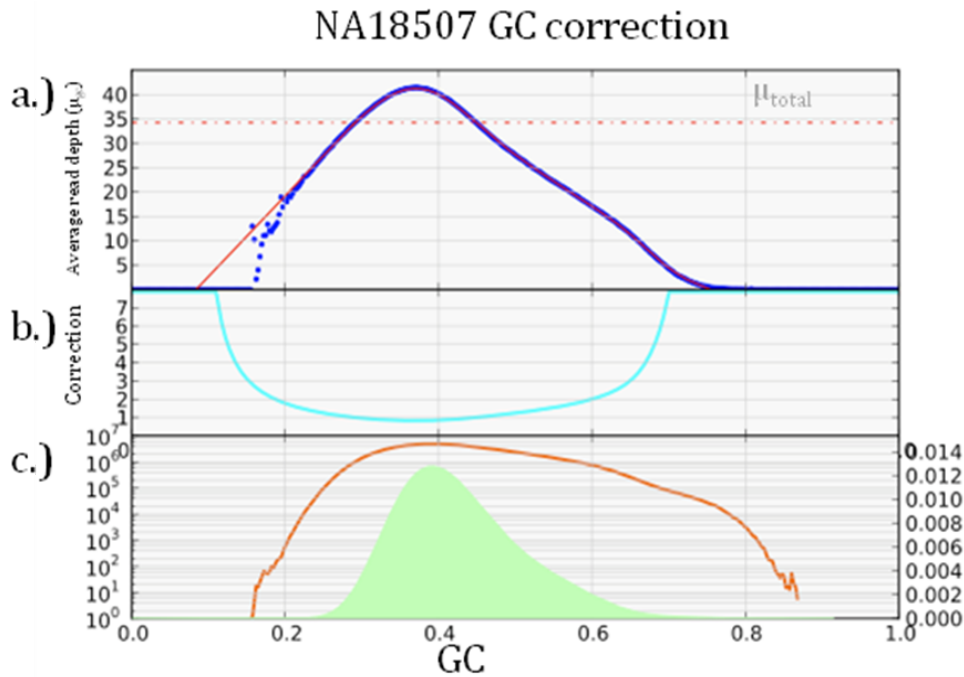


Figure S2. Multiplicative G+C correction for a high coverage genome NA18507. (a) Average read depth for specific G+C% is shown at blue points and fitted with orange line. Total average read depth (μ_{total}) is shown as a red dotted line. (b) The correction factor k_{gc} is shown (truncated at 8) for each individual %(G+C) bin. (c) The fraction of the genome at specific G+C% is shown as a histogram, in green, and overlaid in orange is the log transform of this histogram.

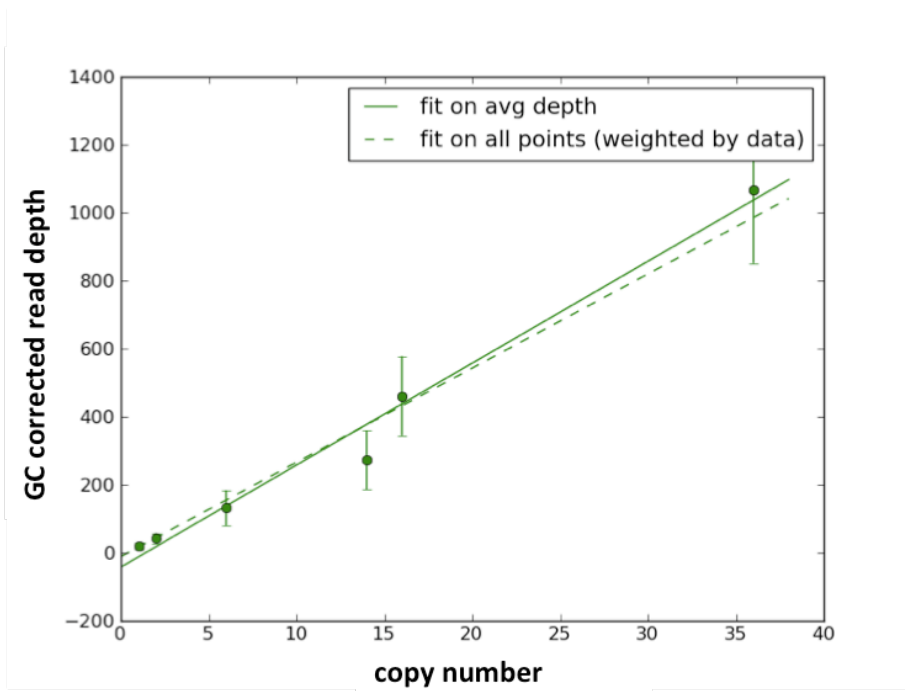


Figure S3. (G+C)-corrected read depth and copy number are linearly correlated.

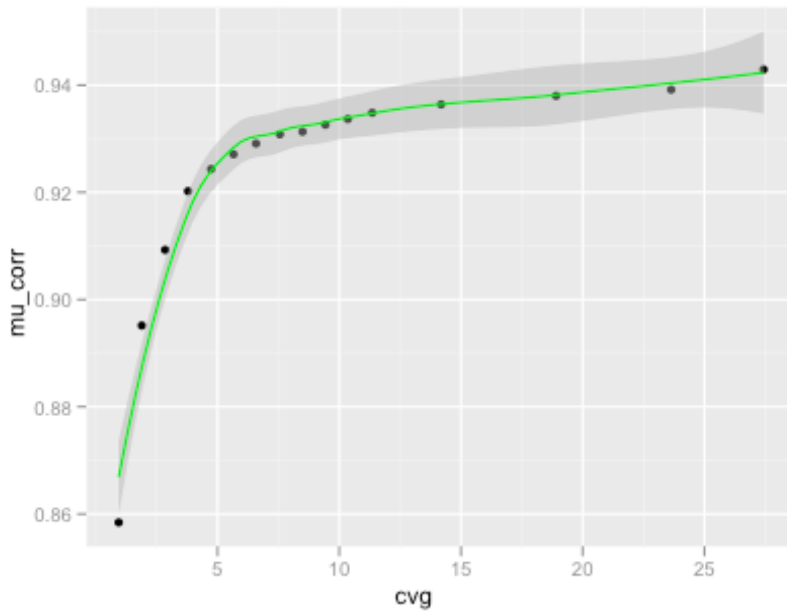


Figure S4. Correlation (μ_{corr}) of read depth to copy in control regions subsampled at different coverages (genome NA18507).

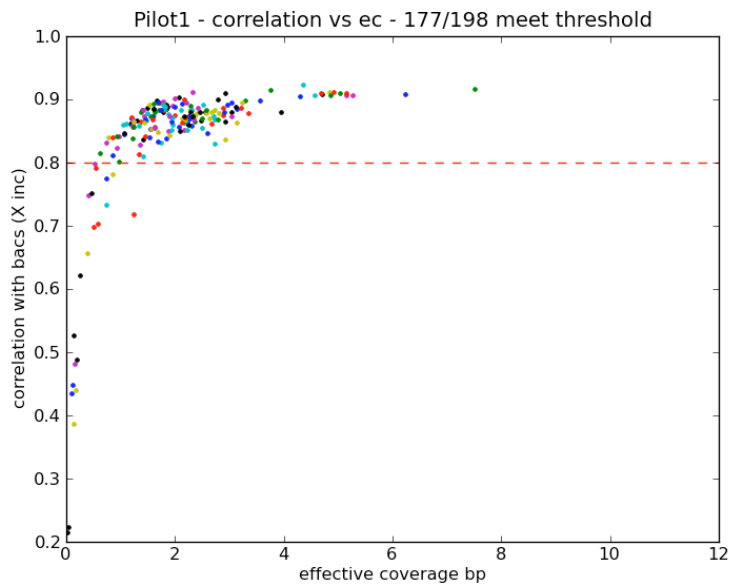


Figure S5. Correlation of read depth to copy in regions of known copy for 1000 Genomes *Pilot 1* libraries (n = 198) colored by genome, plotted as a function of effective coverage.

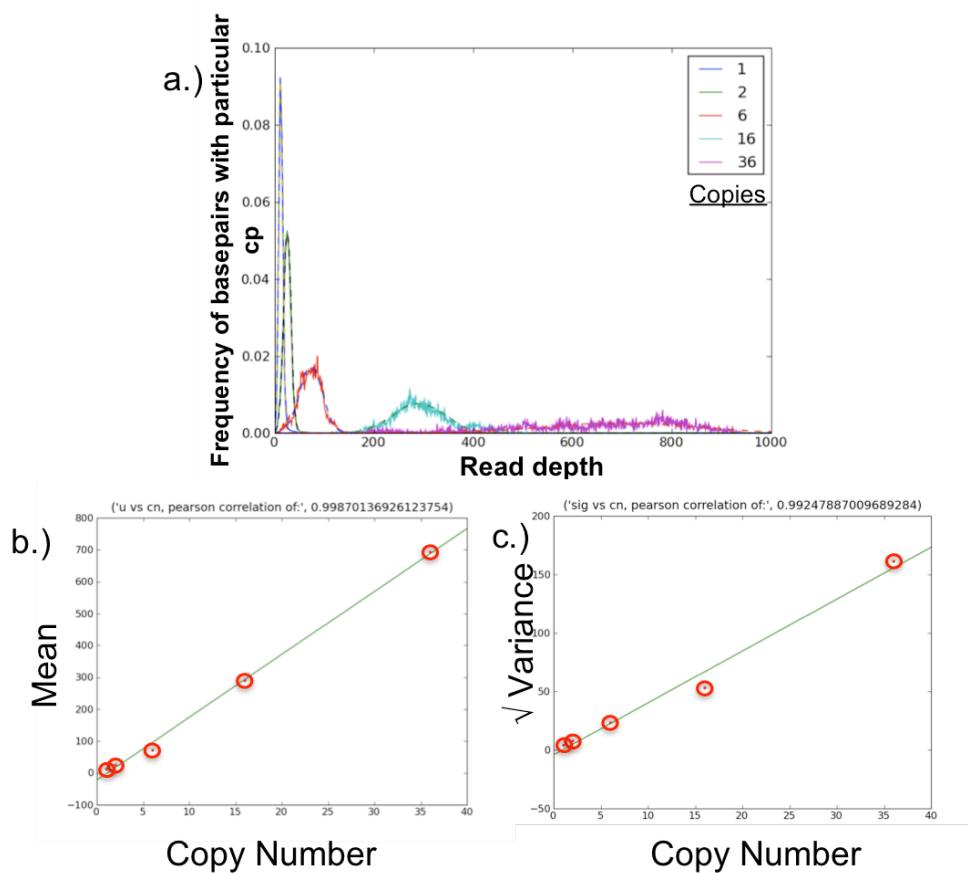


Figure S6. a) Histograms of read depth for regions of known copy with fit Gaussians overlaid. b) Mean of fit Gaussians versus copy number of known regions. c) Square-root of variance versus copy number of known

regions.

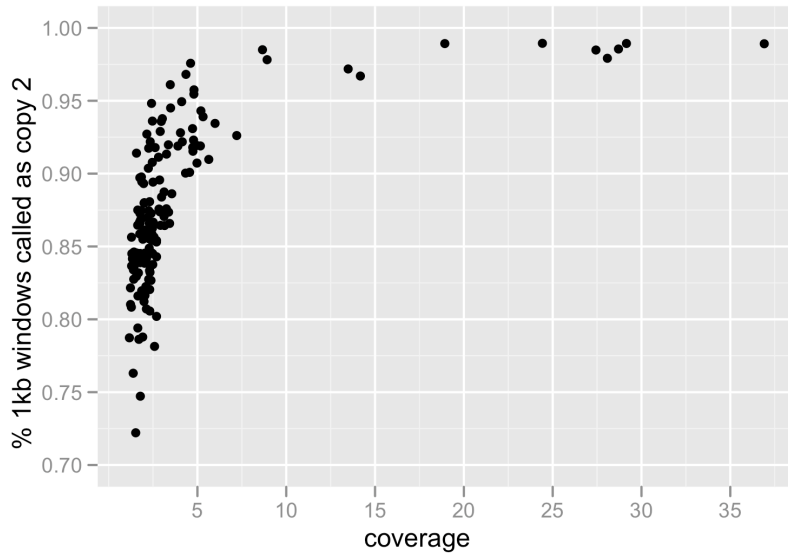


Figure S7. Percentage of 1-kb windows classified as diploid in invariant regions of the genome as a function of varying coverage.

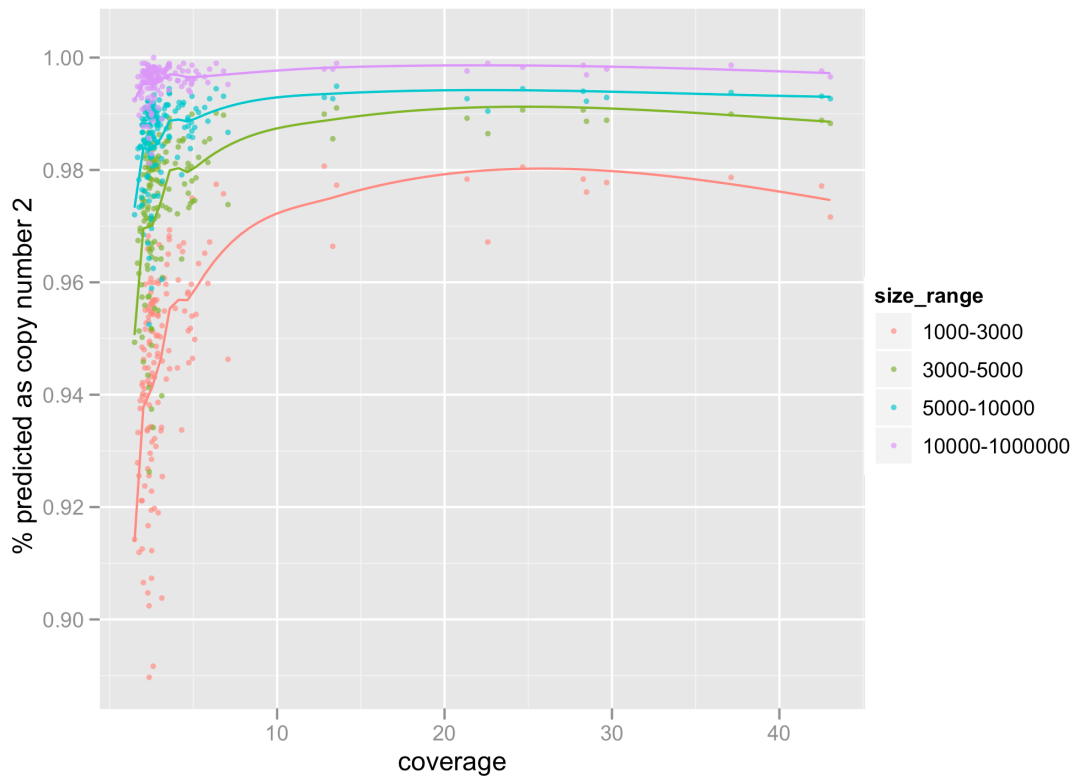


Figure S8. False discovery rate (FDR) as a function of coverage and size. At 3-kb resolution the FDR is <5% for genomes at >1.5X sequencing coverage.

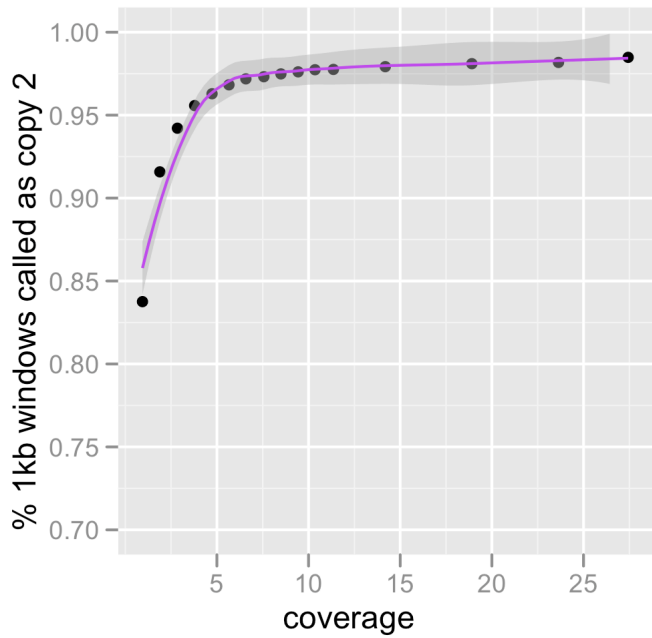


Figure S9. False discovery rate (FDR) estimated based on 1-kb diploid invariant regions from one genome (NA18507) subsampled at various coverages.

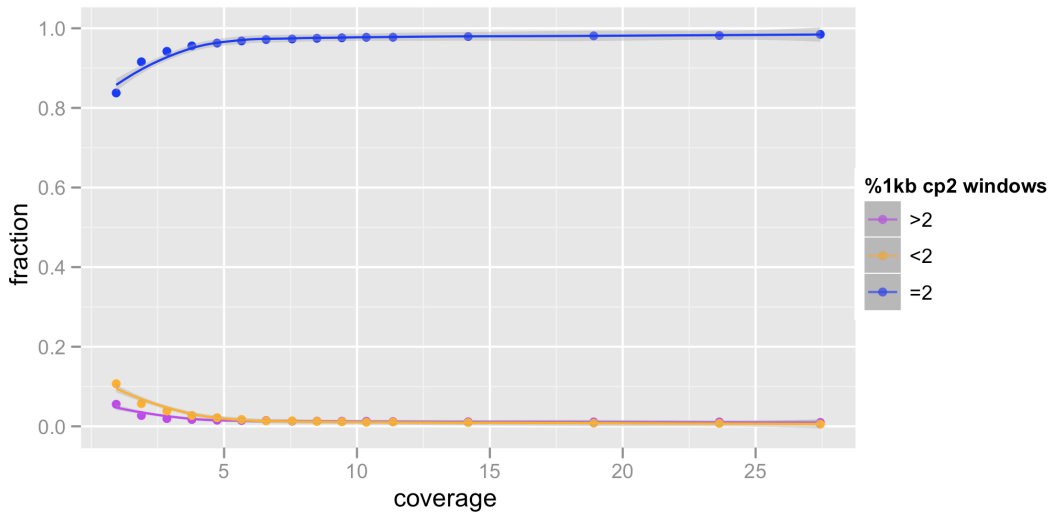


Figure S10. Fraction of 1-kb windows detected as copy number 2, less than 2, or greater than 2 as a function of coverage for one genome (NA18507) subsampled at various coverages.

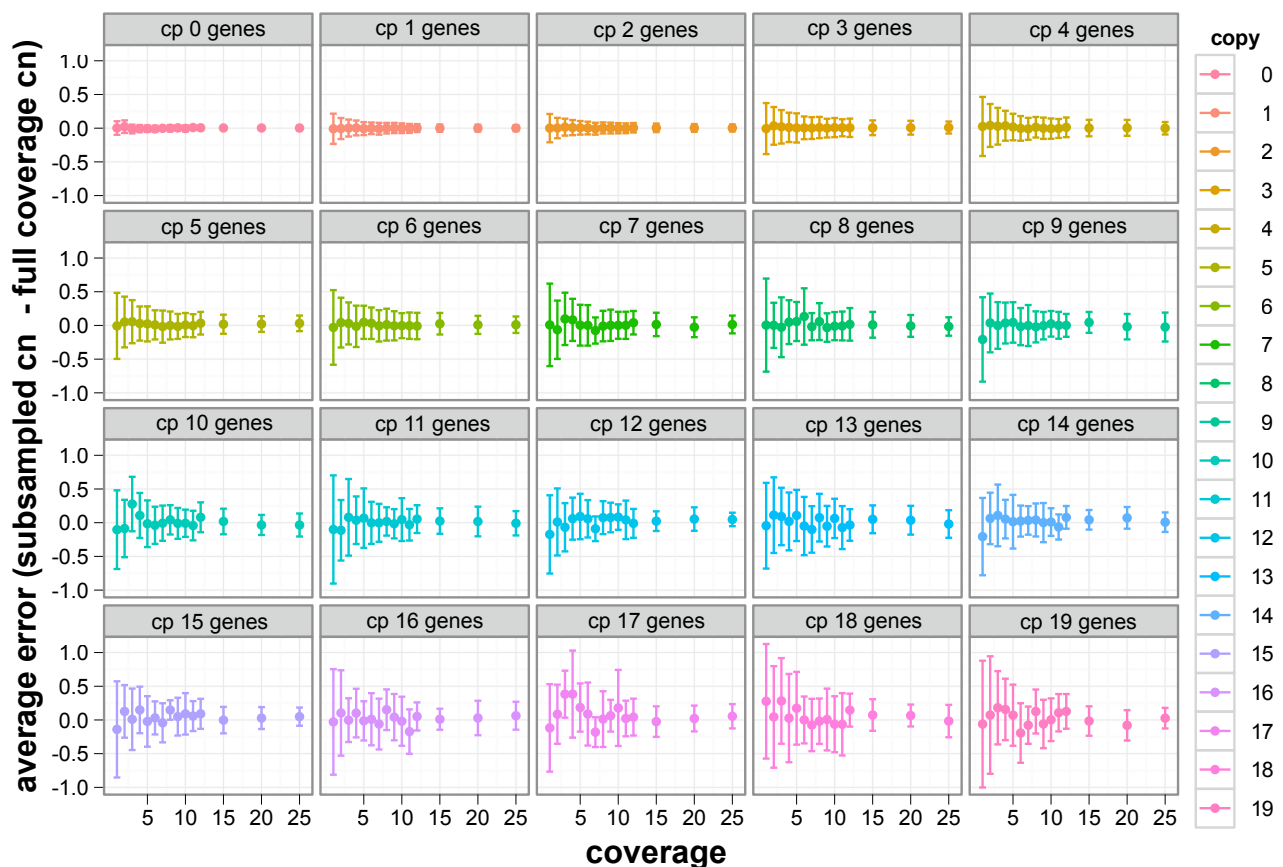


Figure S11. The deviation (mean error \pm 1 s.d.) in copy number prediction for genes ≥ 3 kb in length is shown for genomes subsampled from NA18507 to various coverages, using the full coverage ($\sim 43\times$) genome as a gold standard. Genes are stratified by their predicted copy number in the full-coverage genome. As expected, the accuracy rate is higher for genomes of increased coverage. Higher copy number states are increasingly difficult to predict accurately with low-coverage genomes.

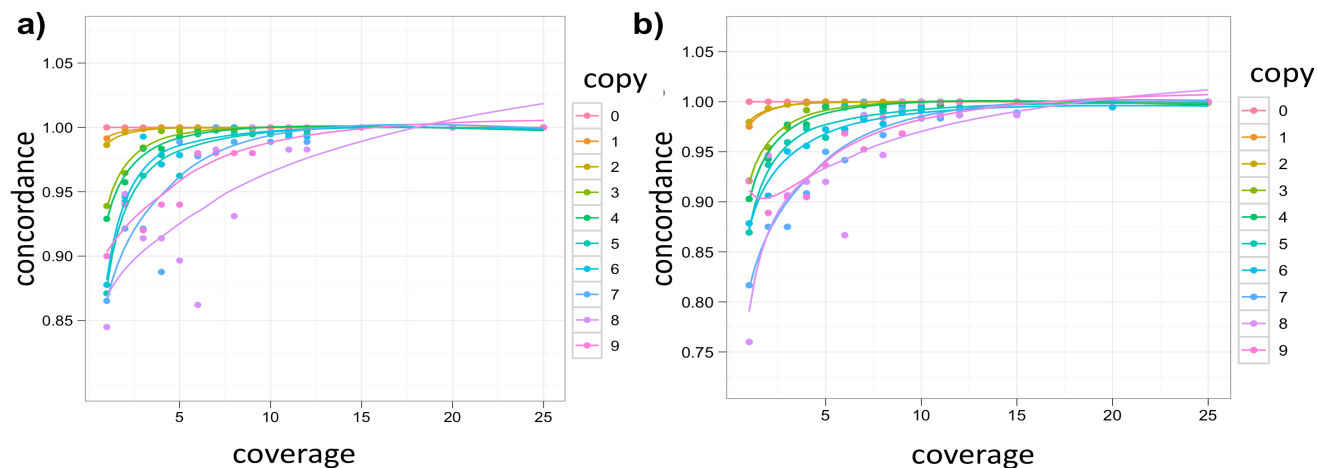


Figure S12. Concordance of copy number estimates versus sequence coverage for all genes a) >3 kb and b) >1 kb. A gene's copy number estimate was considered concordant following subsampling if it was within ± 0.5 copies of the full-coverage estimate.

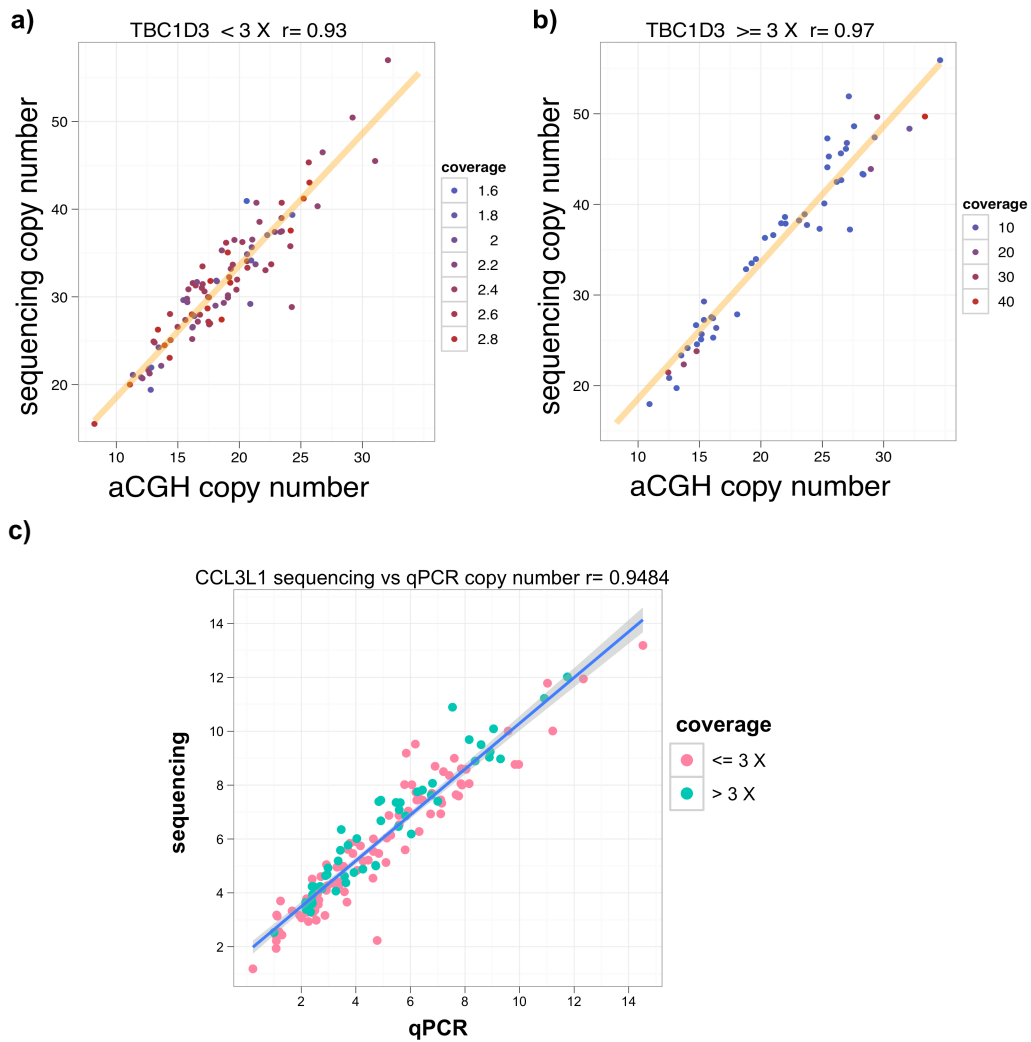


Figure S13. Independent validation of read depth-based copy number predictions for (A,B) TBC1D3 and (C) CCL3L1, stratified by depth of sequence coverage used for read depth-based genotyping.

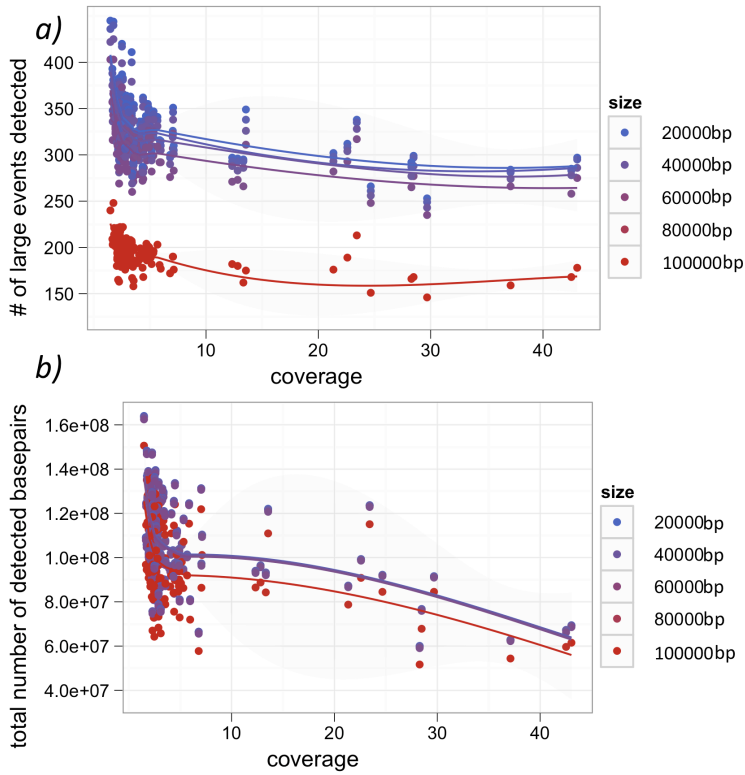


Figure S14. a) The number of events and b) total number of base pairs encompassed by events detected by dCGH in each genome as a function of overall coverage (x-axis). Colors indicate different minimum event-size thresholds. More events are detected in low coverage genomes due to their increased variability.

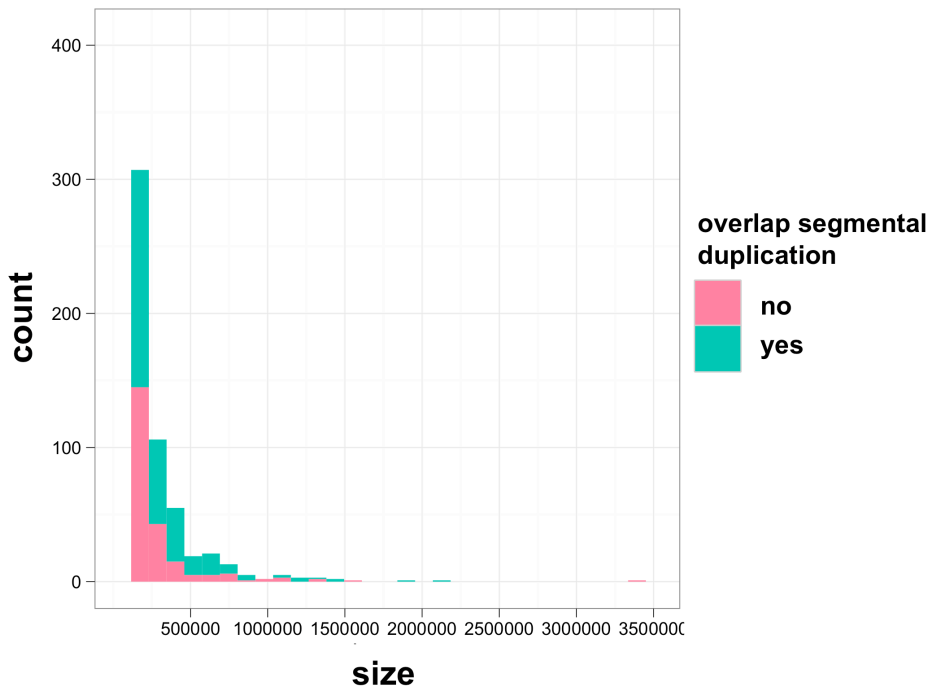


Figure S15. Size distribution of copy number polymorphisms (CNPs) >50 kb colored by segmental duplication overlap (>20% overlapping segmental duplications).

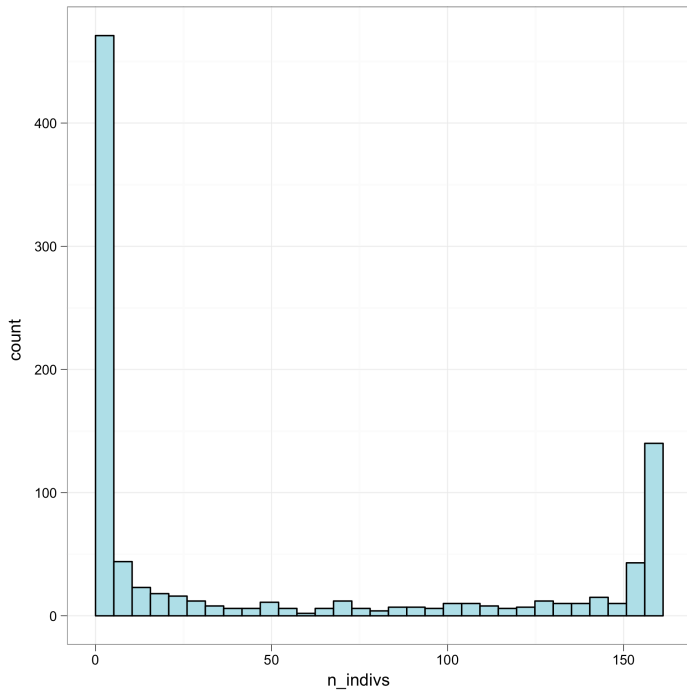


Figure S16. Frequency spectrum of copy number polymorphisms >50 kb shows the majority of events are rare.

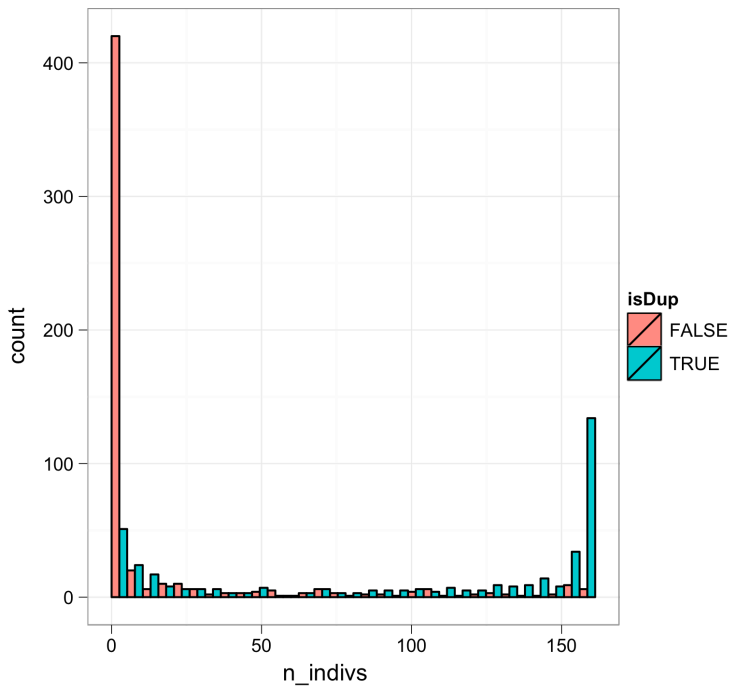


Figure S17. Stratifying the frequency spectrum of events by overlap with segmental duplications (50% overlap) shows CNPs not overlapping segmental duplications are overwhelmingly rare in frequency.

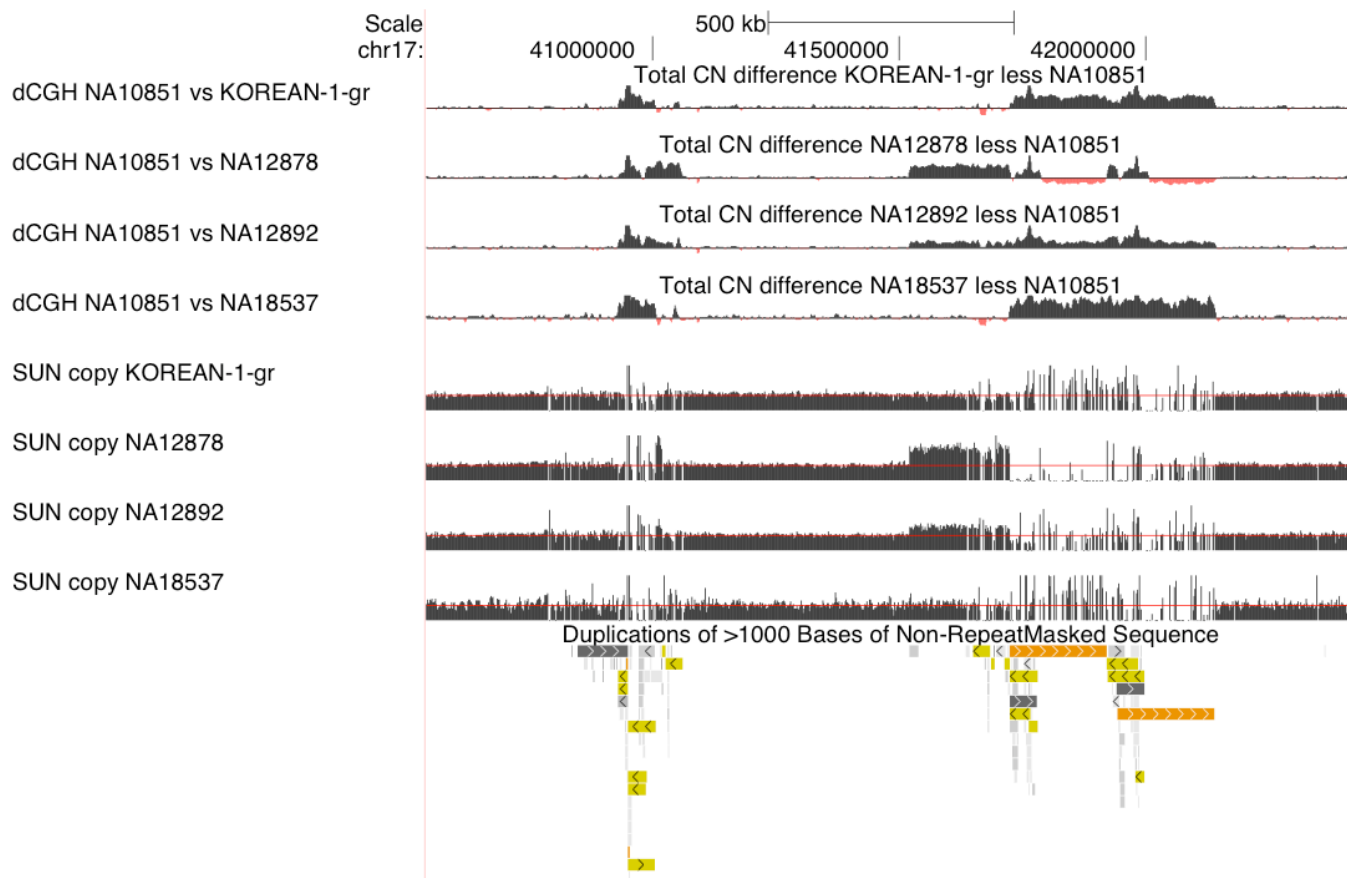


Figure S18. Digital CGH and SUN-based copy number maps demarcate two distinct duplication events on chromosome 17q21.31, for four representative individuals. Digital CGH signal is colored by call: black bars (above the baseline) indicate relative copy number gain; red bars indicate relative loss. Red line on SUN copy tracks indicate predicted paralog-specific copy number of 2.

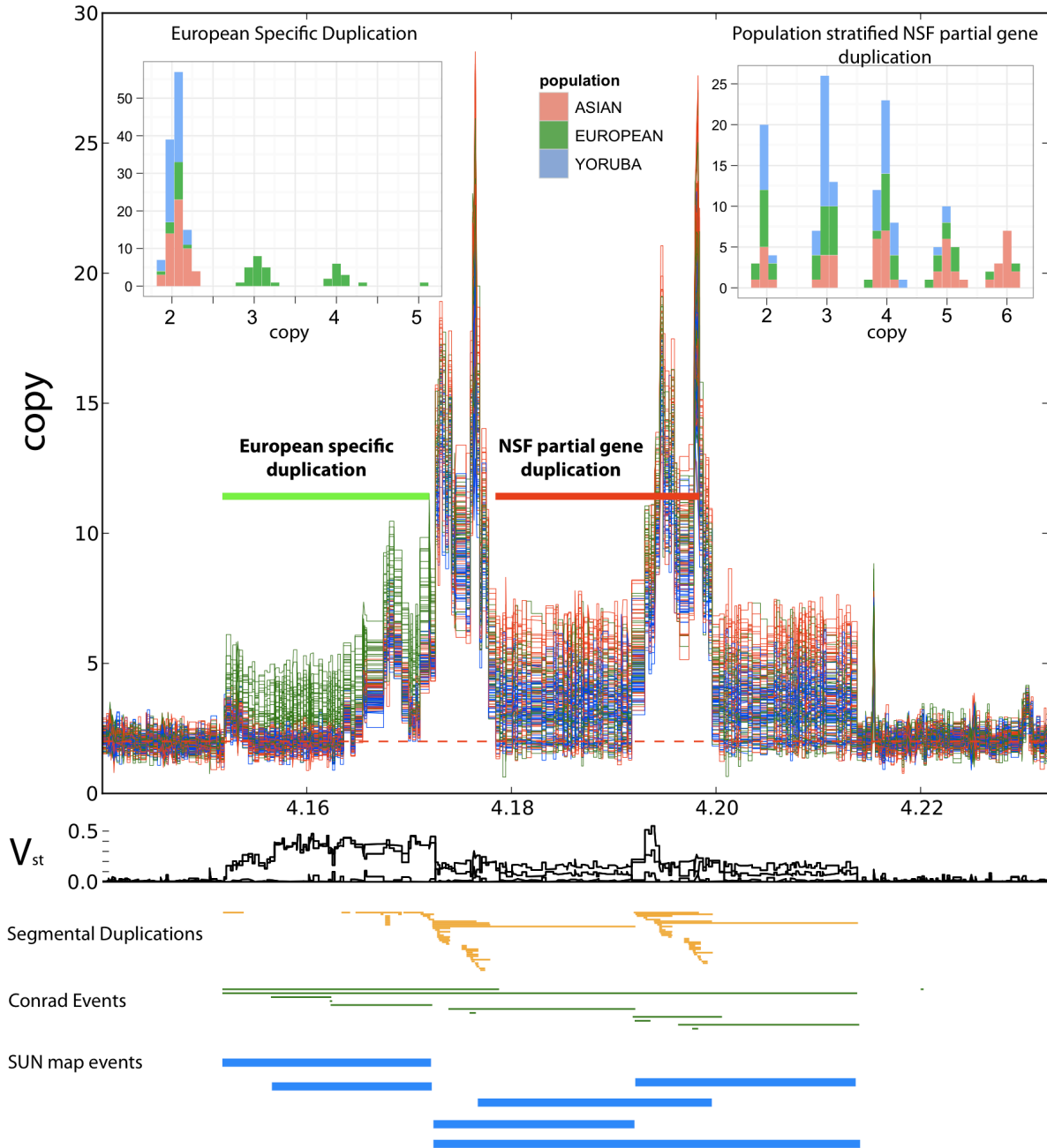


Figure S19. Population stratification in CNVs on chromosome 17q21.31. Total copy number estimates for 154 individuals are overlaid (related and duplicate individuals excluded) and shaded by population. Note the apparent increase in copy among Europeans (green) and Asians (red). Inset are the copy number estimates for two highly population stratified duplications of 210 and 205 kbp. Discrete copy number estimations were confirmed by FISH (see below). The V_{st} metric (as described below) is an indication of population stratification. Strong signals of population stratification are seen for both European-specific and *NSF* partial gene duplications.

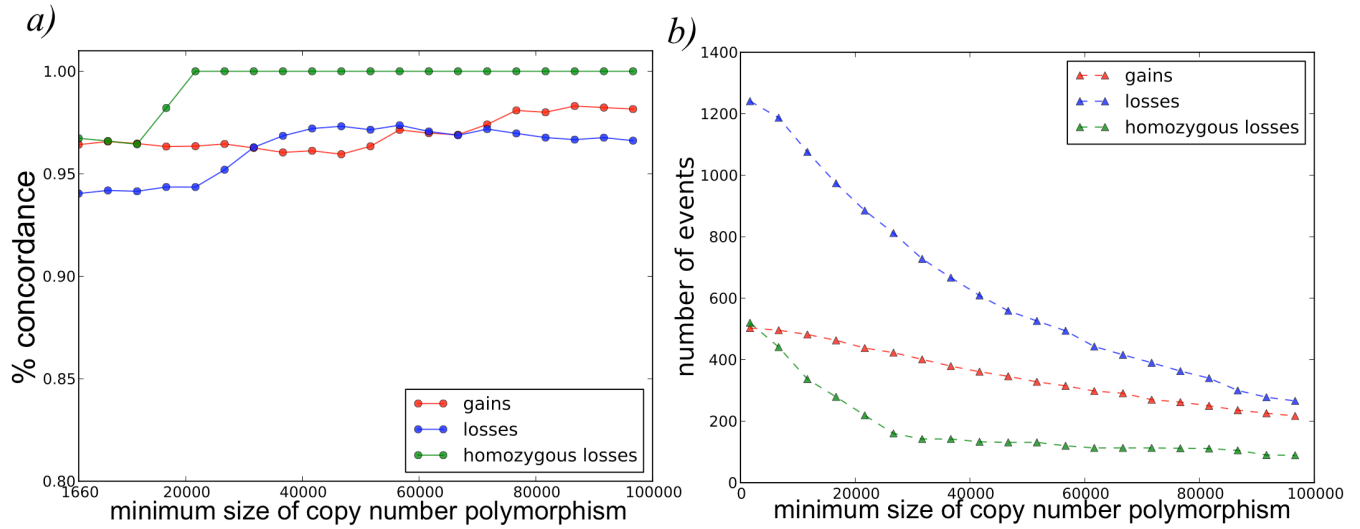


Figure S20. a) Percent concordance between read depth-based copy number genotypes and previous genotype calls made using the Illumina 1M Duo SNP genotyping platform on the same DNA samples (S7). Plotted cumulatively by event size (x-axis) b) The number of events assayed at each cumulative size range.

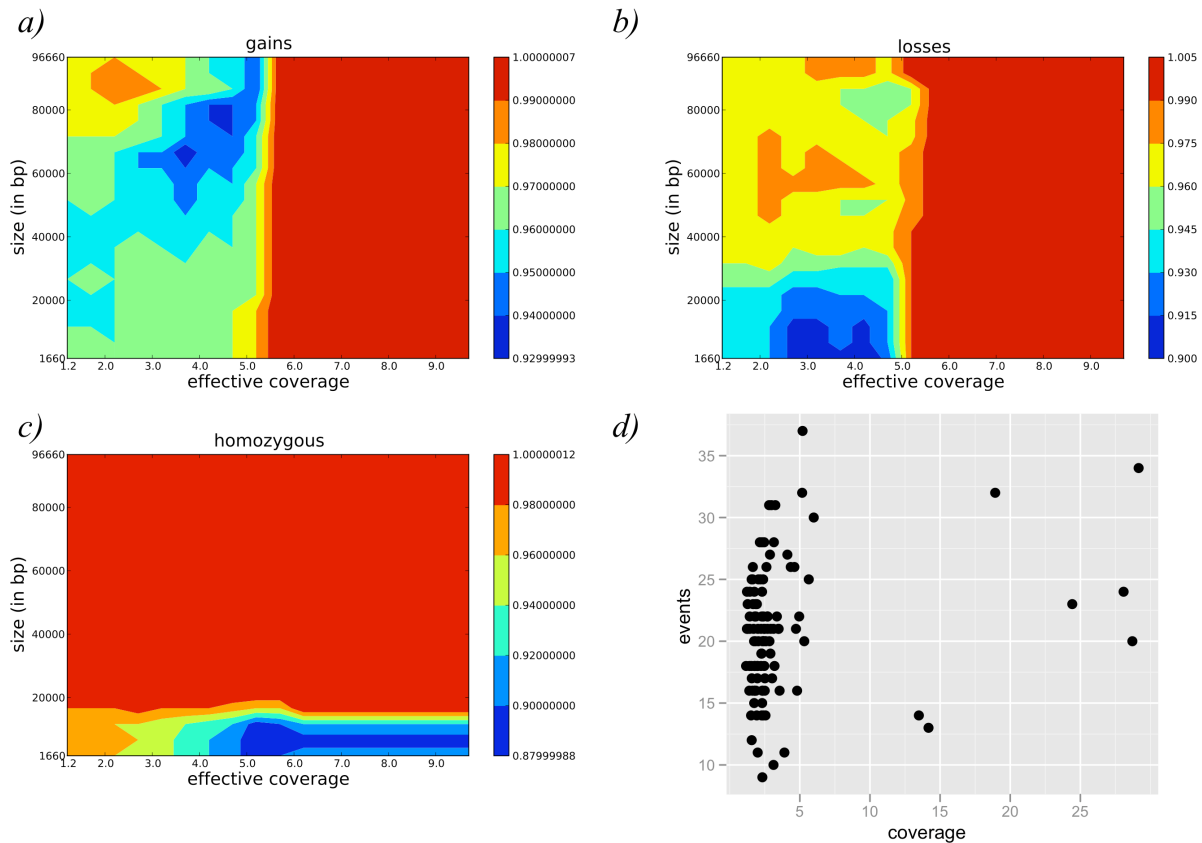


Figure S21. Contour plots showing the effect of sequencing coverage and event size on concordance (ranging from 0 to 1) for a) gains, b) losses and c) homozygous losses. Size and effective coverage both contribute to the overall concordance of events. d). The sudden jump in concordance at ~5X is an artifact of the paucity of genomes between ~5X to ~12X.

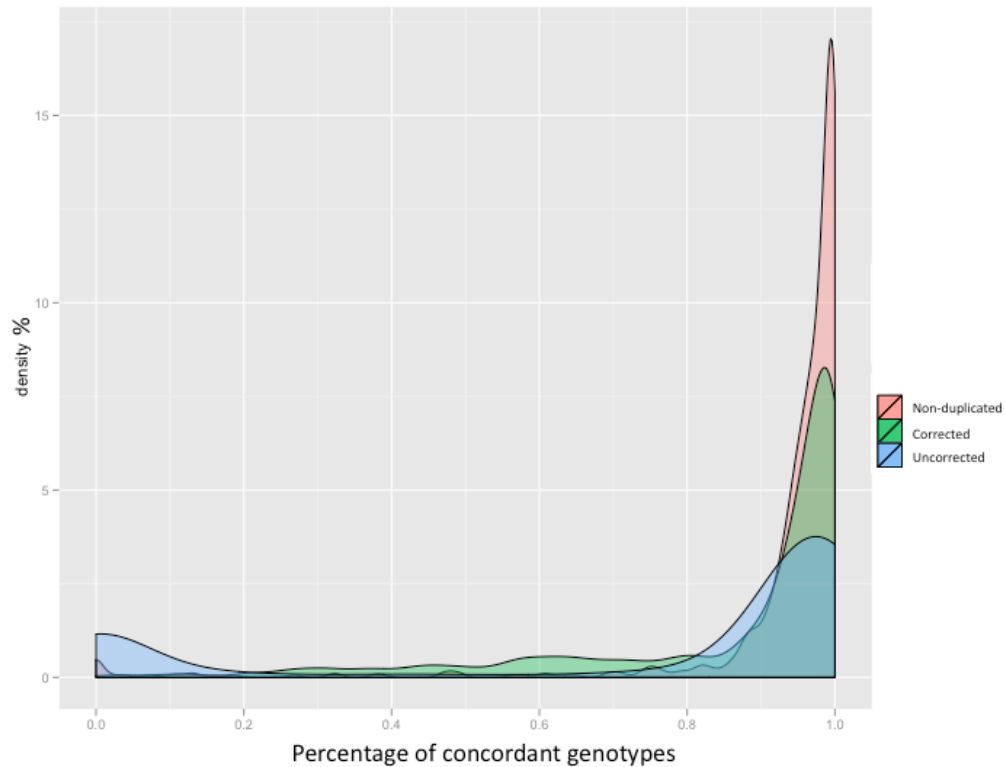


Figure S22. Genotype copy number concordance. Fraction of genotyping calls concordant between absolute read depth estimates and the Affymetrix SNP 6.0 Microarray (S21). 21% of calls show significant departure (uncorrected) but these map overwhelmingly to duplicated sequence. If adjusted (corrected) by a fixed integer amount reflective of the true population average copy number, the fraction of correct genotypes increases. Within unique regions, 94% of genotype calls are in accord.

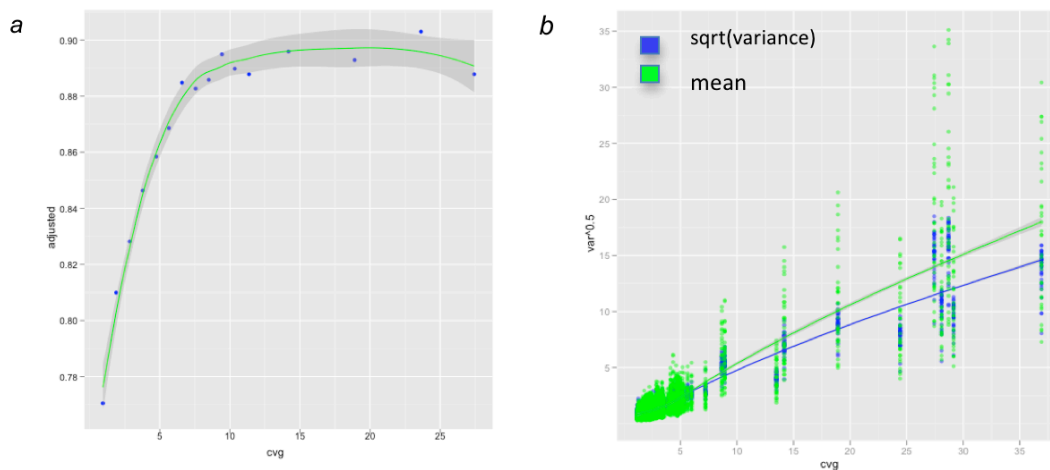
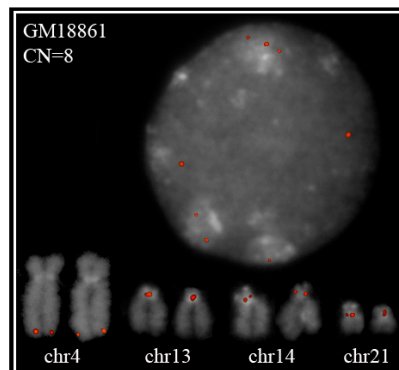
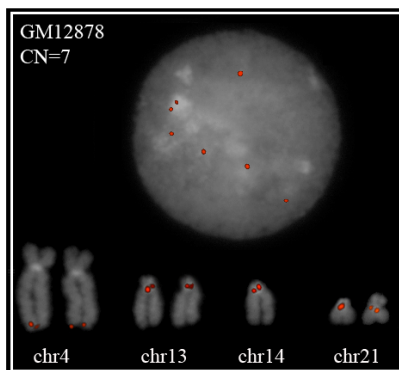
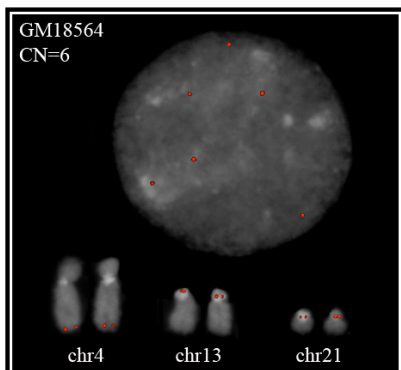
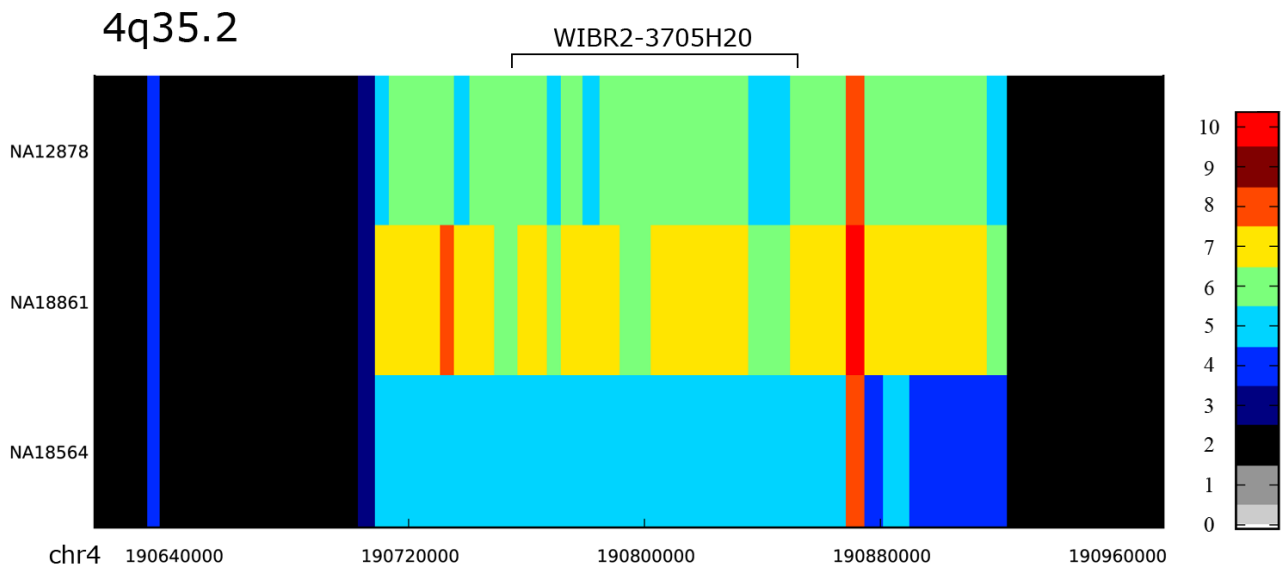
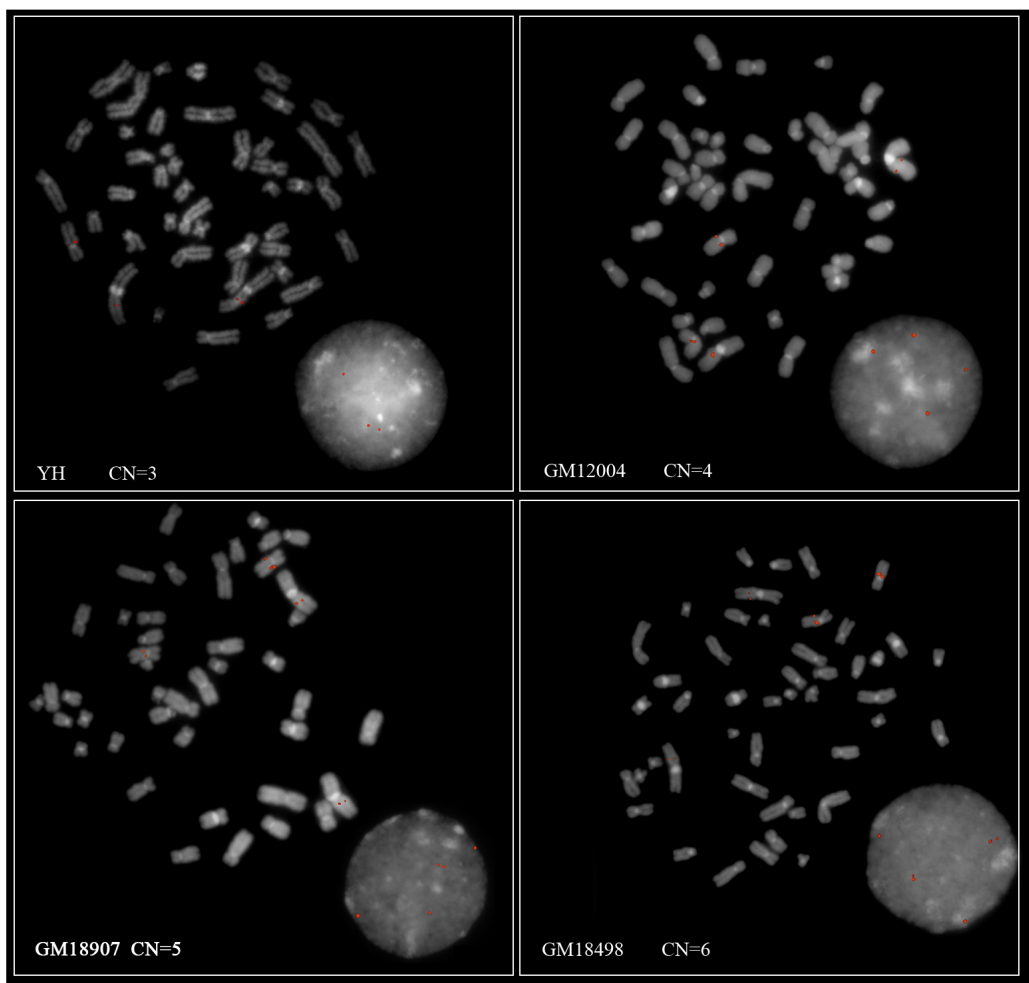
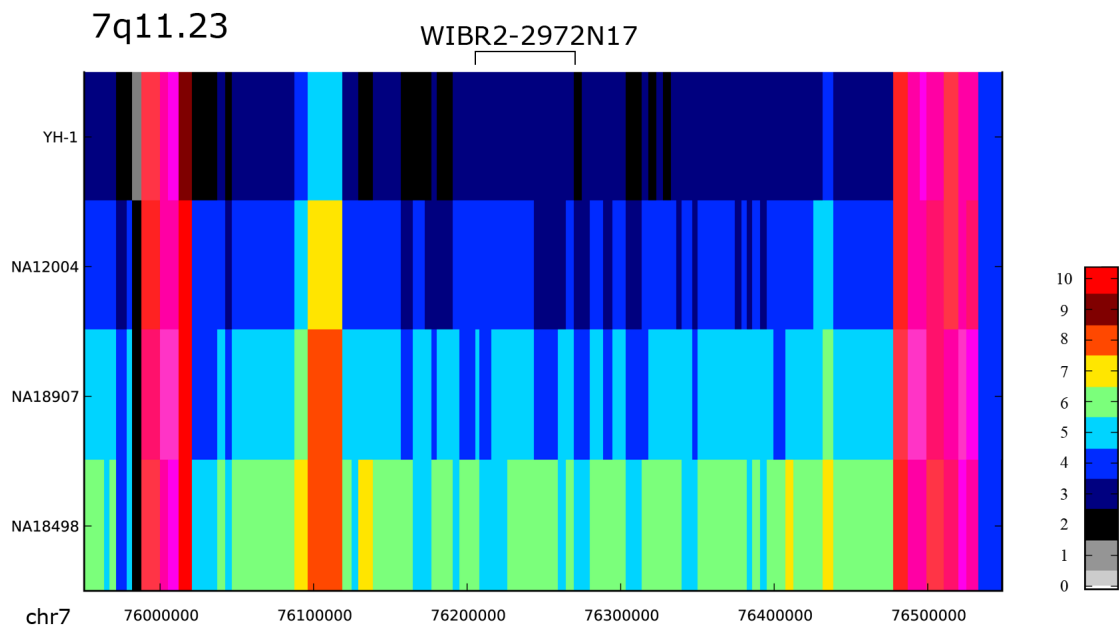


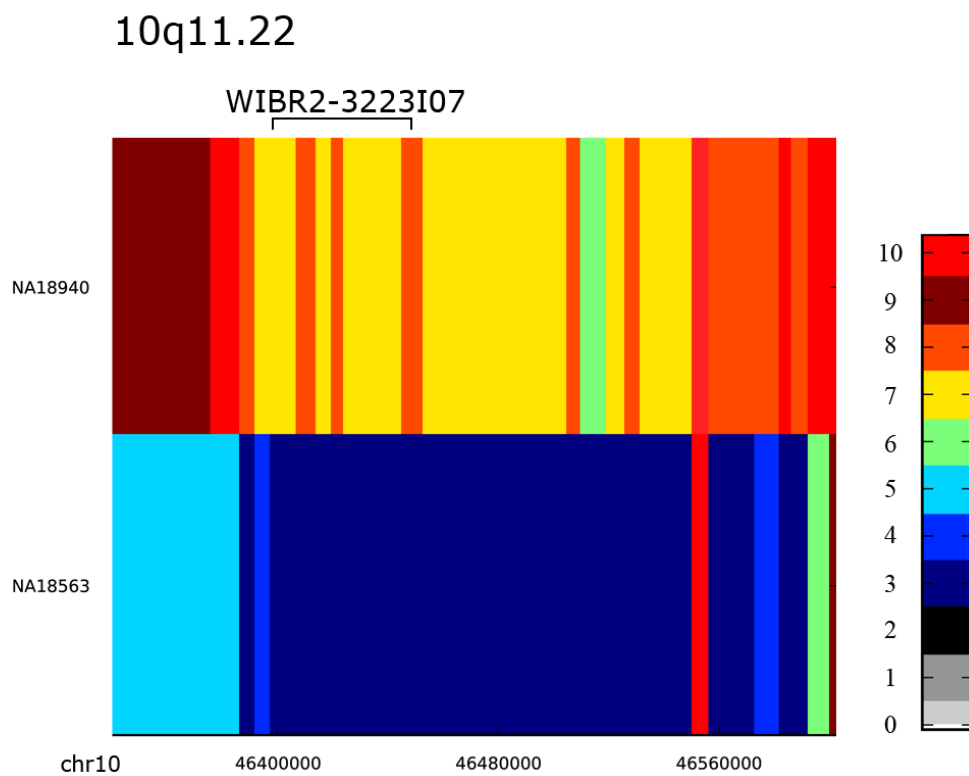
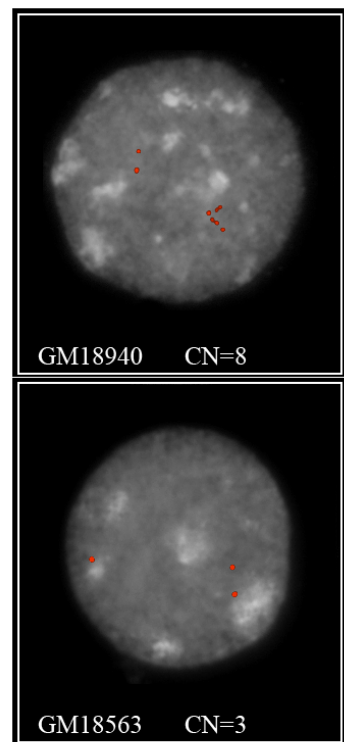
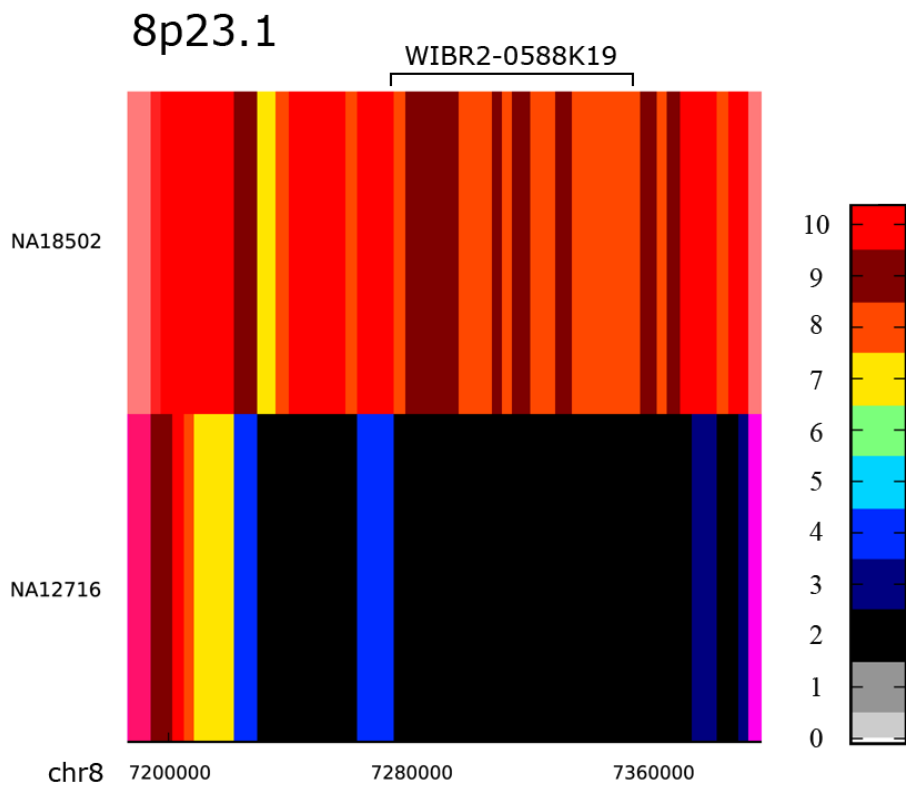
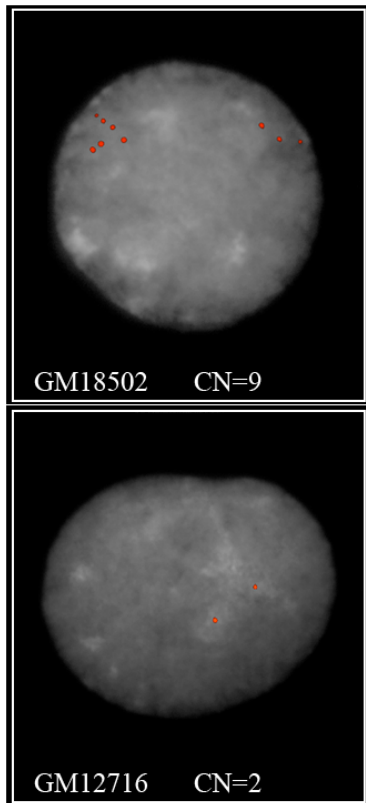
Figure S23. a) Concordance of genotypes with McCarroll *et al.* calls (S21) for NA18507 subsampled to various depths of coverage. b) Mean and standard deviation of read depth in 1-kb windows across copy number invariant (copy 2) region plotted as a function of coverage.

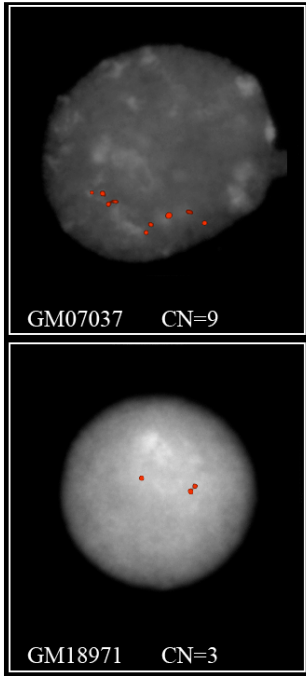
Figure S24. FISH validation (below). **a)** 12 examples of computational prediction and FISH validation of rare events seen in a single individual and duplications predicted to be duplicated in all individuals but represented as single copy in the human reference. Each line in the heat maps represents a sample for which the computational predictions were generated. **b)** Read depth-based predictions and FISH images of two highly copy number variable regions where individuals with the most extreme copy number difference were selected. In these cases FISH failed to accurately estimate copy number differences between individuals since the signals visualized on interphase nuclei were too numerous to provide an exact number of copies.

a)



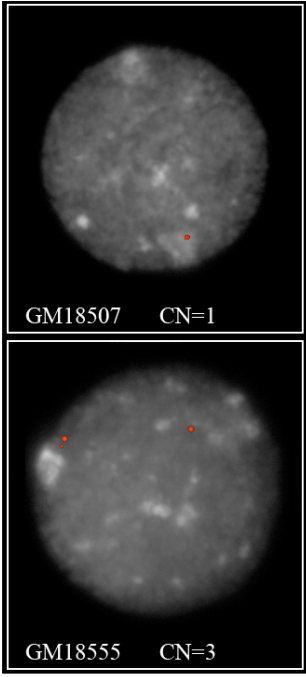
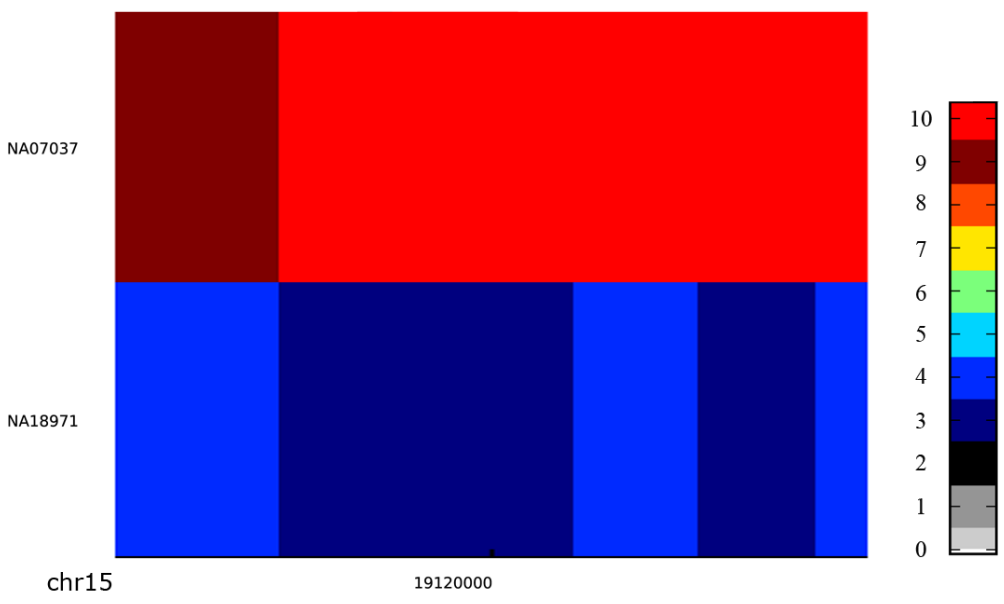






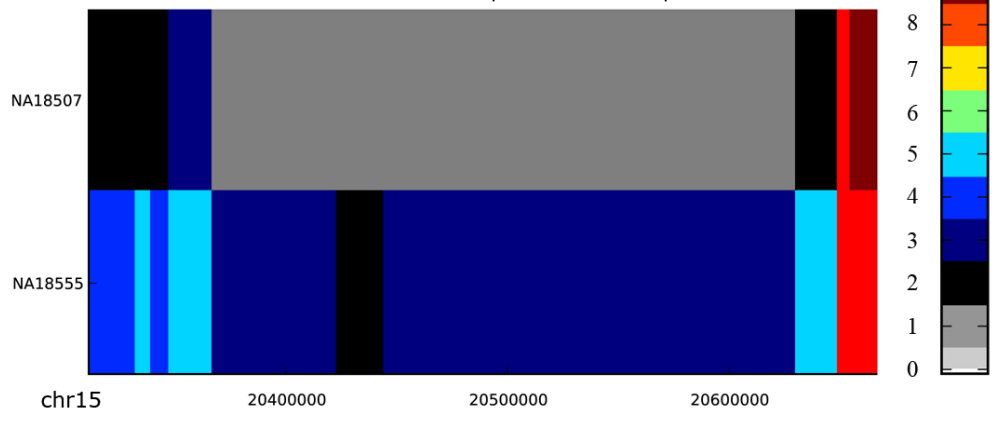
15q11.2

WIBR2-3778G07

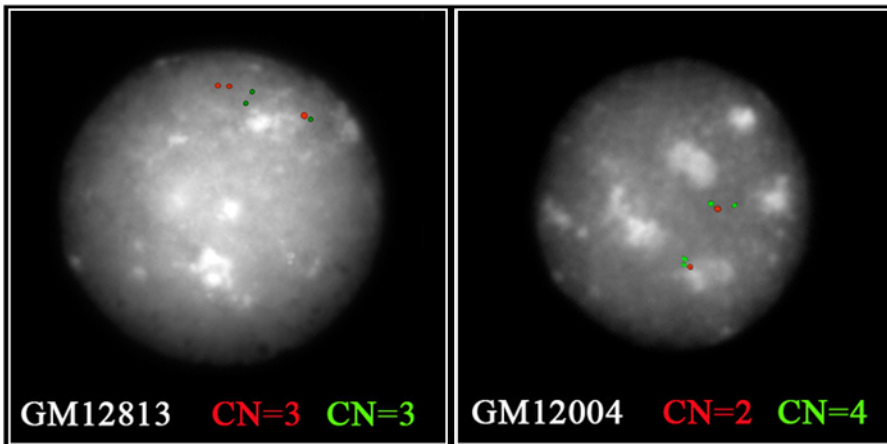
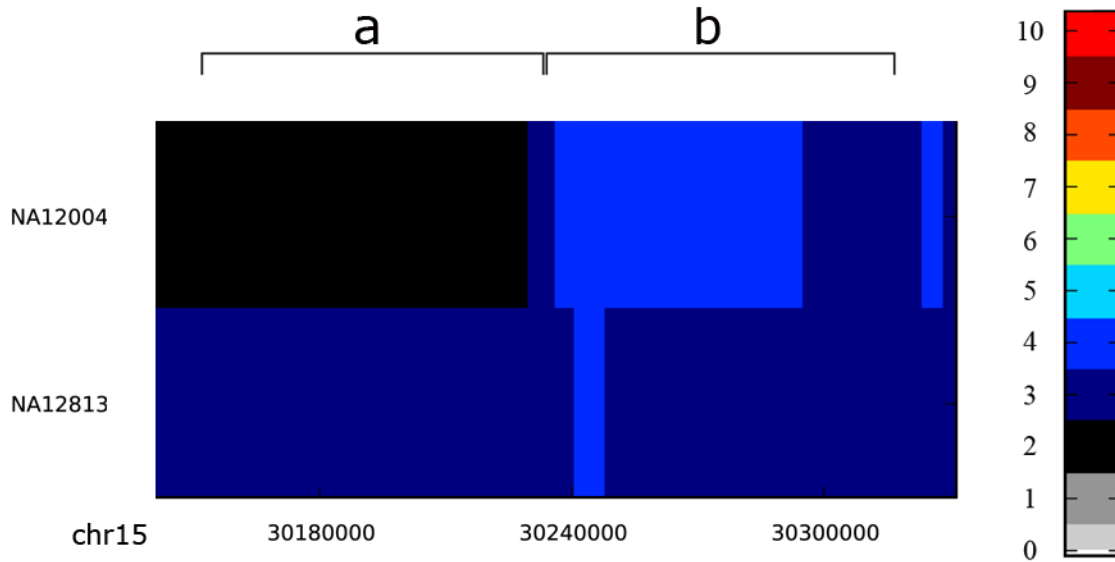


15q11.2

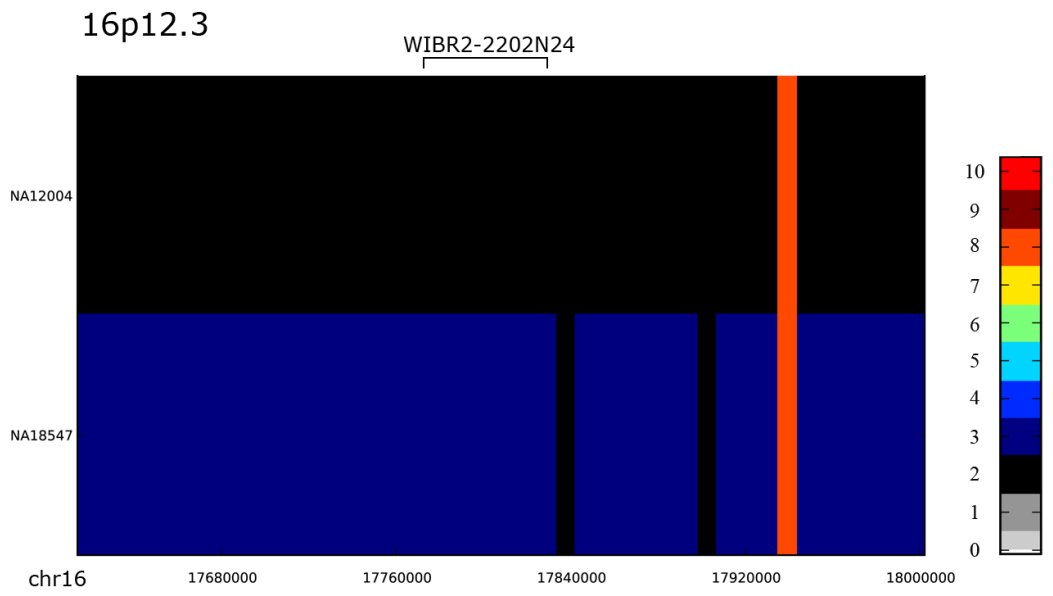
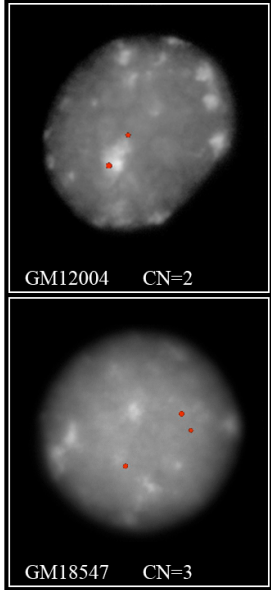
WIBR2-0866L18

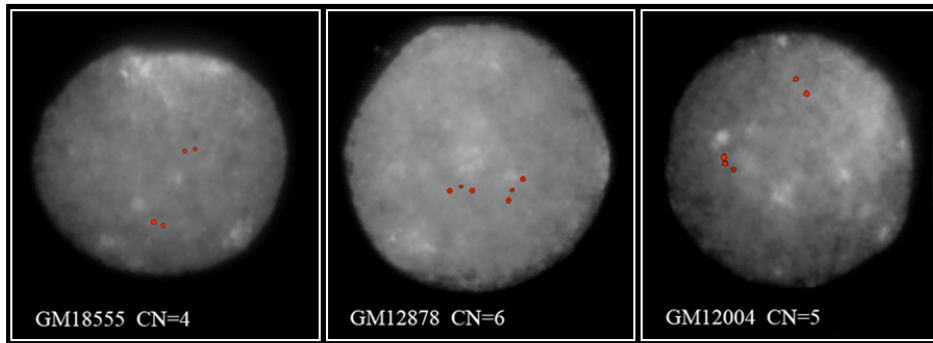
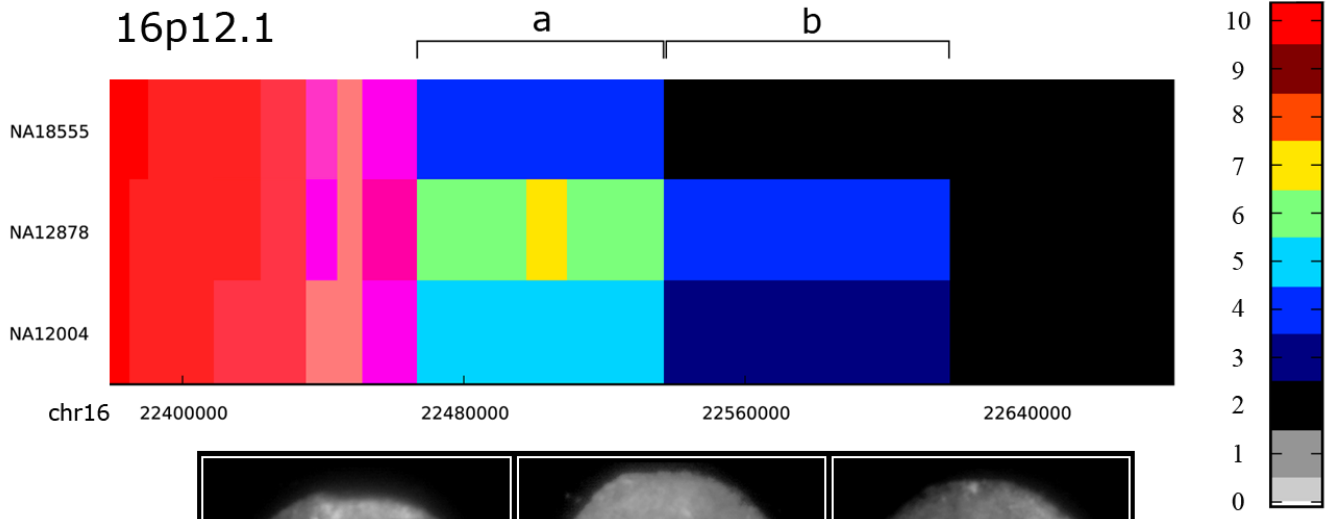


15q13.3

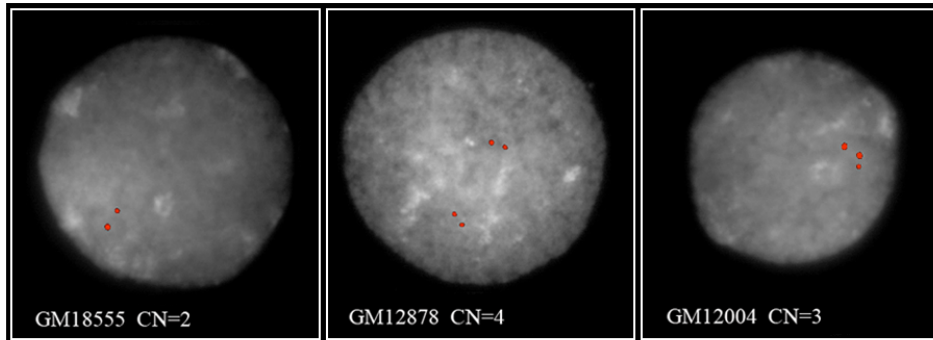


- a) WIBR2-1228I16 (red)
- b) WIBR2-2841A03 (green)





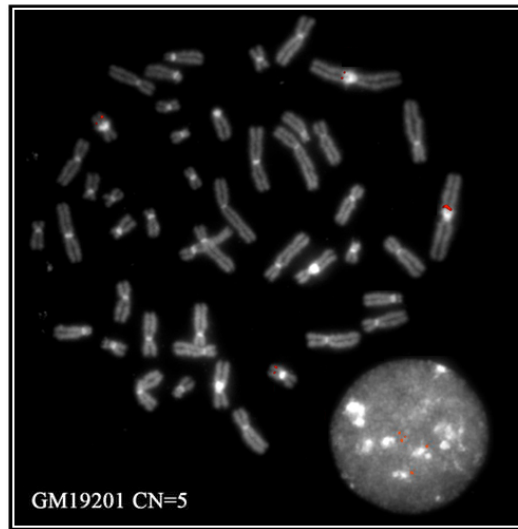
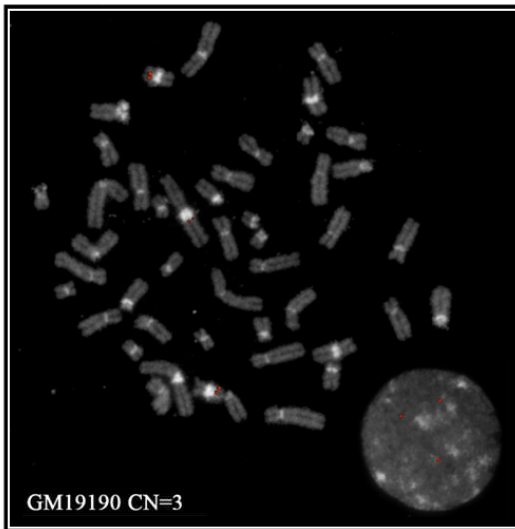
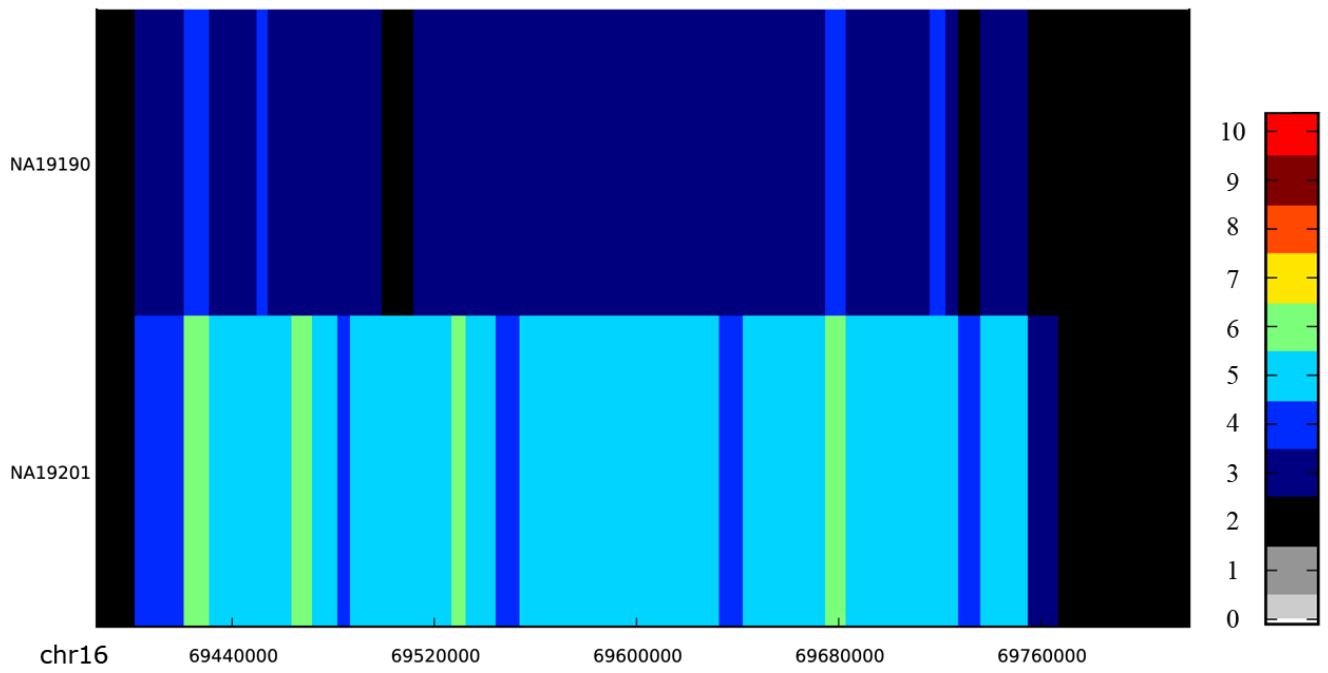
a)WIBR2-2031K01

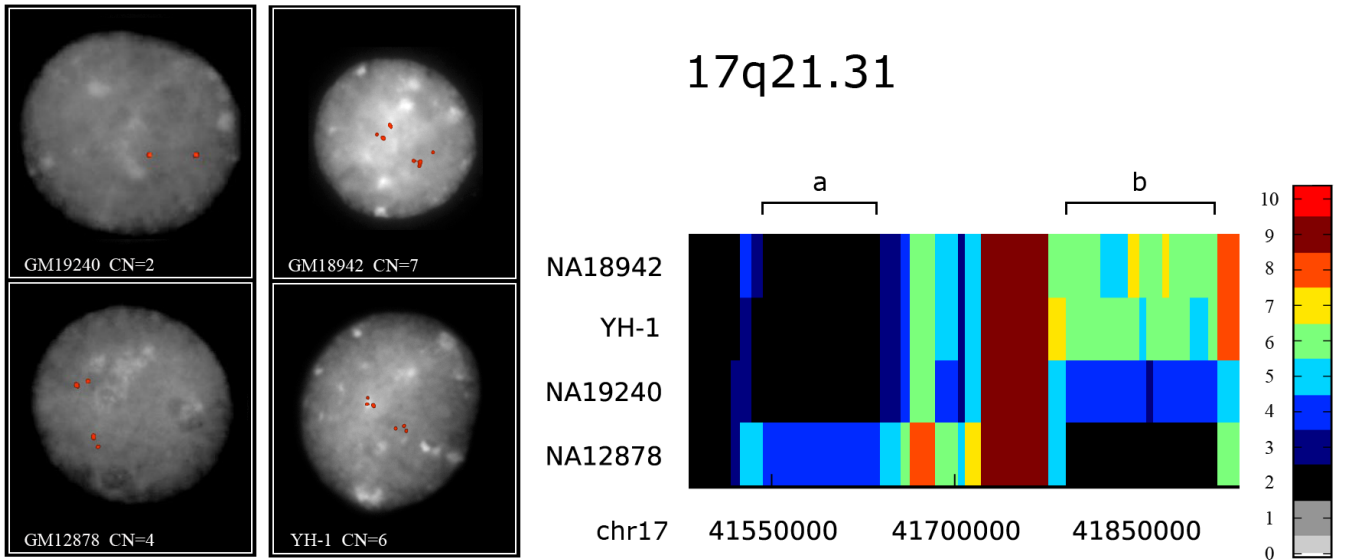
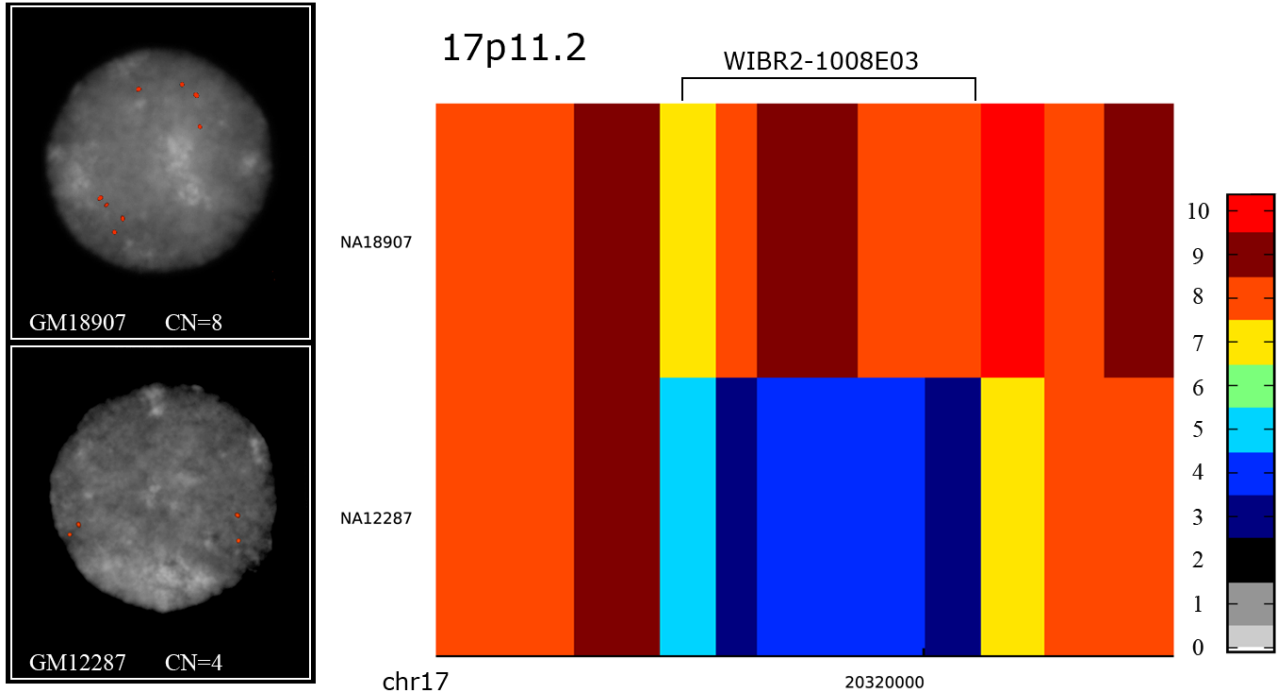


b)WIBR2-3608M06

16q22.2

WIBR2-3823N03



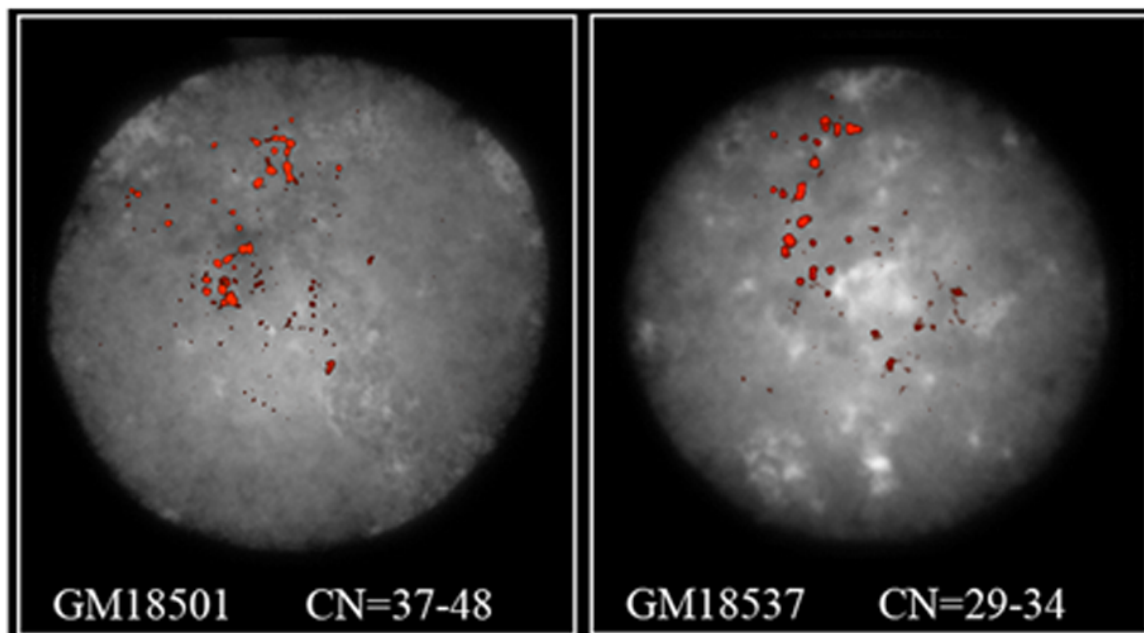
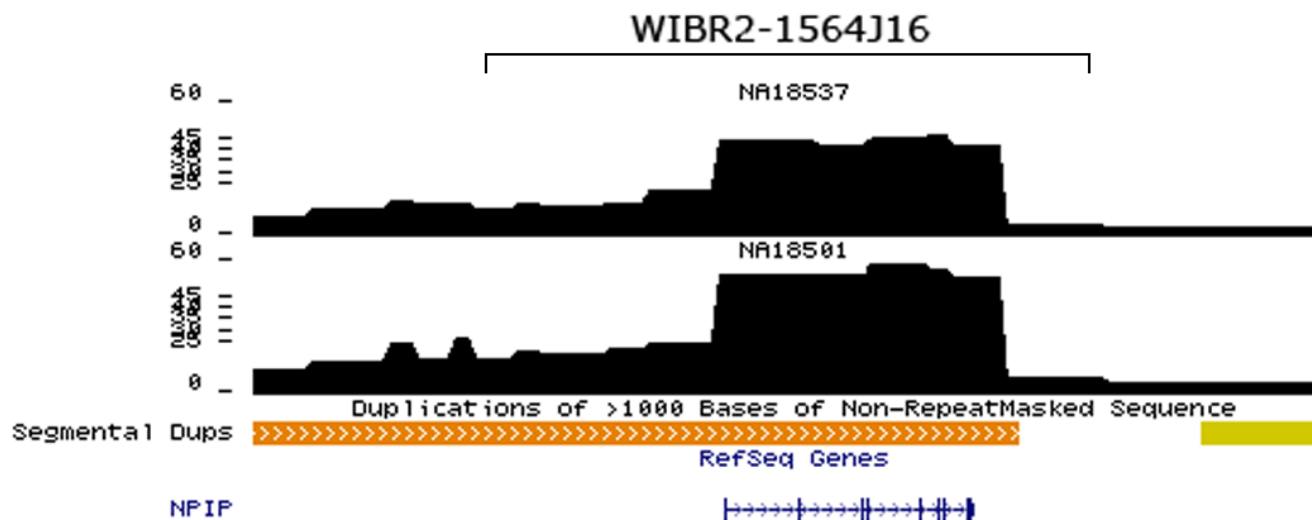


a) WIBR2-2004M23 b) WIBR2-1857K04

b)

16p13.11

chr16: 14920000| 14930000| 14940000| 14950000| 14960000| 14970000|



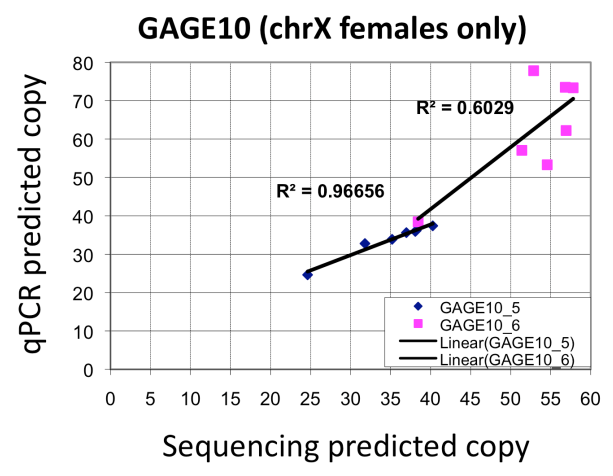
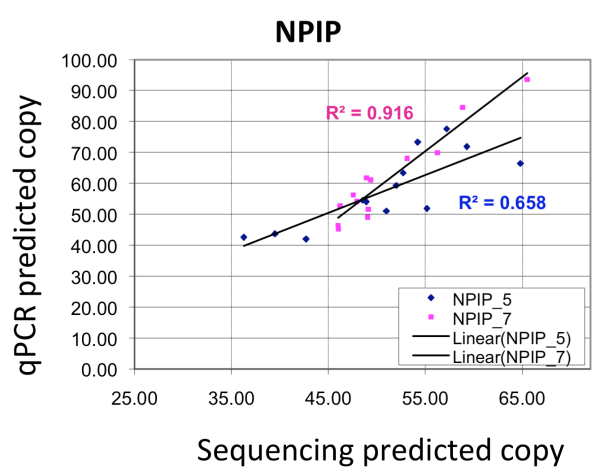
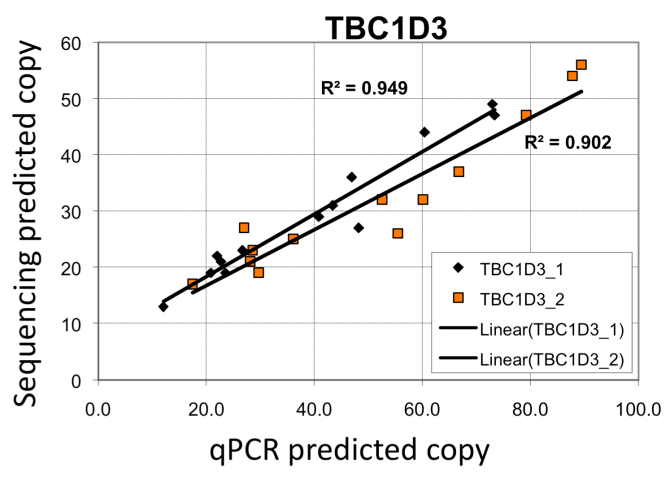
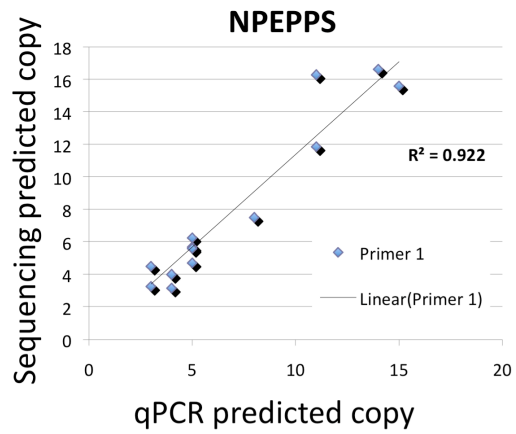


Figure S25. qPCR validation experiments for four genic loci showing strong correlation with sequencing-based copy number for a variety of ranges. Seven out of nine experiments had $R^2 > 0.84$.

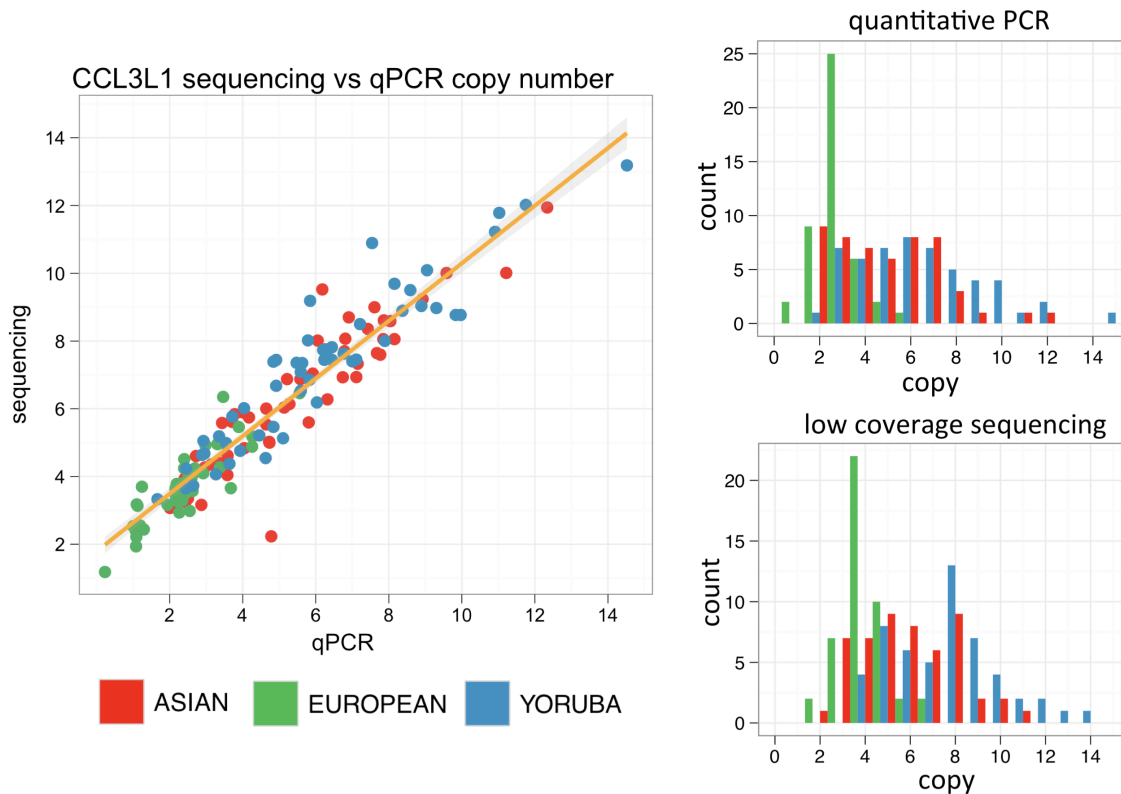


Figure S26. qPCR-based copy number genotyping is highly correlated with sequencing-based copy number estimates ($r = 0.95$). Sequencing-based copy number genotypes capture both the specific copy number range and population stratification of this locus.

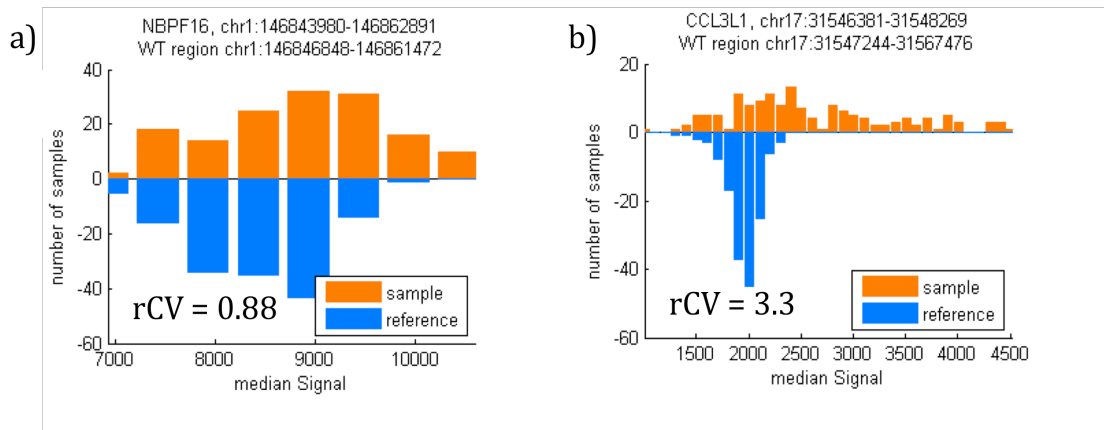


Figure S27. The median single-channel intensity signals for the sample and reference are shown for two loci. Copy number variation of an individual sample can only be detected if the signal variation between the sample and reference is detected. The rCV statistic summarizes the amount of variation observed between the sample and reference genomes.

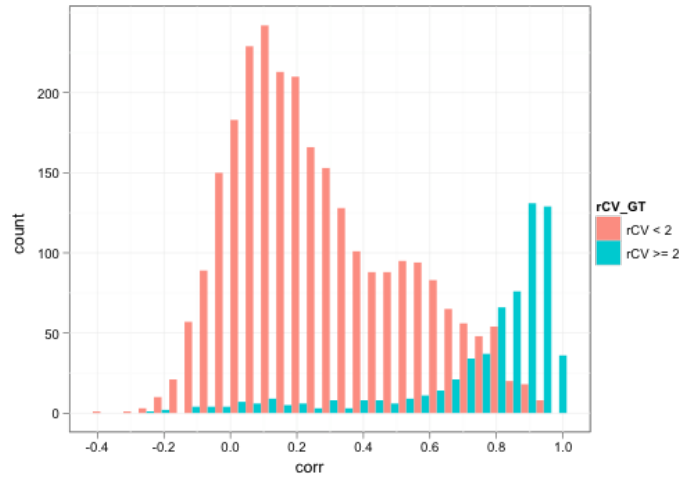


Figure S28. The rCV metric indicates if a region on the array can accurately assay the spectrum of copy number variation at that locus. Histogram of correlations shown between sequencing and array genotypes for rCV <2 and rCV >=2. Above rCV 2, 77% of loci assayed have a correlation >0.7.

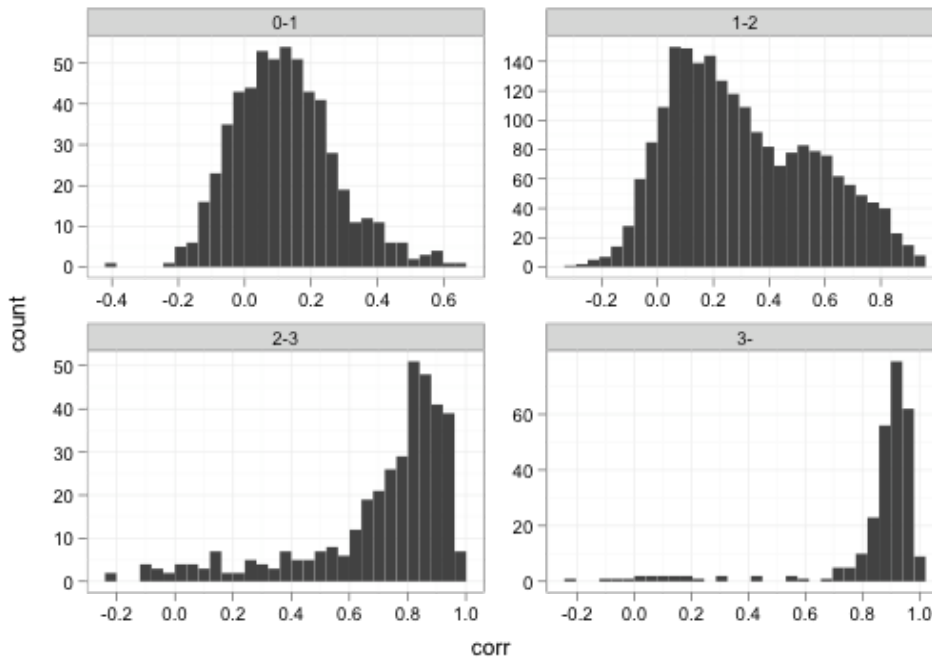


Figure S29. Histograms of correlation between sequencing-based copy number predictions and array-based copy number predictions for different ranges of rCV. At rCV >2 most regions have correlations of >=0.7.

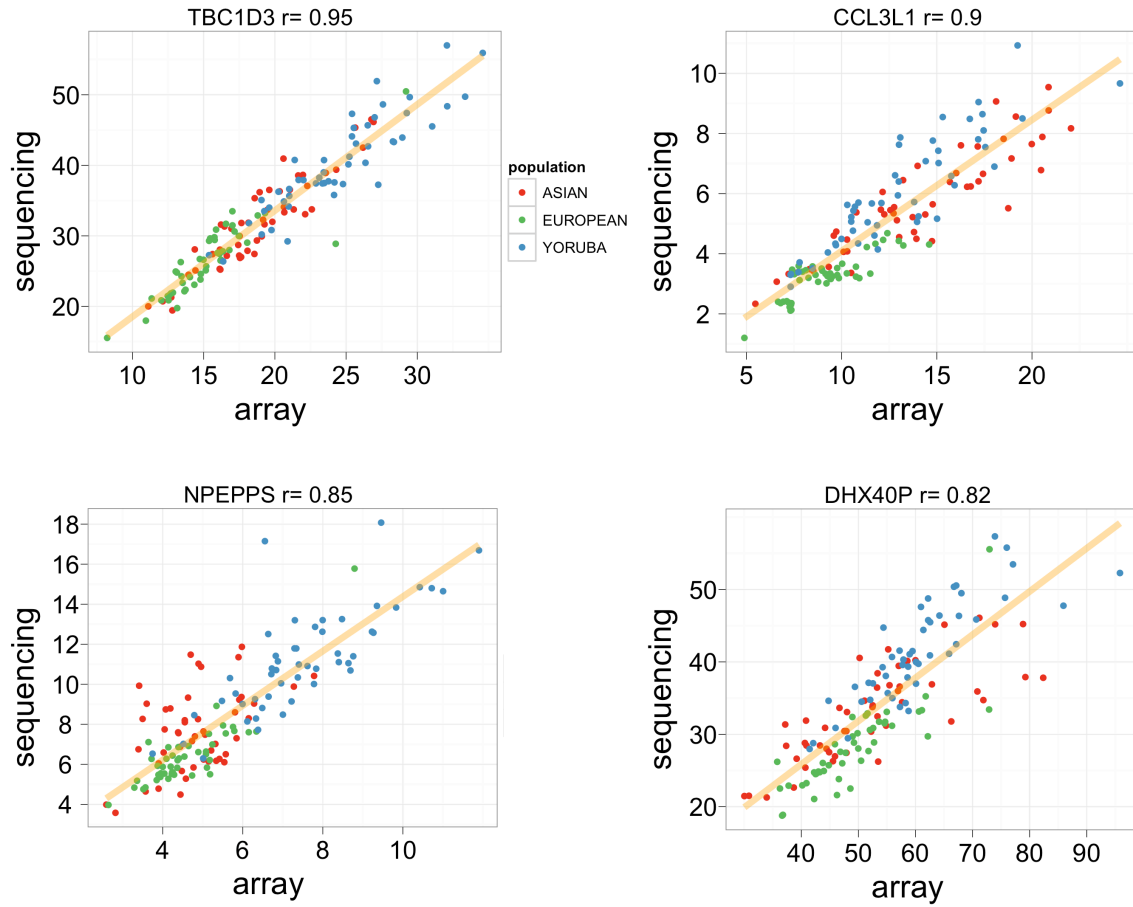


Figure S30. CGH-based copy number estimates in many regions are highly correlated with our read depth-based copy predictions. Read depth-based copy number estimates can then be used to calibrate CGH copy predictions on a per locus basis.

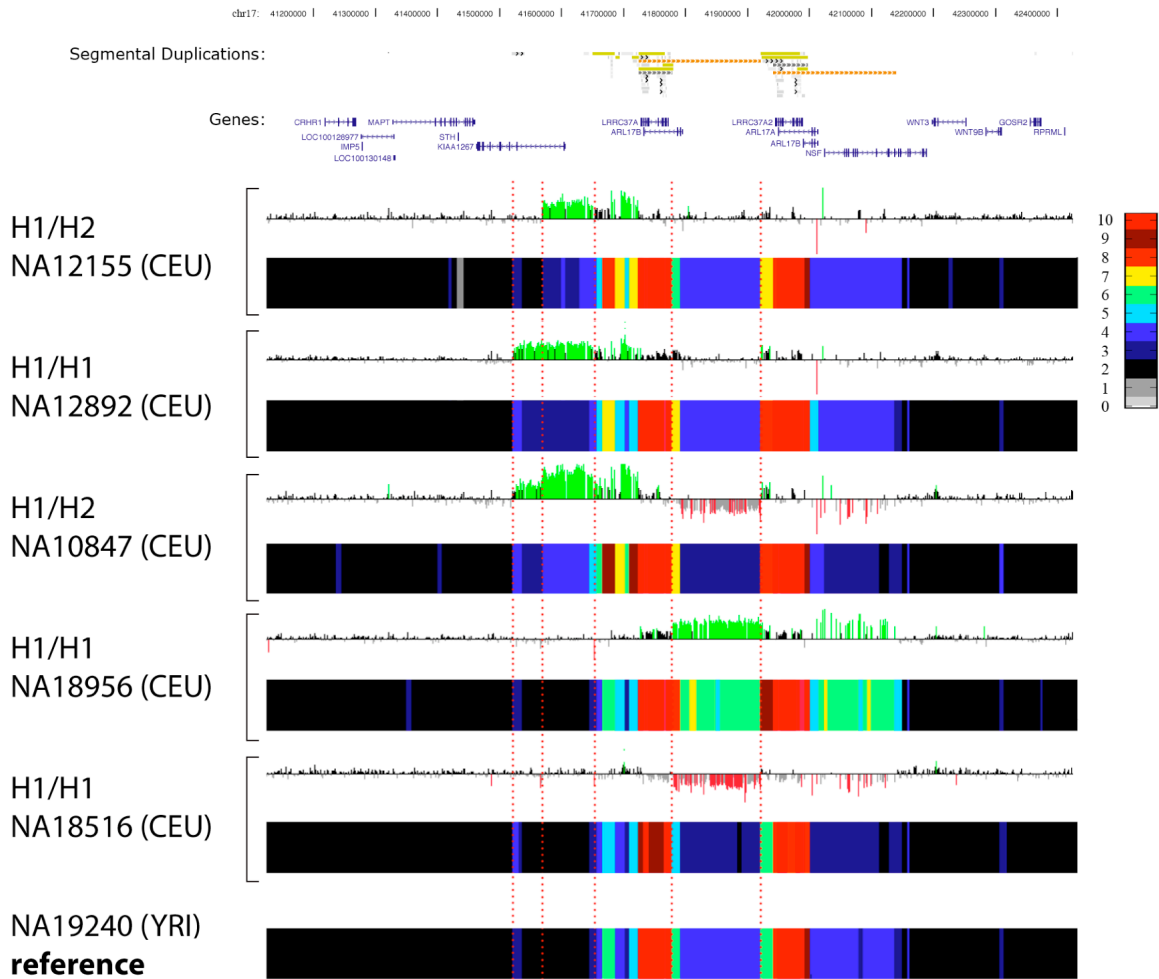


Figure S31.

Analysis of 159 individual genomes used to select a reference individual nearest the population-wide median copy number distribution to maximize discriminatory power in array CGH experiments. Array CGH using NA19240 as a reference sample is shown above read depth-based copy number estimates for chromosome 17q21.

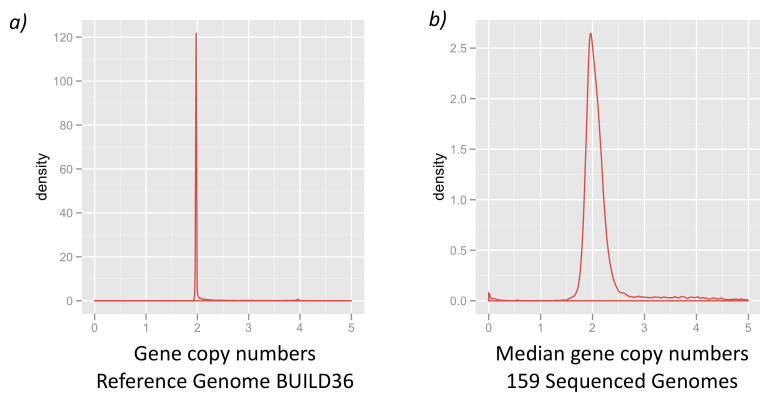


Figure S32. Density plots of the estimated copy number of all genes in the human genome for a) the reference genome Build36 versus b) the median over 159 sequenced individual genomes.

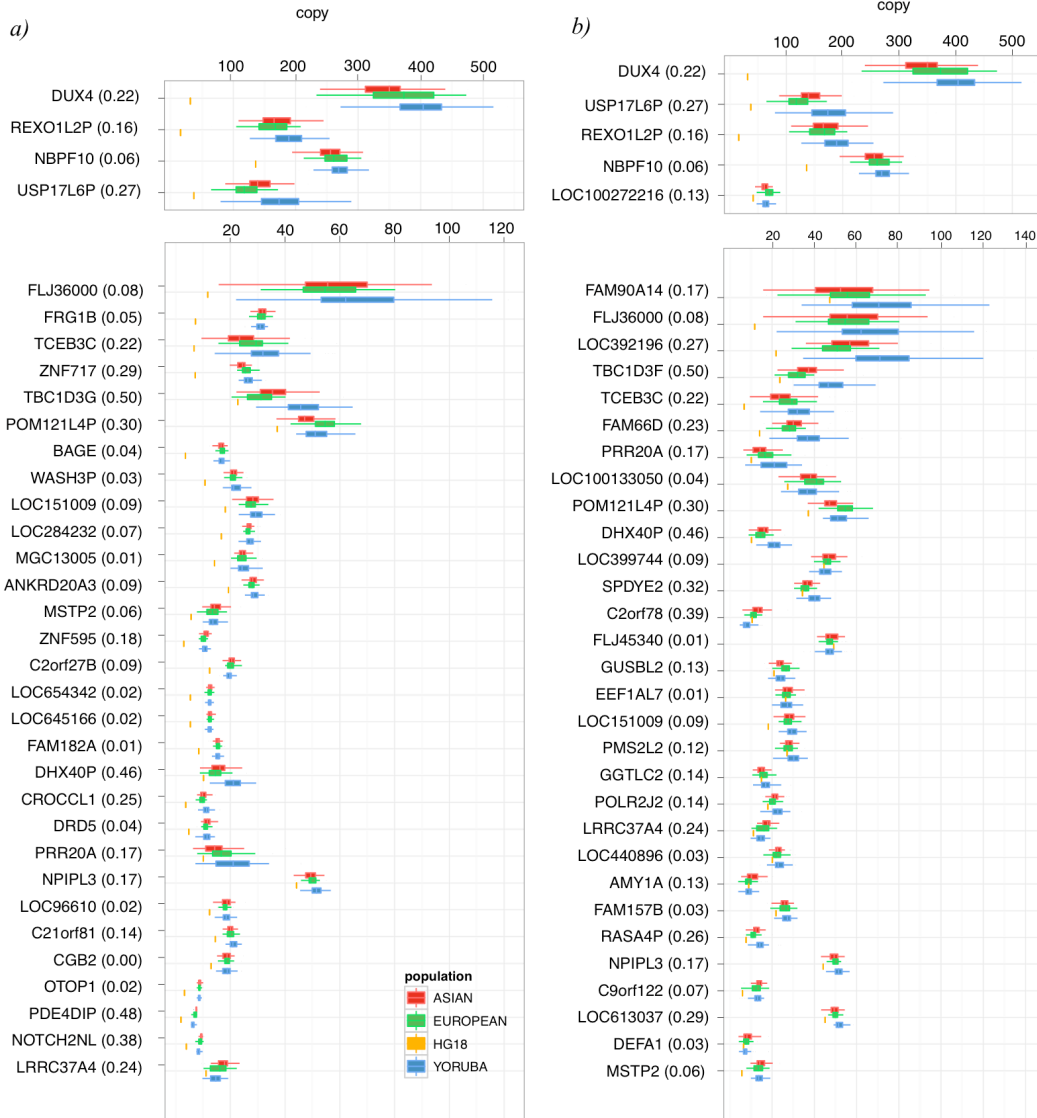


Figure S33. a) Genes missing from the human reference genome (Build36) and b) the most variable genes found among humans. Note that many of the genes with missing copies in the human reference are among the most variable genes within the human species.

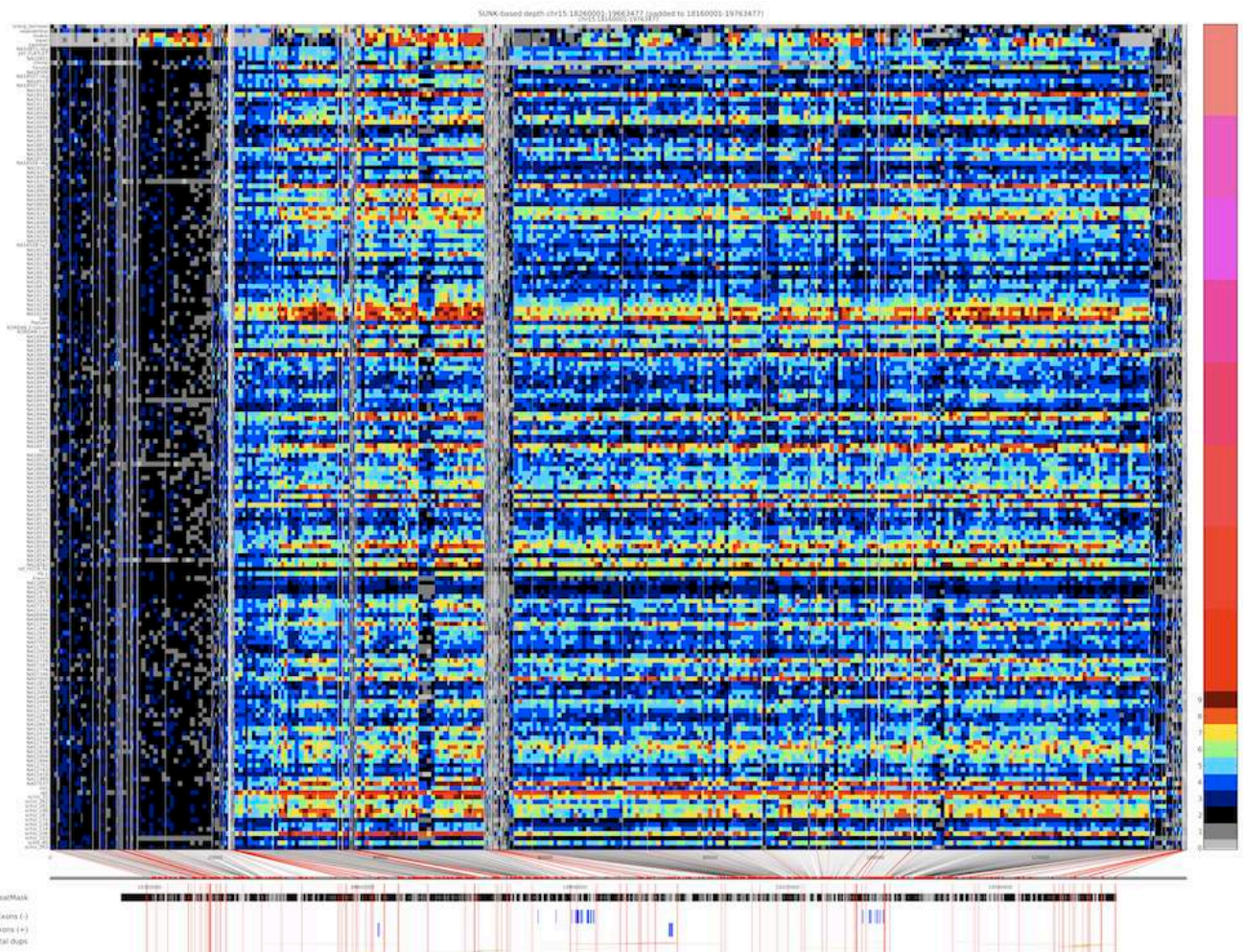


Figure S34. Complex patterns of variation in the copy number landscape of chromosome 15q11 and 15q12. A highly copy-number polymorphic, ~900-kbp portion of this region (chr15:18160001-19763477) ranges from 2-11 copies among individuals sampled. The individuals at the extremes of this range thus differ by ~8.1 Mbp of euchromatic DNA harboring brain- and testis-expressed genes.

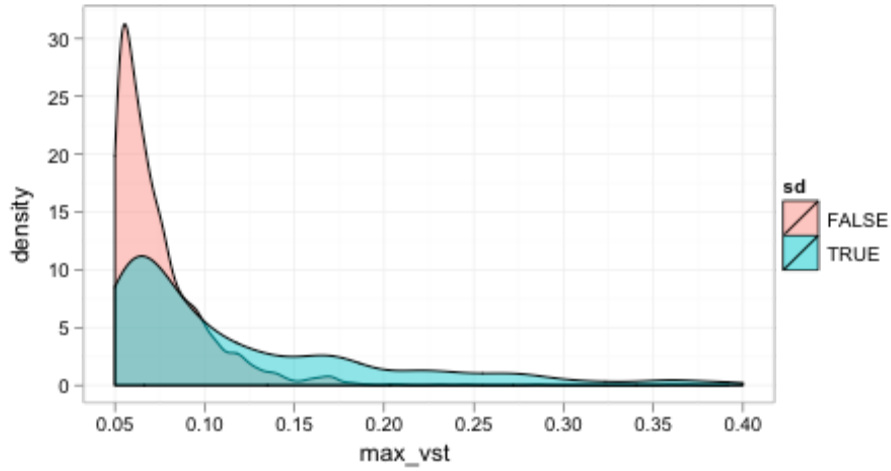


Figure S35. Distribution of the V_{st} statistic for genes with a $V_{st} > 0.05$ classified by segmental duplication (sd) overlap. Genes of increased V_{st} are enriched for segmental duplications.

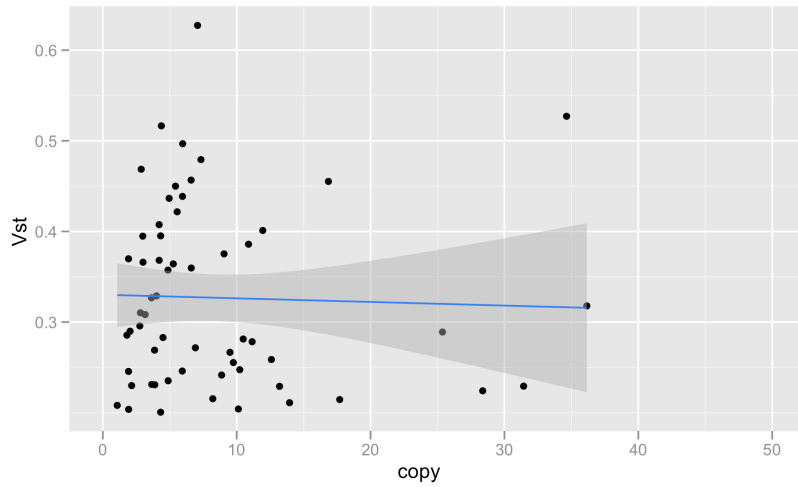


Figure S36. V_{st} plotted against copy. No correlation was observed between V_{st} and copy ($r^2 = 0.01$, $P = 0.2$).

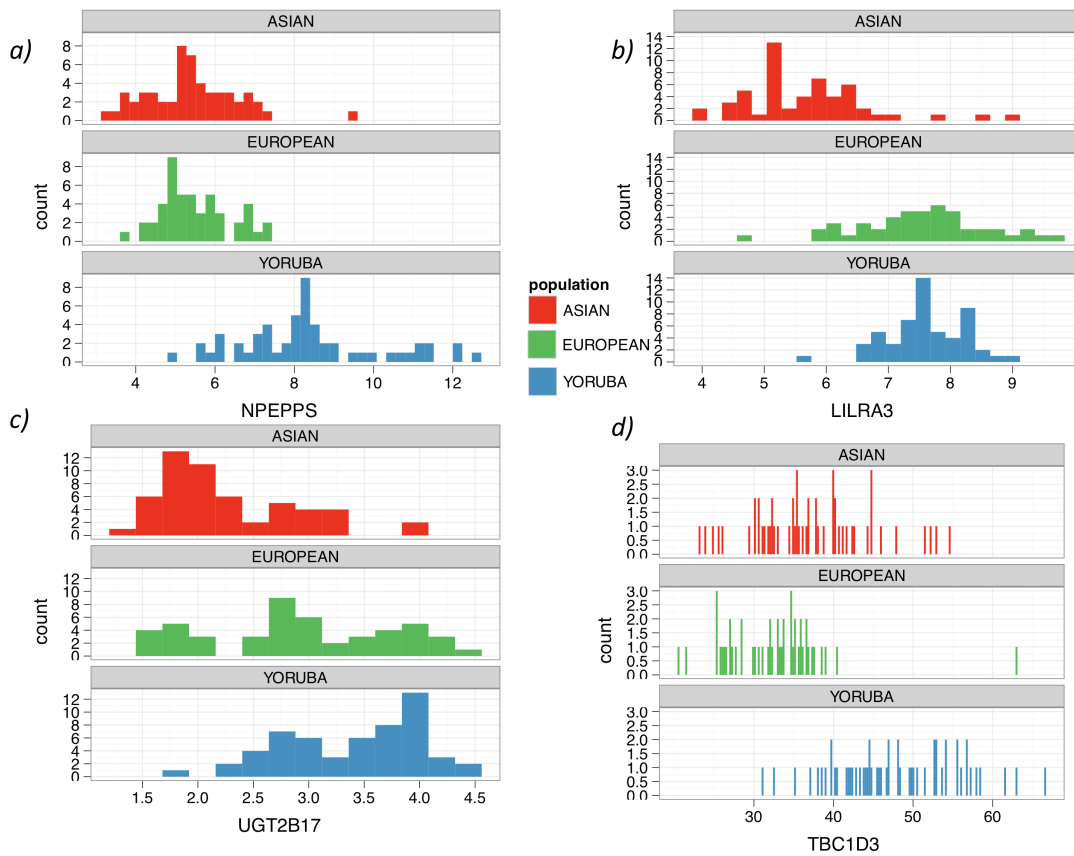


Figure S37. Population-specific copy number distributions for four genes with extreme population stratification of copy number.

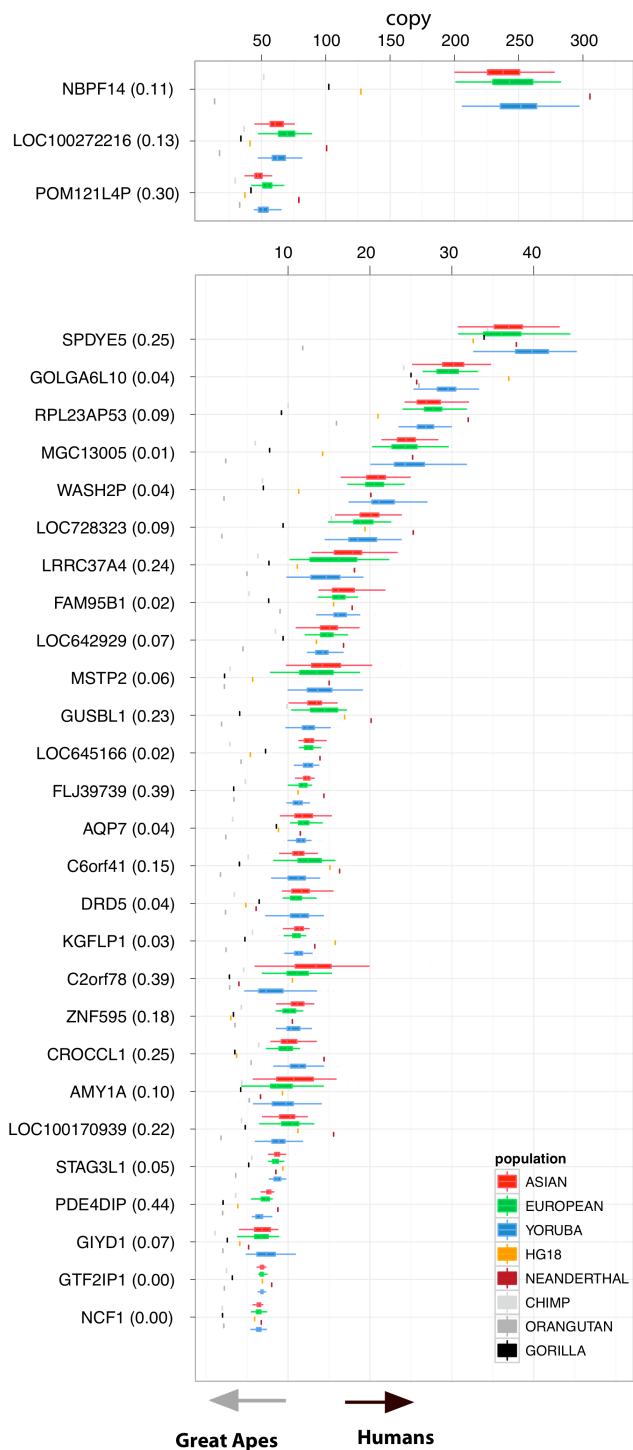


Figure S38. Box plots of 30 gene families identified as copy number expanded within the human lineage. Among these genes we identified several implicated in brain function, including *DRD5* and *GTF2L1P1*. Additionally, we confirm known human-specific expansions such as those at the amylase gene cluster (S34).

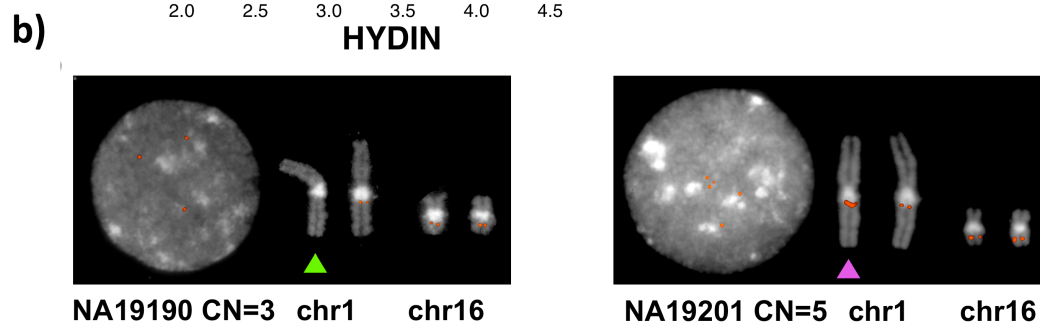
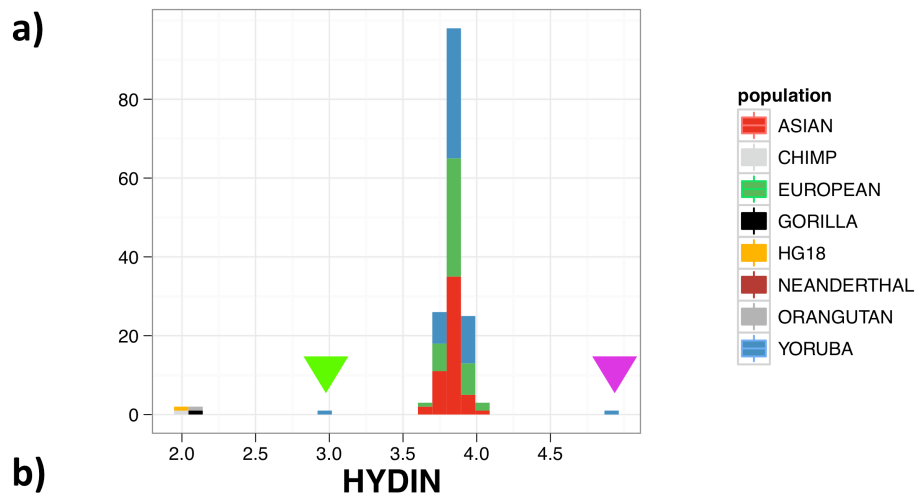


Figure S39. a) Copy number histogram of the *Hydin* gene which is specifically duplicated in the human lineage. b) Two rare *Hydin* duplications/deletions are confirmed by FISH

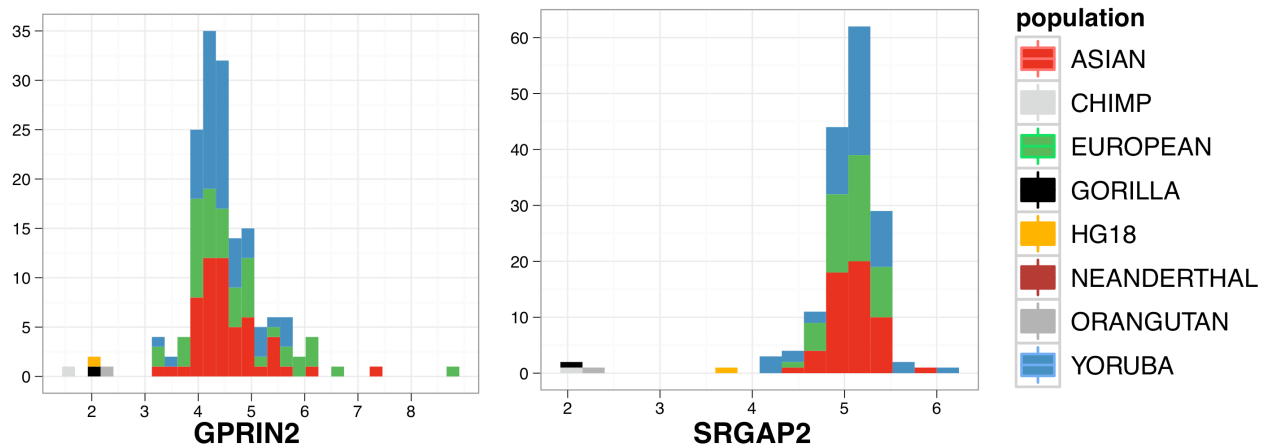


Figure S40. Copy number histograms of the human specific duplicated genes *GPRIN2* and *SRGAP2*

copy1 ATGCTAGGCATATAATATCCGACGATATACATATAGATGTTAG...
copy2 ATGCTAGGCATAGAATATCCGACGATATACATATACATGTTAG...
copy3 ATGCTACGCATAGAATATCCACGATATACATATACATGTTAG...
copy4 ATGCTACGCATATAATATCCGACGATATACATATACATGTTAG.

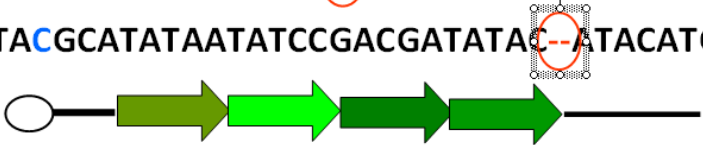


Figure S41. Schematic depicting singly unique nucleotide (SUN) positions used to distinguish one paralog among highly identical duplicates.

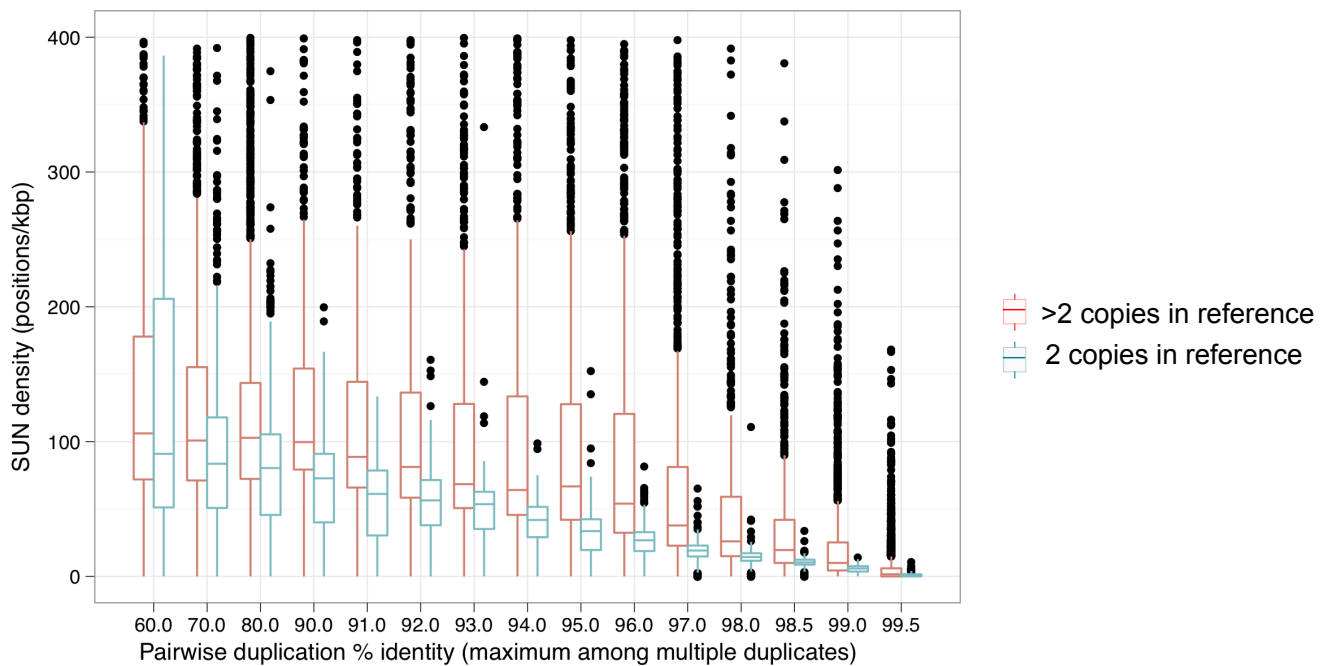


Figure S42. Density of SUN positions (average per 1 kbp) within segmental duplications as a function of duplication percent identity. Segmental duplications were divided into nonoverlapping windows of equal content (2500 bp unmasked sequence, variable physical width). Within each window, the number of SUNs was counted. Local percent identity was computed by extracting each window from the pairwise global alignment(s) between that window and the one or more duplicate copies elsewhere in the genome. When a given window has more than one duplicate locus, the highest percentage pairwise identity between it and its multiple paralogs is taken. Duplicated windows are stratified based on whether they are duplicated exactly once (blue) or at multiple copies in the reference genome (red).

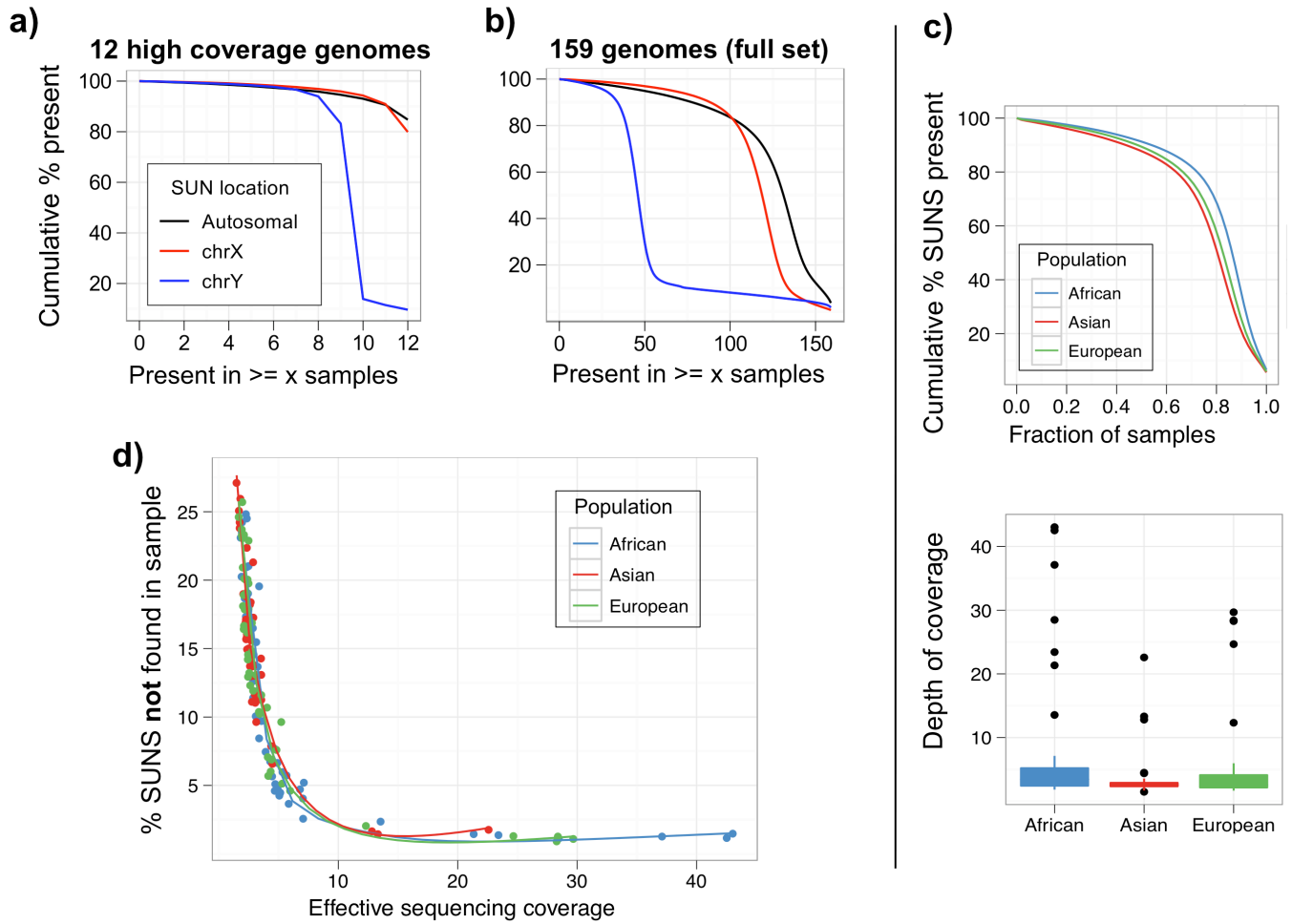


Figure S43. Fraction of SUN positions found within a) genomes of 12 unrelated humans sequenced to high coverage (each >10X, mean = 25.3X) and b) the full set of 159 genomes, including 144 of lower coverage (mean = 3.1X). c). Proportions of SUNs observed among each of the three populations are similar, with a slight excess of markers observed among the African genomes, which were the most deeply sequenced. d). Lower coverage genomes show higher rates of SUN absence to a similar extent among the three populations sampled.

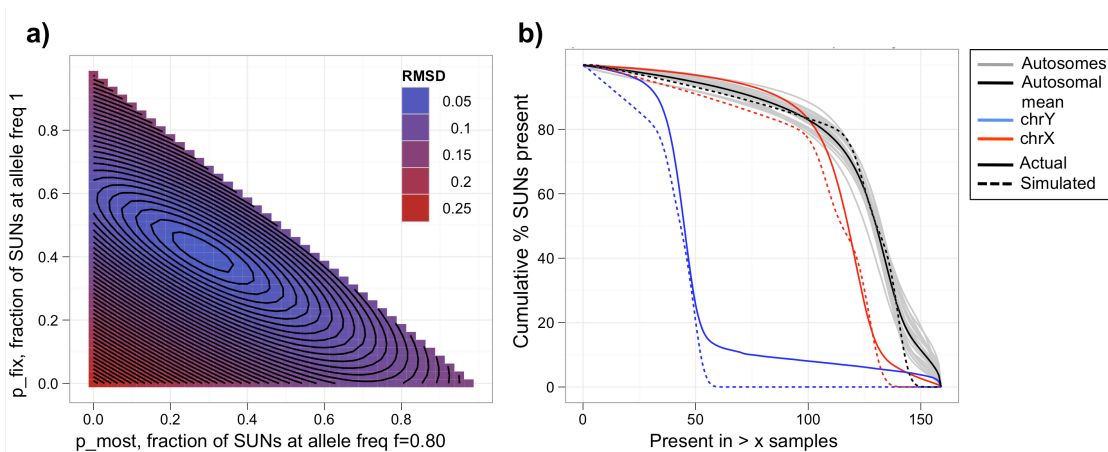
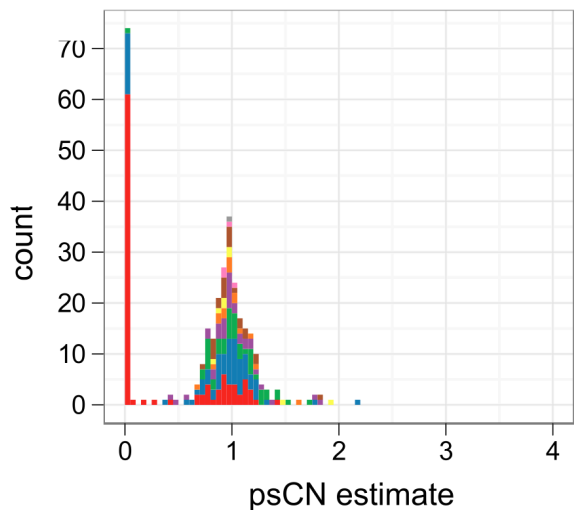


Figure S44. Simulation of SUN presence/absence. a) RMSD between simulated and observed rates of SUN presence, shown for varying p_{fix} and p_{most} with $f=0.800$. RMSD is minimized at $(p_{fix}, p_{most}, f) = (0.425, 0.275, 0.800)$. b). Simulated and observed rates of SUN presence across 159 samples using best-fit parameters.

a. Not overlapping segdups (n=310)



b. Overlapping segdups (n=73)

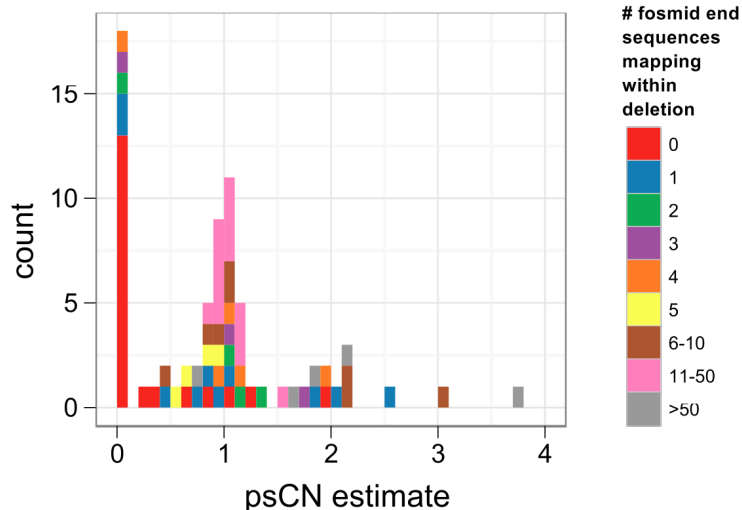


Figure S45. Predicted psCN of validated deletions within the same individual where each deletion was initially mapped. Bars are colored by the number of other fosmid end sequences mapping within each deletion interval (excluding the clone that was sequenced to resolve the deletion). The majority of deletions called as homozygous (psCN = 0) are overlapped by no other clone end sequences (red). a) Of the 310 validated deletions not overlapping segmental duplications, 81 (26.1%) are homozygous (psCN<0.5), 220 (71.0%) are hemizygous (0.5<=psCN<1.5), and 9 (2.9%) are called as diploid or greater (psCN>=1.5). b) Of the 73 deletions overlapping segmental duplications, 22 (30.1%) are homozygous deleted, 37 (50.7%) are hemizygous deleted, and 14 (19.2%) are diploid or greater, the latter possibly being nested within larger blocks of amplification.

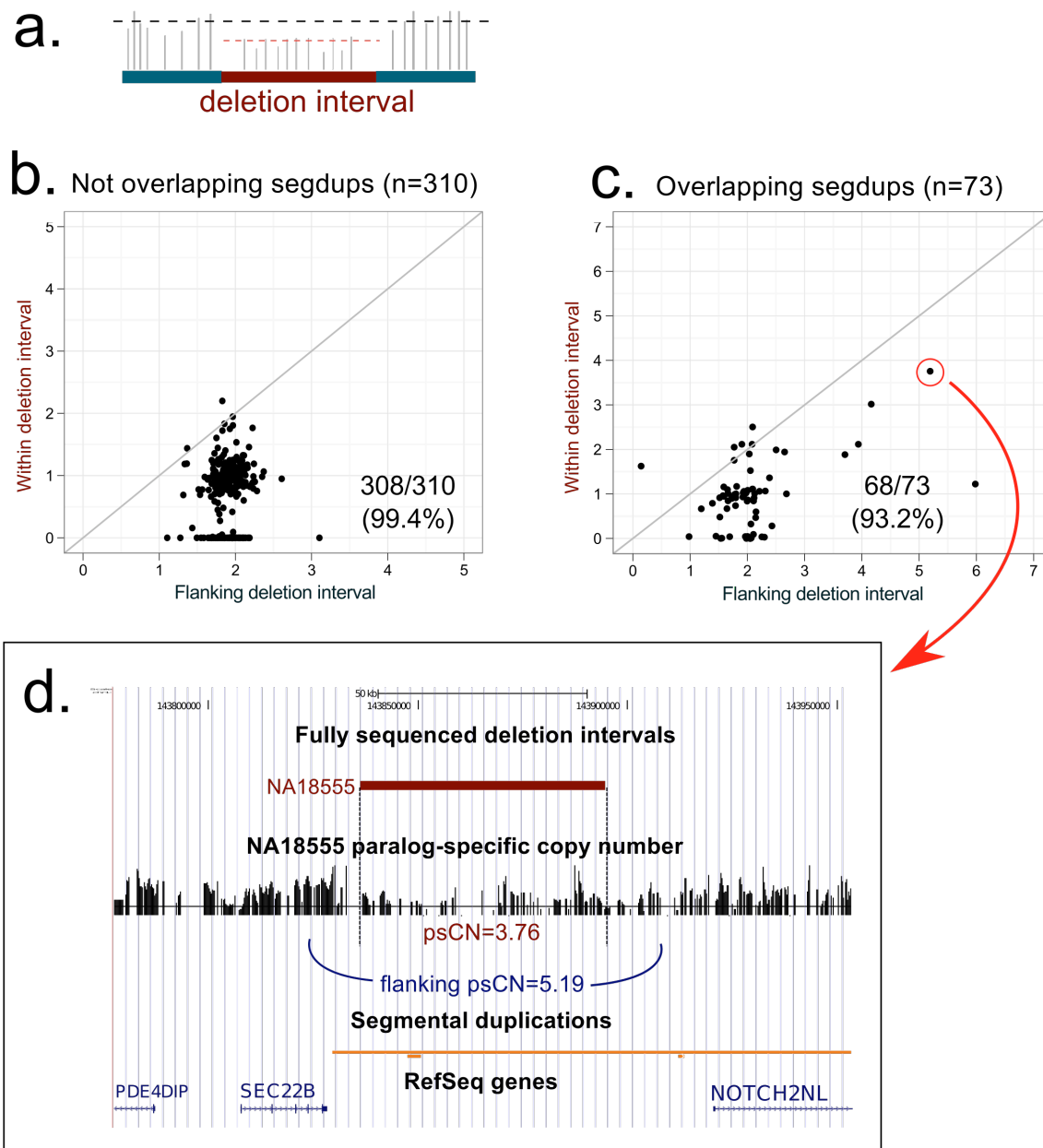


Figure S46. Comparison of psCN estimates within deletion intervals to flanking regions. a) psCN is compared between the deletion interval and for 10 kbp of flanking sequence. If genotyping is accurate, the psCN estimate in the deletion interval should be lower than that of the corresponding flanks. b) For 308/310 deletions that do not overlap segmental duplications, the psCN estimate within the deletion interval was less than that of the flanking regions (below the diagonal). c) For 68/73 sequenced deletions overlapping segmental duplications, the deletion interval shows lower psCN than the flanking regions. d) For some deletions, such as this example on 1q21.1, the psCN estimate is ≥ 2 and depressed relative to flanking regions, indicating that the deletion is within an amplified region.

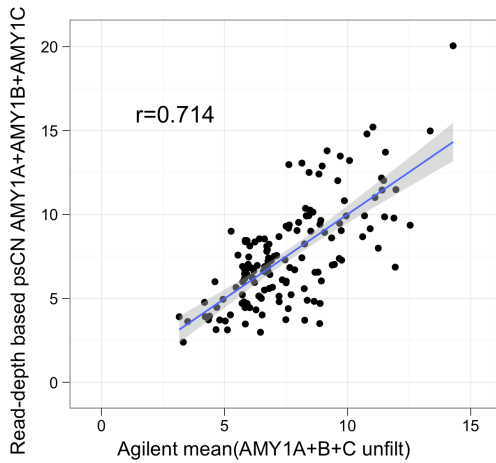


Figure S47. Comparison across 145 individuals of aggregate psCN genotypes of *AMY1* paralogs with estimates obtained from single-channel Agilent microarrays (S6).

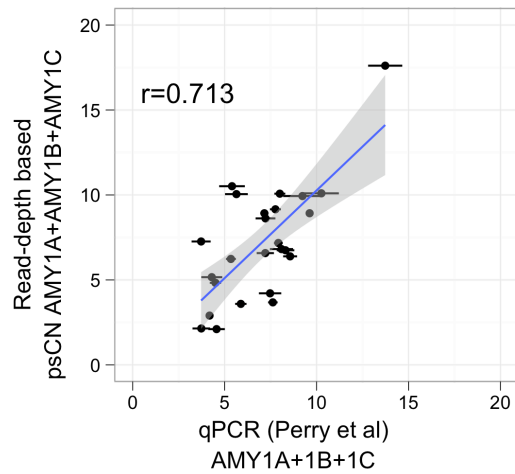


Figure S48. Comparison across 25 Japanese individuals of aggregate psCN genotypes of *AMY1* paralogs with estimates obtained by quantitative PCR directed at the three functional *AMY1* copies (S34). Error bars, s.e.m.

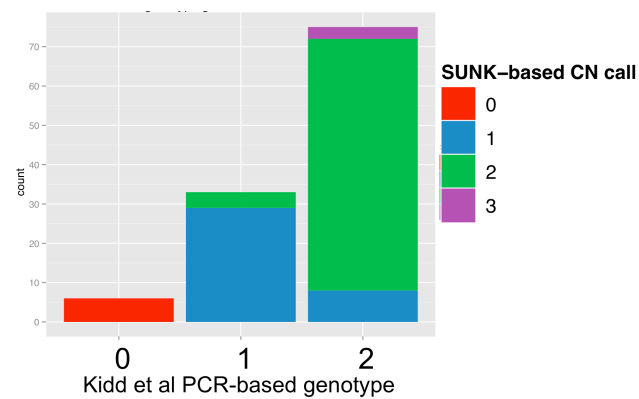


Figure S49. Comparison across 114 individuals of psCN genotypes and breakpoint PCR-based genotypes for *APOBEC3B* deletion.

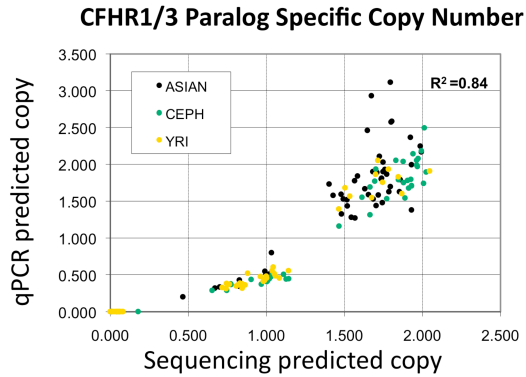


Figure S50. Confirmation of psCN genotyping of *CFHR3* by a paralog-specific qPCR assay.

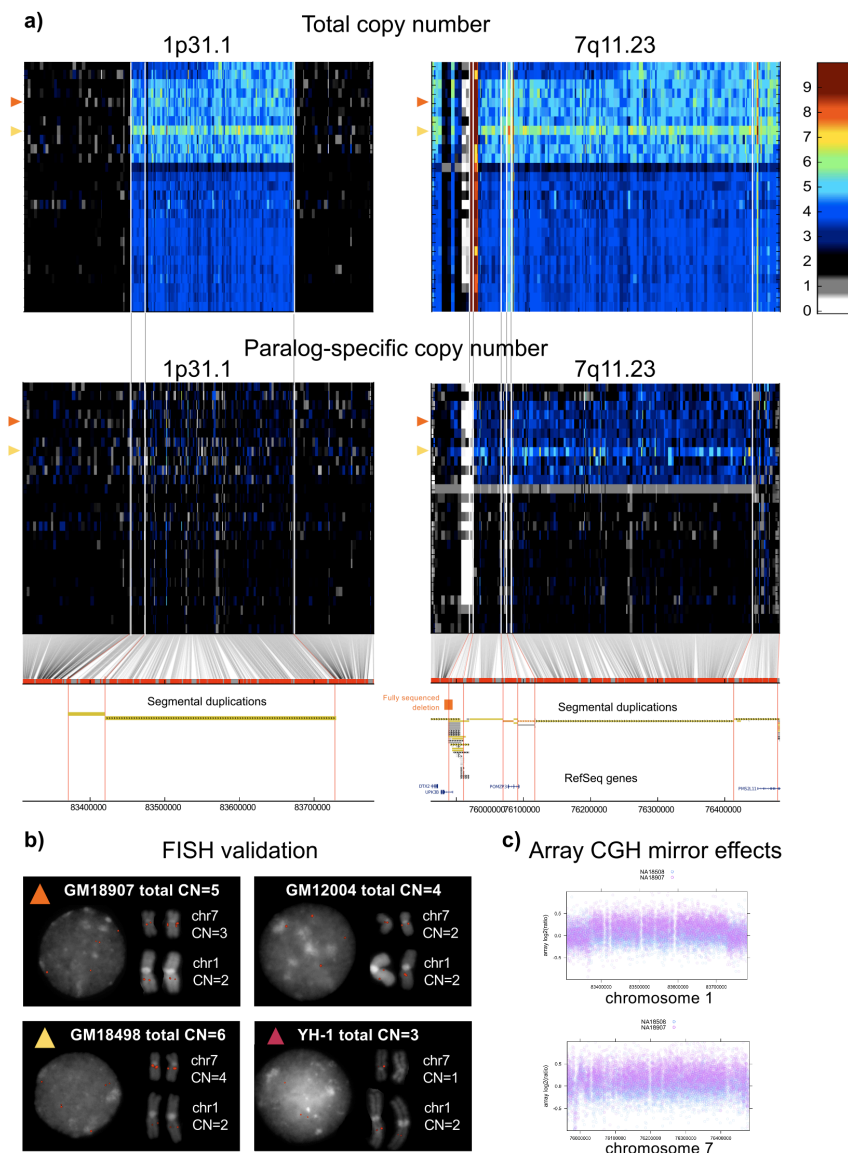


Figure S51. Resolving the true sites of copy number variation with paralog-specific genotyping. a) Apparent CNV within a ~360-kbp segmental duplication on chromosomes 1 and 7 detected by total copy number mapping. Paralog-specific copy number mapping reveals this variation is exclusive to chromosome 7. b) FISH analysis of

interphase nuclei confirms the total copy number change, and analysis of metaphase nuclei confirms the variable copy is on chromosome 7. c) Array CGH (S5) shows a similar effect at both loci in one of the copy-number amplified individuals (NA18907, purple relative to the individual without this amplification, NA18508, blue).

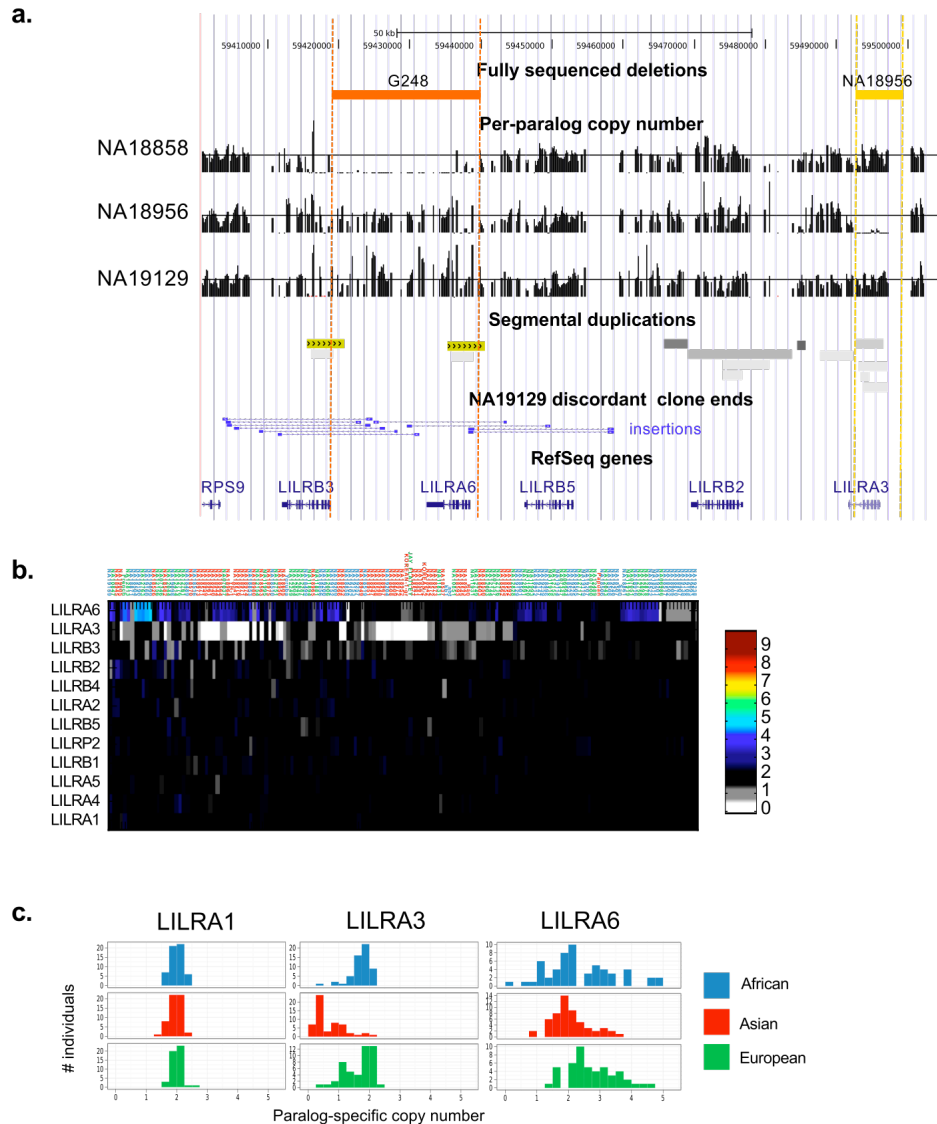


Figure S52. Paralog-specific copy number variation among *LILR* genes

a) Genomic view of paralog-specific copy number at one of the *LILR* gene clusters. Known, fully-sequenced deletions covering *LILRA6* and *LILRA3* are recapitulated, and a reciprocal, two-fold amplification of *LILRA6* supported by fosmid end sequence pair data is found in sample NA19129. b) Gene family heat map of paralog-specific copy number, with genes on the rows and hierarchically clustered individuals on the columns. c) Population comparison of *LILR* family CNV.

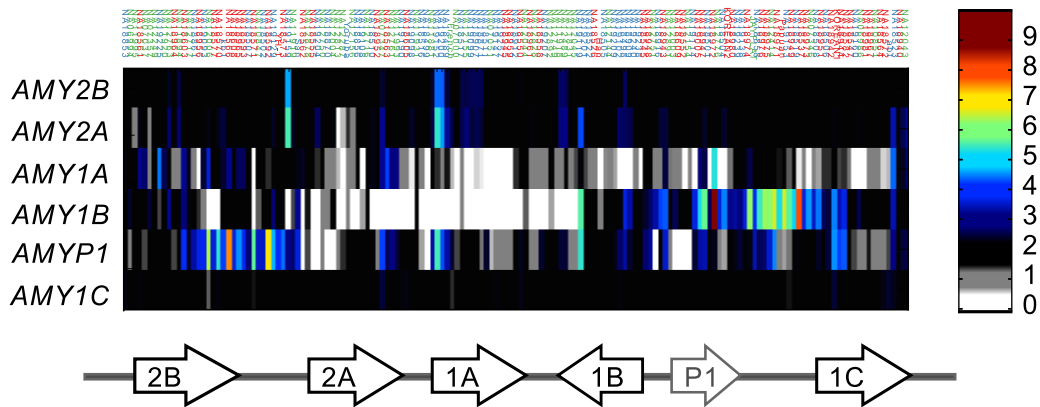


Figure S53. Copy number variation of the both pancreatic and salivary amylase genes shows that the pseudogene *AMY1P* and salivary genes *AMY1A* and *AMY1B* are much more variable in copy when compared to pancreatic amylase genes, *AMY2A* and *AMY2B*, especially among Asians. Expansion of these loci represents a potential adaptation to digestion of starch-rich diets among human populations (S34).

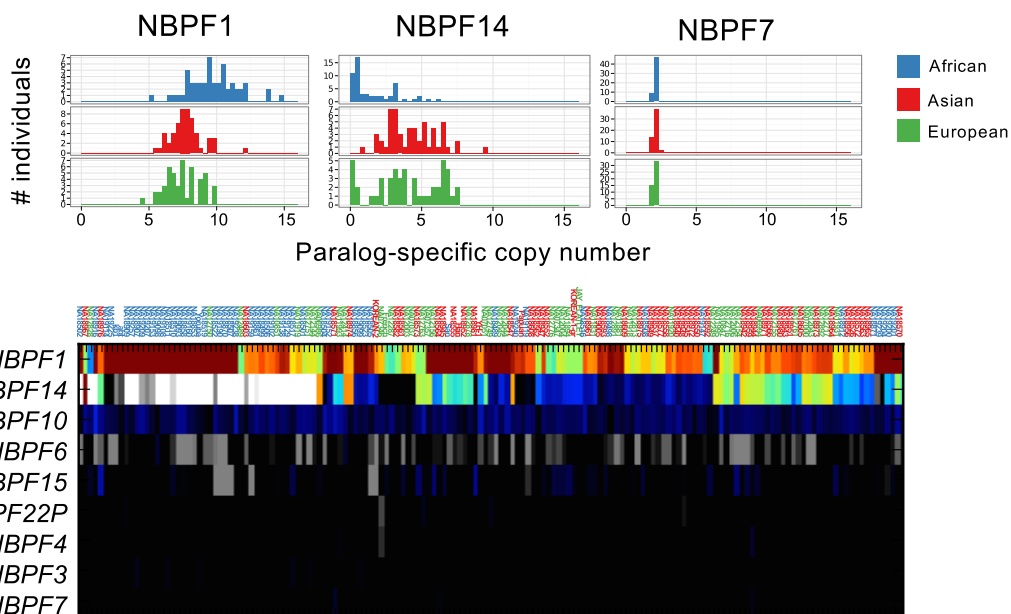


Figure S54. Population stratification and paralog-specific copy variability of a human expanded gene family of unknown function, *NBPF* (neuroblastoma breakpoint gene family). Certain paralogs (e.g., *NBPF1*) are highly amplified, extremely variable, and population stratified while others are nearly fixed and diploid (e.g., *NBPF7*).

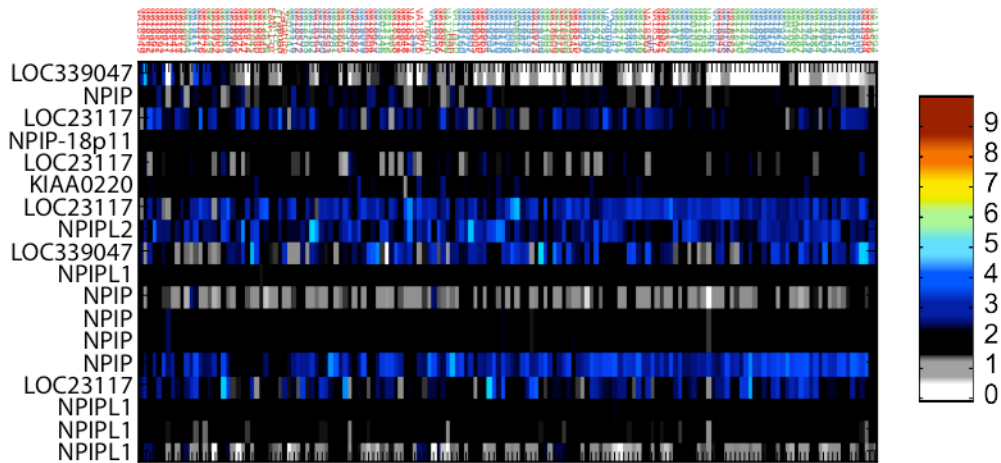


Figure S55. *NPIP* gene family. Each row represents a distinct paralog (e.g., *NPIP*, for which 5 distinct loci on chromosome 16 are shown). Individual genomes are hierarchically clustered (names shaded by population) and shown as columns.

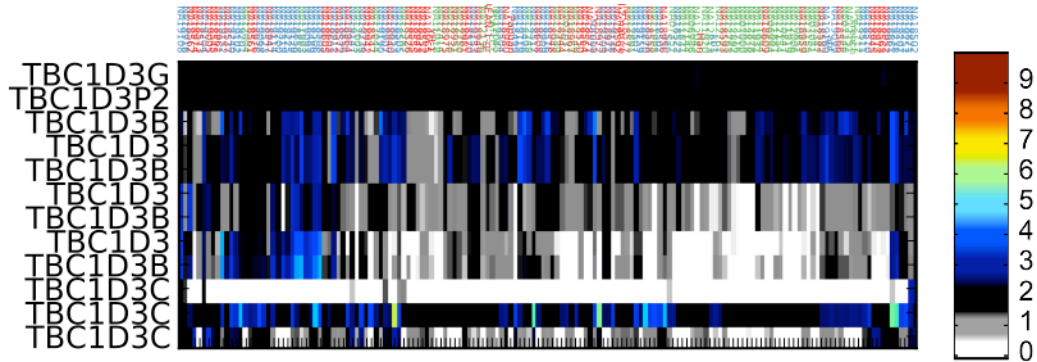


Figure S56. *TBC1D3* gene family.

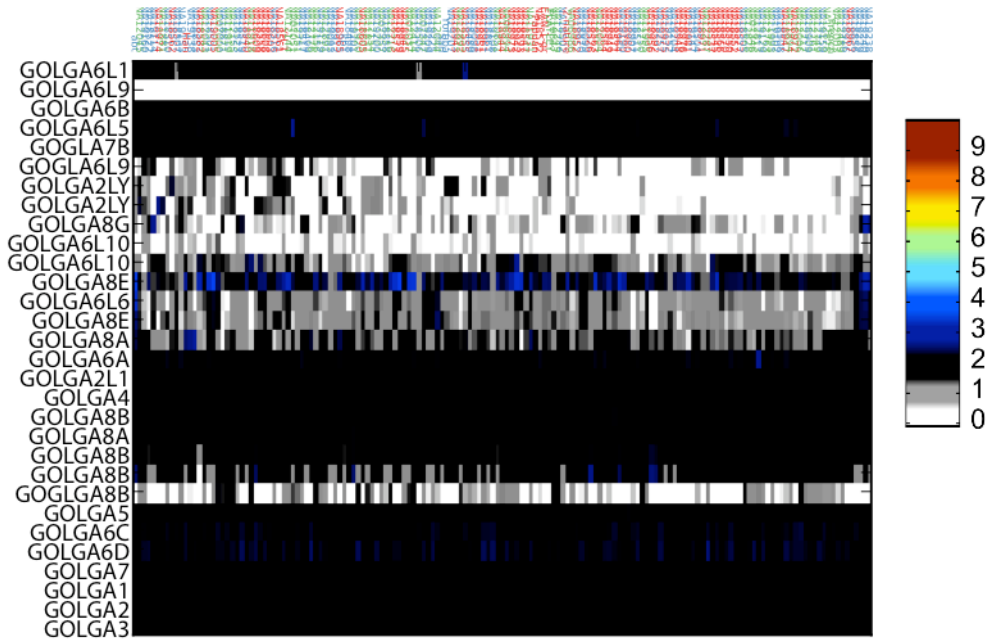


Figure S57. *GOLGA* gene family.

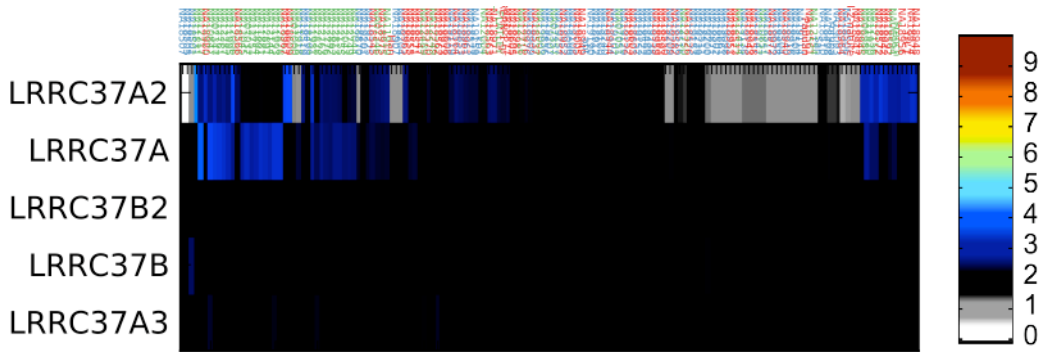
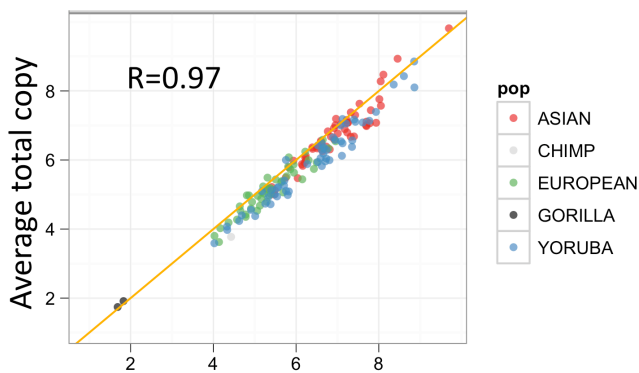


Figure S58. *LRRC37* gene family.



Sum of paralog specific copy numbers

Figure S59. The duplicated gene *ESPN* and the paralogous pseudogene *ESPNP*. Average total copy number estimated is plotted against the sum of the paralog-specific copy numbers of each gene. Orange line denotes $y=x$.

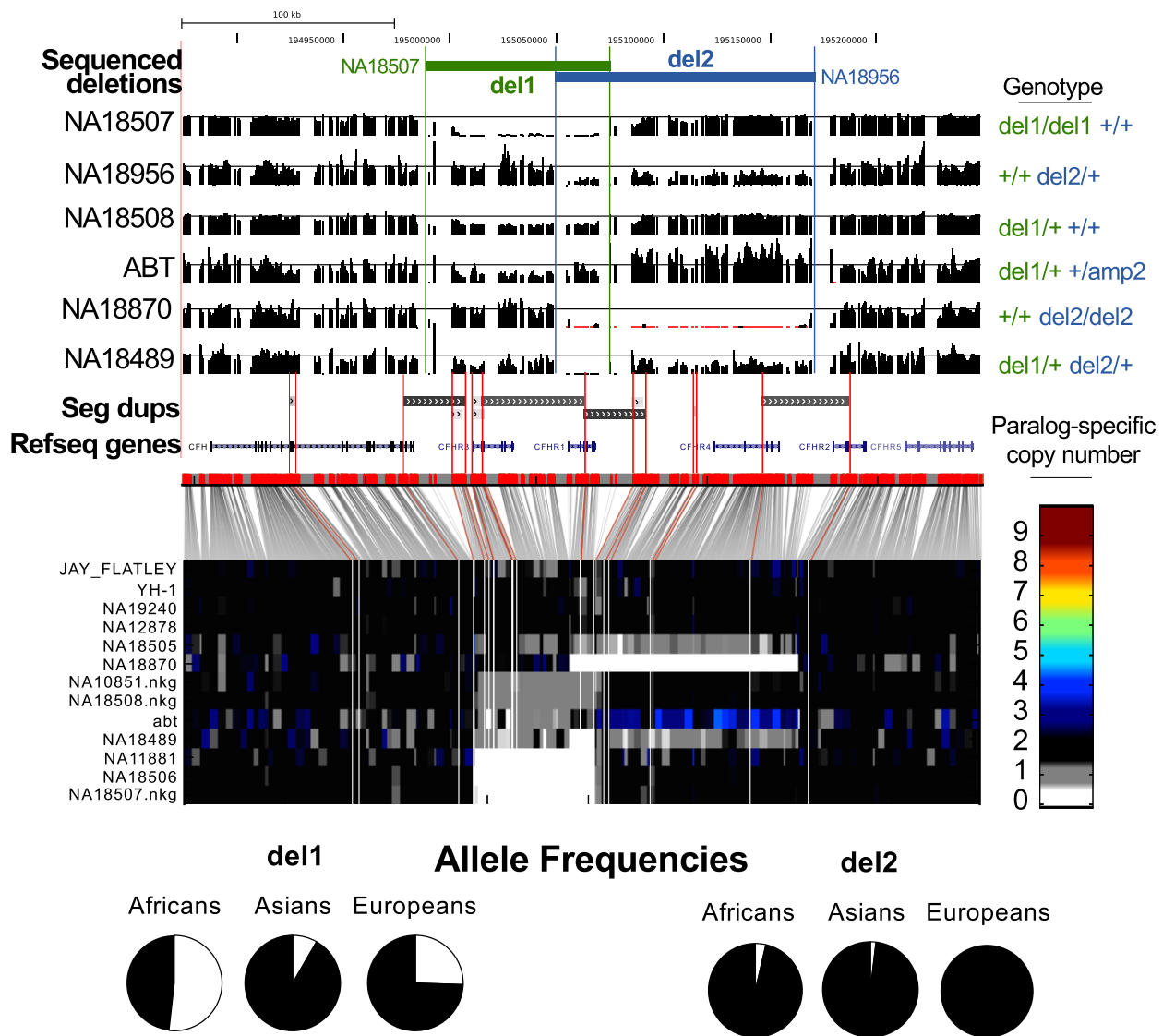
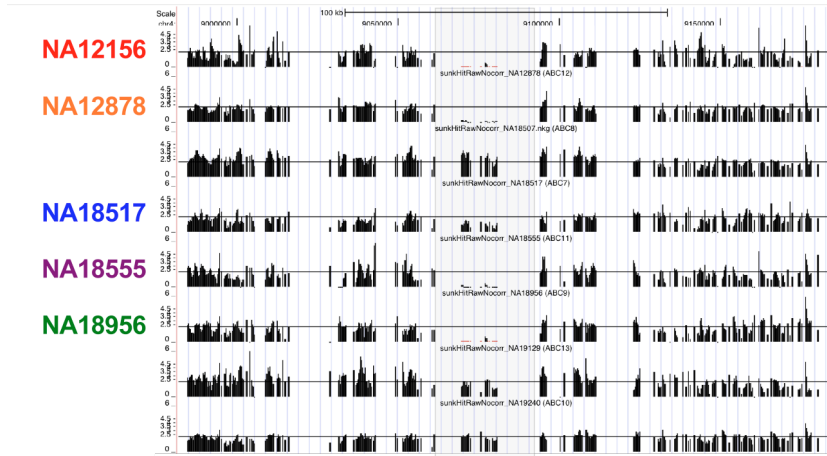
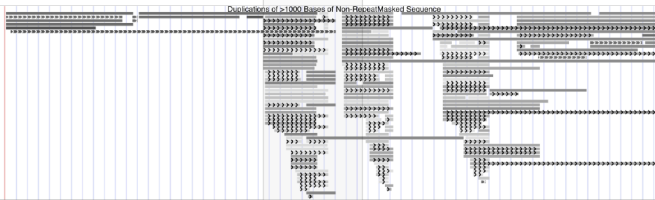


Figure S60. Paralog-specific genotyping of the complement factor H receptor (CFHR) gene family on chromosome 1q31.3. SUN read depth (histograms) across the region delineates the boundaries of two known deletions as defined by clone sequencing. Genotyping paralog-specific copy number reveals population stratification in allele frequencies both for del1 (CFHR3/1) and the less common del2 (CFHR1/4). Rarer states, such as the reciprocal duplication of del2 found only in Archbishop Desmond Tutu (ABT) are also detected.

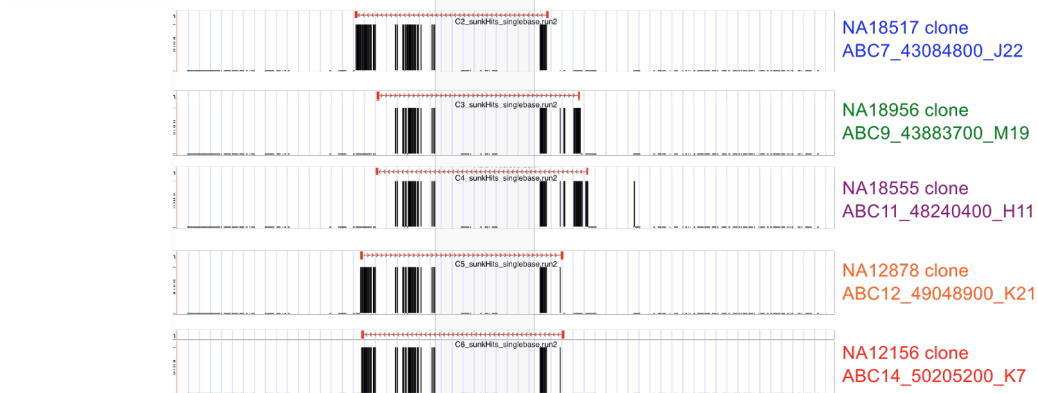
Paralog specific copy number (baseline: ppCN=2)



Segmental duplications



Selected clones (end-sequence pair mapping) and Illumina Unique SUNK hits



RepeatMasked elements



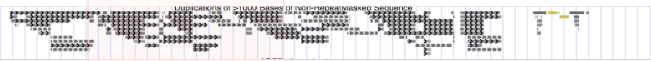
Figure S61. Deletion on chromosome 4p16 confirmed in 5/5 individuals based on fosmid resequencing.

Paralog specific copy number (baseline: ppCN=2)

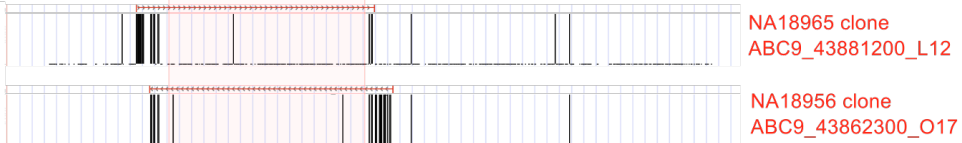


NA18956

Segmental duplications



Selected clones (end-sequence pair mapping) and Illumina Unique SUNK hits



NA18965 clone
ABC9_43881200_L12
NA18956 clone
ABC9_43862300_O17

RefSeq genes



Figure S62. Deletion within the *PSG* gene cluster on chromosome 19q13.31 confirmed in NA18956 (2/2 clones).

Paralog specific copy number (baseline: ppCN=2)

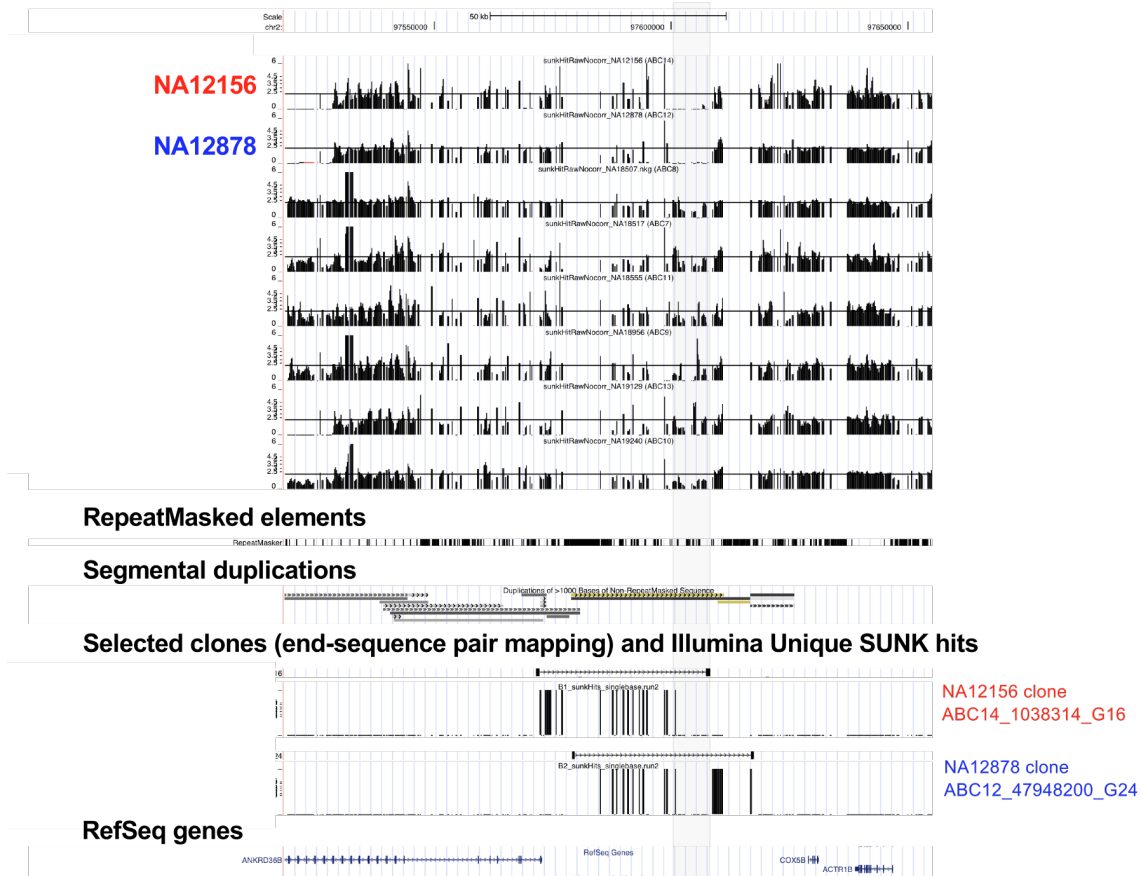


Figure S63. Small (~8-kbp) deletion upstream of *ANKRD36B* confirmed in 2/2 individuals.

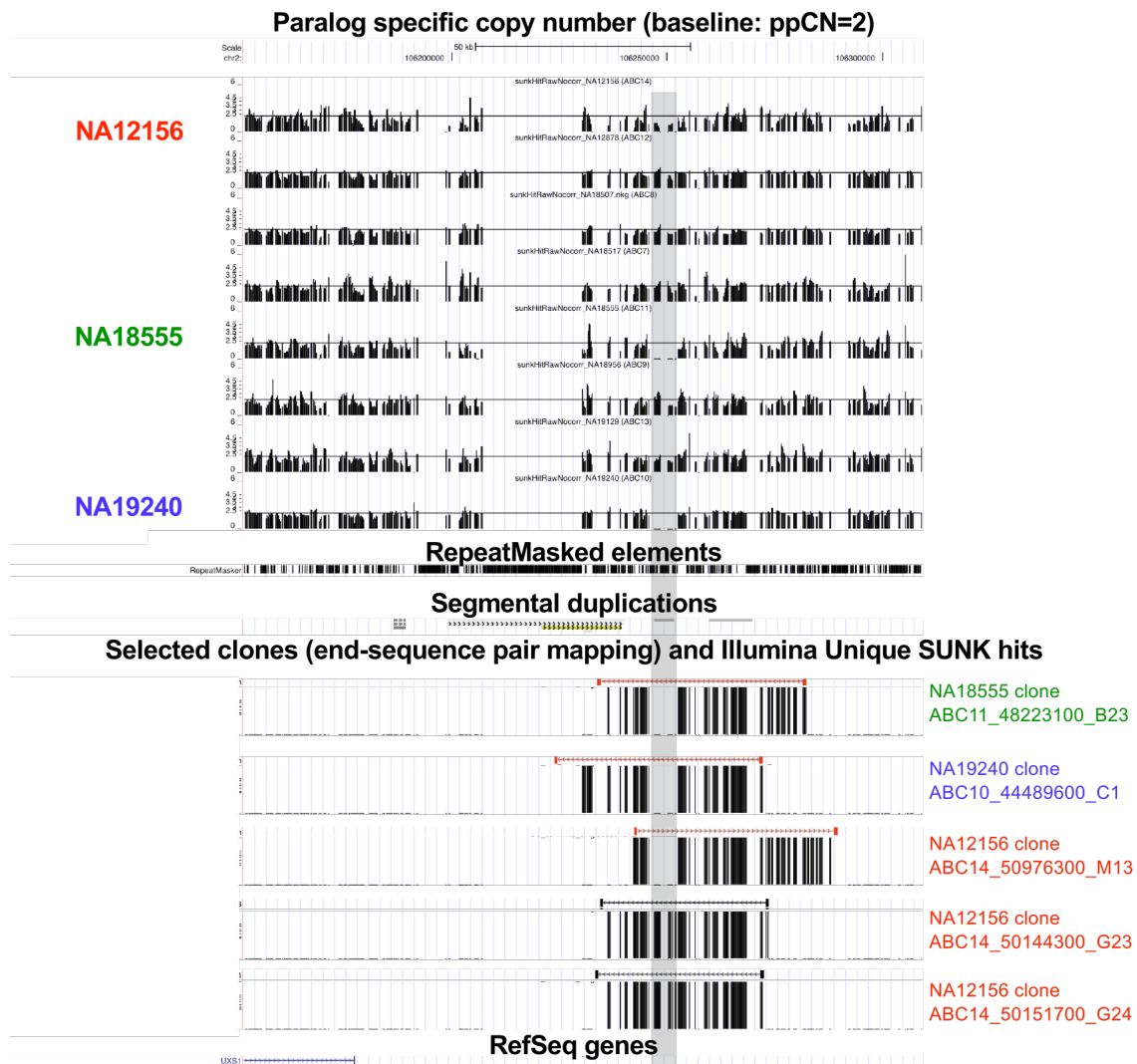


Figure S64. Deletion on chromosome 2q12.2 confirmed in 3/3 samples (deletion allele confirmed in 3/3 clones, nondeletion allele in 2/2).

Paralog specific copy number (baseline: ppCN=2)

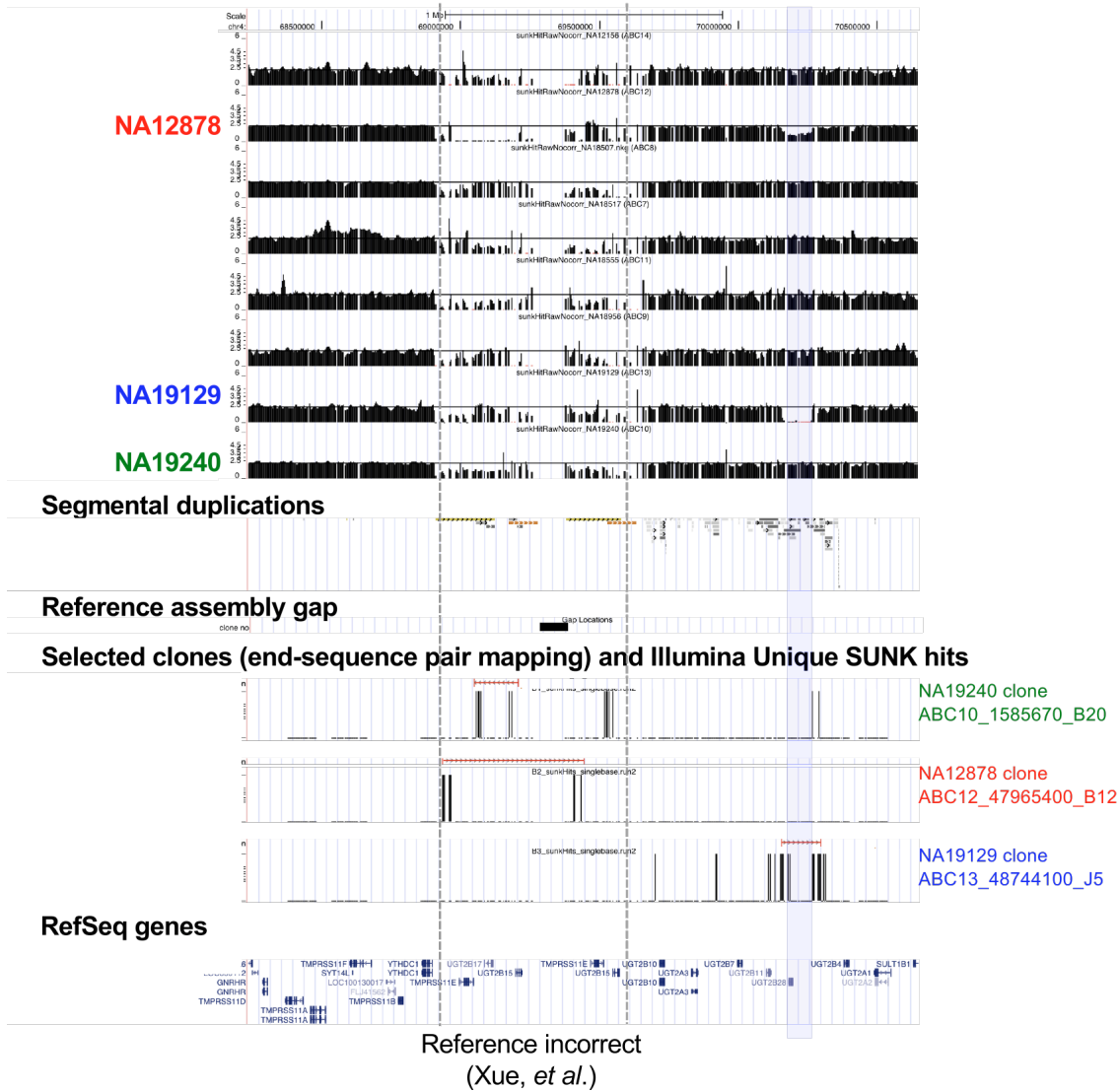
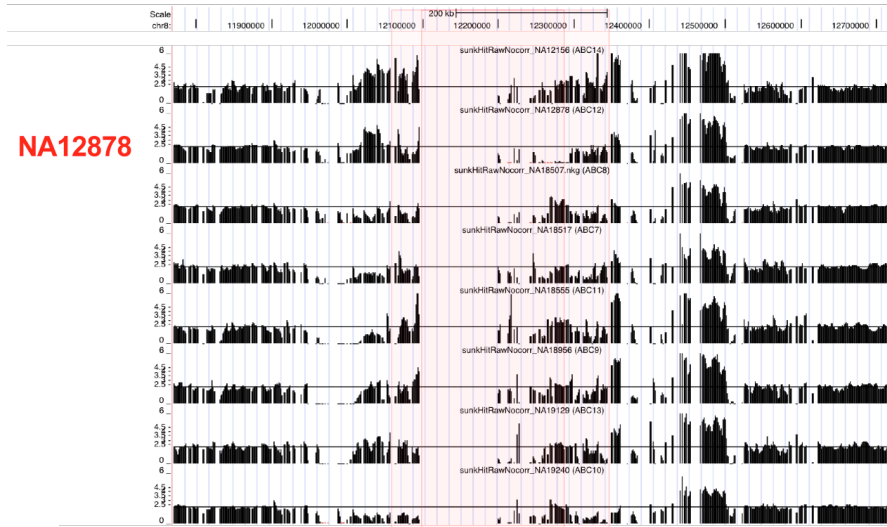


Figure S65. Deletion of *UGT2B28* on chromosome 4q13 confirmed in NA19129. Proximal deletion clones and whole-genome SUNK data reflect apparent deletion generated by spurious segmental duplication in the reference assembly (S31).

Paralog specific copy number (baseline: ppCN=2)

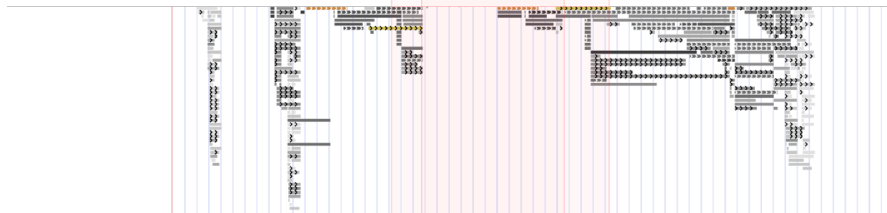


NA12878

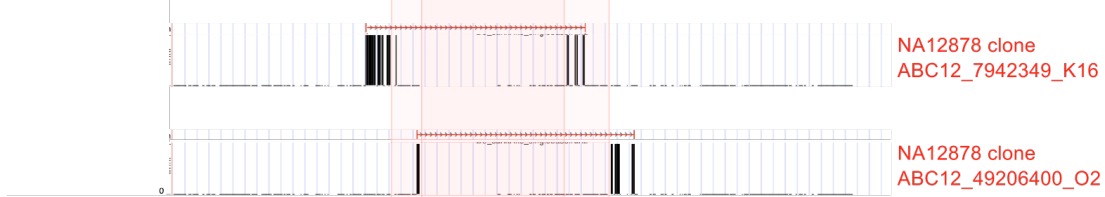
Assembly Gaps



Segmental duplications



Selected clones (end-sequence pair mapping) and Illumina Unique SUNK hits



RefSeq genes



Figure S66. Nested deletions of defensin genes on chromosome 8p23.1 confirmed in NA12878.

Paralog specific copy number (baseline: ppCN=2)

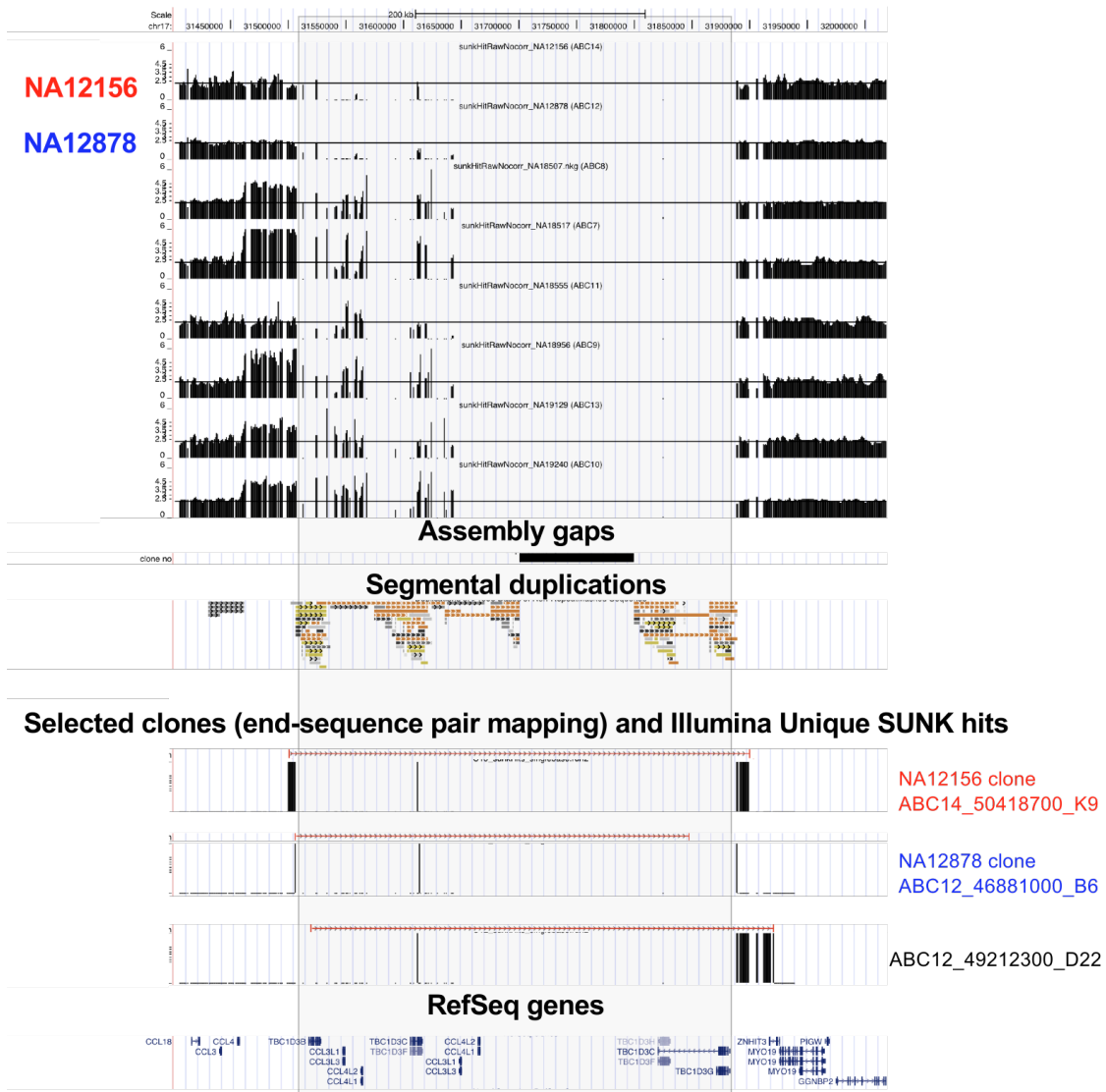


Figure S67. Nested deletions of chemokine ligand genes on chromosome 17q12 confirmed based on fosmid sequencing individuals NA12878 and NA12156.

Paralog specific copy number (baseline: ppCN=2)

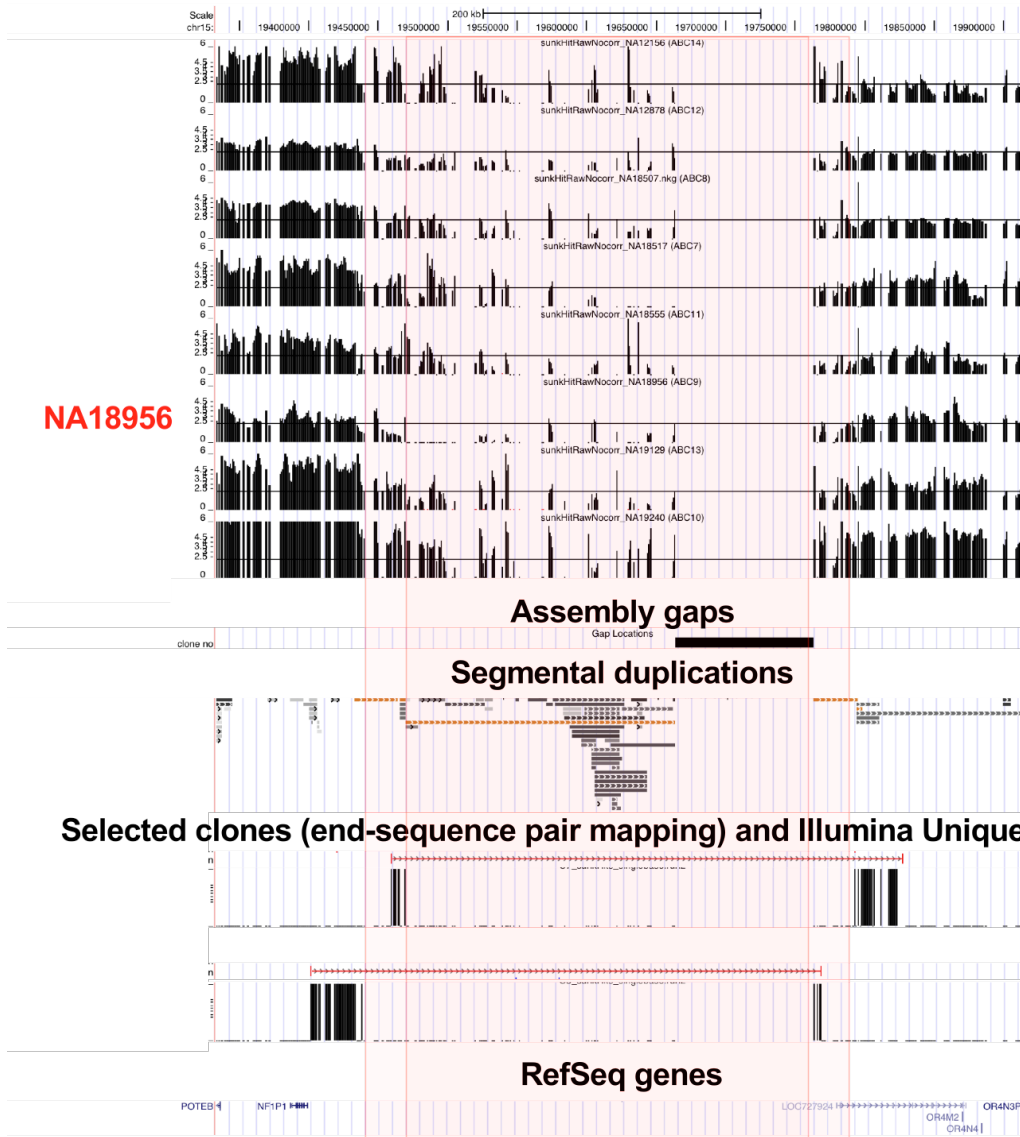


Figure S68. Nested deletions on chromosome 15q11.2 confirmed in NA18956.

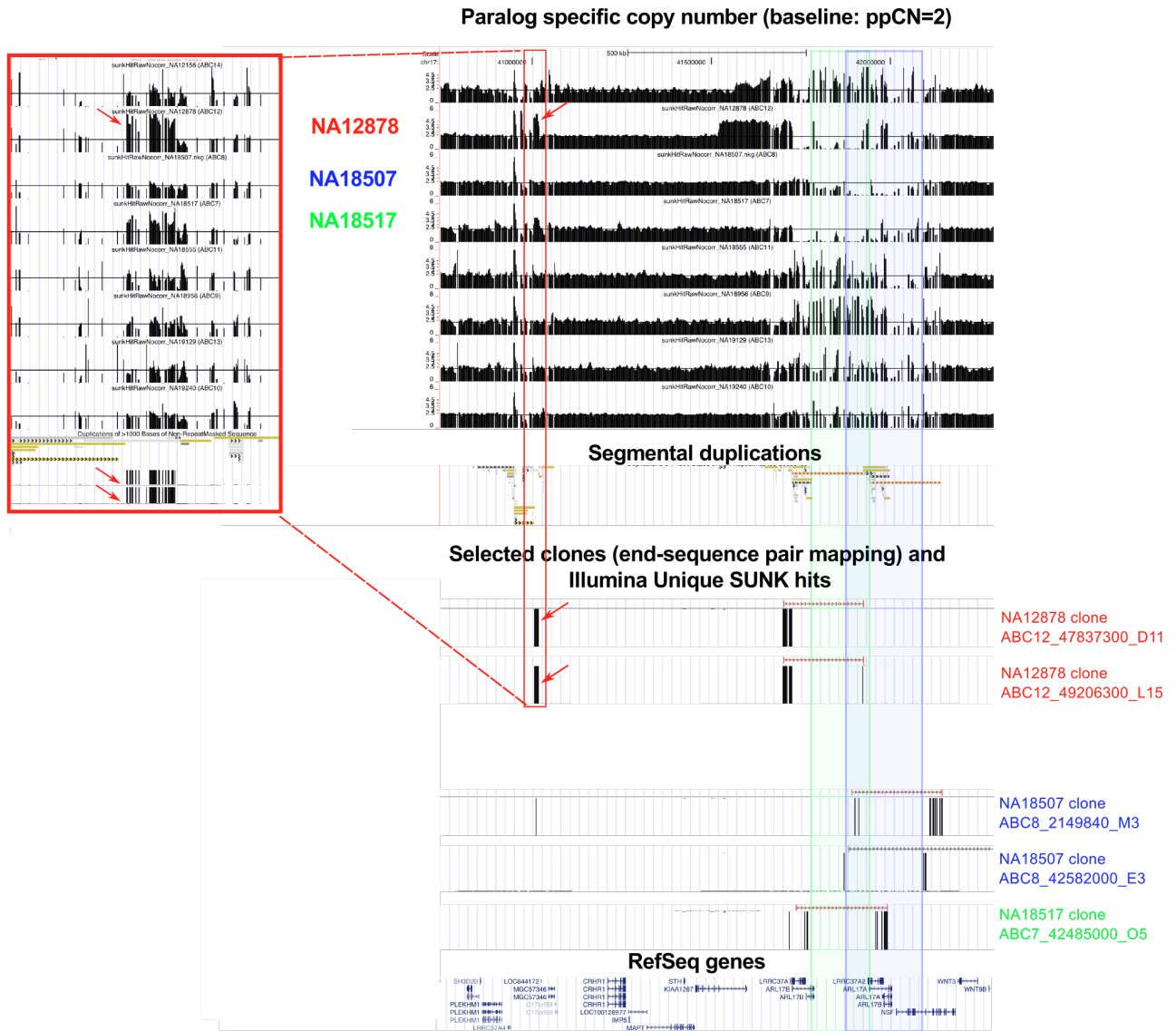


Figure S69. Deletions on 17q21.31 confirmed in 2/3 individuals. In the other individual, NA12878, 2/2 clones suggest this deletion is nested within an amplified, alternative structural configuration including a segmental duplication found in the reference genome ~600 kbp proximal to the end sequence mapping location (inset).

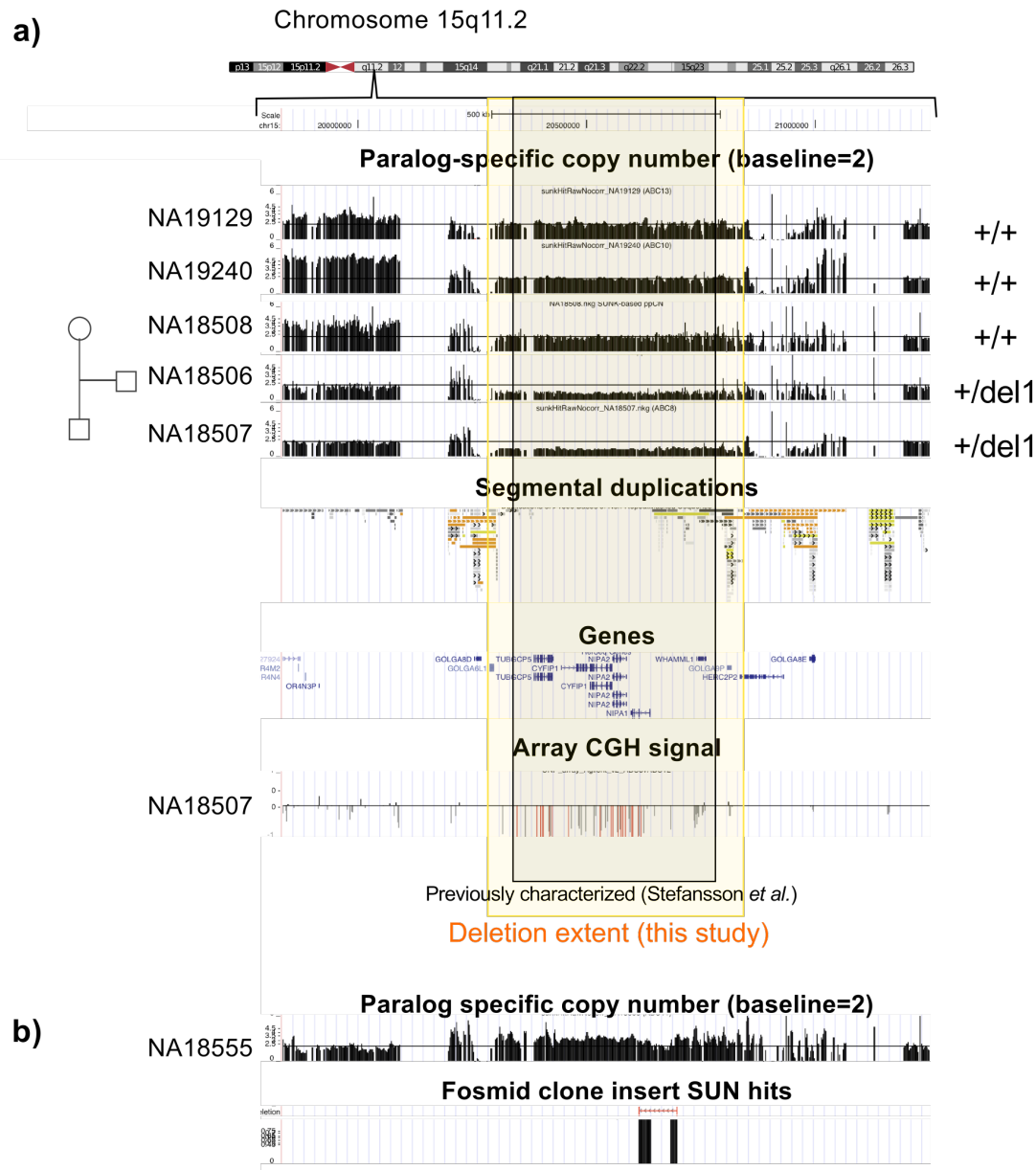


Figure S70. Breakpoint refinement at 15q11.2 schizophrenia-associated deletion

a) Two individuals carry a deletion previously associated with schizophrenia (S36). Paralog-specific copy number tracks show the true extent of this deletion into flanking segmental duplications, encompassing additional genes relative to its previously characterized boundaries. This deletion is inherited by NA18506 and is confirmed in the father (NA18507) by array CGH and FISH. b) A novel deletion within the reciprocal single-copy amplification is confirmed by fosmid resequencing in individual NA18555.

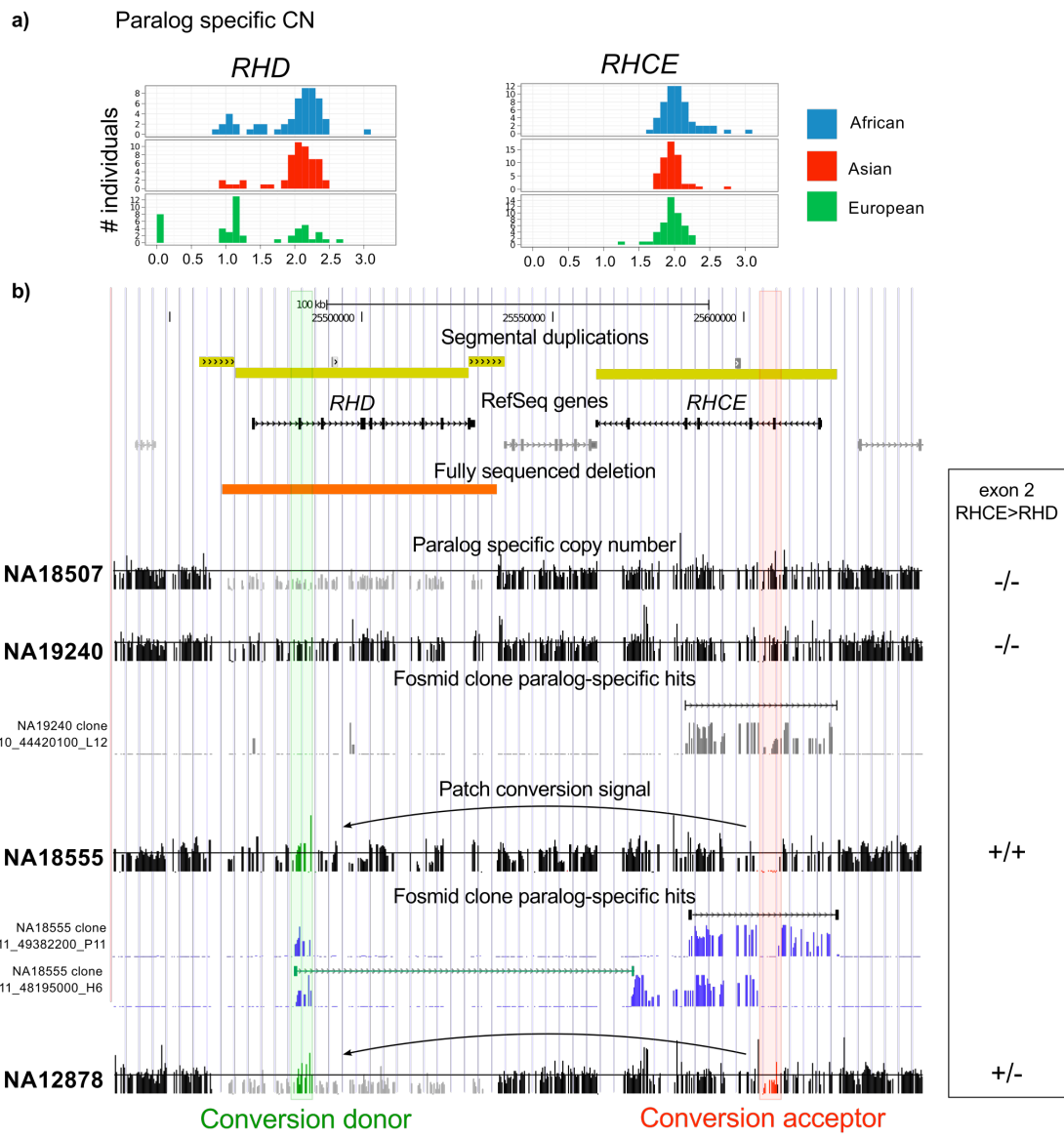


Figure S71. Signatures for gene conversion at the Rh blood group locus
 Rhesus (Rh) blood group genes *RHD* and *RHCE* lie within tandem segmental duplications in inverted orientation on chromosome 1. a) *RHD*, but not *RHCE*, is commonly deleted among Europeans. b) Signatures of patch gene conversion detected using whole-genome psCN maps and validated by fosmid clone insert short read sequencing.

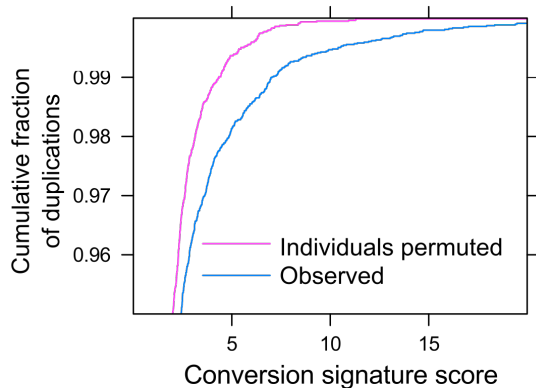


Figure S72. Significant excess of conversion score ($P < 2.2 \times 10^{-16}$) relative to permuted controls suggests widespread signatures of recent interlocus gene among duplicate alleles segregating within the population.

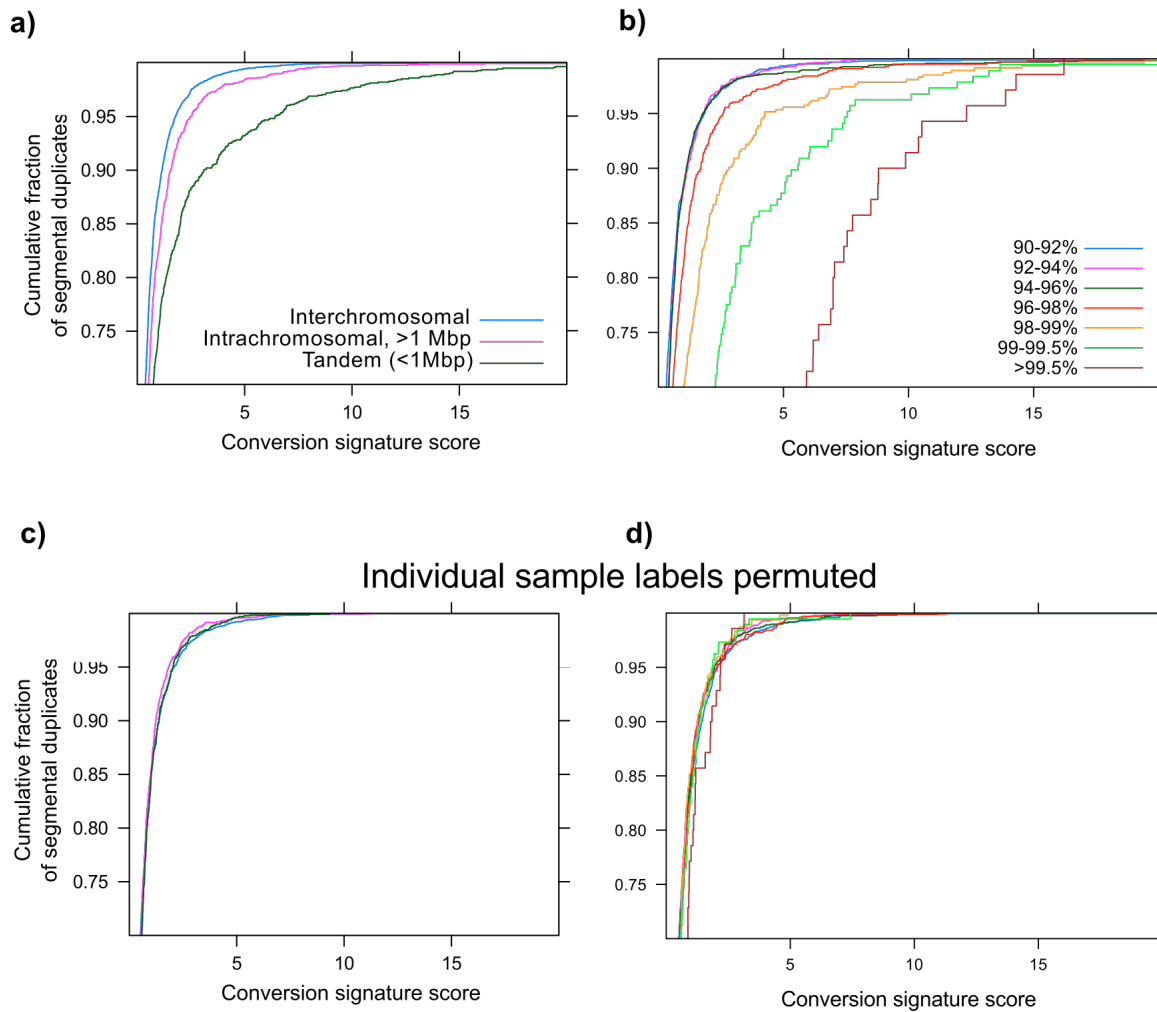


Figure S73. Conversion signature scores for segmental duplications binned by a) arrangement and b) homology show strong preferences for nearby, highly identical duplications. c,d) In permuted controls no significant difference is observed.

SUPPORTING TABLES

Table S1 – Summary of human genomes analyzed

| | Population | # of individuals | Effective Coverage | Read Placements | Source |
|------------------------|-------------------------------|------------------|--------------------|-----------------------|------------------------------------------------------|
| Low Coverage | CEU | 43 | 1.7-6.4X | 5.37X10 ⁹ | 1000 Genomes Pilot1 |
| | YRI | 46 | 1.8-7.1X | 7.11X10 ⁹ | 1000 Genomes Pilot1 |
| | ASN | 50 | 1.5-3.6X | 5.40X10 ⁹ | 1000 Genomes Pilot1 |
| High Coverage | | | | | |
| NA12878 trio | CEU | 3 | 24.7-29.7X | 3.52X10 ⁹ | 1000 Genomes Pilot2 |
| NA19239 trio | YRI | 3 | 13.6-28.5X | 2.71X10 ⁹ | 1000 Genomes Pilot2 |
| NA18507 trio | YRI | 3 | 37.1-43.0X | 5.26X10 ⁹ | (S53) |
| JF | European | 1 | 12.3X | 5.00X10 ⁸ | Illumina |
| NA10851 | CEU | 1 | 28.4X | 1.23X10 ⁹ | (S10) |
| AK1 | Korean | 1 | 22.6X | 9.32X10 ⁸ | (S11) |
| SJK | Korean | 1 | 12.8X | 5.25X10 ⁸ | (S12) |
| YH-1 | CHB | 1 | 13.3X | 6.00X10 ⁸ | (S14) |
| South Africans | Bantu, Kalahari Bushman | 2 | 7.0-23.4X | 1.36X10 ⁹ | (S13) |
| HGDP | Han, Papuan, San, Yoruba | 4 | 4.4-7.1X | 9.16X10 ⁸ | (S54) |
| HUMAN TOTAL | | 159 | 1145X | 3.54x10 ¹⁰ | |
| Chimpanzee | | 1 | 6.7X | 2.78x10 ⁸ | (S55) |
| Orangutan | | 1 | 8.5X | 3.25x10 ⁸ | (S56) |
| Gorilla | | 1 | 4.5X | 2.08x10 ⁸ | GenBank Short Read Archive accession SRP002878 |
| NON-HUMAN TOTAL | | 162 | 1165X | 3.62x10 ¹⁰ | |

Data were downloaded from the NCBI Sequence Read Archive (<http://www.ncbi.nlm.nih.gov/sra>) and mapped to a human reference genome (Build36) using the *mrsFAST* aligner.

Table S2 – Singly Unique Nucleotide (SUN) identifier summary

| | All SD | Genic | Protein-coding |
|---------------|--------------------------|--------------------------|--------------------------|
| Total Size | 1.539x10 ⁸ bp | 4.080x10 ⁷ bp | 2.125x10 ⁶ bp |
| Non-repeat | 5.480x10 ⁷ bp | 1.726x10 ⁷ bp | 1.638x10 ⁶ bp |
| Indels | 1,898,842 | 619,484 | 37,476 |
| Transitions | 1,283,671 | 398,268 | 37,852 |
| Transversions | 782,245 | 236,094 | 19,996 |
| Multi-SUNs | 111,502 | 28,400 | 2,205 |
| Total bp | 4,076,260 | 1,282,246 | 97,529 |
| SUNs/kbp | 74.4 | 74.3 | 59.5 |
| Ti/Tv | 1.64 | 1.69 | 1.89 |

Insertions were counted along their entire lengths; deletions were counted only at junction positions. Multi-SUNs are positions with distinct, uniquely identifying differences to more than one paralog, and, therefore, could be classified into more than one category (e.g., a transition with respect to one duplicate and a transversion with respect to another). Repeats as determined by RepeatMasker and TandemRepeatsFinder, padded +/- 36bp. Genic and protein-coding as defined by RefSeq. SD segmental duplications Ti/Tv, transition to transversion ratio.

| Table S3: Summary of FISH validation | | | | | |
|--------------------------------------------------|------------|-------------------------------------------|-----------|------------------------------|------------------------------|
| Region | Prediction | FISH result | Cell Line | Clone used as probe | Clone location (Build 36) |
| 4q35.2 | 5 | 6 (chr4; chr13; chr21) | GM18564 | WIBR2-3705H20_G248P801591D10 | chr4:190,776,606-190,812,852 |
| | 7 | 8 (chr4; chr13; chr14; chr21) | YH-1 | | |
| | 7 | 8 (chr4; chr13; chr14; chr21) | GM18861 | | |
| | 6 | 7 (chr4; chr13; one chr14; chr21) | GM12878 | | |
| 7q11.23 | 6 | 6 (chr1; duplicated on both chr7) | GM18498 | WIBR2-2972N17_G248P88301G9 | chr7:76,229,616-76,269,483 |
| | 5 | 5 (chr1; chr7 and duplicated on one chr7) | GM18907 | | |
| | 3 | 3 (chr1; one chr7) | YH-1 | | |
| | 4 | 4 (chr1; chr7) | GM12004 | | |
| 7q11.32 | 6 | 6 (chr1; duplicated on both chr7) | GM18498 | WIBR2-2854G20_G248P88510D10 | chr7:76,144,718-76,182,710 |
| | 4 | 4 (chr1; chr7) | GM12004 | | |
| 7q35 | 8 | 5 | GM07347 | WIBR2-3730J21_G248P801433E11 | chr7:143,542,919-143,583,088 |
| | 2 | 2 | GM18853 | | |
| 8p23.1 | 2 | 2 | GM12716 | WIBR2-0588K19_G248P80424F10 | chr8:7,286,320-7,324,309 |
| | 9 | 9 | GM18502 | | |
| 10q11.22 | 3 | 3 | GM18563 | WIBR2-3223I07_G248P802188E4 | chr10:46,393,134-46,431,399 |
| | 8 | 8 | GM18940 | | |
| 15q11.2 | 3 | 3 | GM18555 | WIBR2-0866L18_G248P8587F9 | chr15:20,506,689-20,550,176 |
| | 1 | 1 | GM18507 | | |
| 15q11.2 | 3 | 3 | GM18971 | WIBR2-3778G07_G248P802620D4 | chr15:19,107,938-19,144,015 |
| | 9 | 9 | GM07037 | | |
| 15q13.3 | 3 | 3 | GM12813 | WIBR2-1228I16_G248P83186E8 | chr15:30,068,513-30,106,795 |
| | 2 | 2 | GM12004 | | |
| | 3 | 3 | GM12813 | WIBR2-2841A03_G248P88552A2 | chr15:30,235,939-30,278,876 |
| | 4 | 4 | GM12004 | | |
| 15q13.3 | 3 | 3 | GM12813 | WIBR2-0682I02_G248P80054E1 | chr15:29,881,242-29,924,151 |
| | 2 | 2 | GM12004 | | |
| 16p12.3 | 3 | 3 | GM18547 | WIBR2-2202N24_G248P86159G12 | chr16:17,788,765-17,827,587 |
| | 2 | 2 | GM12004 | | |
| 16p12.1 | 5 | 5 | GM12004 | WIBR2-2031K01_G248P86836F1 | chr16:22,477,340-22,523,335 |
| | 6 | 6 | GM12878 | | |
| | 6 | 6 | GM18861 | | |
| | 4 | 4 | GM18555 | | |
| | 5 | 5 | GM18947 | | |
| | 5 | 5 | GM18956 | | |
| 16p12.1 | 3 | 3 | GM12004 | WIBR2-3608M06_G248P801202G3 | chr16:22,542,172-22,579,566 |
| | 4 | 4 | GM12878 | | |
| | 4 | 4 | GM18861 | | |
| | 2 | 2 | GM18555 | | |
| | 3 | 3 | GM18947 | | |
| | 3 | 3 | GM18956 | | |
| 16p12.1 | 6 | 7 | GM12004 | WIBR2-0590C03_G248P80492B2 | chr16:21,423,413-21,466,802 |
| | 8 | 8 | GM12878 | | |
| | 7 | 8 | GM18861 | | |
| | 6 | 6 | GM18555 | | |
| | 7 | 7 | GM18947 | | |
| | 6 | 7 | GM18956 | | |
| 16q22.2 | 5 | 5 (duplicated on one chr1) | GM19201 | WIBR2-3823N03_G248P802141G2 | chr16:69,580,922-69,624,156 |
| | 3 | 3 (missing copy on one chr1) | GM19190 | | |
| 17p11.2 | 4 | 4 | GM12287 | WIBR2-1008E03_G248P84612C2 | chr17:20,284,001-20,322,710 |
| | 8 | 8 | GM18907 | | |
| 17q21.31 | 2 | 2 | GM19240 | WIBR2-2004M23_G248P85784G12 | chr17:41,545,746-41,586,094 |
| | 4 | 4 | GM12878 | | |
| 17q21.32 | 6 | 7 | GM18942 | WIBR2-1857K04_G248P85426F2 | chr17:42,055,759-42,093,331 |
| | 6 | 6 | YH-1 | | |
| | 4 | 4 | GM19129 | | |
| | 4 | 4 | GM12892 | | |
| 17q21.32 | 3 | 5 | GM18555 | WIBR2-0946N09_G248P801829G5 | chr17:42,986,667-43,025,934 |
| | 4 | 6 | GM12813 | | |
| | 10 | 12 | GM18502 | | |
| 5q31.3* | 3 | 2 | GM07347 | WIBR2-3662F08_G248P801703C4 | chr5:140,151,286-140,188,584 |
| | 6 | 2 | GM18545 | | |
| | 4 | 2 | GM07347 | WIBR2-2544K17_G248P86728F9 | chr5:140,526,836-140,569,682 |
| | 8 | 2 | GM18545 | | |
| 7q11.23** | 19 | 18-20 | GM18552 | WIBR2-0458K07_G248P8624F4 | chr7:73,942,214-73,981,865 |
| | 19 | 16-20 | GM12004 | | |
| | 30 | 22-26 | GM18861 | | |
| 16p13.11** | 33 | 29-34 | GM18537 | WIBR2-1564J16_G248P83059E8 | chr16:14,921,937-14,963,735 |
| | 49 | 37-48 | GM18501 | | |
| 17q12** | 18 | 14-18 | GM18573 | WIBR2-1684A05_G248P81904A3 | chr17:33,536,637-33,572,498 |
| | 48 | 19-26 | GM18861 | | |
| *prediction not validated | | | | | |
| **FISH failed to accurately estimate copy-number | | | | | |

Table S4. Quantitative PCR validation targets and assay selection

| TARGET | Primer-set #1 r ² | Primer-set #2 r ² | on X chromosome | Primer-set 1 product size | Primer-set 2 product size |
|---------------------------------------------|---------------------------------|---------------------------------|--------------------|------------------------------|------------------------------|
| <i>OPN1MW2</i> | 0.76 | NA | No | 100 | NA |
| <i>NPEPPS</i> | 0.92 | NA | no | 120 | NA |
| <i>TBC1D3</i> | 0.95 | 0.9 | no | 103 | 103 |
| <i>CCL3L1</i> | 0.93 | 0.90 | no | 119 | 117 |
| <i>NBPF16</i> | 0.93 | 0.66 | no | 108 | 112 |
| <i>GAGE10</i> (males and females) | 0.98 | 0.91 | yes | 130 | 204 |
| <i>GAGE10</i> (males) | 0.37 | 0.63 | yes | 130 | 204 |
| <i>GAGE10</i> (females) | 0.97 | 0.6 | yes | 130 | 204 |
| <i>PRSSI</i> | 0.14 | 0.59 | no | 201 | 115 |
| <i>NPIP</i> | 0.66 | 0.92 | no | 231 | 374 |
| <i>CFHR3</i> (paralog-specific assay) | 0.84 | NA | no | 159 | NA |

Table S5. Quantitative PCR primer sets

| Primer Target | Primer Name | Primer Sequence |
|---------------------------|-------------|-----------------------|
| Albumin (diploid control) | alb_HSA_3F | TTGTGGGCTGTAATCATCG |
| | alb_HSA_3R | TGCTGGTTCTCTTTCACTGAC |
| <i>CCL3L1</i> | CCL3L1_2F | GGGTCCAGAAATACGTCAGT |
| | CCL3L1_2R | CATGTTCCCAAGGCTCAG |
| | CCL3L1_4F | TGGGGTCTGTTCTTCACTCT |
| | CCL3L1_4R | CGGTTCAAGAAGTCATACCC |
| <i>NBPF16</i> | NBPF16_1F | AAACGTCAGCATGGTGGTAT |
| | NBPF16_1R | TGTTTCTTCTCTGCCAACTG |
| | NBPF16_3F | TGCAGGACTCACTGGATAGA |
| | NBPF16_3R | CCAAGGTAAGTTCCTCAA |
| <i>NPEPPS</i> | NPEPPS_2F | CCGCACACCTGTTATGTCTA |
| | NPEPPS_2R | CAGGAGTGTAACACGGACA |
| | NPEPPS_3F | AGAGCATCCACCAGTACCTC |
| | NPEPPS_3R | TTCGGTAGCTCCACCTTATC |
| <i>OPN1MW2</i> | OPN1MW2_1F | ACCATGAAGTTCAAGAAGCTG |
| | OPN1MW2_1R | CTGGTTCACAACGCTGATAG |
| | OPN1MW2_4F | TGATGCTTCAGTGCCTCTT |
| | OPN1MW2_4R | ATGGGTACCGGAGTTCATTA |
| <i>TBC1D3</i> | TBC1D3_1F | TAGCCTGTCTGCTCTCTG |
| | TBC1D3_1R | TCAGGGAGAAAACCTTTGAG |
| | TBC1D3_2F | CATAGATCGAGCGTACAAGG |
| | TBC1D3_2R | GTATCTTCCGGGGTTTTTC |

| | | |
|---------------|-----------|-------------------------|
| <i>GAGE10</i> | GAGE10_5F | GAGTTGGAGAAATGTCTTTAGGC |
| | GAGE10_5R | AATACGTGTGGGTGTGCAAAAT |
| | GAGE10_6F | ATTGTCACATAGGAGGAAGAGG |
| | GAGE10_6R | TTTTATCATAGTGAGGGATTTGC |
| <i>NPIP</i> | NPIP_5F | GAAGCTGTCTGAACTTACTCAGG |
| | NPIP_5R | TGAATAGCGTGGGATTTCTC |
| | NPIP_7F | TTTTCAGATCACCCCTTCTG |
| | NPIP_7R | AGCCGTAGGGAGAAAAATGT |
| <i>CFHR3</i> | CFHR3.2F* | TGATTCTGGACGTTTTGCTG |
| | CFHR3.2R* | AAGGGAACGAAAGGCTTCTG |

* CFHR3.2 primers were designed to overlap SUNs specific to the *CFHR1/3* deletion

Table S6: Missing Paralogous Genes in Reference Genome

Gene families were identified which are missing from the human reference genome if the reference genome had 5 or more fewer copies than the median number of copies among 159 individuals. Genes were clustered into families. One representative gene from each family (*) is displayed. Locations are with respect to BUILD36 of the human reference.

| gene name* | location | median copy number (159 individuals) | BUILD36 reference genome copy | Vst |
|------------|---------------------------|-----------------------------------------|----------------------------------|------|
| WASH3P | chr15:100318538-100334331 | 21.23 | 10.75 | 0.03 |
| NBPF10 | chr1:144004727-144080039 | 262.88 | 135.86 | 0.06 |
| CROCCL1 | chr1:16817339-16829988 | 10.23 | 3.72 | 0.25 |
| MSTP2 | chr1:16844655-16849501 | 13.85 | 5.67 | 0.06 |
| PDE4DIP | chr1:143663117-143787436 | 7.33 | 1.98 | 0.48 |
| NOTCH2NL | chr1:143920467-143997269 | 9.06 | 3.92 | 0.38 |
| LOC645166 | chr1:147194909-147218219 | 12.5 | 5.4 | 0.02 |
| DUX4 | chr10:135330357-135335265 | 367.72 | 31.34 | 0.22 |
| LOC284232 | chr13:18306542-18344109 | 26.81 | 16.72 | 0.07 |
| PRR20A | chr13:56619622-56622644 | 16.42 | 10.13 | 0.17 |
| NPIPL3 | chr16:22432384-22455362 | 50.29 | 44.18 | 0.17 |
| FLJ36000 | chr17:21828188-21837197 | 58.28 | 11.76 | 0.08 |
| TBC1D3G | chr17:31871287-31882216 | 37.3 | 22.75 | 0.5 |
| LRRC37A4 | chr17:40939892-40948305 | 16.22 | 11.13 | 0.24 |
| DHX40P | chr17:55408166-55451118 | 16.85 | 10.22 | 0.46 |
| TCEB3C | chr18:42808570-42810447 | 28.38 | 6.74 | 0.22 |
| CGB2 | chr19:54226941-54228307 | 18.62 | 12.96 | 0 |
| LOC654342 | chr2:91188435-91211702 | 12.52 | 5.4 | 0.02 |
| ANKRD20A3 | chr9:67516580-67560113 | 28.15 | 19.31 | 0.09 |
| LOC151009 | chr2:110567223-110576650 | 28.33 | 18.16 | 0.09 |
| MGC13005 | chr2:114073074-114075623 | 24.27 | 14.22 | 0.01 |
| C2orf27B | chr2:132269003-132275704 | 19.98 | 12.44 | 0.09 |
| FAM182A | chr20:25983249-26015552 | 15.48 | 8.47 | 0.01 |
| FRG1B | chr20:28225539-28247668 | 31.33 | 7.28 | 0.05 |
| BAGE | chr21:10079666-10120808 | 16.89 | 3.59 | 0.04 |
| C21orf81 | chr21:14237966-14274636 | 20.21 | 14.52 | 0.14 |
| POM121L4P | chr22:19373842-19376009 | 50.71 | 37.14 | 0.3 |
| LOC96610 | chr22:20982462-21007324 | 18.33 | 12.42 | 0.02 |
| ZNF717 | chr3:75868718-75916945 | 25.37 | 7.18 | 0.29 |
| ZNF595 | chr4:43226-78099 | 10.63 | 3.01 | 0.18 |

| | | | | |
|--------------|--------------------------|--------|-------|------|
| OTOP1 | chr4:4241430-4279522 | 8.73 | 3.25 | 0.02 |
| USP17L6P | chr4:8978697-8979892 | 142.64 | 37.04 | 0.27 |
| DRD5 | chr4:9392355-9394731 | 11.34 | 4.84 | 0.04 |
| LOC100133050 | chr5:99743108-99751857 | 38.23 | 27.39 | 0.04 |
| FLJ35390 | chr7:44045591-44048606 | 13.16 | 7.75 | 0.21 |
| SPDYE5 | chr7:74962234-74971564 | 37.77 | 32.6 | 0.25 |
| MUC12 | chr7:100399623-100448949 | 12.21 | 5.82 | 0.04 |
| RPL23AP53 | chr8:148346-172318 | 27.08 | 20.99 | 0.09 |
| FAM90A14 | chr8:7124701-7127711 | 59.59 | 47.29 | 0.17 |
| FAM66D | chr8:12010699-12046107 | 31.43 | 14.06 | 0.23 |
| REXO1L2P | chr8:86884283-86885361 | 171.62 | 15.68 | 0.16 |
| C9orf122 | chr9:38611084-38613275 | 13.11 | 5.96 | 0.07 |
| LOC442421 | chr9:66234088-66242849 | 15.42 | 6.1 | 0.16 |
| CCDC29 | chr9:68715487-68738681 | 32.2 | 22.03 | 0.03 |

Table S7: Most Variable Copy Number Human Genes

The most variable gene families among all individuals were identified by selecting for those with a variance >3 among all 159 individuals. Genes were clustered into families. One representative gene from each family (*) is displayed. Locations are with respect to BUILD36 of the human reference.

| gene name* | location | variance | median copy | Vst | segmentally duplicated base-pairs |
|--------------|---------------------------|----------|-------------|------|-----------------------------------|
| DUX4 | chr10:135330357-135335265 | 3858 | 367.72 | 0.22 | 4908 |
| USP17L6P | chr4:8978697-8979892 | 2031.71 | 142.64 | 0.27 | 1195 |
| REXO1L2P | chr8:86884283-86885361 | 1055.8 | 171.62 | 0.16 | 1078 |
| NBPF10 | chr1:144004727-144080039 | 629.65 | 262.88 | 0.06 | 75312 |
| FAM90A14 | chr8:7124701-7127711 | 488.38 | 59.59 | 0.17 | 3010 |
| FLJ36000 | chr17:21828188-21837197 | 334.11 | 58.28 | 0.08 | 2218 |
| LOC392196 | chr8:12022775-12024213 | 331.86 | 58.27 | 0.27 | 1438 |
| TBC1D3F | chr17:33591517-33602455 | 93.2 | 38.13 | 0.5 | 10938 |
| TCEB3C | chr18:42808570-42810447 | 86.57 | 28.38 | 0.22 | 1877 |
| LOC100272216 | chr5:68962735-68964784 | 82.32 | 63.64 | 0.13 | 2049 |
| FAM66D | chr8:12010699-12046107 | 79.24 | 31.43 | 0.23 | 35408 |
| PRR20A | chr13:56619622-56622644 | 45.67 | 16.42 | 0.17 | 3022 |
| LOC100133050 | chr5:99743108-99751857 | 44.5 | 38.23 | 0.04 | 8749 |
| POM121L4P | chr22:19373842-19376009 | 33.18 | 50.71 | 0.3 | 2167 |
| DHX40P | chr17:55408166-55451118 | 18.4 | 16.85 | 0.46 | 42952 |
| LOC399744 | chr10:38757079-38781086 | 16.09 | 46.03 | 0.09 | 24007 |
| SPDYE2 | chr7:101983351-101989853 | 14.05 | 37.2 | 0.32 | 6502 |
| C2orf78 | chr2:73864823-73897782 | 14.03 | 10.88 | 0.39 | 14288 |
| FLJ45340 | chr7:128068530-128088288 | 12.1 | 47.2 | 0.01 | 18612 |
| GUSBL2 | chr6:58354117-58395683 | 9.95 | 24.47 | 0.13 | 41566 |
| EEF1AL7 | chr4:106625311-106626956 | 9.69 | 26.85 | 0.01 | 1645 |
| LOC151009 | chr2:110567223-110576650 | 9.53 | 28.33 | 0.09 | 9427 |
| PMS2L2 | chr7:72114615-72152902 | 9.49 | 28.15 | 0.12 | 38287 |
| GGTLC2 | chr22:21318781-21320368 | 9.36 | 15.62 | 0.14 | 1587 |
| POLR2J2 | chr7:102064709-102099418 | 9.34 | 21.26 | 0.14 | 34709 |
| LRRC37A4 | chr17:40939892-40948305 | 9.3 | 16.22 | 0.24 | 8413 |
| LOC440896 | chr9:68464033-68470861 | 8.99 | 22.6 | 0.03 | 6828 |
| AMY1A | chr1:104031562-104040435 | 8.31 | 9.49 | 0.13 | 8873 |
| FAM157B | chr9:140226457-140253993 | 8.04 | 26.46 | 0.03 | 27536 |

| | | | | | |
|-----------|--------------------------|------|-------|------|-------|
| RASA4P | chr7:44035010-44046747 | 7.94 | 12.59 | 0.26 | 11737 |
| NPIPL3 | chr16:22432384-22455362 | 7.57 | 50.29 | 0.17 | 22978 |
| C9orf122 | chr9:38611084-38613275 | 7.08 | 13.11 | 0.07 | 2191 |
| LOC613037 | chr16:30141850-30164433 | 6.61 | 50.59 | 0.29 | 22583 |
| DEFA1 | chr8:6822580-6825012 | 6.55 | 7.58 | 0.03 | 2432 |
| MSTP2 | chr1:16844655-16849501 | 6.45 | 13.85 | 0.06 | 4846 |
| UPLP | chr7:102064709-102070474 | 6.38 | 9.5 | 0.27 | 5765 |
| GOLGA9P | chr15:20806682-20814184 | 6.15 | 34.92 | 0.11 | 7502 |
| RPL23AP82 | chr22:49569022-49584930 | 5.67 | 31.94 | 0.09 | 15908 |
| MGC13005 | chr2:114073074-114075623 | 5.51 | 24.27 | 0.01 | 2549 |
| WASH1 | chr9:4510-19739 | 5.09 | 22.42 | 0.05 | 15229 |
| MUC12 | chr7:100399623-100448949 | 5.02 | 12.21 | 0.04 | 7765 |
| MBD3L3 | chr19:7007216-7009645 | 5.01 | 10.03 | 0 | 2429 |
| ZNF717 | chr3:75868718-75916945 | 4.94 | 25.37 | 0.29 | 36176 |
| CCL3L1 | chr17:31546381-31548269 | 4.87 | 4.95 | 0.44 | 1888 |
| DGCR9 | chr22:17385346-17387760 | 4.81 | 17.06 | 0.1 | 0 |
| TP53TG3 | chr16:32592349-32595554 | 4.78 | 8.55 | 0.21 | 3205 |
| POTEH | chr22:14636331-14667937 | 4.36 | 19.26 | 0.07 | 31606 |
| CCDC29 | chr9:68715487-68738681 | 4.39 | 32.2 | 0.03 | 23194 |
| FRG1B | chr20:28225539-28247668 | 4.39 | 31.33 | 0.05 | 22129 |
| PRAMEF15 | chr1:13514559-13521574 | 4.39 | 17.83 | 0.01 | 7015 |
| LOC728323 | chr2:242679516-242751142 | 4 | 19.18 | 0.09 | 40432 |
| OR7E91P | chr2:71104712-71110568 | 3.82 | 34.64 | 0.12 | 5856 |
| ANKRD20A1 | chr9:67516580-67559660 | 3.51 | 28.11 | 0.08 | 43080 |
| NPEPPS | chr17:42963442-43055641 | 3.5 | 5.97 | 0.5 | 62993 |
| FKSG73 | chr2:91492885-91494221 | 3.35 | 14.23 | 0.03 | 1336 |
| CGB2 | chr19:54226941-54228307 | 3.32 | 18.62 | 0 | 1366 |

Table S8: Population Stratified Human Gene Families

The most population stratified genes among 159 individuals in three representative populations (European, Yoruba and Asian) were determined using the Vst metric. Genes were clustered into families. One representative gene from each family (*) is displayed. Locations are with respect to BUILD36 of the human reference.

| gene name* | location | Vst | median copy | variance | segmentally duplicated basepairs | mean European copy | mean Asian copy | mean Yoruba copy |
|--------------|--------------------------|------|-------------|----------|----------------------------------|--------------------|-----------------|------------------|
| TBC1D3C | chr17:31820231-31882204 | 0.53 | 34.64 | 83.69 | 61973 | 29.28 | 34.17 | 43.86 |
| TRY6 | chr7:142158331-142161973 | 0.52 | 4.37 | 1.18 | 3642 | 4.39 | 3.51 | 5.17 |
| NPEPPS | chr17:42963442-43055641 | 0.5 | 5.97 | 3.5 | 62993 | 5.5 | 5.42 | 8.27 |
| PDE4DIP | chr1:143663117-143787436 | 0.48 | 7.33 | 0.47 | 34281 | 7.08 | 7.49 | 6.51 |
| UGT2B17 | chr4:69085497-69116840 | 0.47 | 2.87 | 0.69 | 31343 | 3.03 | 2.25 | 3.41 |
| ESPNP | chr1:16890299-16919239 | 0.46 | 6.6 | 1.95 | 28940 | 5.62 | 7.47 | 6.6 |
| DHX40P | chr17:55408166-55451118 | 0.46 | 16.85 | 18.4 | 42952 | 14.81 | 15.79 | 21.02 |
| LOC644172 | chr17:41033278-41035531 | 0.45 | 5.43 | 0.76 | 2253 | 6.56 | 5.36 | 5.26 |
| LGALS9B | chr17:20293767-20311440 | 0.44 | 5.95 | 1.11 | 17673 | 5.07 | 6.04 | 6.68 |
| CCL3L1 | chr17:31546381-31548269 | 0.44 | 4.95 | 4.87 | 1888 | 3.44 | 5.93 | 6.44 |
| PRSS3 | chr9:33785558-33789229 | 0.42 | 5.55 | 0.51 | 3671 | 5.58 | 5.02 | 5.97 |
| LILRA3 | chr19:59491666-59496050 | 0.63 | 7.08 | 1.61 | 3358 | 7.62 | 5.62 | 7.62 |
| DHFRL1 | chr3:95259455-95264350 | 0.41 | 4.21 | 0.32 | 3434 | 4.55 | 4.37 | 3.79 |
| LOC100286793 | chr1:142510035-142536110 | 0.4 | 11.95 | 0.64 | 26075 | 11.79 | 12.31 | 11.25 |

| | | | | | | | | |
|-----------|---------------------------|------|--------|---------|--------|--------|--------|--------|
| CCL4 | chr17:31455332-31457127 | 0.4 | 4.32 | 1.63 | 1795 | 3.49 | 4.92 | 5.13 |
| CFHR1 | chr1:195055483-195067942 | 0.39 | 2.99 | 0.48 | 7459 | 3.09 | 3.32 | 2.43 |
| C2orf78 | chr2:73864823-73897782 | 0.39 | 10.88 | 14.03 | 14288 | 11.07 | 13.52 | 8.24 |
| NOTCH2NL | chr1:143920467-143997269 | 0.38 | 9.06 | 0.53 | 76802 | 8.96 | 9.38 | 8.5 |
| HLA-DPA1 | chr6:33140771-33149356 | 0.37 | 1.92 | 0.05 | 0 | 2.06 | 1.91 | 1.79 |
| RDM1 | chr17:31269199-31281416 | 0.37 | 4.21 | 0.1 | 1741 | 4.43 | 4.26 | 4.02 |
| OCLN | chr5:70405174-70424776 | 0.37 | 3 | 0.68 | 19602 | 3.06 | 2.53 | 3.56 |
| KRT14 | chr17:36992058-36996673 | 0.36 | 5.26 | 0.57 | 4615 | 5.74 | 5.18 | 4.75 |
| PGA5 | chr11:60765244-60775491 | 0.36 | 6.61 | 2.26 | 0 | 5.62 | 7.52 | 6.18 |
| PDXDC1 | chr16:14976333-15039053 | 0.36 | 4.87 | 0.53 | 55935 | 5.31 | 4.38 | 5.23 |
| PCDHB13 | chr5:140573692-140577177 | 0.33 | 4.01 | 1.13 | 2510 | 4.94 | 4.42 | 3.77 |
| MGC57346 | chr17:41053494-41071110 | 0.33 | 3.64 | 0.48 | 7454 | 4.4 | 3.52 | 3.49 |
| SPDYE6 | chr7:101772912-101783609 | 0.32 | 36.15 | 12.86 | 10697 | 34.67 | 35.78 | 39.05 |
| ACOT2 | chr14:73105524-73112112 | 0.31 | 2.81 | 0.64 | 6588 | 2.92 | 2.38 | 3.31 |
| C17orf58 | chr17:63417678-63420227 | 0.31 | 3.16 | 0.3 | 2549 | 3.76 | 3.11 | 3.07 |
| POM121L4P | chr22:19373842-19376009 | 0.3 | 50.71 | 33.18 | 2167 | 54.28 | 47.69 | 51.89 |
| PRH1 | chr12:10924826-11215477 | 0.3 | 2.76 | 0.03 | 29610 | 2.75 | 2.9 | 2.7 |
| KIAA1267 | chr17:41463128-41605371 | 0.29 | 2.04 | 0.17 | 18638 | 2.55 | 2.02 | 2.02 |
| ZNF717 | chr3:75868718-75916945 | 0.29 | 25.37 | 4.94 | 36176 | 25.78 | 24.08 | 26.64 |
| ADAM5P | chr8:39291338-39379532 | 0.29 | 1.8 | 0.13 | 0 | 1.4 | 1.68 | 1.79 |
| GTF2H2 | chr5:68891809-68924334 | 0.28 | 4.49 | 0.84 | 32525 | 4.76 | 3.88 | 4.86 |
| USP17 | chr4:8969206-8970799 | 0.28 | 139.97 | 1990.24 | 1593 | 120.79 | 137.72 | 175.41 |
| FAM157A | chr3:199363633-199392125 | 0.28 | 10.49 | 0.84 | 28492 | 10.55 | 10.17 | 11.16 |
| LRRC37A3 | chr17:60280949-60345365 | 0.28 | 11.15 | 1.36 | 64416 | 11.48 | 11.56 | 10.4 |
| UBE2QP2 | chr15:80820827-80881396 | 0.27 | 6.92 | 0.55 | 60569 | 6.45 | 7.2 | 7.01 |
| LOC392196 | chr8:12022775-12024213 | 0.27 | 58.27 | 331.86 | 1438 | 51.8 | 57.94 | 73.51 |
| RHD | chr1:25471567-25529523 | 0.27 | 3.86 | 0.39 | 57956 | 3.15 | 3.85 | 3.81 |
| UPLP | chr7:102064709-102070474 | 0.27 | 9.5 | 6.38 | 5765 | 8.29 | 9.53 | 11.08 |
| NPIPL3 | chr16:21320950-21344159 | 0.26 | 50.36 | 6.58 | 23209 | 49.78 | 49.33 | 52.09 |
| RASA4P | chr7:44035010-44046747 | 0.26 | 12.59 | 7.94 | 11737 | 11.22 | 12.33 | 14.37 |
| POLR2J4 | chr7:43947018-44025273 | 0.26 | 9.75 | 5.46 | 57818 | 8.62 | 9.53 | 11.18 |
| CROCCL1 | chr1:16817339-16829988 | 0.25 | 10.23 | 2.57 | 12649 | 9.64 | 10.23 | 11.34 |
| ARHGEF5 | chr7:143683421-143708658 | 0.25 | 5.94 | 2.44 | 21888 | 6.09 | 6.71 | 5.15 |
| ERC2 | chr3:55517375-56477431 | 0.25 | 1.92 | 0 | 0 | 1.94 | 1.9 | 1.92 |
| PLA2G10 | chr16:14673905-14696027 | 0.24 | 8.87 | 2.09 | 7521 | 8.15 | 9.17 | 9.64 |
| RFPL4A | chr19:60962318-60966351 | 0.24 | 4.88 | 1.75 | 4033 | 4.86 | 5.93 | 4.59 |
| FAM103A1 | chr15:81445998-81450427 | 0.23 | 3.66 | 0.26 | 0 | 3.87 | 3.81 | 3.36 |
| CES1 | chr16:54394264-54424576 | 0.23 | 3.88 | 0.1 | 5672 | 3.86 | 4.07 | 3.76 |
| HEATR4 | chr14:73014944-73095404 | 0.23 | 2.15 | 0.02 | 3215 | 2.2 | 2.08 | 2.22 |
| FAM66D | chr8:12010699-12046107 | 0.23 | 31.43 | 79.24 | 35408 | 28.55 | 30.6 | 38.22 |
| GUSBL1 | chr6:26947244-27032312 | 0.23 | 13.2 | 2.62 | 85068 | 14.25 | 13.27 | 12.53 |
| TCEB3C | chr18:42808570-42810447 | 0.22 | 28.38 | 86.57 | 1877 | 29.01 | 24.38 | 33.91 |
| DUX4 | chr10:135330357-135335265 | 0.22 | 367.72 | 3858 | 4908 | 368.19 | 341.29 | 401.37 |
| TP53TG3 | chr16:33112480-33115680 | 0.22 | 8.22 | 4.31 | 3200 | 8.34 | 7.32 | 9.35 |
| ANKRD36 | chr2:97142959-97279633 | 0.21 | 17.69 | 2.2 | 136674 | 17.22 | 17.6 | 18.69 |
| NBPF3 | chr1:21639217-21683980 | 0.21 | 13.95 | 1.49 | 44763 | 13.52 | 14.13 | 14.74 |
| GSTT1 | chr22:22706138-22714284 | 0.21 | 1.08 | 0.76 | 0 | 1.61 | 0.79 | 0.86 |
| ZNF705D | chr8:11984255-12010434 | 0.2 | 10.12 | 1.31 | 26179 | 10.15 | 9.87 | 10.91 |
| SMARCA2 | chr9:2005341-2183623 | 0.2 | 1.93 | 0 | 0 | 1.96 | 1.92 | 1.93 |

Table S9: Most Copy-number Stratified Genomic Regions

The most population stratified genomic regions were detected by computing the Vst statistic in 1kb unmasked sequence windows across the entire genome and merging contiguous windows of increased signal (Vst >0.2).

| chr | start | end | size | average Vst | segmentally duplicated base-pairs | genes |
|-------|-----------|-----------|--------|-------------|-----------------------------------|------------------------------------------------------------------------------------------------------------------|
| chr1 | 16855269 | 16994786 | 139517 | 0.23697321 | 137937 | ESPNP MSTP9 LOC728855 LOC728855 LOC728875 |
| chr1 | 143012474 | 143113897 | 101423 | 0.223041759 | 100627 | PPIAL4B PPIAL4C PPIAL4A |
| chr1 | 143660533 | 143789398 | 128865 | 0.306949902 | 36865 | PDE4DIP PDE4DIP PDE4DIP PDE4DIP |
| chr1 | 143899256 | 144016942 | 117686 | 0.21152537 | 117294 | NOTCH2NL NBPF10 |
| chr1 | 147777596 | 147882589 | 104993 | 0.239820262 | 104993 | PPIAL4A PPIAL4C LOC728855 |
| chr1 | 194978578 | 195093170 | 114592 | 0.333558968 | 109991 | CFH CFHR3 CFHR1 PRR4 PRH1 TAS2R19 TAS2R31 TAS2R46 TAS2R43 TAS2R30 |
| chr12 | 11061265 | 11184215 | 122950 | 0.210171532 | 66557 | |
| chr16 | 14938245 | 15039130 | 100885 | 0.221789233 | 83352 | NP1P PDXDC1 CCL3 CCL4 TBC1D3B CCL3L3 CCL3L1 CCL4L1 CCL4L2 TBC1D3C TBC1D3F CCL3L3 CCL3L1 CCL4L1 CCL4L2 |
| chr17 | 31427937 | 31700688 | 272751 | 0.426615843 | 229816 | |
| chr17 | 33287994 | 33430126 | 142132 | 0.389299833 | 139484 | LOC284100 |
| chr17 | 33529126 | 33664621 | 135495 | 0.481327103 | 135390 | TBC1D3 TBC1D3F TBC1D3F LRRC37A4 LOC644172 MGC57346 MGC57346 |
| chr17 | 40927933 | 41058228 | 130295 | 0.264685464 | 129472 | KIAA1267 LRRC37A ARL17 ARL17 LRRC37A2 ARL17P1 ARL17P1 ARL17 |
| chr17 | 41521057 | 42155415 | 634358 | 0.22378867 | 510129 | NSF |
| chr19 | 15254 | 135375 | 120121 | 0.205305338 | 120121 | FAM138A FAM138C FAM138F OR4F17 |
| chr22 | 17035971 | 17269339 | 233368 | 0.223062515 | 233368 | USP18 GGT3P |
| chr22 | 19794051 | 20040977 | 246926 | 0.224252184 | 245294 | POM121L8P |
| chr3 | 163996489 | 164111761 | 115272 | 0.418407193 | 0 | |
| chr4 | 8932857 | 9033353 | 100496 | 0.208320107 | 98308 | USP17 USP17 USP17 USP17 USP17 USP17 USP17 USP17 USP17 USP17L6P |
| chr4 | 69049455 | 69237384 | 187929 | 0.429892723 | 135770 | UGT2B17 UGT2B15 POLR2J POLR2J3 SPDYE2 RASA4 |
| chr7 | 101901561 | 102120157 | 218596 | 0.224432098 | 218596 | RASA4 UPLP POLR2J2 CTAGE4 ARHGEF5L OR2A42 OR2A1 OR2A9P OR2A20P OR2A7 CTAGE4 OR2A20P OR2A9P OR2A1 OR2A42 |
| chr7 | 143508006 | 143705630 | 197624 | 0.21453918 | 195288 | ARHGEF5 |
| chr8 | 39345036 | 39508331 | 163295 | 0.243867462 | 2922 | ADAM5P ADAM3A ADAM3A ADAM3A |

Table S10: Human Specific Duplications

Genes duplicated specifically in the human lineage (diploid in great apes). Genes were clustered into families. One representative gene from each family (*) is displayed. Locations are with respect to BUILD36 of the human reference.

| gene name* | locus | variance | median copy | Vst |
|------------|--------------------------|----------|-------------|------|
| FCGR1A | chr1:148020873-148030698 | 0.28 | 5.64 | 0.03 |
| RBM8A | chr1:144218994-144222801 | 0.12 | 3.52 | 0.04 |
| HIST2H2BF | chr1:148020868-148050552 | 0.21 | 5.49 | 0.02 |
| SRGAP2 | chr1:204582822-204696128 | 0.08 | 5.36 | 0.01 |
| PTPN20A | chr10:45970128-46040064 | 0.07 | 3.68 | 0.01 |
| FRMPD2 | chr10:49034611-49053173 | 0.08 | 3.94 | 0 |

| | | | | |
|------------|--------------------------|------|------|------|
| GPRIN2 | chr10:46413551-46420574 | 0.53 | 4.4 | 0.02 |
| C10orf57 | chr10:81828405-81842287 | 0.38 | 4.06 | 0.09 |
| CHRFAM7A | chr15:28440734-28473156 | 0.42 | 4.27 | 0.05 |
| ARHGAP11A | chr15:30694982-30715674 | 0.08 | 3.55 | 0 |
| ARHGEF5 | chr7:143683421-143708658 | 2.44 | 5.94 | 0.25 |
| HYDIN | chr16:69398789-69822070 | 0.02 | 3.85 | 0.01 |
| NCRNA00152 | chr2:87536088-87602143 | 0.06 | 3.88 | 0.03 |
| GTF2H2 | chr5:68891809-68924334 | 0.84 | 4.49 | 0.28 |
| SERF1A | chr5:69356827-69364281 | 0.51 | 3.6 | 0.01 |
| SMN1 | chr5:69381105-69409172 | 0.36 | 3.58 | 0.02 |
| NAIP | chr5:70300065-70356697 | 0.66 | 5.04 | 0.1 |
| DUSP22 | chr6:237100-296355 | 0.33 | 3.98 | 0.1 |
| NCF1 | chr7:73826244-73841595 | 0.18 | 6.43 | 0 |
| GTF2IRD2 | chr7:73848419-73905777 | 0.09 | 5.5 | 0.02 |
| FAM115C | chr7:143028166-143053109 | 0.61 | 3.93 | 0.07 |
| LOC154761 | chr7:143139993-143164743 | 0.63 | 3.94 | 0.06 |
| ZNF322B | chr9:98999357-99001731 | 0.16 | 3.85 | 0.04 |

Table S11: Gene Families Expanded in the Human Lineage

Genes which are duplicated in the primate lineage and have continued to expand along the human lineage. Genes were clustered into families. One representative gene from each family (*) is displayed. Locations are with respect to BUILD36 of the human reference.

| gene name* | location | variance | median copy number | Vst |
|--------------|--------------------------|----------|--------------------|------|
| WASH2P | chr2:114057699-114073081 | 3.97 | 20.81 | 0.04 |
| CROCCL1 | chr1:16817339-16829988 | 2.57 | 10.23 | 0.25 |
| MSTP2 | chr1:16844655-16849501 | 6.45 | 13.85 | 0.06 |
| AMY1A | chr1:103999663-104008696 | 7.74 | 9.91 | 0.1 |
| FLJ39739 | chr1:142510035-142535980 | 0.61 | 11.83 | 0.39 |
| PDE4DIP | chr1:143663117-143706379 | 0.52 | 7.2 | 0.44 |
| NBPF14 | chr1:146470265-146492472 | 521.77 | 244.66 | 0.11 |
| NCF1 | chr7:73826244-73841595 | 0.18 | 6.43 | 0 |
| LOC645166 | chr1:147194909-147218219 | 0.56 | 12.5 | 0.02 |
| GOLGA6L10 | chr15:80420179-80428761 | 3.91 | 29.81 | 0.04 |
| GIYD1 | chr16:29373375-29377041 | 1.83 | 6.9 | 0.07 |
| LRRC37A4 | chr17:40939892-40948305 | 9.3 | 16.22 | 0.24 |
| C2orf78 | chr2:73864823-73897782 | 14.03 | 10.88 | 0.39 |
| MGC13005 | chr2:114073074-114075623 | 5.51 | 24.27 | 0.01 |
| RPL23AP53 | chr8:148346-172318 | 3.15 | 27.08 | 0.09 |
| LOC728323 | chr2:242679516-242751142 | 4 | 19.18 | 0.09 |
| POM121L4P | chr22:19373842-19376009 | 33.18 | 50.71 | 0.3 |
| ZNF595 | chr4:43226-78099 | 1.27 | 10.63 | 0.18 |
| DRD5 | chr4:9392355-9394731 | 2.13 | 11.34 | 0.04 |
| LOC100272216 | chr5:68962735-68964784 | 82.32 | 63.64 | 0.13 |
| GUSBL1 | chr6:26947244-27032312 | 2.62 | 13.2 | 0.23 |
| LOC100170939 | chr5:69459044-69557378 | 2.1 | 9.73 | 0.22 |
| C6orf41 | chr6:27032750-27099731 | 2.99 | 11.48 | 0.15 |
| STAG3L1 | chr7:74826382-74834926 | 0.34 | 8.61 | 0.05 |
| GTF2IP1 | chr7:72206961-72259270 | 0.09 | 6.81 | 0 |
| SPDYE5 | chr7:74962234-74971564 | 12.4 | 37.77 | 0.25 |

| | | | | |
|-----------|------------------------|------|-------|------|
| AQP7 | chr9:33374947-33392517 | 1.19 | 11.7 | 0.04 |
| KGFLP1 | chr9:44185511-44246438 | 1.16 | 11.29 | 0.03 |
| FAM95B1 | chr9:42458584-42464233 | 2.26 | 16.19 | 0.02 |
| LOC642929 | chr9:43130534-43135480 | 1.94 | 14.62 | 0.07 |

Table S12. Summary of SUN markers found among:

a) 12 high coverage genomes

| Chromosome(s) | # SUNs | Found in at least one sample | | Found in >=66.7% of samples | | Found in >=83% of samples | | Found in all, or all but one | | Found in all samples | |
|-------------------|---------|------------------------------|--------|-----------------------------|--------|---------------------------|--------|------------------------------|--------|----------------------|--------|
| chrY (of males) | 163025 | 162715 | 99.81% | ND* | | ND* | | 152857 | 93.76% | 134327 | 82.40% |
| chrX (of females) | 257838 | 257350 | 99.81% | ND* | | ND* | | 234449 | 90.93% | 205982 | 79.89% |
| Autosomal | 3606033 | 3594803 | 99.69% | 3458149 | 95.90% | 3410578 | 94.58% | 3267192 | 90.60% | 3058924 | 84.83% |
| TOTAL | 4026896 | 4014868 | 99.70% | 3861166 | 95.88% | 3793594 | 94.21% | 3654498 | 90.75% | 3399233 | 84.41% |

b) full set of 159 genomes

| | | Found in at least one sample | | Found in >=66.7% of samples | | Found in >=80% of samples | | Found in >=90% of samples | | Found in all samples | |
|-------------------|---------|------------------------------|--------|---------------------------------|--------|---------------------------------|--------|--------------------------------|--------|----------------------------|-------|
| chrY (of males) | 163025 | 162838 | 99.89% | 59245 (in >= 48/70 males) | 36.34% | 16835 (in >= 57/70 males) | 10.33% | 10456 (in >= 64/70 males) | 6.41% | 4586 (in 70/70 males) | 2.81% |
| chrX (of females) | 257838 | 257623 | 99.92% | 225691 (in >= 59/88 females) | 87.53% | 154755 (in >= 59/88 females) | 60.02% | 33279 (in >= 59/88 females) | 12.91% | 3698 (in 88/88 females) | 1.43% |
| Autosomal | 3606033 | 3602413 | 99.90% | 2926297 | 81.15% | 2079587 | 57.67% | 679645 | 18.85% | 137371 | 3.81% |

*too few high coverage genomes (9 males, 3 females) to accurately estimate

Table S13. List of regions showing mirror effects, in which copy number variation occurring at a distant but homologous locus is detected by CGH or total read depth mapping but not supported by paralog-specific markers.

| | | | |
|--------------------------|---------------------------|--------------------------|--------------------------|
| chr1:499-77296 | chr14:105144802-105204335 | chr19:8702687-8764008 | chr6:170773423-170896961 |
| chr1:511276-713934 | chr14:105490591-105556154 | chr19:38161285-38223951 | chr7:34035-137131 |
| chr1:1557744-1673566 | chr14:105641496-105701838 | chr19:45054470-45098440 | chr7:32965738-33019569 |
| chr1:6399215-6437393 | chr15:18260050-18809093 | chr19:48308360-48340279 | chr7:38251784-38277712 |
| chr1:12764158-12975525 | chr15:19578259-19663437 | chr19:59931522-60053003 | chr7:43965543-44053608 |
| chr1:13025610-13071008 | chr15:19806304-19928954 | chr2:13417182-13484676 | chr7:56838073-56864221 |
| chr1:13084311-13122423 | chr15:20197602-20373877 | chr2:87420977-87476309 | chr7:57672481-57704192 |
| chr1:13218798-13328379 | chr15:20769611-20821158 | chr2:87905414-88068613 | chr7:57737500-57780156 |
| chr1:16927478-16998216 | chr15:20851648-21066254 | chr2:88989402-89325711 | chr7:62844017-62869981 |
| chr1:17048309-17078719 | chr15:21881157-21930208 | chr2:89613729-89913809 | chr7:64533527-64579210 |
| chr1:21608059-21687749 | chr15:26121887-26158746 | chr2:91225894-91333583 | chr7:64612545-64679379 |
| chr1:25457726-25537715 | chr15:26201755-26321807 | chr2:94689944-94777006 | chr7:64731171-64785020 |
| chr1:25561153-25626740 | chr15:26321862-26378706 | chr2:94819012-94932793 | chr7:64873354-64944010 |
| chr1:83370836-83727807 | chr15:26478772-26742609 | chr2:95955189-96011581 | chr7:66373788-66405265 |
| chr1:103937591-104122772 | chr15:26748304-26774250 | chr2:97162966-97221482 | chr7:71941119-71987586 |
| chr1:108580035-108656868 | chr15:28228784-28467743 | chr2:106451384-106498087 | chr7:73939568-73973416 |
| chr1:108714654-108802432 | chr15:29695733-29806849 | chr2:110044996-110090723 | chr7:74953470-74982353 |
| chr1:116931478-117008345 | chr15:30233196-30468651 | chr2:111724832-112091455 | chr7:75898835-75982664 |
| chr1:120328470-120498601 | chr15:30485989-30532114 | chr2:112191129-112298835 | chr7:75983944-76011844 |
| chr1:120556168-120738188 | chr15:45148821-45183847 | chr2:112309964-112357768 | chr7:76473646-76538796 |
| chr1:120788253-120844675 | chr15:70719459-70750478 | chr2:113884126-113983779 | chr7:99734143-99775575 |
| chr1:142004106-142090112 | chr15:72145166-72169711 | chr2:114038741-114078338 | chr7:100420686-100434591 |
| chr1:142186144-142261160 | chr15:73329425-73378675 | chr2:165536378-165565682 | chr7:101758653-101785006 |
| chr1:142436913-142562145 | chr15:75990028-76014384 | chr20:25679048-25714947 | chr7:101895096-102119471 |
| chr1:142612557-142700497 | chr15:80420460-80468829 | chr20:25821968-25860787 | chr7:126301909-126340192 |
| chr1:143020644-143113064 | chr15:80676755-80851330 | chr20:26007188-26064813 | chr7:142926646-143202833 |
| chr1:143522106-143594098 | chr15:80906475-81016695 | chr20:28141672-28267541 | chr7:143508477-143539578 |
| chr1:143900869-144094047 | chr15:82739692-82775415 | chr21:13852857-14025073 | chr7:143651360-143665697 |
| chr1:146349000-146460960 | chr15:83523026-83609542 | chr21:14205091-14274260 | chr7:143667406-143705959 |
| chr1:146542717-146712807 | chr15:100209456-100338508 | chr22:15241227-15430274 | chr7:151535690-151621339 |
| chr1:146840958-146919371 | chr16:2528458-2575015 | chr22:17038085-17257818 | chr7:157813890-157828171 |
| chr1:147004580-147090322 | chr16:5069063-5132802 | chr22:18706922-18750743 | chr8:73-160197 |
| chr1:147776324-148071632 | chr16:16320057-16404187 | chr22:18939490-19056898 | chr8:6813531-6866789 |
| chr1:194976941-195091905 | chr16:18343176-18418104 | chr22:19369489-19424484 | chr8:7047741-7461991 |
| chr1:195146908-195191991 | chr16:18775476-18844533 | chr22:19791592-20008506 | chr8:7606131-7904302 |

| | | | |
|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| chr1:204239359-204384868 chr1:204620809-204650800 chr1:205762356-205821816 chr1:220705662-220760196 chr1:222163349-222246574 chr1:232984323-233024889 chr1:241220014-241331740 chr1:246670883-246701528 chr1:50073-125923 chr10:37480058-37535779 chr10:38733715-38812330 chr10:38928865-39116929 chr10:47442802-47486994 chr10:47725753-47774013 chr10:48365756-48489335 chr10:48935328-49060727 chr10:51536286-51588070 chr10:81128690-81224323 chr10:81495386-81567953 chr10:89179686-89265924 chr11:50059-206243 chr11:3224806-3315713 chr11:3382657-3631565 chr11:4195474-4343961 chr11:60723617-60777607 chr11:71176703-71281616 chr11:88254398-88307369 chr12:17370-64628 chr12:8205041-8481089 chr12:9524006-9609667 chr12:31116113-31245092 chr12:52754366-52806269 chr13:18065871-18110282 chr13:34562609-34609933 chr13:56609653-56650416 chr13:92080803-92145168 chr14:18070219-18431517 chr14:19264340-19377293 chr14:27280845-27401766 chr14:73064493-73121425 | chr16:21634210-21717977 chr16:21775793-21853183 chr16:29290525-29560625 chr16:30107412-30212745 chr16:31867401-32004307 chr16:32151318-32204882 chr16:32755320-32946521 chr16:33004830-33206817 chr16:33508063-33725429 chr16:68528337-68601385 chr16:68791706-68842482 chr16:72917919-73021767 chr16:86067717-86121360 chr16:86446982-86517872 chr16:88689291-88822221 chr17:4965165-4989740 chr17:18869145-19081192 chr17:20284345-20432628 chr17:21473391-21507142 chr17:22981956-23027009 chr17:23091385-23119190 chr17:31430268-31469339 chr17:31505984-31699929 chr17:33324909-33429029 chr17:33529260-33666666 chr17:33869857-33942552 chr17:40929257-41067483 chr17:41762682-42142846 chr17:42446144-42532220 chr17:42966030-43026172 chr17:55434740-55460046 chr17:55527349-55559551 chr17:57650372-57720112 chr17:60263711-60377923 chr17:63378002-63419729 chr18:657-95963 chr18:14421322-14589321 chr18:42796008-42816438 chr19:11070-79285 chr19:143071-196320 | chr22:20008734-20127283 chr22:20142750-20247196 chr22:21979375-22002856 chr22:22133474-22158812 chr22:22963691-23005073 chr22:23334960-23387034 chr3:587318-627139 chr3:75464498-75899889 chr3:131246442-131405726 chr3:197157403-197201638 chr3:199317200-199446787 chr4:8933780-9021628 chr4:9117494-9354801 chr4:48975057-49025982 chr4:70060340-70266462 chr4:119557627-119586189 chr4:132778147-133121338 chr4:191032481-191137850 chr4:191182042-191222546 chr5:17561678-17668989 chr5:21296394-21429236 chr5:21487748-21556948 chr5:21952601-22061452 chr5:34204972-34303863 chr5:49921606-50025172 chr5:61546963-61591513 chr5:68865176-69131061 chr5:69319374-69433460 chr5:69569367-69817074 chr5:69865639-70533749 chr5:70567964-70697786 chr5:85602854-85633701 chr5:180643835-180763081 chr6:26831072-26883319 chr6:26950195-27097376 chr6:32066995-32093119 chr6:32099567-32124504 chr6:46903506-46939935 chr6:58352136-58382297 chr6:160940570-160990675 | chr8:7907662-7927204 chr8:8024623-8146538 chr8:11937321-11977411 chr8:12023029-12099266 chr8:12199406-12597502 chr8:86737456-86763666 chr8:86851058-86913407 chr9:485-199354 chr9:38754947-39054160 chr9:39134780-39270784 chr9:39703727-39800094 chr9:40014824-40222966 chr9:40465886-40617095 chr9:40752326-40823000 chr9:41405823-41469268 chr9:41801881-41910251 chr9:41961482-42332096 chr9:43150072-43203636 chr9:44011107-44055066 chr9:44084665-44118479 chr9:44191033-44375254 chr9:45267857-45616738 chr9:46261321-46351007 chr9:65207526-65658150 chr9:65708247-65927481 chr9:67007950-67056088 chr9:67106152-67379756 chr9:67627850-68003967 chr9:68054080-68142779 chr9:68278801-68379303 chr9:69797264-69975213 chr9:83717389-83756427 chr9:114860491-114896671 chr9:140102069-140273191 chrX:49061057-49129879 chrX:114866273-114919727 chrX:119890554-119948231 |
|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|

Table S14. GenBank accessions for 383 fosmid clone insert sequences corresponding to deletions used to validate SUN-based paralog-specific genotyping.

| | | | | | | | |
|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| AC208148 AC210994 AC217141 AC210971 AC211947 AC210431 AC225590 AC216084 AC211782 AC213505 AC208591 AC210708 AC209280 AC208065 AC213246 AC209539 AC226589 AC212295 AC209547 AC231268 AC225832 AC225576 AC208179 AC213260 AC207977 AC203608 AC209239 AC225033 AC209237 AC214014 AC208167 AC216809 AC220942 AC214174 AC207984 | AC216799 AC225381 AC225383 AC225622 AC207974 AC212911 AC210894 AC226182 AC215706 AC229892 AC216804 AC225578 AC210709 AC213232 AC208867 AC208868 AC216746 AC206480 AC226386 AC213294 AC210432 AC213727 AC213729 AC213752 AC215707 AC213273 AC216132 AC216796 AC203645 AC214985 AC208598 AC212262 AC207302 AC225998 AC210882 | AC209543 AC231957 AC207996 AC205871 AC214989 AC216897 AC215924 AC221040 AC213282 AC208872 AC225589 AC223410 AC207435 AC217015 AC208867 AC216230 AC215932 AC213290 AC231961 AC214991 AC203652 AC226178 AC206483 AC217013 AC208052 AC212591 AC209550 AC215990 AC213291 AC209317 AC206478 AC207447 AC226624 AC210913 AC226004 | AC225597 AC209553 AC203586 AC210769 AC215277 AC212839 AC208949 AC212900 AC215999 AC210963 AC208006 AC208062 AC215330 AC213467 AC225633 AC226631 AC216284 AC226066 AC210900 AC207579 AC209198 AC209561 AC217055 AC203664 AC206485 AC213469 AC207580 AC212908 AC225642 AC226103 AC216007 AC210408 AC207431 AC208070 AC210965 | AC210423 AC210988 AC217060 AC206486 AC225702 AC207974 AC213535 AC210436 AC209204 AC207593 AC216693 AC215336 AC226126 AC226739 AC206489 AC210996 AC216026 AC213259 AC219165 AC213059 AC226743 AC213061 AC225708 AC215337 AC210440 AC225581 AC207595 AC217139 AC209277 AC214988 AC210993 AC208151 AC216741 AC225327 AC214075 | AC214076 AC229602 AC205938 AC216064 AC215704 AC211406 AC215523 AC225767 AC229864 AC209285 AC225386 AC210702 AC207606 AC216967 AC226158 AC217409 AC212750 AC225553 AC225769 AC226160 AC212494 AC216748 AC216080 AC213215 AC216116 AC206020 AC209287 AC207054 AC214158 AC216816 AC229890 AC217411 AC215696 AC207711 AC210710 | AC214160 AC217516 AC216810 AC216091 AC226063 AC230047 AC210755 AC214162 AC213243 AC203589 AC211949 AC209297 AC206358 AC225779 AC216896 AC208299 AC225573 AC215521 AC215698 AC216995 AC206361 AC217544 AC203590 AC216959 AC207169 AC214170 AC207780 AC210761 AC231120 AC211952 AC216993 AC209303 AC213255 AC225574 AC226186 | AC213263 AC217324 AC216970 AC211957 AC218915 AC206436 AC226450 AC207966 AC210773 AC209340 AC217952 AC216119 AC231522 AC203603 AC220967 AC225385 AC226451 AC203620 AC210879 AC211963 AC208508 AC223425 AC225914 AC225585 AC216124 AC216973 AC214811 AC206438 AC216131 AC225387 AC231952 AC208584 AC216976 AC206469 AC215795 |
|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|

| | | | | | | | |
|----------|----------|----------|----------|----------|----------|----------|----------|
| AC209048 | AC215924 | AC208877 | AC209200 | AC226134 | AC226172 | AC217189 | AC226526 |
| AC213220 | AC206475 | AC208059 | AC204959 | AC208159 | AC216238 | AC207170 | AC207179 |
| AC213737 | AC209541 | AC217514 | AC207367 | AC211321 | AC213239 | AC231265 | AC212261 |
| AC213744 | AC226595 | AC223412 | AC215331 | AC226156 | AC225571 | AC206416 | AC210881 |
| AC213749 | AC221032 | AC214993 | AC225695 | AC225708 | AC208185 | AC213261 | AC213272 |
| AC207778 | AC216989 | AC209545 | AC225038 | AC211237 | AC216133 | AC217956 | AC209540 |
| AC225916 | AC216991 | AC209552 | AC216022 | AC215328 | AC216082 | AC203598 | AC214980 |
| AC216921 | AC203647 | AC216232 | AC216115 | AC210539 | AC225098 | AC210764 | AC225586 |
| AC210542 | AC206476 | AC226630 | AC209202 | AC213116 | AC225778 | AC225580 | AC207985 |
| AC214809 | AC212490 | AC226061 | AC204970 | AC215344 | AC230043 | AC208389 | AC203622 |
| AC206736 | AC225588 | AC207578 | AC213226 | AC216800 | AC209289 | AC226371 | AC207301 |
| AC214182 | AC207429 | AC215992 | AC208107 | AC207604 | AC203588 | AC225582 | AC221030 |
| AC212258 | AC216179 | AC210916 | AC226695 | AC217405 | AC206356 | AC207221 | |

Table S15. Clones resequenced overlapping predicted duplication-embedded deletions

| Supple- mentary Note Figure (if shown) | Band and approximate location | Genes | Genome-wide psCN map | Fosmid end sequence mapping | Individual | Clone ID | Bar- code | Reads on- target | | Notes | | | | | | | | |
|-------------------------------------------------------|------------------------------------------|------------------------------------------------------|--------------------------------|--------------------------------------------------|------------------------|------------------------|--------------|---------------------|-----------------------------------------------------------|----------------------------------------------------------------|--------------------------------|--------------------------------------------------|------------------------|------------------------|-----|-----|-----------------------------------------------------------|-------------------|
| | | | | | | | | Run A | Run B | | | | | | | | | |
| Fig. S61 | 4p16, 9.05-9.10 Mbp | <i>DEFB131</i> | Hemizygous deletion | Deletion | NA18517 | ABC7_43084 800_J22 | C2 | Yes | Yes | Supports predicted deletion | | | | | | | | |
| | | | | | | ABC9_43883 700_M19 | C3 | Yes | Yes | Supports predicted deletion | | | | | | | | |
| | | | | | | ABC11_4824 0400_H11 | C4 | Yes | Yes | Supports predicted deletion | | | | | | | | |
| | | | | | | ABC12_4904 8900_K21 | C5 | Yes | Yes | Supports predicted deletion | | | | | | | | |
| | | | | | | ABC14_5020 5200_K7 | C6 | Yes | Yes | Supports predicted deletion | | | | | | | | |
| Fig. S62 | 19q13.31, 47.9-48.3 Mb | <i>PSG3, PSG8, PSG1, PSG7, PSG11</i> | Hemizygous deletion | Deletion | NA18956 | ABC9_43881 200_L12 | G10 | Yes | Yes | Supports predicted deletion | | | | | | | | |
| | | | | | | ABC9_43862 300_O17 | G11 | Yes | Yes | Supports predicted deletion | | | | | | | | |
| Fig. S63 | 2q11.2, ~97.61 Mbp | <i>ANKRD36B</i> | Homozygous deletion (~8 kb) | Normal (deletion < detection threshold) | NA12156 | ABC14_1038 314_G16 | B1 | Yes | Yes | Supports predicted deletion but does not confirm distal end | | | | | | | | |
| | | | | | | | | | | | Homozygous deletion (~8 kb) | Normal (deletion < detection threshold) | NA12878 | ABC12_4794 8200_G24 | B2 | Yes | Yes | Supports deletion |
| | | | | | | | | | | | | | | | | | | |
| | Deletion, or interlocus conversion | | Inversion | NA12878 | ABC12_4698 4600_N4 | B4 | Yes | No | No coverage | | | | | | | | | |
| | | | | | | | | | | Deletion, or interlocus conversion | Inversion | NA18517 | ABC7_42377 600_B5 | B5 | Yes | No | Supports deletion or interlocus conversion (not shown) | |
| | Deletion, or interlocus conversion | | Normal | NA18517 | ABC7_42407 100_N7 | B6 | Yes | Yes | Supports deletion or interlocus conversion (not shown) | | | | | | | | | |
| Fig. S64 | | 2q12.2, ~106.25 Mb | | | | | | | | Homozygous deleted | deletion | NA18555 | ABC11_4822 3100_B23 | D1 | yes | no | Supports deletion | |
| | Homozygous deleted | | deletion | NA19240 | ABC10_4448 9600_C1 | D2 | yes | yes | Supports deletion | | | | | | | | | |
| | Hemizygous deleted | | deletion | NA12156 | ABC14_5097 6300_M13 | D3 | yes | yes | Supports deletion | | | | | | | | | |

| | | | | | | | | | | |
|----------|------------------------------|---------------------------------------------------------|---------------------------------------------------|---------------------------------|---------|------------------------|-----|-----|-----|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| | | | Hemizygous deleted | concordant (non-deleted allele) | NA12156 | ABC14_5014 4300_G23 | D4 | yes | yes | Supports non-deletion allele |
| | | | Hemizygous deleted | concordant (non-deleted allele) | NA12156 | ABC14_5015 1700_G24 | D5 | yes | yes | Supports non-deletion allele |
| Fig. S65 | 4q13, ~68.81-70.38 Mbp | <i>UGT2B17</i> , <i>UGT2B15</i> , <i>TMPRSS1E</i> | Hemizygous deletion in vicinity of <i>UGT2B17</i> | deletion | NA19240 | ABC10_1585 670_B20 | B1 | Yes | Yes | Misassembly in reference causes apparent deletion |
| | | | Hemizygous deletion in vicinity of <i>UGT2B17</i> | deletion | NA12878 | ABC12_4796 5400_B12 | B2 | Yes | Yes | Misassembly in reference causes apparent deletion |
| | | | Homozygous deletion in vicinity of <i>UGT2B28</i> | deletion | NA19129 | ABC13_4874 4100_J5 | B3 | No | Yes | Supports deletion |
| Fig. S66 | 8p23.1 | <i>FAM86B1</i> , <i>DEFB130</i> | Homozygous deletion within amplified region | Deletion | NA12878 | ABC12_7942 349_K16 | D8 | Yes | Yes | Supports predicted deletion; difference in breakpoints between clones plus genome-wide psCN signal suggests two distinct nested deletions within an amplified region with psCN = 4 flanking these deletions |
| | | | Homozygous deletion within amplified region | Deletion | NA12878 | ABC12_4920 6400_O2 | D9 | Yes | Yes | Supports predicted deletion |
| Fig. S67 | 17q12, 31.40-32.02 Mbp | <i>CCL3L1/4L/4L2/3L3</i> , <i>TBC1D3C/F/G</i> | Homozygous deleted, except for small patches | deletion | NA12156 | ABC14_1185 722_G2 | C9 | No | No | No coverage |
| | | | Homozygous deleted, except for small patches | deletion | NA12156 | ABC14_5041 8700_K9 | C10 | Yes | Yes | Supports deletion |
| | | | Homozygous deleted, except for small patches | deletion | NA12878 | ABC12_4688 1000_B6 | C11 | No | Yes | Supports deletion |
| | | | Homozygous deleted, except for small patches | deletion | NA12878 | ABC12_4921 2300_D22 | C12 | Yes | Yes | Supports deletion - different breakpoints |
| Fig. S68 | 15q11.2, 19.33-19.92 Mbp | | Deletion(s) within amplified region | deletion | NA18956 | ABC9_43887 900_I3 | C7 | Yes | Yes | Supports deletion |
| | | | Deletion(s) within amplified region | deletion | NA18956 | ABC9_43885 500_N14 | C8 | Yes | Yes | Supports deletion |
| Fig. S69 | 17q21.31-32, 41.71-41.97 Mbp | <i>LRRC37A/A2</i> , <i>ARL14/17/17B</i> | Deletions within amplified region | Deletion | NA12878 | ABC12_4783 7300_D11 | E1 | Yes | No | Supports deletion; does not confirm distal end. Patch of hits to proximal duplication at proximal duplication (~41.000 Mbp) supports alternative architecture bringing that patch in proximity to this deletion. |
| | | | Deletions within amplified region | Deletion | NA12878 | ABC12_4920 6300_L15 | E2 | Yes | Yes | Same |
| | | | Homozygous deletion | Deletion | NA18507 | ABC8_43243 500_A3 | E3 | No | No | No coverage |
| | | | Homozygous deletion | Deletion | NA18507 | ABC8_21498 40_M3 | E4 | Yes | Yes | Supports deletion |
| | | | Partially homozygous (i.e. | Inversion | NA18507 | ABC8_42582 000_E3 | E5 | Yes | Yes | Supports deletion |

| | | | | | | | | | | |
|-----------|----------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------|-----------------------------|---------|------------------------|-----|-----|-----|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| | | | Homozygous deletion | Deletion | NA18517 | ABC7_43063 500_H12 | E6 | No | No | No coverage |
| | | | Homozygous deletion | Deletion | NA18517 | ABC7_42485 000_O5 | E7 | No | Yes | Supports deletion |
| Fig. S70 | 15q11.2, 20.254-21.009 Mbp | <i>TUBGCP5,</i> <i>CYFIP1,</i> <i>NIPA2,</i> <i>NIPA1,</i> <i>WHAMML,</i> <i>HERC2P2,</i> <i>GOLGA9P/</i> <i>8D/8E/6L1</i> <i>NIPA1</i> | Hemizygous deletion | Deletion | NA18507 | ABC8_41095 200_D24 | D11 | No* | No* | *No coverage among SUNKs, but non-unique positions at locus are covered. Deletion haplotype may be contained on unanchored contigs (chr15_random) which were excluded from SUN mapping. Deletion confirms by array CGH, FISH, and Mendelian consistency. |
| | 15q11.2, 20.632-20.690 Mbp | | Single-copy deletion (to diploid) nested psCN = 3 amplification | Deletion | NA18555 | ABC11_4959 7500_C7 | D12 | Yes | Yes | Supports predicted deletion |
| Not shown | 9p12, ~39.16-39.74 Mbp | | Hemizygous deletion | Deletion | NA19240 | ABC10_4458 8000_L12 | A1 | Yes | Yes | Supports predicted deletion |
| | | | Hemizygous deletion | Normal (non-deleted allele) | NA19240 | ABC10_4362 3500_O17 | A2 | Yes | Yes | Supports presence of non-deleted allele |
| | | | Hemizygous deletion | Normal (non-deleted allele) | NA19240 | ABC10_4447 7800_E4 | A3 | Yes | Yes | Supports presence of non-deleted allele |
| | | | Homozygous deletion | deletion | NA12878 | ABC12_4902 8400_G17 | A4 | No | No | No coverage |
| Not shown | 9p11, ~44.16-46.87 Mbp | | Deletion, partially homozygous | deletion | NA18507 | ABC8_41061 800_A22 | H10 | Yes | Yes | Supports predicted deletion |
| | | | Deletion, partially homozygous | inversion | NA19240 | ABC10_4447 2100_I19 | H11 | Yes | Yes | Supports deletion and/or inversion |
| | | | Deletion, partially homozygous | inversion | NA18555 | ABC11_4804 3400_I14 | H12 | Yes | Yes | Ambiguous - other end may be in gap |
| Not shown | 22q11.21 | | Hemizygous deletion with other nested events | deletion | NA12156 | ABC14_5041 7400_I12 | D10 | yes | yes | Supports deletion |
| Not shown | 12p13.31 | <i>KLRC2</i> <i>KLRC3</i> | Hemizygous deletion, ~10.449-10.475 Mbp (~26 kbp) | Deletion | NA19129 | ABC13_4861 1600_I24 | G5 | Yes | Yes | Ambiguous |
| | | | Hemizygous deletion, ~10.449-10.475 Mbp (~26 kbp) | Deletion | NA19129 | ABC13_4888 8000_M12 | G6 | Yes | Yes | Unclear |
| | 12p13.2 | | Hemizygous deletion ~11.111-11.138 Mbp (~37 kbp) | Deletion | NA19129 | ABC13_4872 9400_D23 | G9 | Yes | Yes | Supports deletion |

| | | | | | | | | | | |
|-----------|-----------------------------|----------------------------|---------------------------------------------------|----------|---------|------------------------|-----|-----|-----|-------------------|
| Not shown | | <i>PRB1</i> <i>PRB2</i> | Hemizygous deletion, ~11.397-11.433 Mbp (~46 kbp) | Deletion | NA12878 | ABC12_4928 3600_H13 | G7 | Yes | Yes | Supports deletion |
| | | | Hemizygous deletion, ~11.397-11.433 Mbp (~46 kbp) | Deletion | NA12878 | ABC12_4665 6200_N8 | G8 | Yes | Yes | Supports deletion |
| Not shown | 15q11.2 | | Hemizygous deletion, ~22.204-22.276 Mbp | Deletion | NA18517 | ABC7_43086 600_F18 | F10 | Yes | Yes | Supports deletion |
| | | | Hemizygous deletion, ~22.204-22.276 Mbp | Deletion | NA18507 | ABC8_40979 300_O18 | F11 | Yes | Yes | Supports deletion |
| | | | Hemizygous deletion, ~22.204-22.276 Mbp | Deletion | NA19129 | ABC13_9266 22_P22 | F12 | Yes | Yes | Supports deletion |
| Not shown | 1q23.3 | | Hemizygous deletion | Deletion | NA18517 | ABC7_42430 700_E21 | F5 | Yes | Yes | Supports deletion |
| Not shown | 7p22.1, ~68.54-68.84 Mbp | | Hemizygous deletion | Deletion | NA18507 | ABC8_40884 700_F19 | F8 | Yes | Yes | Supports deletion |
| | | | | | | ABC8_40870 800_M21 | F9 | Yes | Yes | Supports deletion |

Table S16. Selected high-scoring loci for interlocus conversion signatures

| Duplication coordinates | | Percent identity | Conversion signature score | Gene(s) | Reference(s) |
|-------------------------|---------------------|------------------|----------------------------|------------------------------------------------------------------------------------------------------------------------|--------------|
| chr8 | chr8 | 98.55% | 19.39 | <i>DUB3, FAM86B1</i> | |
| 12027750-12099344 | 12276250-12348580 | | | | |
| chr8 | chr8 | 98.93% | 16.02 | <i>DEFA10P, DEFA1B, DEFA1, DEFA3</i> | (S57) |
| 6815345-6834499 | 6834500-6853602 | | | | |
| chr5 | chr5 | 97.52% | 14.72 | <i>ZDHHC11</i> | |
| 783526-820318 | 874953-904001 | | | | |
| chr10 | chr10 | 99.79 | 14.69 | <i>MRC, MRC1L1</i> | |
| 17817718-18014681 | 18064682-18261586 | | | | |
| chr1 | chr1 | 98.61% | 12.64 | <i>OR2T34, OR2T29, OR2T3, OR2T5</i> | |
| 246689711-246749190 | 246758720-246818330 | | | | |
| chr2 | chr2 | 99.24% | 10.83 | <i>CFC1, CFC1B</i> | |
| 130923990-131193978 | 130875596-131143174 | | | | |
| chr8 | chr8 | 99.10% | 10.11 | <i>DEFB103B, SPAG11B, DEFB104A, DEFB106A, DEFB105A, DEFB107A, HE2, DEFB4, DEFB109, TRIM49, TRIM53, TRIM64B, TRIM64</i> | (S58, 59) |
| 7121221-7431099 | 7629812-7929091 | | | | |
| chr11 | chr11 | 99.60% | 9.88 | | |
| 89115225-89293439 | 89296665-89470331 | | | | |
| chr10 | chr10 | 99.78% | 9.06 | <i>PTPN20A, PTPN20B</i> | |
| 45896971-46325469 | 48284311-48715542 | | | | |
| chr5 | chr5 | 99.14% | 7.45 | <i>THOC3, NY-REN-7</i> | (S60) |
| 175282971-175491278 | 177066105-177280072 | | | | |
| chr16 | chr16 | 98.06% | 6.68 | <i>SLC6A10P, IGHV</i> | |
| 32717851-32943799 | 33539405-33771971 | | | | |
| chr7 | chr7 | 99.07% | 13.22 | <i>RSPH10B, PMS2, PMS2CL</i> | (S61) |
| 5899795-5995202 | 6741801-6839043 | | | | |

SUPPORTING REFERENCES

- S1. F. Hach *et al.*, *Nat Methods* **7**, 576 (Aug).
- S2. M. A. Quail *et al.*, *Nat Methods* **5**, 1005 (Dec, 2008).
- S3. C. Alkan *et al.*, *Nat Genet* **41**, 1061 (Oct, 2009).
- S4. P. Lichter *et al.*, *Science* **247**, 64 (Jan 5, 1990).
- S5. D. F. Conrad *et al.*, *Nature*, (Oct 7, 2009).
- S6. N. Craddock *et al.*, *Nature* **464**, 713 (Apr 1, 2010).
- S7. A. Itsara *et al.*, *Am J Hum Genet* **84**, 148 (Feb, 2009).
- S8. S. A. McCarroll, D. M. Altshuler, *Nat Genet* **39**, S37 (Jul, 2007).
- S9. J. M. Kidd *et al.*, *Nature* **453**, 56 (May 1, 2008).
- S10. H. Park *et al.*, *Nat Genet* **42**, 400 (May, 2010).
- S11. J. I. Kim *et al.*, *Nature* **460**, 1011 (Aug 20, 2009).
- S12. S. M. Ahn *et al.*, *Genome Res* **19**, 1622 (Sep, 2009).
- S13. S. C. Schuster *et al.*, *Nature* **463**, 943 (Feb 18, 2010).
- S14. J. Wang *et al.*, *Nature* **456**, 60 (Nov 6, 2008).
- S15. R. E. Green *et al.*, *Nature* **444**, 330 (Nov 16, 2006).
- S16. G. Benson, *Nucleic Acids Res* **27**, 573 (Jan 15, 1999).
- S17. A. J. Iafrate *et al.*, *Nat Genet* **36**, 949 (Sep, 2004).
- S18. J. A. Bailey *et al.*, *Science* **297**, 1003 (Aug 9, 2002).
- S19. J. A. Bailey, A. M. Yavor, H. F. Massa, B. J. Trask, E. E. Eichler, *Genome Res* **11**, 1005 (Jun, 2001).
- S20. W. J. Kent *et al.*, *Genome Res* **12**, 996 (Jun, 2002).
- S21. S. A. McCarroll *et al.*, *Nat Genet* **40**, 1166 (Oct, 2008).
- S22. J. P. Noonan *et al.*, *Am J Hum Genet* **72**, 621 (Mar, 2003).
- S23. T. Bhattacharya *et al.*, *Nat Med* **15**, 1112 (Oct, 2009).
- S24. S. F. Field *et al.*, *Nat Med* **15**, 1115 (Oct, 2009).
- S25. E. Gonzalez *et al.*, *Science* **307**, 1434 (Mar 4, 2005).
- S26. W. He *et al.*, *Nat Med* **15**, 1117 (Oct, 2009).
- S27. T. J. Urban *et al.*, *Nat Med* **15**, 1110 (Oct, 2009).
- S28. R. Redon *et al.*, *Nature* **444**, 444 (Nov 23, 2006).
- S29. D. Brown, J. Trowsdale, R. Allen, *Tissue Antigens* **64**, 215 (Sep, 2004).
- S30. K. Hirayasu *et al.*, *Am J Hum Genet* **82**, 1075 (May, 2008).
- S31. Y. Xue *et al.*, *Am J Hum Genet* **83**, 337 (Sep, 2008).
- S32. D. M. Altshuler *et al.*, *Nature* **467**, 52 (Sep 2, 2010).
- S33. M. E. Johnson *et al.*, *Nature* **413**, 514 (Oct 4, 2001).
- S34. G. H. Perry *et al.*, *Nat Genet* **39**, 1256 (Oct, 2007).
- S35. J. M. Kidd, T. L. Newman, E. Tuzun, R. Kaul, E. E. Eichler, *PLoS Genet* **3**, e63 (Apr 20, 2007).
- S36. H. Stefansson *et al.*, *Nature* **455**, 232 (Sep 11, 2008).
- S37. M. Doornbos *et al.*, *Eur J Med Genet* **52**, 108 (Mar-Jun, 2009).
- S38. S. K. Murthy *et al.*, *Cytogenet Genome Res* **116**, 135 (2007).
- S39. K. Vandepoele, V. Andries, F. van Roy, *Mol Biol Evol* **26**, 1321 (Jun, 2009).
- S40. I. Moore *et al.*, *Blood* **115**, 379 (Jan 14).

- S41. M. C. Popesco *et al.*, *Science* **313**, 1304 (Sep 1, 2006).
- S42. B. B. Rosenblum, A. D. Merritt, *Am J Hum Genet* **30**, 434 (Jul, 1978).
- S43. E. Tuzun *et al.*, *Nat Genet* **37**, 727 (Jul, 2005).
- S44. G. Q. Zhou *et al.*, *Placenta* **18**, 491 (Sep, 1997).
- S45. J. Sambrook, D. W. Russell, *Molecular cloning : a laboratory manual*. (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, N.Y., ed. 3rd, 2001).
- S46. J. M. Chen, D. N. Cooper, N. Chuzhanova, C. Ferec, G. P. Patrinos, *Nat Rev Genet* **8**, 762 (Oct, 2007).
- S47. N. D. Avent *et al.*, *Blood* **89**, 1779 (Mar 1, 1997).
- S48. H. Innan, *Proc Natl Acad Sci U S A* **100**, 8793 (Jul 22, 2003).
- S49. T. Kitano, N. Saitou, *J Mol Evol* **49**, 615 (Nov, 1999).
- S50. F. F. Wagner, W. A. Flegel, *Blood* **95**, 3662 (Jun 15, 2000).
- S51. G. P. Patrinos, P. Kollia, A. Loutradi-Anagnostou, D. Loukopoulos, M. N. Papadakis, *Hum Genet* **102**, 629 (Jun, 1998).
- S52. K. Ezawa, O. O. S, N. Saitou, *Mol Biol Evol* **23**, 927 (May, 2006).
- S53. D. R. Bentley *et al.*, *Nature* **456**, 53 (Nov 6, 2008).
- S54. R. E. Green *et al.*, *Science* **328**, 710 (May 7, 2010).
- S55. L. W. H. Tarjei S. Mikkelsen, Evan E. Eichler, Michael C. Zody, David B. Jaffe, Shiaw-Pyng Yang, Wolfgang Enard, Ines Hellmann, Kerstin Lindblad-Toh, Tasha K. Altheide, Nicoletta Archidiacono, Peer Bork, Jonathan Butler, Jean L. Chang, Ze Cheng, Asif T. Chinwalla, Pieter deJong, Kimberley D. Delehaunty, Catrina C., L. L. F. Fronick, Yoav Gilad, Gustavo Glusman, Sante Gnerre, Tina A. Graves, Toshiyuki Hayakawa, Karen E. Hayden, H. J. Xiaoqiu Huang, W. James Kent, Mary-Claire King, Edward J. Kulbokas III, Ming K. Lee, Ge Liu, Carlos Lopez-Otin, O. M. Kateryna D. Makova, Elaine R. Mardis, Evan Mauceli, Tracie L. Miner, William E. Nash, Joanne O. Nelson, Svante Pa`a`bo, Nick J. Patterson, Craig S. Pohl, Katherine S. Pollard, Kay Pru`fer, Xose S. Puente, David Reich, Mariano Rocchi, Kate Rosenbloom, Maryellen Ruvolo, Daniel J. Richter, Stephen F. Schaffner, Arian F. A. Smit, Scott M. Smith, Mikita Suyama, James Taylor, David Torrents, Eray Tuzun, Ajit Varki, Gloria Velasco, Mario Ventura, John W. Wallis, Michael C. Wendl, Richard K. Wilson, Eric S. Lander & Robert H. Waterston, *Nature* **437**, 69 (Sep 1, 2005).
- S56. T. Marques-Bonet *et al.*, *Nature* **457**, 877 (Feb 12, 2009).
- S57. E. J. Hollox, J. C. Barber, A. J. Brookes, J. A. Armour, *Genome Res* **18**, 1686 (Nov, 2008).
- S58. C. Nusbaum *et al.*, *Nature* **439**, 331 (Jan 19, 2006).
- S59. S. Taudien *et al.*, *BMC Genomics* **5**, 92 (2004).
- S60. N. Kurotaki, P. Stankiewicz, K. Wakui, N. Niikawa, J. R. Lupski, *Hum Mol Genet* **14**, 535 (Feb 15, 2005).
- S61. J. Auclair *et al.*, *Hum Mutat* **28**, 1084 (Nov, 2007).

THE 1000 GENOMES CONSORTIUM

Participants are arranged by project role, then by institution alphabetically, and finally alphabetically within institutions except for Principal Investigators and Project Leaders, as indicated.

Steering Committee: David L. Altshuler (Co-Chair)²⁻⁴, Richard M. Durbin (Co-Chair)¹, Gonçalo R. Abecasis⁵, David R. Bentley⁶, Aravinda Chakravarti⁷, Andrew G. Clark⁸, Francis S. Collins⁹, Francisco M. De La Vega¹⁰, Peter Donnelly¹¹, Michael Egholm¹², Paul Flicek¹³, Stacey B. Gabriel², Richard A. Gibbs¹⁴, Bartha M. Knoppers¹⁵, Eric S. Lander², Hans Lehrach¹⁶, Elaine R. Mardis¹⁷, Gil A. McVean^{11,18}, Debbie A. Nickerson¹⁹, Leena Peltonen*, Alan J. Schafer²⁰, Stephen T. Sherry²¹, Jun Wang^{22,23}, Richard K. Wilson¹⁷

Production Group: Baylor College of Medicine Richard A. Gibbs (Principal Investigator)¹⁴, David Deiros¹⁴, Mike Metzker¹⁴, Donna Muzny¹⁴, Jeff Reid¹⁴, David Wheeler¹⁴ **BGI-Shenzhen** Jun Wang (Principal Investigator)^{22,23}, Jingxiang Li²², Min Jian²², Guoqing Li²², Ruiqiang Li^{22,23}, Huiqing Liang²², Geng Tian²², Bo Wang²², Jian Wang²², Wei Wang²², Huanming Yang²², Xiuqing Zhang²², Huisong Zheng²² **Broad Institute of MIT and Harvard** Eric S. Lander (Principal Investigator)², David L. Altshuler²⁻⁴, Lauren Ambrogio², Toby Bloom², Kristian Cibulskis², Tim J. Fennell², Stacey B. Gabriel (Co-Chair)², David B. Jaffe², Erica Shefler², Carrie L. Sougnez² **Illumina** David R. Bentley (Principal Investigator)⁶, Niall Gormley⁶, Sean Humphray⁶, Zoya Kingsbury⁶, Paula Koko-Gonzales⁶, Jennifer Stone⁶ **Life Technologies** Kevin J. McKernan (Principal Investigator)²⁴, Gina L. Costa²⁴, Jeffry K. Ichikawa²⁴, Clarence C. Lee²⁴ **Max Planck Institute for Molecular Genetics** Ralf Sudbrak (Project Leader)¹⁶, Hans Lehrach (Principal Investigator)¹⁶, Tatiana A. Borodina¹⁶, Andreas Dahl²⁵, Alexey N. Davydov¹⁶, Peter Marquardt¹⁶, Florian Mertes¹⁶, Wilfried Nietfeld¹⁶, Philip Rosenstiel²⁶, Stefan Schreiber²⁶, Aleksey V. Soldatov¹⁶, Bernd Timmermann¹⁶, Marius Tolzmann¹⁶ **Roche Applied Science** Michael Egholm (Principal Investigator)¹², Jason Affourtit²⁷, Dana Ashworth²⁷, Said Attiya²⁷, Melissa Bachorski²⁷, Eli Buglione²⁷, Adam Burke²⁷, Amanda Caprio²⁷, Christopher Celone²⁷, Shauna Clark²⁷, David Connors²⁷, Brian Desany²⁷, Lisa Gu²⁷, Lorri Guccione²⁷, Calvin Kao²⁷, Andrew Kebbel²⁷, Jennifer Knowlton²⁷, Matthew Labrecque²⁷, Louise McDade²⁷, Craig Mealmaker²⁷, Melissa Minderman²⁷, Anne Nawrocki²⁷, Faheem Niazi²⁷, Kristen Pareja²⁷, Ravi Ramenani²⁷, David Riches²⁷, Wanmin Song²⁷, Cynthia Turcotte²⁷, Shally Wang²⁷ **Washington University in St. Louis** Elaine R. Mardis (Co-Chair) (Co-Principal Investigator)¹⁷, Richard K. Wilson (Co-Principal Investigator)¹⁷, David Dooling¹⁷, Lucinda Fulton¹⁷, Robert Fulton¹⁷, George Weinstock¹⁷ **Wellcome Trust Sanger Institute** Richard M. Durbin (Principal Investigator)¹, John Burton¹, David M. Carter¹, Carol Churcher¹, Alison Coffey¹, Anthony Cox¹, Aarno Palotie^{1,28}, Michael Quail¹, Tom Skelly¹, James Stalker¹, Harold P. Swerdlow¹, Daniel Turner¹

Analysis Group: Agilent Technologies Anniek De Witte²⁹, Shane Giles²⁹ **Baylor College of Medicine** Richard A. Gibbs (Principal Investigator)¹⁴, David Wheeler¹⁴, Matthew Bainbridge¹⁴, Danny Challis¹⁴, Aniko Sabo¹⁴, Fuli Yu¹⁴, Jin Yu¹⁴ **BGI-Shenzhen** Jun Wang (Principal Investigator)^{22,23}, Xiaodong Fang²², Xiaosen Guo²², Ruiqiang Li^{22,23}, Yingrui Li²², Ruibang Luo²², Shuaishuai Tai²², Honglong Wu²², Hancheng Zheng²², Xiaole Zheng²², Yan Zhou²², Guoqing Li²², Jian Wang²², Huanming Yang²² **Boston College** Gabor T. Marth (Principal Investigator)³⁰, Erik P. Garrison³⁰, Weichun Huang³¹, Amit Indap³⁰, Deniz Kural³⁰, Wan-Ping Lee³⁰, Wen Fung Leong³⁰, Aaron R. Quinlan³², Chip Stewart³⁰, Michael P.

Stromberg³³, Alistair N. Ward³⁰, Jiantao Wu³⁰ **Brigham and Women's Hospital** Charles Lee (Principal Investigator)³⁴, Ryan E. Mills³⁴, Xinghua Shi³⁴ **Broad Institute of MIT and Harvard** Mark J. Daly (Principal Investigator)², Mark A. DePristo (Project Leader)², David L. Altshuler^{2,4}, Aaron D. Ball², Eric Banks², Toby Bloom², Brian L. Browning³⁵, Kristian Cibulskis², Tim J. Fennell², Kiran V. Garimella², Sharon R. Grossman^{2,36}, Robert E. Handsaker², Matt Hanna², Chris Hartl², David B. Jaffe², Andrew M. Kernysky², Joshua M. Korn², Heng Li², Jared R. Maguire², Steven A. McCarroll^{2,4}, Aaron McKenna², James C. Nemesh², Anthony A. Philippakis², Ryan E. Poplin², Alkes Price³⁷, Manuel A. Rivas², Pardis C. Sabeti^{2,36}, Stephen F. Schaffner², Erica Shefler², Ilya A. Shlyakhter^{2,36} **Cardiff University, The Human Gene Mutation Database** David N. Cooper (Principal Investigator)³⁸, Edward V. Ball³⁸, Matthew Mort³⁸, Andrew D. Phillips³⁸, Peter D. Stenson³⁸ **Cold Spring Harbor Laboratory** Jonathan Sebat (Principal Investigator)³⁹, Vladimir Makarov⁴⁰, Kenny Ye⁴¹, Seungtae C. Yoon⁴² **Cornell and Stanford Universities** Carlos D. Bustamante (Co-Principal Investigator)⁴³, Andrew G. Clark (Co-Principal Investigator)⁸, Adam Boyko⁴³, Jeremiah Degenhardt⁸, Simon Gravel⁴³, Ryan N. Gutenkunst⁴⁴, Mark Kaganovich⁴³, Alon Keinan⁸, Phil Lacroute⁴³, Xin Ma⁸, Andy Reynolds⁸ **European Bioinformatics Institute** Laura Clarke (Project Leader)¹³, Paul Flicek (Co-Chair, DCC) (Principal Investigator)¹³, Fiona Cunningham¹³, Javier Herrero¹³, Stephen Keenen¹³, Eugene Kulesha¹³, Rasko Leinonen¹³, William M. McLaren¹³, Rajesh Radhakrishnan¹³, Richard E. Smith¹³, Vadim Zalunin¹³, Xiangqun Zheng-Bradley¹³ **European Molecular Biology Laboratory** Jan O. Korbel (Principal Investigator)⁴⁵, Adrian M. Stütz⁴⁵ **Illumina** Sean Humphray (Project Leader)⁶, Markus Bauer⁶, R. Keira Cheetham⁶, Tony Cox⁶, Michael Eberle⁶, Terena James⁶, Scott Kahn⁶, Lisa Murray⁶ **Johns Hopkins University** Aravinda Chakravarti⁷ **Leiden University Medical Center** Kai Ye⁴⁶ **Life Technologies** Francisco M. De La Vega (Principal Investigator)¹⁰, Yutao Fu²⁴, Fiona C.L. Hyland¹⁰, Jonathan M. Manning²⁴, Stephen F. McLaughlin²⁴, Heather E. Peckham²⁴, Onur Sakarya¹⁰, Yongming A. Sun¹⁰, Eric F. Tsung²⁴ **Louisiana State University** Mark A. Batzer (Principal Investigator)⁴⁷, Miriam K. Konkel⁴⁷, Jerilyn A. Walker⁴⁷ **Max Planck Institute for Molecular Genetics** Ralf Sudbrak (Project Leader)¹⁶, Marcus W. Albrecht¹⁶, Vyacheslav S. Amstislavskiy¹⁶, Ralf Herwig¹⁶, Dimitri V. Parkhomchuk¹⁶ **US National Institutes of Health** Stephen T. Sherry (Co-Chair, DCC) (Principal Investigator)²¹, Richa Agarwala²¹, Hoda M. Khouri²¹, Aleksandr O. Morgulis²¹, Justin E. Paschall²¹, Lon D. Phan²¹, Kirill E. Rotmistrovsky²¹, Robert D. Sanders²¹, Martin F. Shumway²¹, Chunlin Xiao²¹ **Oxford University** Gil A. McVean (Co-Chair) (Co-Chair, Population Genetics) (Principal Investigator)^{11,18}, Adam Auton¹¹, Zamin Iqbal¹¹, Gerton Lunter¹¹, Jonathan L. Marchini^{11,18}, Loukas Moutsianas¹⁸, Simon Myers^{11,18}, Afidalina Tumian¹⁸ **Roche Applied Science** Brian Desany (Project Leader)²⁷, James Knight²⁷, Roger Winer²⁷ **The Translational Genomics Research Institute** David W. Craig (Principal Investigator)⁴⁸, Steve M. Beckstrom-Sternberg⁴⁸, Alexis Christoforides⁴⁸, Ahmet A. Kurdoglu⁴⁸, John V. Pearson⁴⁸, Shripad A. Sinari⁴⁸, Waibhav D. Tembe⁴⁸ **University of California, Santa Cruz** David Haussler (Principal Investigator)⁴⁹, Angie S. Hinrichs⁴⁹, Sol J. Katzman⁴⁹, Andrew Kern⁴⁹, Robert M. Kuhn⁴⁹ **University of Chicago** Molly Przeworski (Co-Chair, Population Genetics) (Principal Investigator)⁵⁰, Ryan D. Hernandez⁵¹, Bryan Howie⁵², Joanna L. Kelley⁵², S. Cord Melton⁵² **University of Michigan** Gonçalo R. Abecasis (Co-Chair) (Principal Investigator)⁵, Yun Li (Project Leader)⁵, Paul Anderson⁵, Tom Blackwell⁵, Wei Chen⁵, William O. Cookson⁵³, Jun Ding⁵, Hyun Min Kang⁵, Mark Lathrop⁵⁴, Liming Liang⁵⁵, Miriam F. Moffatt⁵³, Paul Scheet⁵⁶, Carlo Sidore⁵, Matthew Snyder⁵, Xiaowei Zhan⁵, Sebastian Zöllner⁵ **University of Montreal** Philip Awadalla (Principal Investigator)⁵⁷, Ferran Casals⁵⁸, Youssef Idaghdour⁵⁸, John Keebler⁵⁸, Eric A. Stone⁵⁸, Martine Zilversmit⁵⁸ **University of Utah** Lynn Jorde (Principal Investigator)⁵⁹, Jinchuan Xing⁵⁹ **University of Washington** Evan E. Eichler (Principal Investigator)⁶⁰, Gozde Aksay¹⁹, Can Alkan⁶⁰, Iman Hajirasouliha⁶¹, Fereydoon Hormozdiari⁶¹, Jeffrey M. Kidd^{19,43}, S. Cenk Sahinalp⁶¹, Peter H. Sudmant¹⁹ **Washington University in St. Louis** Elaine R.

Mardis (Co-Principal Investigator)¹⁷, Ken Chen¹⁷, Asif Chinwalla¹⁷, Li Ding¹⁷, Daniel C. Koboldt¹⁷, Mike D. McLellan¹⁷, David Dooling¹⁷, George Weinstock¹⁷, John W. Wallis¹⁷, Michael C. Wendl¹⁷, Qunyuan Zhang¹⁷ **Wellcome Trust Sanger Institute** Richard M. Durbin (Principal Investigator)¹, Cornelis A. Albers⁶², Qasim Ayub¹, Senduran Balasubramanian¹, Jeffrey C. Barrett¹, David M. Carter¹, Yuan Chen¹, Donald F. Conrad¹, Petr Danecek¹, Emmanouil T. Dermitzakis⁶³, Min Hu¹, Ni Huang¹, Matt E. Hurles¹, Hanjun Jin⁶⁴, Luke Jostins¹, Thomas M. Keane¹, Si Quang Le¹, Sarah Lindsay¹, Quan Long¹, Daniel G. MacArthur¹, Stephen B. Montgomery⁶³, Leopold Parts¹, James Stalker¹, Chris Tyler-Smith¹, Klaudia Walter¹ Yujun Zhang¹ **Yale and Stanford Universities** Mark B. Gerstein (Co-Principal Investigator)^{65,66}, Michael Snyder (Co-Principal Investigator)⁴³, Alexej Abyzov⁶⁵, Suganthi Balasubramanian⁶⁷, Robert Bjornson⁶⁶, Jiang Du⁶⁶, Fabian Grubert⁴³, Lukas Habegger⁶⁵, Rajini Haraksingh⁶⁵, Justin Jee⁶⁵, Ekta Khurana⁶⁷, Hugo Y.K. Lam⁴³, Jing Leng⁶⁵, Ximeng Jasmine Mu⁶⁵, Alexander E. Urban^{43,68}, Zhengdong Zhang⁶⁷

Structural Variation Group: BGI-Shenzhen Yingrui Li²², Ruibang Luo²² **Boston College** Gabor T. Marth (Principal Investigator)³⁰, Erik P. Garrison³⁰, Deniz Kural³⁰, Aaron R. Quinlan³², Chip Stewart³⁰, Michael P. Stromberg³³, Alistair N. Ward³⁰, Jiantao Wu³⁰ **Brigham and Women's Hospital** Charles Lee (Co-Chair) (Principal Investigator)³⁴, Ryan E. Mills³⁴, Xinghua Shi³⁴ **Broad Institute of MIT and Harvard** Steven A. McCarroll (Project Leader)^{2,4}, Eric Banks², Mark A. DePristo², Robert E. Handsaker², Chris Hartl², Joshua M. Korn², Heng Li², James C. Nemesh² **Cold Spring Harbor Laboratory** Jonathan Sebat (Principal Investigator)³⁹, Vladimir Makarov⁴⁰, Kenny Ye⁴¹, Seungtae C. Yoon⁴² **Cornell and Stanford Universities** Jeremiah Degenhardt⁸, Mark Kaganovich⁴³ **European Bioinformatics Institute** Laura Clarke (Project Leader)¹³, Richard E. Smith¹³, Xiangqun Zheng-Bradley¹³ **European Molecular Biology Laboratory** Jan O. Korbel⁴⁵ **Illumina** Sean Humphray (Project Leader)⁶, R. Keira Cheetham⁶, Michael Eberle⁶, Scott Kahn⁶, Lisa Murray⁶ **Leiden University Medical Center** Kai Ye⁴⁶ **Life Technologies** Francisco M. De La Vega (Principal Investigator)¹⁰, Yutao Fu²⁴, Heather E. Peckham²⁴, Yongming A. Sun¹⁰ **Louisiana State University** Mark A. Batzer (Principal Investigator)⁴⁷, Miriam K. Konkel⁴⁷, Jerilyn A. Walker⁴⁷ **US National Institutes of Health** Chunlin Xiao²¹ **Oxford University** Zamin Iqbal¹¹ **Roche Applied Science** Brian Desany²⁷ **University of Michigan** Tom Blackwell (Project Leader)⁵, Matthew Snyder⁵ **University of Utah** Jinchuan Xing⁵⁹ **University of Washington** Evan E. Eichler (Co-Chair) (Principal Investigator)⁶⁰, Gozde Aksay¹⁹, Can Alkan⁶⁰, Iman Hajirasouliha⁶¹, Fereydoun Hormozdiari⁶¹, Jeffrey M. Kidd^{19,43} **Washington University in St. Louis** Ken Chen¹⁷, Asif Chinwalla¹⁷, Li Ding¹⁷, Mike D. McLellan¹⁷, John W. Wallis¹⁷ **Wellcome Trust Sanger Institute** Matt E. Hurles¹ (Co-Chair) (Principal Investigator), Donald F. Conrad¹, Klaudia Walter¹, Yujun Zhang¹ **Yale and Stanford Universities** Mark B. Gerstein (Co-Principal Investigator)^{65,66}, Michael Snyder (Co-Principal Investigator)⁴³, Alexej Abyzov⁶⁵, Jiang Du⁶⁶, Fabian Grubert⁴³, Rajini Haraksingh⁶⁵, Justin Jee⁶⁵, Ekta Khurana⁶⁷, Hugo Y.K. Lam⁴³, Jing Leng⁶⁵, Ximeng Jasmine Mu⁶⁵, Alexander E. Urban^{43,68}, Zhengdong Zhang⁶⁷

Exon Pilot Group: Baylor College of Medicine Richard A. Gibbs (Co-Chair) (Principal Investigator)¹⁴, Matthew Bainbridge¹⁴, Danny Challis¹⁴, Cristian Coafra¹⁴, Huyen Dinh¹⁴, Christie Kovar¹⁴, Sandy Lee¹⁴, Donna Muzny¹⁴, Lynne Nazareth¹⁴, Jeff Reid¹⁴, Aniko Sabo¹⁴, Fuli Yu¹⁴, Jin Yu¹⁴ **Boston College** Gabor T. Marth (Co-Chair) (Principal Investigator)³⁰, Erik P. Garrison³⁰, Amit Indap³⁰, Wen Fung Leong³⁰, Aaron R. Quinlan³², Chip Stewart³⁰, Alistair N. Ward³⁰, Jiantao Wu³⁰ **Broad Institute of MIT and Harvard** Kristian Cibulskis², Tim J. Fennell², Stacey B. Gabriel², Kiran V. Garimella², Chris Hartl², Erica Shefler², Carrie L. Sougnez², Jane Wilkinson² **Cornell and Stanford Universities** Andrew G. Clark (Co-Principal Investigator)⁸, Simon Gravel⁴³, Fabian Grubert⁴³

European Bioinformatics Institute Laura Clarke (Project Leader)¹³, Paul Flicek (Principal Investigator)¹³, Richard E. Smith¹³, Xiangqun Zheng-Bradley¹³ **US National Institutes of Health** Stephen T. Sherry (Principal Investigator)²¹, Hoda M. Khouri²¹, Justin E. Paschall²¹, Martin F. Shumway²¹, Chunlin Xiao²¹ **Oxford University** Gil A. McVean^{11,18} **University of California, Santa Cruz** Sol J. Katzman⁴⁹ **University of Michigan** Gonçalo R. Abecasis (Co-Chair) (Principal Investigator)⁵, Tom Blackwell⁵ **Washington University in St. Louis** Elaine R. Mardis (Principal Investigator)¹⁷, David Dooling¹⁷, Lucinda Fulton¹⁷, Robert Fulton¹⁷, Daniel C. Koboldt¹⁷ **Wellcome Trust Sanger Institute** Richard M. Durbin (Principal Investigator)¹, Senduran Balasubramaniam¹, Allison Coffey¹, Thomas M. Keane¹, Daniel G. MacArthur¹, Aarno Palotie^{1,28}, Carol Scott¹, James Stalker¹, Chris Tyler-Smith¹ **Yale University** Mark B. Gerstein (Principal Investigator)^{65,66}, Suganthi Balasubramanian⁶⁷

Samples and ELSI Group: Aravinda Chakravarti (Co-Chair)⁷, Bartha M. Knoppers (Co-Chair)¹⁵, Leena Peltonen (Co-Chair)*, Gonçalo R. Abecasis⁵, Carlos D. Bustamante⁴³, Neda Gharani⁶⁹, Richard A. Gibbs¹⁴, Lynn Jorde⁵⁹, Jane S. Kaye⁷⁰, Alastair Kent⁷¹, Taosha Li²², Amy L. McGuire⁷², Gil A. McVean^{11,18}, Pilar N. Ossorio⁷³, Charles N. Rotimi⁷⁴, Yeyang Su²², Lorraine H. Toji⁶⁹, Chris Tyler-Smith¹

Scientific Management: Lisa D. Brooks⁷⁵, Adam L. Felsenfeld⁷⁵, Jean E. McEwen⁷⁵, Assya Abdallah⁷⁶, Christopher R. Juenger⁷⁷, Nicholas C. Clemm⁷⁵, Francis S. Collins⁹, Audrey Duncanson²⁰, Eric D. Green⁷⁸, Mark S. Guyer⁷⁵, Jane L. Peterson⁷⁵, Alan J. Schafer²⁰

Writing Group: Gonçalo R. Abecasis⁵, David L. Altshuler²⁻⁴, Adam Auton¹¹, Lisa D. Brooks⁷⁵, Richard M. Durbin¹, Richard A. Gibbs¹⁴, Matt E. Hurles¹, Gil A. McVean^{11,18}

- 1 Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Cambridge, CB10 1SA, UK.
- 2 The Broad Institute of MIT and Harvard, 7 Cambridge Center, Cambridge, Massachusetts 02142, USA.
- 3 Center for Human Genetic Research, Massachusetts General Hospital, Boston, Massachusetts 02114, USA.
- 4 Dept of Genetics, Harvard Medical School, Cambridge, Massachusetts 02115, USA.
- 5 Center for Statistical Genetics and Biostatistics, University of Michigan, Ann Arbor, Michigan 48109, USA.
- 6 Illumina Cambridge Ltd., Chesterford Research Park, Little Chesterford, Nr Saffron Walden, Essex CB10 1XL, UK.
- 7 McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University School of Medicine, Baltimore, Maryland 21205, USA.
- 8 Center for Comparative and Population Genomics, Cornell University, Ithaca, New York 14850, USA.
- 9 US National Institutes of Health, 1 Center Drive, Bethesda, Maryland 20892, USA.
- 10 Life Technologies, Foster City, California 94404, USA.
- 11 Wellcome Trust Centre for Human Genetics, Roosevelt Drive, Oxford OX3 7BN, UK.
- 12 Pall Corporation, 25 Harbor Park Drive, Port Washington, New York 11050 USA.
- 13 European Bioinformatics Institute, Wellcome Trust Genome Campus, Cambridge, CB10 1SD, UK.

- 14 Human Genome Sequencing Center, Baylor College of Medicine, 1 Baylor Plaza, Houston, Texas 77030, USA.
- 15 Centre of Genomics and Policy, McGill University, Montréal, Québec H3A 1A4, Canada.
- 16 Max Planck Institute for Molecular Genetics, D-14195 Berlin-Dahlem, Germany.
- 17 The Genome Center, Washington University School of Medicine, St Louis, Missouri 63108, USA.
- 18 Dept of Statistics, University of Oxford, Oxford OX1 3TG, UK.
- 19 Dept of Genome Sciences, University of Washington School of Medicine, Seattle, Washington 98195, USA.
- 20 Wellcome Trust, Gibbs Building, 215 Euston Road, London NW1 2BE, UK.
- 21 US National Institutes of Health, National Center for Biotechnology Information, 45 Center Drive, Bethesda, Maryland 20892, USA.
- 22 BGI-Shenzhen, Shenzhen 518083, China.
- 23 Dept of Biology, University of Copenhagen, Denmark.
- 24 Life Technologies, Beverly, Massachusetts 01915, USA.
- 25 Deep Sequencing Group, Biotechnology Center TU Dresden, Tatzberg 47/49, 01307, Dresden, Germany.
- 26 Institute of Clinical Molecular Biology, Christian-Albrechts-University Kiel, Kiel, Germany.
- 27 Roche Applied Science, 20 Commercial Street, Branford, Connecticut 06405, USA.
- 28 Department of Medical Genetics, Institute of Molecular Medicine (FIMM) of the University of Helsinki and Helsinki University Hospital, Helsinki, Finland.
- 29 Agilent Technologies Inc., Santa Clara, California 95051, USA.
- 30 Dept of Biology, Boston College, Chestnut Hill, Massachusetts 02467, USA.
- 31 US National Institutes of Health, National Institute of Environmental Health Sciences, 111 T W Alexander Drive, Research Triangle Park, North Carolina 27709, USA.
- 32 Dept of Biochemistry and Molecular Genetics, University of Virginia School of Medicine, Charlottesville, Virginia 22908, USA.
- 33 Illumina, San Diego, California 92121, USA.
- 34 Dept of Pathology, Brigham and Women's Hospital and Harvard Medical School, Boston, Massachusetts 02115, USA.
- 35 Dept of Medicine, Division of Medical Genetics, University of Washington, Seattle, Washington 98195, USA.
- 36 Center for Systems Biology, Dept Organismic and Evolutionary Biology, Harvard University, Cambridge, Massachusetts 02138, USA.
- 37 Dept of Epidemiology, Harvard School of Public Health, Boston, Massachusetts 02115, USA.
- 38 Institute of Medical Genetics, Cardiff University, Heath Park, Cardiff CF14 4XN, UK.
- 39 Depts of Psychiatry and Cellular and Molecular Medicine, University of California San Diego, 9500 Gilman Dr, La Jolla, California 92093, USA.
- 40 Seaver Autism Center and Department of Psychiatry, Mount Sinai School of Medicine, New York, New York 10029, USA.
- 41 Dept of Epidemiology and Population Health, Albert Einstein College of Medicine, Bronx, New York 10461, USA.
- 42 Dept of Genetics and Genomic Sciences, Mount Sinai School of Medicine, New York, New York 10029, USA.
- 43 Dept of Genetics, Stanford University, Stanford, California 94305, USA.
- 44 Dept of Molecular and Cellular Biology, University of Arizona, Tucson, Arizona 85721, USA.

- 45 European Molecular Biology Laboratory, Genome Biology Research Unit, Meyerhofstr. 1, Heidelberg, Germany.
- 46 Molecular Epidemiology Section, Medical Statistics and Bioinformatics, Leiden University Medical Center, 2333 ZA, The Netherlands.
- 47 Dept of Biological Sciences, Louisiana State University, Baton Rouge, Louisiana 70803, USA.
- 48 The Translational Genomics Research Institute, 445 N Fifth Street, Phoenix, Arizona 85004, USA.
- 49 Center for Biomolecular Science and Engineering, University of California Santa Cruz, Santa Cruz, California 95064, USA.
- 50 Dept of Human Genetics and Howard Hughes Medical Institute, University of Chicago, Chicago, Illinois 60637, USA.
- 51 Dept of Bioengineering and Therapeutic Sciences, University of California San Francisco, San Francisco, California 94158, USA.
- 52 Dept of Human Genetics, University of Chicago, Chicago, Illinois 60637, USA.
- 53 National Heart and Lung Institute, Imperial College London, London SW7 2, UK.
- 54 Centre Nationale de Génotypage, Evry, France.
- 55 Depts of Epidemiology and Biostatistics, Harvard School of Public Health, Boston, Massachusetts 02115, USA.
- 56 Dept of Epidemiology, University of Texas MD Anderson Cancer Center, Houston, Texas 77030, USA.
- 57 Dept of Pediatrics, Faculty of Medicine, University of Montréal, Ste. Justine Hospital Research Centre, Montréal, Québec H3T 1C5, Canada.
- 58 Dept of Medicine, Centre Hospitalier de l'Université de Montréal Research Center, Université de Montréal, Montréal, Québec H2L 2W5, Canada.
- 59 Eccles Institute of Human Genetics, University of Utah School of Medicine, Salt Lake City, Utah 84112, USA.
- 60 Dept of Genome Sciences, University of Washington School of Medicine and Howard Hughes Medical Institute, Seattle, Washington 98195, USA.
- 61 Dept of Computer Science, Simon Fraser University, Burnaby, British Columbia V5A 1S6, Canada.
- 62 Dept of Haematology, University of Cambridge and National Health Service Blood and Transplant, Cambridge CB2 1TN, UK.
- 63 Dept of Genetic Medicine and Development, University of Geneva Medical School, Geneva, 1211 Switzerland.
- 64 Center for Genome Science, Korea National Institute of Health, 194, Tongil-Lo, Eunpyung-Gu, Seoul, 122-701, Korea.
- 65 Program in Computational Biology and Bioinformatics, Yale University, New Haven, Connecticut 06520, USA.
- 66 Dept of Computer Science, Yale University, New Haven, Connecticut 06520, USA.
- 67 Dept of Molecular Biophysics and Biochemistry, Yale University, New Haven, Connecticut 06520, USA.
- 68 Dept of Psychiatry and Behavioral Studies, Stanford University, Stanford, California 94305, USA.
- 69 Coriell Institute, 403 Haddon Avenue, Camden, New Jersey 08103, USA.
- 70 Centre for Health, Law and Emerging Technologies, University of Oxford, Oxford OX3 7LF, UK.

- 71 Genetic Alliance, 436 Essex Road, London, N1 3QP, UK.
- 72 Center for Medical Ethics and Health Policy, Baylor College of Medicine, 1 Baylor Plaza, Houston, Texas 77030, USA.
- 73 Dept of Medical History and Bioethics, University of Wisconsin--Madison, Madison, Wisconsin 53706, USA.
- 74 US National Institutes of Health, Center for Research on Genomics and Global Health, 12 South Drive, Bethesda, Maryland 20892, USA.
- 75 US National Institutes of Health, National Human Genome Research Institute, 5635 Fishers Lane, Bethesda, Maryland 20892, USA.
- 76 The George Washington University School of Medicine and Health Sciences, Washington, DC 20037, USA.
- 77 US Food and Drug Administration, 11400 Rockville Pike, Rockville, Maryland 20857, USA.
- 78 US National Institutes of Health, National Human Genome Research Institute, 31 Center Drive, Bethesda, Maryland 20892, USA.

* Deceased