Supplementary Information

Personalized Copy-Number and Segmental Duplication Maps using Next-Generation Sequencing

Can Alkan[1,6], Jeffrey M. Kidd[1], Tomas Marques-Bonet[1,2,] Gozde Aksay[1], Francesca Antonacci[1], Fereydoun Hormozdiari[3], Jacob O. Kitzman[1], Carl Baker[1], Maika Malig[1], Onur Mutlu[4], S. Cenk Sahinalp[3], Richard A. Gibbs[5], Evan E. Eichler[1,6]

[1]Department of Genome Sciences, University of Washington School of Medicine, Seattle, WA, USA, [2]Institut de Biologia Evolutiva (UPF-CSIC), Barcelona, Catalonia, Spain, [3]School of Computing Science, Simon Fraser University, Burnaby, BC, Canada, [4]Department of Electrical and Computer Engineering, Carnegie Mellon University, Pittsburgh, PA, USA, [5]Baylor College of Medicine, Houston, TX, USA, [6]Howard Hughes Medical Institute, Seattle, WA, USA

CONTENTS

Note: Supplementary Tables 4 and 5 are provided in separate spreadsheet files on Nature Genetics web site.

Nature Genetics: doi:10.1038/ng.437

Nature Genetics: doi:10.1038/ng.437

Nature Genetics: doi:10.1038/ng.437

I. *mrFAST* Algorithm

mrFAST (micro-read fast alignment search tool) implements a collision-free hash table to create indices of the reference genome that can efficiently utilize the main memory of the system. A collision-free hash table is a hash table for storing strings, where no two different strings can be assigned the same hash value. The hash function we use for the *mrFAST* tool basically packs the four possible base pairs in two bits. We encode *A* with *00*, *C* with *01*, *G* with *10*, and *T* with *11*. In this way, we can encode a string of length *k* in *2k* bits. If $k \leq 16$ is used, we can represent this encoding of *k*-mers as the *unsigned integer* data type (one integer can store 32 bits). We then simply use the integer value of this encoding as the hash value of the interrogated *k*-mer.

The starting locations of strings in the text to be indexed are stored in the hash table entries. The choice of using a hash table for indexing *k*-mers offers a number of advantages. Looking up keys in a hash table is extremely fast. Worst case look-up time for a key of length *k* (*k*-mer) is O(k) because there are no key collisions. The worst-case memory requirement of a collision-free hash table to index a genomic sequence of length *L*, alphabet $\sum = \{A, C, G, T\}$, window size *k*, and slide size 1 (overlaps of length $k - 1$) is $O(4^k + L)$.

The memory requirement of the hash table as described above prohibits indexing the whole human reference genome. To tackle this problem, we partition the human genome reference sequence into smaller genome contigs, (optionally) first broken by the existing sequence gaps, and we further divide the large contigs to sub-contigs of lengths up to 30 Mb. In the human genome build36 assembly, this results in 361 contigs, and each of these contigs is indexed separately. This approach makes it possible to fit the index of a contig (longest contig is about 30 Mb), for example, under 1.5 GB of memory. If the available amount of memory is less than 1.5 GB, we can simply divide the genome to smaller contigs. Both orientations of the query sequences are then searched in each contig and the results are merged.

We can exploit some of the properties of the reads generated with the Illumina sequencer. First, the micro-reads are significantly shorter than 454 and capillary-based reads. Second, we can also exploit the fact that we will allow less base pair mismatches (or gaps); we typically allow 2−3 bp along a micro-read. Furthermore, Illumina sequencer generates uniform read-length in the same run. We made three major enhancements for faster and accurate search of micro-read sequences: a) optimizations for the shorter read-length and allowing 2−3 mismatches and gaps; b) optimizations for fast extension of placed seeds; and c) optimizations exploiting uniform read-length within a single machine run.

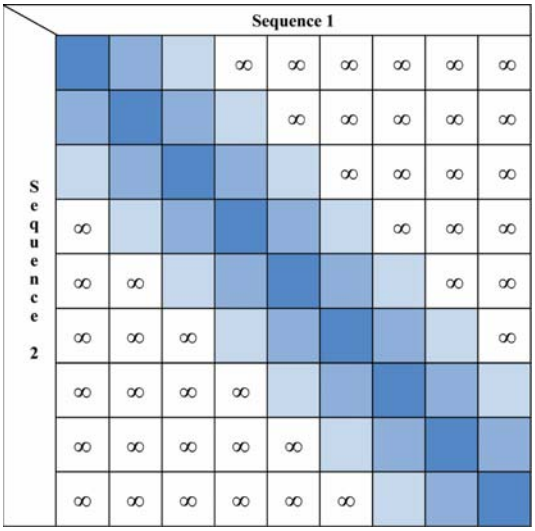
Nature Genetics: doi:10.1038/ng.437

a) Optimizations for the shorter read-length and allowing 2−3 mismatches and gaps

Most common sequence search methods, such as BLAST [1], BLAT [2], SHRiMP [3] and many others[4-7], implement the so-called "seed-and-extend" methods to achieve faster sequence comparison. In this method, one or more exactly matching "seeds", or *k*-mers, are first mapped to the reference genome, and then the initial seed matches are extended using the Smith-Waterman algorithm [8] to find and return a mapping location for high sequence similarity. The *mrFAST* tool also uses the seed-and-extend method, and the seed step employs the same collision-free hash table described above. We implement further improvements in the "extend" step. Realizing the short length of Illumina reads and relatively low sequencing error rate, rather than extending the alignment with Smith-Waterman algorithm [8], we calculate the simpler Levenshtein distance [9] (also called "edit distance"). Levenshtein distance basically assigns the same score for matching base pairs and the same penalty for mismatching base pairs and gaps. This basic scheme can be accurately applied to short-read mapping since we are interested in finding only very highly similar sequences (edit distance 0−3). One property of the Levenshtein distance is especially beneficial for such an alignment; the objective of edit distance computation is *minimization* of the sequence differences, where the goal of Smith-Waterman algorithm is the *maximization* of alignment score. This property enables us to stop the alignment step as soon as we detect that the edit distance value grows to be more than the allowed threshold. Thus, we save computation time for "non-extendable hot spots", with a technique that can not be applied to maximization algorithms such as Smith-Waterman.

We can further accelerate the alignment step by taking advantage of the very low threshold of allowed edit distance for read mappings (2−3 bp). Ukkonen improved run time and space complexity of the Levenshtein algorithm to $O(s \cdot min(m, n))$ (as opposed to $O(mn)$), where s is the maximum allowed Levenshtein distance between the aligned sequences and m and n are the length of the sequences [10]. Ukkonen's algorithm uses the fact that if s is small, calculating all of the cells in the dynamic programming matrix is not necessary (see Supplementary Note Figure 1). Instead, the algorithm starts computing the scores at the main diagonal, ($M[i, j]$, where $i = j$). It is also observed that substitution errors can be counted at the current diagonal, and extending the score calculation to the neighbor diagonals are necessitated only by gaps (indels). Consequently, if the number of maximum allowed gaps is at most t, then the total number of diagonals one needs to compute is at most $2t−1$. Thus in the worst case, scores of at most $(2t−1) \cdot min(m, n)$ cells are calculated during the edit distance computation, reducing the run time significantly for small values of t. Ukkonen's algorithm was recently applied to very fast all-against-all alignments of small RNA sequences by G. Cozen [11], and together with some other optimizations, 82.8-fold acceleration against Clustal W [12] was achieved.

Nature Genetics: doi:10.1038/ng.437



Supplementary Note Figure 1. Improvement of Levenshtein distance computation by Ukkonen's algorithm [10]. We need to calculate only the color-marked cells in the dynamic programming matrix if the maximum allowed number of gaps is bounded by a small value *t*. This figure shows the diagonals to be calculated when $t = 2$.

Implementation enhancements for faster computation of Levenshtein distance. Algorithms such as Ukkonen's [10] help accelerate the edit distance computation. However, we can improve the speed even more through the use of specific instruction sets in the CPUs without the need of specialized hardware. Previously, Wozniak [13] proposed to use the specialized video-oriented instructions implemented in SPARC (Sun Microsystems) and Pentium Pro (Intel) processors to speed up sequence comparison. Liu *et al.* accelerated the Smith-Waterman alignment using the GPU (graphics processing unit) chips on video cards [14]. This method enhances the speed of Smith-Waterman computation (two-fold), however it requires the installation of powerful video-cards, which may not be cost-efficient in a multi-node computer cluster. Another method was recently described by Farrar [15] that utilizes the SSE2 instruction set extensions in Intel-based CPUs extending on a similar method by Rognes and Seeberg [16]. The SSE2 instruction set was first implemented in the Intel Pentium IV processor in 2001 and subsequently adopted by AMD in Opteron line of CPUs in 2003. SSE2 adds the SIMD (single instruction, multiple data) feature to the processors and is available in all current Intel- and AMD-based commodity hardware, without any extra cost. Basically, through the use of SSE2, we can calculate values for multiple data points with only a single instruction. The SIMD feature can thus be used for "vectored-alignments", and Farrar [15] showed that instead of calculating the values of each cell of dynamic programming matrix one-by-one, we can simultaneously calculate the values of multiple cells in a single pass (8 cells if 64-bit SSE2 registers are used, 16 cells can be filled with 128-bit registers). This approach was implemented in the SHRiMP package [3] to speed-up the Smith-Waterman computation. We also added the use of SSE2 instructions to accelerate the Levenshtein distance calculation step in *mrFAST* and achieved significant speed enhancement. For example, if we use a 64-bit SSE2 register, we can simultaneously fill 8 cells in the dynamic programming, giving 8-fold speed enhancement in edit distance

Nature Genetics: doi:10.1038/ng.437

computation. However, in practice we gain less acceleration when we also apply Ukkonen's method (which also improves the speed by $n/(2t-1)$-fold) and need to calculate only $2t-1$ diagonals around the main diagonal. For $t = 2$, the resulting speed enhancement achieved by SSE2 implementation is three-fold ($2 \times 2 - 1 = 3$). The advantage of using SSE2 instructions is more significant for larger values of t, the dynamic programming matrix calculation will take the same amount of time: up to $t = 4$ with 64-bit SSE2 (8-fold acceleration) and $t = 8$ with 128-bit SSE2 (16-fold acceleration).

b) Optimizations for fast extension of placed seeds

mrFAST also features a heuristic method that helps extend the initially placed seed (k-mer) without the extra cost of alignment computation. We pre-calculate a look-up table that keeps the hash value of the substring of length t starting from each base position of the indexed reference sequence (or contig). While searching for reads in the genome, after the initial placement of k-mers, we use the look-up table to extend the seed; in case this extension is not possible, we switch to the edit distance calculation. For example, assume the first k-mer of a read s is mapped to a location in the genome starting at position p. We calculate the hash value of the subsequence $h = s[k + 1, ..., k + t + 1]$ and compare h with the value in the look-up table at position $L[p+k]$. If they are equal, we extend the placed seed; otherwise, we invoke the alignment function. This heuristic enhances the speed of perfect-match placements as well as for non-perfect matches if the substitution (or gap) place is towards the end of the sequence by shortening the length of the subsequence to be aligned.

c) Optimizations exploiting uniform read-length within a single machine run

We added an optional speed-up heuristic to *mrFAST* that assumes identical read-length ℓ of query sequences. When building the indices and the extension look-up table for the reference genome, we also create three additional pruning tables, *PA, PC, PG*, that keep the numbers of *A, C* and *G* characters in subsequences of length ℓ starting from each base pair position of the contig, i.e. $PA[i] = count_A(S[i] \ldots S[i + \ell])$, $PC[i] = count_A(S[i] \ldots S[i + \ell])$, and $P_G[i] = count_G(S[i] \ldots S[i + \ell])$. These pruning tables are utilized during search: we first calculate the *A,C,G* content of the reads, and after the initial k-mer match, we check whether the number of *A,C,G* characters of the read is within two errors of the *A,C,G* counts of the genome starting at the same k-mer location. If the difference in counts is higher than the maximum allowed edit distance, we simply discard that "hot spot" as "non-extendable", saving time from the unnecessary alignment step.

Nature Genetics: doi:10.1038/ng.437

Build36	MAQ (v 0.7.1)	mrFAST	Mosaik (0.9.0855)
Run time	255m 31s	299m 4s (236m 31s CPU time)	7m 44s (wall time) 56m 9s (CPU time)
Placed reads	1,836,870	1,857,449	1,857,180
Map locations	1,836,870	3,257,162,476	27,640,876

Supplementary Note Table 1. Alignment statistics of two million 36-bp Illumina reads to human reference genome (build36). Note that by default, *MAQ* returns only one placement per sequence. This table lists the run time and the total number of placed reads. Mosaik was run with multi-threaded option (8 threads) and was tested on different hardware due to higher memory requirements (20-GB memory usage vs. ≤2 GB memory requirements for both *MAQ* and *mrFAST*).

Extension of *mrFAST* to enable more mismatches and gaps for longer reads in the future is straightforward. The current implementation allows for two errors, but this is simply a parameter that can be updated when longer reads are available or when there are alignments with more errors are of interest. The speed enhancement algorithms and heuristics enabled *mrFAST* to be considerably faster than other tools like SHRiMP [3] and comparable to *MAQ* [17] without sacrificing sensitivity. *mrFAST* also has the ability to record all possible map locations. Our goal was to capture the maximum amount of information from next-gen sequencing technology while still being competitive in terms of speed performance.

We experimented with a set of two million 36-bp sequences generated with an Illumina Genome Analyzer and aligned to the human genome (build36) with *MAQ* [17] (version 0.7.1), Mosaik[18] (version 0.9.0855) and *mrFAST*. Supplementary Note Table 1 compares the run times, number of placed reads, and the total number of reported mapping locations with these tools. In order to better understand the performance comparison, we emphasize that *MAQ* is not capable of performing gapped alignments for single fragment reads (i.e. indels are not permitted), and ungapped alignments (calculating Hamming distance) [19] can be finished faster than gapped alignments: O(n) worst case running time vs. O(mn). *MAQ* can find short indels only if paired-end information is available, and this also slows down the computation. Furthermore, *MAQ* reports only one map location per read by default. In contrast, both Mosaik and *mrFAST* allow indels, and both can report multiple map locations for repetitive sequences. Gapped alignments increase the run time since they are more computationally intensive than ungapped alignments, and reporting multiple sites causes more I/O operations, contributing to the run time increase. *mrFAST* can return all locations without extra memory usage, however this also increases the I/O operations.

In our benchmarks (Supplementary Note Table 1), *mrFAST* was able to record more than three billion map locations in total, proving its power in copy-number detection while outperforming all the other search tools for which we are currently aware. *mrFAST* also guarantees to find all possible map locations of reads of length 36 bp within edit distance 2, when $k = 12$ is used. This is because in each query sequence we interrogate the first, middle, and last k-mers for alignment. In this case, since the query length is 36 bp and we set $k = 12$, no interrogated k-mer overlaps with another one, and since two errors can

Nature Genetics: doi:10.1038/ng.437

cause the loss of at most two *k*-mers, at least one of the *k*-mers will be found in the reference genome and the rest of the read sequence will be mapped through dynamic programming, allowing mismatches and indels. Similarly, we can select $k = 11$ to ensure only one base pair overlaps in query sequences of length ≥32 bp, thus recovering all exact and inexact matches with ≤2 errors.

d) Comparison of segmental duplication detection power.

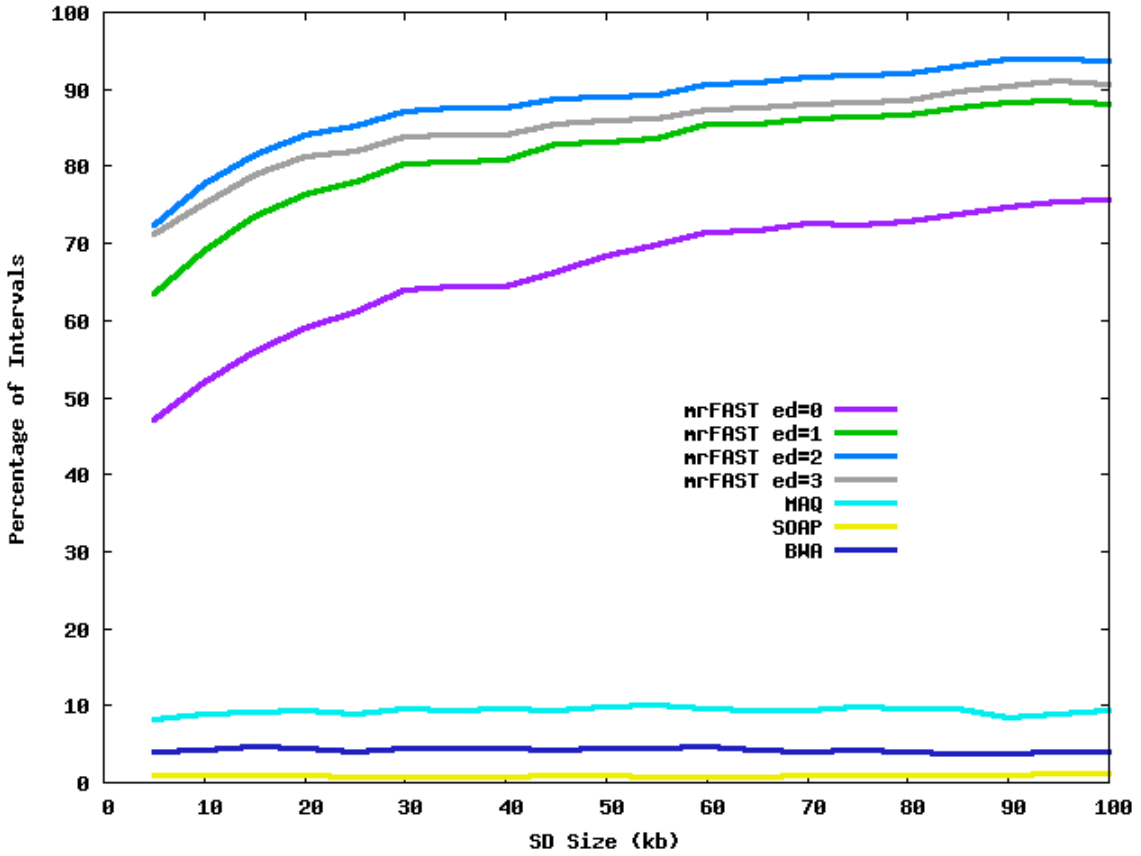
We performed benchmark analyses to compare the segmental duplication detection power of *mrFAST* with different edit distance parameters, as well as against other tools that map next-generation sequencing data. We aligned all the reads in the NA18507 genome to the control set with known copy number with *mrFAST* allowing edit distances of 0, 1, 2, and 3 and computed the average depth and standard deviation for X and autosomal loci (see Supplementary Note Table 2). While the tightest distribution in read-depth is obtained with perfect matches (edit distance=0), we find that there is a tradeoff in our ability to detect more divergent duplications and reduction in variation. We find that our optimal power to detect known segmental duplications (>94% sequence identity) occurs with an edit distance of 2 (Supplementary Note Figure 2). This is the parameter that was used in this manuscript. We also included *MAQ*[17], *BWA*[20] and *SOAP*[21] statistics for comparison (using standard settings with identical read set and repeat masking parameters). While *BWA* is significantly faster than *mrFAST*, *MAQ* and *SOAP* (*BWA*: 110 CPU-hours, *mrFAST*: 1350 CPU-hours, *MAQ*: 1930 CPU-hours, *SOAP*: 2536 CPU-hours), the read mapping accuracy of *BWA* is in fact less than *MAQ* as explained in Table 1 of the *BWA* paper [20] and exemplified by Supplementary Note Table 2 below.

Dynamic Range Response Correlation Values

Tool & Parameter	R^2	Average_auto	STD_auto	Average_chrX	STD_chrX
mrFAST ed=0	0.88	1807.45	400.46	1025.434	171.13
mrFAST ed=1	0.88	2192.61	465.16	1239.47	285.38
mrFAST ed=2	0.87	2393.52	542.8	1427.5	615.84
mrFAST ed=3	0.86	2517.43	643.93	1602.46	1099.34
MAQ	0.86	2539.39	686.85	1629.73	1185.26
BWA	0.71	716.66	248.51	425.31	183.78
SOAP	0.8	3100.58	2841.45	2016.12	1610.72

Supplementary Note Table 2. Dynamic range response correlation values for mrFAST, MAQ, BWA, and SOAP. The correlation of dynamic range response with respect to: known copy number; average and standard deviation read-depth values in autosomal and chrX-unique regions using three different tools (mrFAST, MAQ and SOAP); and four different threshold values for mrFAST. (ed = edit distance)

Nature Genetics: doi:10.1038/ng.437



Supplementary Note Figure 2. Segmental duplication detection power of mrFAST (edit distances 0,1,2,3), MAQ, BWA, and SOAP using Illumina NA18507 read data. The most sensitive tool is shown to be mrFAST with edit distance=2 threshold. We also mapped the NA18507 read data with MAQ, BWA and SOAP to the human genome reference assembly and called possible duplications using the same thresholds (6/7 windows with >average+3std read-depth). Since the standard deviation value is particularly high for SOAP (Supplementary Note Table 2), the detection power is the lowest. Note standard settings of MAQ, BWA and SOAP are designed to optimally align reads to a single location in order to identify SNPs and not to detect segmental duplications. An "allhits" option in MAQ shows comparable sensitivity to mrFAST but does not report single nucleotide differences.

II. Data Acquisition

JDW data was downloaded from NCBI Trace Archive (http://www.ncbi.nlm.nih.gov/Traces/trace.cgi) with the query = "*CENTER_NAME = 'CSHL' and CENTER_PROJECT = 'Project Jim'*". Approximately 74 million sequences averaging at 266 bp in length were then broken into 36-bp short sequences (see Data Processing below).

Nature Genetics: doi:10.1038/ng.437

NA18507 WGS was downloaded from NCBK Short Read Archive Provisional FTP site (ftp://ftp.ncbi.nih.gov/pub/TraceDB/ShortRead/SRA000271/). Approximately half of the data was used in this study.

YH WGS was downloaded from EBI European Read Archive FTP site (ftp://ftp.era-xml.ebi.ac.uk/ERA000/ERA000005/). Only the paired-end sequences in this data set (16X coverage) were used.

Details of the input sequences are given in Table 1.

III. Data Processing

a) Repeat masking

First, all known common repeats were masked with RepeatMasker [22] (sensitive option –s), and a second level of masking with Tandem Repeats Finder (TRF) [23] was run to remove short tandem repeats. Finally, WindowMasker [24] (default parameters) was applied to additional low-complexity sequences.

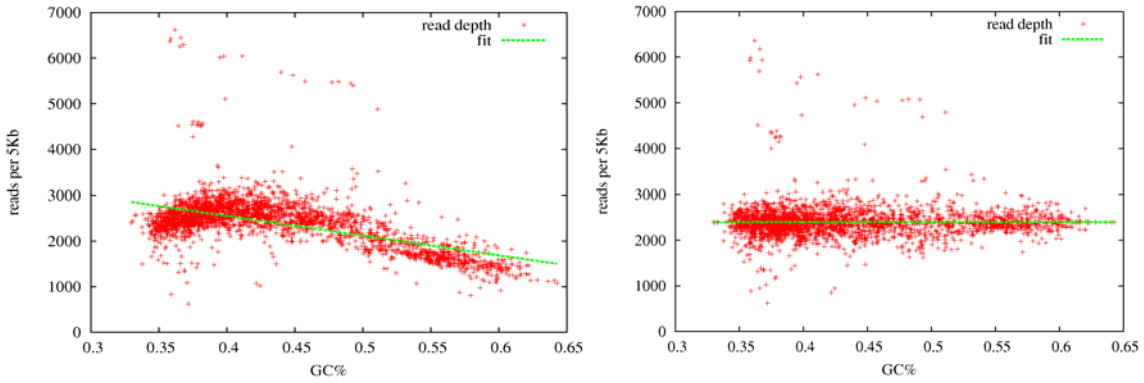
b) 454 read processing

The reads generated by 454 and Illumina technologies show different length properties. The average read-length in the JDW genome was 266 bp, where NA18507 and YH reads averaged at ~36 bp. The difference in the read-length requires different parameters which in turn causes different window borders and mapping artifacts. We therefore processed the longer 454 reads to render Illumina-like sequences. The WGS reads generated from the JDW genome were broken into non-overlapping 36-bp short sequences to obtain a sequence set with mapping biases similar to Illumina-based reads (NA18507, and YH).

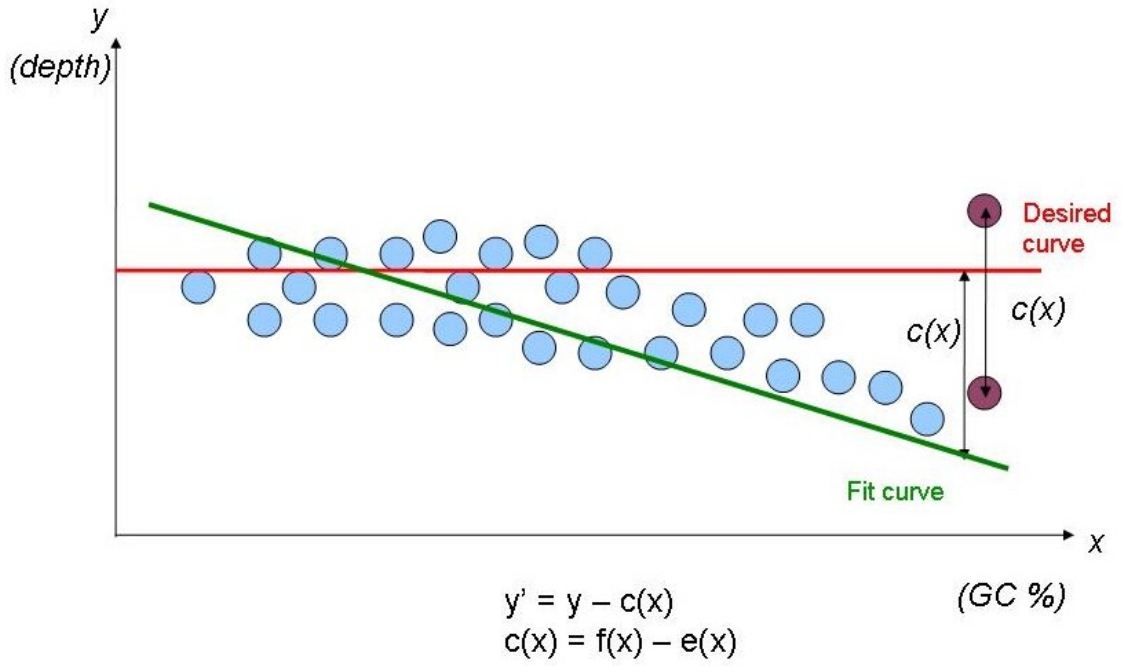
c) GC correction

Due to the known sequencing biases in GC-rich and GC-poor regions with high-throughput sequencing technologies [25], we applied a simple statistical correction method (LOESS) to normalize the read-depth based on the GC content of the windows. GC smoothing starts with calculating the regression curve on the scatter plot. We computed the average read-depth in regions known to have unique sequence in 0.1% increments of GC content. Assume f(g) is the function that returns the average read-depth in the scatter plot for GC percentage g, y(x) returns the read-depth of a point *x* on the scatter plot, and g(x) returns the GC content value of the point *x*. We then adjust the value y'(x) as: $y'(x) = y(x) - [f(x) - e(x)]$, where e(x) is the *expected* value for *x*, which we set to the overall average read-depth in the unique regions. See Supplementary Note Figure 3 for GC vs. read-depth plot in autosomal unique sequences before and after GC normalization and Supplementary Note Figure 4 for the depiction of the LOESS principle.

Nature Genetics: doi:10.1038/ng.437



Supplementary Note Figure 3. GC bias associated with the Illumina technology. Read-depth of 5-kbp windows of unique DNA is not distributed uniformly over increments of 0.1% GC content. A LOESS-based GC normalization is then applied to correct the GC bias. The high-depth points in the graph correspond to smaller (≤1 kbp) duplications that were not detected by the whole-genome assembly comparison (WGAC) method.



Supplementary Note Figure 4. LOESS normalization principle. For each interval in the genome, the c(x) value is computed based on the GC% of the 5-kbp window as the divergence of the fit curve at that GC% value from the desired curve. Then, the depth value of the interval is moved c(x).

Nature Genetics: doi:10.1038/ng.437

d) Removal of short-read mapping artifacts

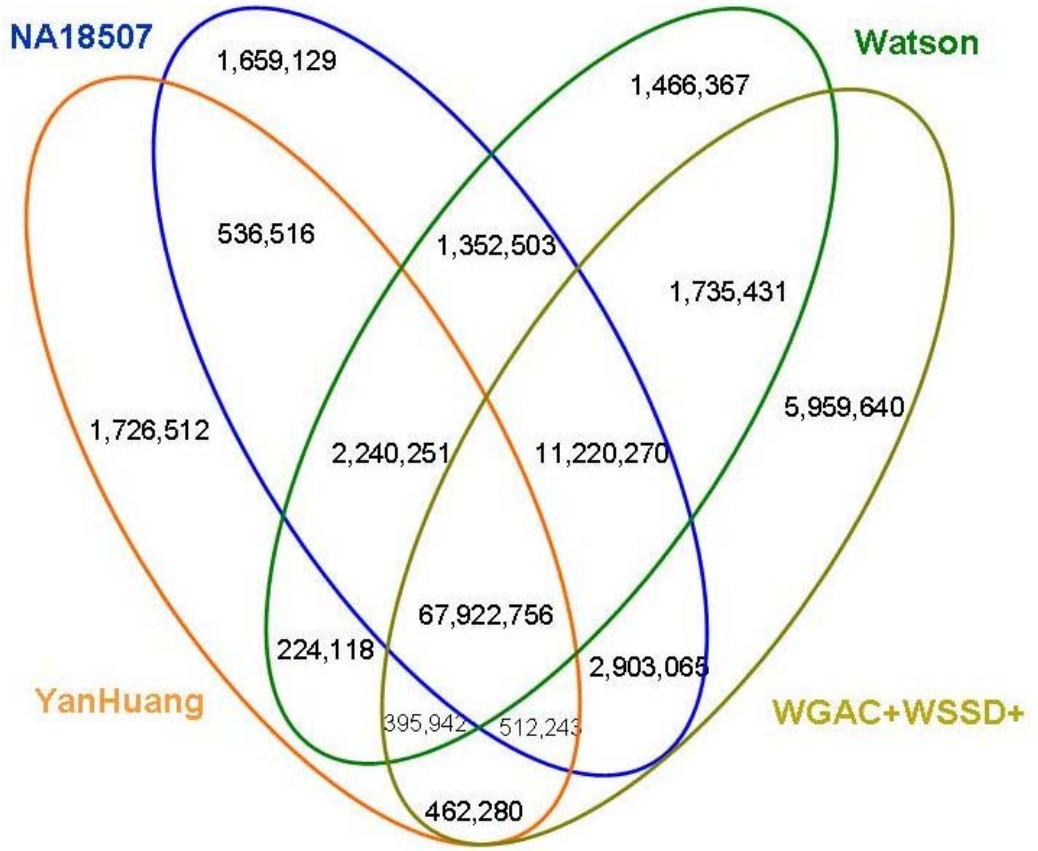
Due to the short sequence read-length with the Illumina (and rendered 454) sequences, in theory, one can detect smaller duplications in the genome. However, in this study we are mainly interested in segmental duplications (specifically ≥20 kbp). The whole-genome shotgun sequence detection (WSSD) heuristic was developed for this purpose, and small duplications and repeats can generate false positive segmental duplication calls when the read-depth values over 5-kbp windows are calculated. To account for this "mapping artifact", we simulated a whole-genome shotgun (WGS) set with the human reference genome. The sequences from build35 were broken into 36-bp reads (sliding size = 1 bp) and remapped back to the same reference genome with *mrFAST*. Segmental duplication intervals were predicted using the same parameters used with real Illumina WGS read sets. We then compared these predictions with the known duplications (both WGAC [26,27] and WSSD [26] positive intervals) and classified any intervals (or subintervals) as short-read mapping "artifacts" if they did not agree with the known duplication set. Such regions were subsequently removed from the segmental duplications predicted in JDW, NA18507 and YH genomes. However, differences in read-depth across such regions likely reflect real variation in the sequence content of the analyzed genomes; accordingly, such segments were not omitted from calculations of absolute copy number. For some comparisons with external data sets (i.e., data from microarrays) that do not directly interrogate such regions, we omitted microduplications from copy-number calculations in order to more directly compare predictions with the sequences assayed by the experimental platforms. For example, VNTR (variable number of tandem repeats) can not be accurately predicted by arrayCGH but may represent an important source of human genetic variation.

IV. WSSD Classification Scheme

a) Duplications

Segmental duplications were predicted for NA18507, JDW, and YH based upon excess depth-of-coverage in 5-kbp sliding windows. Since each individual is male, the sex chromosomes have a lower depth-of-coverage and correspondingly less robust predictions. Therefore, we have limited the analysis to predictions on the autosomes. For NA18507 we predicted 1,369 intervals encompassing 88,345,364 bp; for JDW 1,348 intervals encompassing 86,556,290 bp; and for YH 1,146 intervals encompassing 74,019,472 bp.

Nature Genetics: doi:10.1038/ng.437



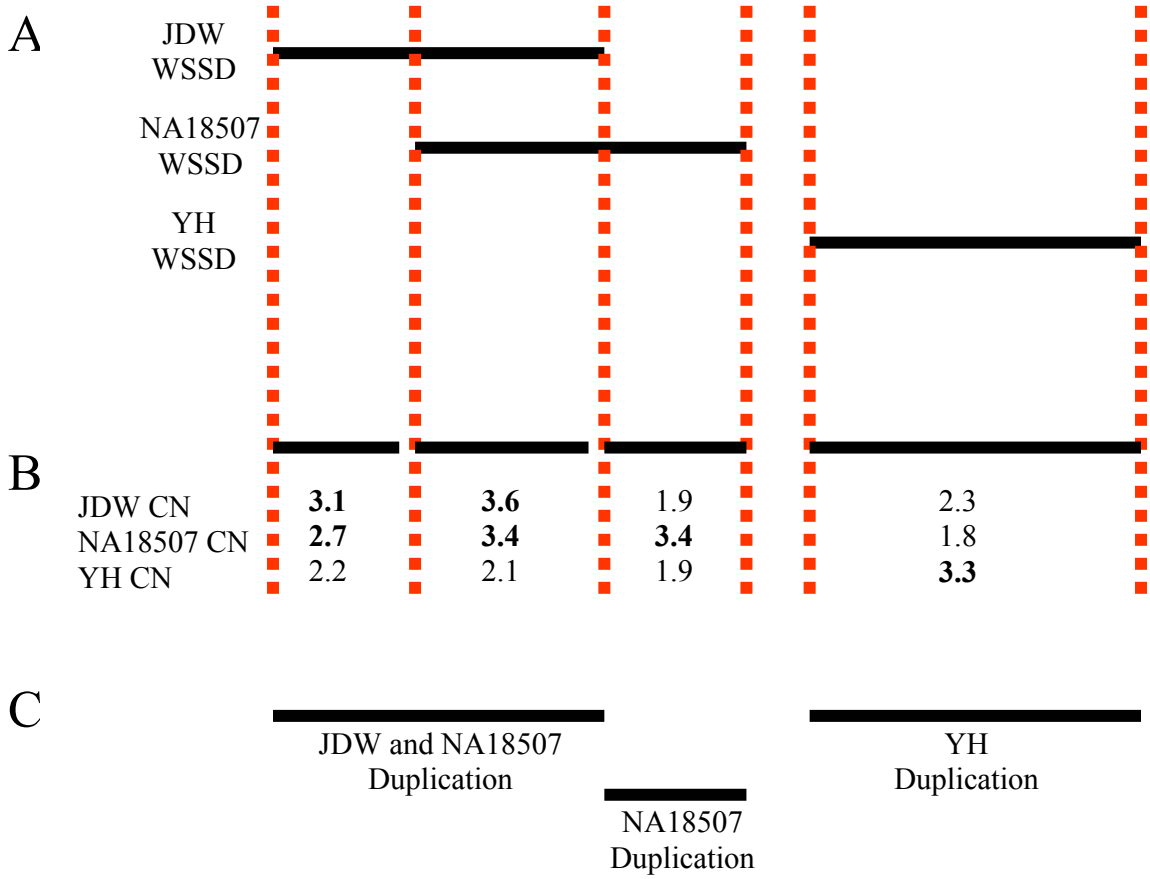
Supplementary Note Figure 5. Venn diagram of shared and individual-specific segmental duplications in contrast with previously identified duplications (WGAC [25] and WSSD [25,27]). Numbers denote the total number of base pairs in the duplicated segment. Only autosomes are shown.

Initial duplication interval predictions were based on the following criteria: 6/7 genome windows having a read-depth at least three standard deviations above the mean depth calculated for the single copy regions. These criteria were previously established to accurately predict segmental duplications using whole-genome shotgun sequence data derived using capillary sequencers and are known to work well for duplication intervals greater than 20 kbp in size [28]. Absolute copy number values were predicted in 1 kb non-overlapping windows by calculating the ratio of read depth of the interrogated window and the average depth of the single copy regions (19 autosomal and 8 chrX regions are benchmarked, copy-number is extrapolated to diploid copy number). Direct comparison of duplication status across samples is complicated by two factors. First, each individual has a different sequence depth and variance. Second, the three standard-deviation (StdDev) threshold is conservatively based on the observed coverage distribution, not directly on the inferred absolute copy number of each segment. The three StdDev cutoffs correspond to a diploid copy number of ~3.5. As a result, some segments that are truly duplicated in an individual may fall below the three standard-deviation threshold. The resulting false negative predictions, a consequence of the comparatively high stringency required for variant discovery, confound comparisons across the samples. Additionally,

Nature Genetics: doi:10.1038/ng.437

we emphasize not just the segmentation of genomic intervals based on depth-of-coverage but also the ability to accurately determine absolute copy number using next-gen sequencing data. Toward this end, we estimated the diploid copy number for each individual in each predicted interval. First, we partitioned the predicted duplication segments into a non-overlapping set of intervals. We used the median copy number predicted in 1-kbp repeat-free non-overlapping windows as the estimate for the entire interval. In order to limit boundary effects, the first and last windows were omitted.

We considered a given segment to be duplicated in an individual if the estimated copy number for that individual was greater than 2.5. In some cases, an interval did not meet this strict criterion even though it was originally classified as duplicated on the three standard-deviation criteria (Supplementary Note Figure 6). Following this reassessment we merged adjacent intervals having the same duplication classifications across the three samples.

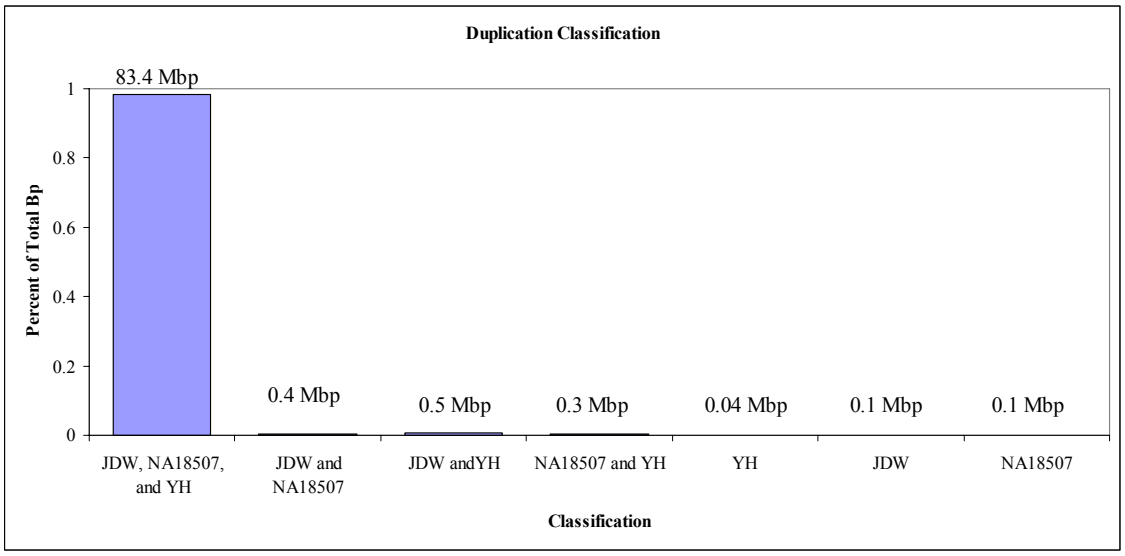


Supplementary Note Figure 6. Refining the duplication predictions. Before this analysis we created a refined, non-redundant set of duplication predictions. We began with the intervals identified as duplicated in each sample by our standard WSSD heuristics (A). We then split the predicted intervals from the three samples into non-overlapping segments. We calculated the median diploid copy number of each segment for each of the three samples (B). We reclassified a segment as being duplicated in a sample if the median copy number was greater than 2.5. Finally, we merged together

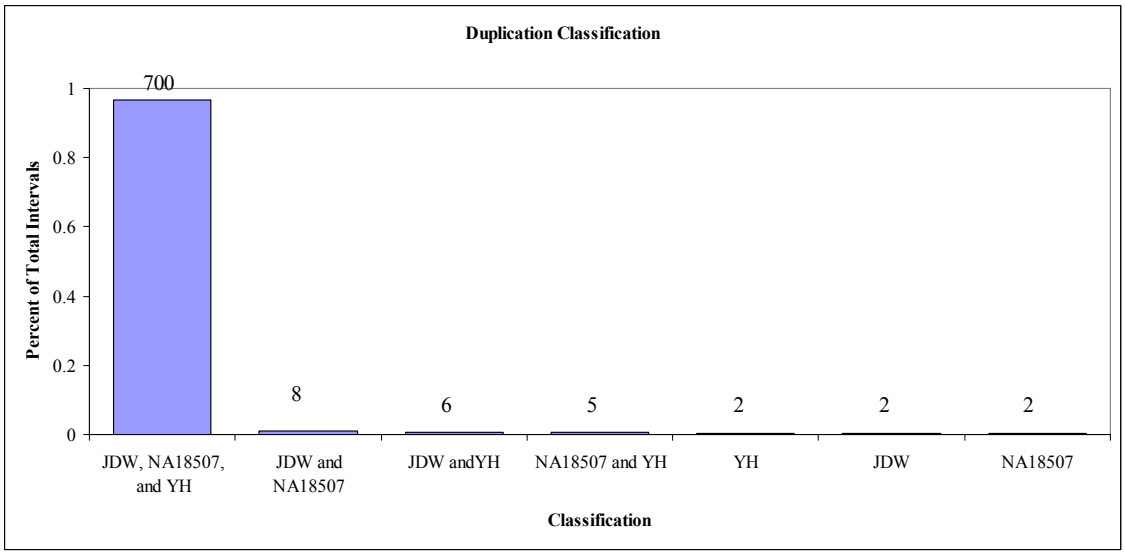
Nature Genetics: doi:10.1038/ng.437

adjacent segments having the same pattern of duplication status to create a refined, non-redundant set of predicted duplication intervals in the three individuals (C).

Predictions smaller than 20 kbp may be unreliable because of the heuristic cutoffs we employed. Additionally, some intervals contain a high fraction of masked bases or of positions associated with mapping artifacts. Such predictions are difficult to validate experimentally and may have unusual coverage characteristics because of the discontinuous nature of the unmasked positions. Therefore, all analyses were restricted to those intervals at least 20 kbp in size where less than 30% of the spanned positions annotated as mapping artifacts and less than 80% repeat masked. With these criteria, we defined a total of 725 non-overlapping intervals across the three individuals encompassing a total of 84.76 Mbp. Most of the predictions (97% of the intervals and 98% of the predicted bp) are predicted to be shared duplications found in all three individuals (Supplementary Note Figure 7, Supplementary Note Table 3). We predicted that each individual carried two duplications not present in either of the other two samples.



Nature Genetics: doi:10.1038/ng.437



Supplementary Note Figure 7. Classification of shared and individual-specific segmental duplications. a) number of base pairs, b) number of intervals.

	Intervals			Length		
	Validated	Total	% Validation	Validated	Total	% Validation
JDW Specific SDs	1	2	50.00%	80,000	108,219	73.92%
NA18507 Specific SDs	1	2	50.00%	31,736	101,156	31.37%
YH Specific SDs	0	2	0.00%	0	43,731	0.00%
JDW/NA18507 Shared SDs	6	8	75.00%	318,831	386,887	82.41%
JDW/YH Shared SDs	5	6	83.33%	431,718	452,516	95.40%
NA18507/YH Shared SDs	4	5	80.00%	269,171	297,005	90.63%
JDW/NA18507/YH Shared SDs		700			83,374,651	

Supplementary Note Table 3. Summary of known and detected autosomal segmental duplications >20 kbp.

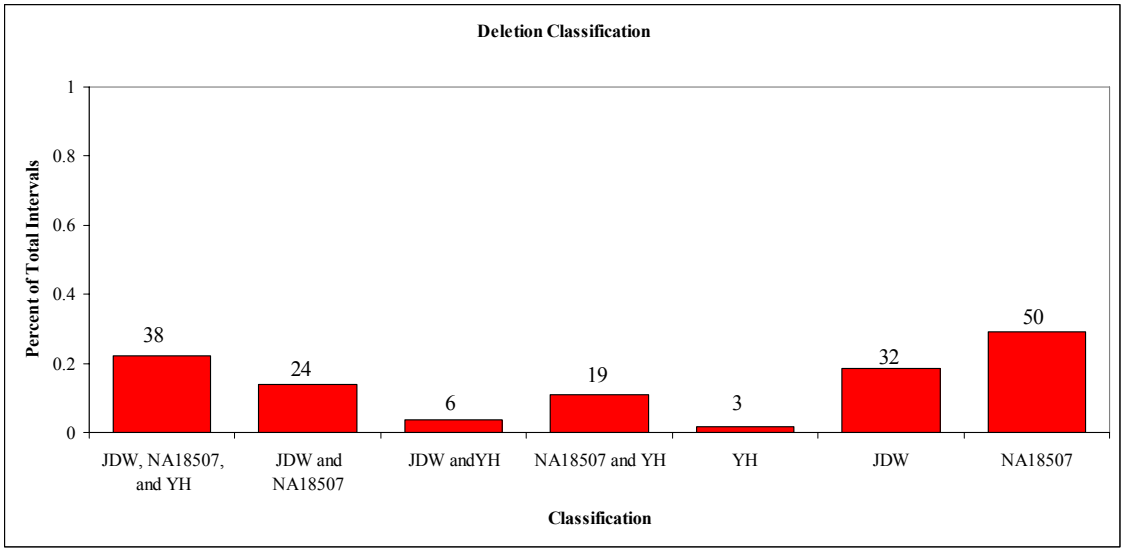
b) Deletions

Just as duplications can be characterized by increased sequence coverage, deletions can be identified based on decreased depth-of-coverage. For NA18507 we initially predicted 276 deletion intervals encompassing 6,850,775 bp; for JDW 430 intervals encompassing 9,091,572 bp; and for YH 240 intervals encompassing 5,374,931 bp. Deletion predictions were initially based on segments having a read-depth at least two standard deviations below the mean. We subjected these predictions to the same cross-sample reassessment as used for the duplications. We reclassified an individual as being deleted if the estimated diploid copy number over the interval was less than 1.5. Limiting analysis to segments 20 kbp and larger, we defined 172 non-overlapping deletion intervals encompassing 5.35 Mbp. In contrast to the duplications, only 22% of the identified intervals are predicted to be shared among all three of the individuals (Supplementary Note Figures 8, 9, 10; Supplementary Note Table 4).

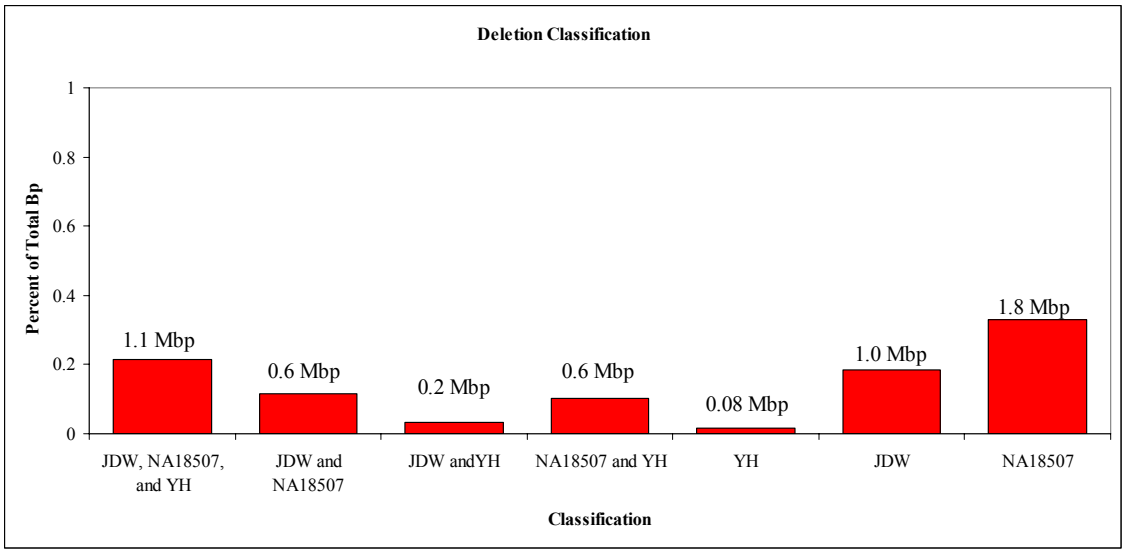
In contrast to the duplication predictions, most of the predicted deletion intervals are not shared among all three individuals. The calling and mapping algorithms we have

Nature Genetics: doi:10.1038/ng.437

implemented are optimized for the detection of large blocks of duplicated sequence and the accurate estimation of absolute copy number, particularly for genomic segments that are highly duplicated. The detection of smaller deletions of unique sequence based on reduced sequence coverage is a problem best solved by other segmentation and analysis approaches.



Nature Genetics: doi:10.1038/ng.437

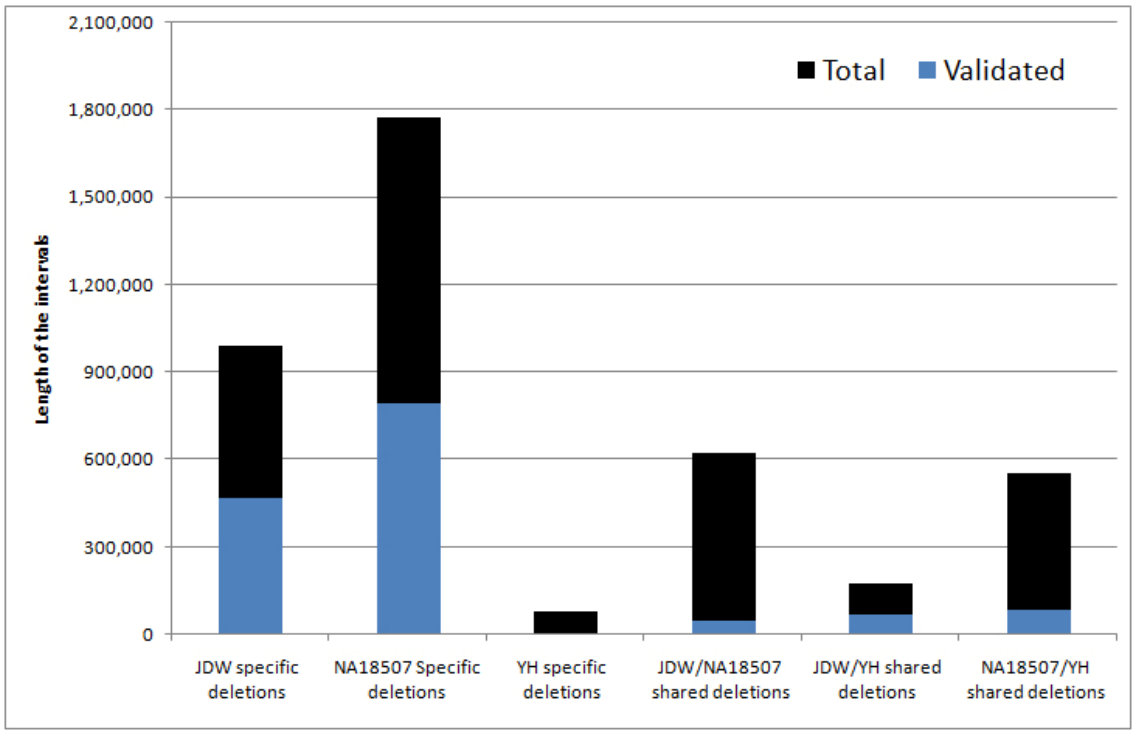


Supplementary Note Figure 8. Classification of shared and individual-specific deletions. a) number of base pairs, b) number of intervals.

	Intervals			Length		
	Validated	**Total**	**% Validation**	**Validated**	**Total**	**% Validation**
JDW Specific Deletions	10	32	31.25%	466,987	991,941	47.08%
NA18507 Specific Deletions	13	50	26.00%	794,150	1,774,821	44.75%
YH Specific Deletions	0	3	0.00%	0	80,773	0.00%
JDW/NA18507 Shared Deletions	2	24	8.33%	49,392	625,333	7.90%
JDW/YH Shared Deletions	2	6	33.33%	71,130	174,642	40.73%
NA18507/YH Shared Deletions	1	19	5.26%	83,844	544,445	15.40%
JDW/NA18507/YH Shared Deletions		38			1,149,623	

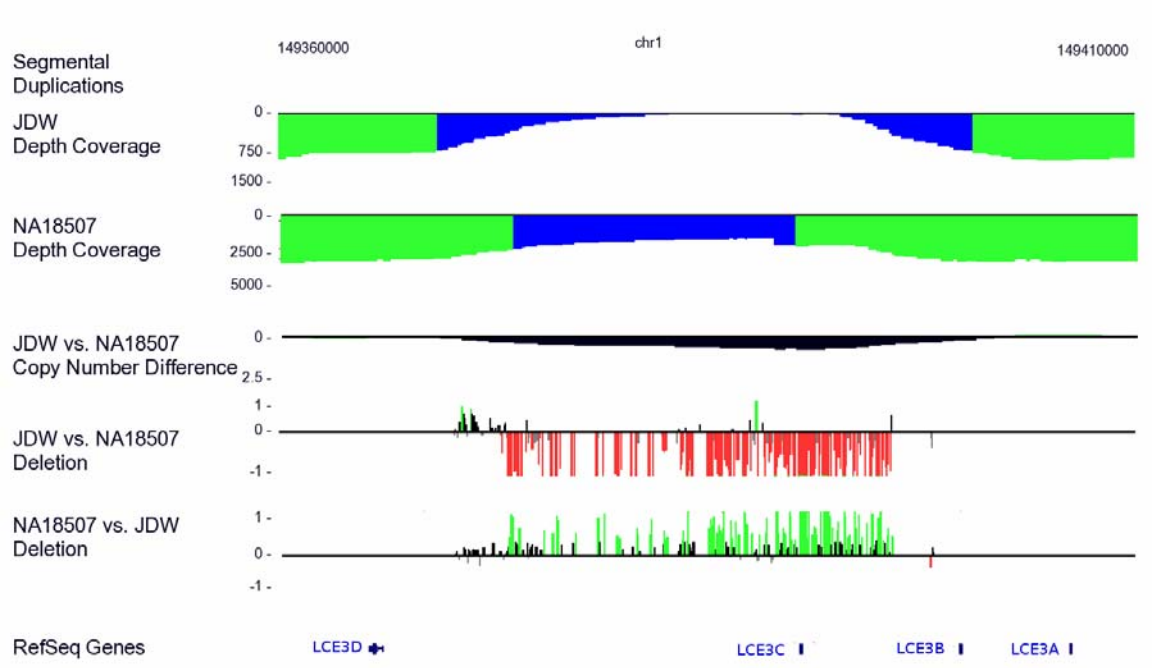
Supplementary Note Table 4. Summary of known and detected autosomal segmental deletions >20 kbp.

Nature Genetics: doi:10.1038/ng.437



Supplementary Note Figure 9. Comparison and validation of detected autosomal deletions. Number of deleted base pairs predicted and validated in NA18507, JDW and YH (autosomes only) are shown. The height of the bars represents the sum of computationally predicted interval lengths, and the blue color bars correspond to the experimentally validated portion. Only deleted intervals >20 kbp were considered for validation.

Nature Genetics: doi:10.1038/ng.437



Supplementary Note Figure 10. Deletion prediction from reduced depth-of-coverage. A homozygous *LCE3C* deletion in JDW genome overlapping with a hemizygous deletion in NA18507 detected from reduced depth-of-coverage and validated with arrayCGH (build35 coordinates chr1:149,360,000-149,410,000).

V. ArrayCGH Validation

We performed array comparative genomic hybridization (arrayCGH) to confirm individual-specific duplications and to confirm copy-number differences in shared duplications. We used two customized oligonucleotide microarrays (NimbleGen, 385,000 isothermal probes). One design was targeted specifically to the primate segmental duplications detected in human, chimpanzee, orangutan and macaque [28]. This covered 180 Mbp of corresponding sequence from the human genome at a density of 1 probe every 525 bp (GEO13934). The second microarray targeted new duplications and deletions identified by our three-way human comparison. This design spanned 88.4 Mb of sequencing, resulting in 1 probe every 230 bp. As part of both designs, we also selected 10 regions (100 kbp each) of single copy DNA to serve as copy-number invariant control regions for the analysis of the hybridizations (9 autosomal and 1 X chromosome regions). For the second microarray, only 8 autosomal regions were considered as a control regions since one of them (in chr10) overlapped with a potential deletion in JDW.

All the results are combined together in a single validation experiment. A total of six intra-specific experiments were performed and the $\log_2$ relative hybridization intensity was calculated for each probe. These experiments included the following genomic DNA comparisons: JDW vs. NA18507, JDW vs. YH, and NA18507 vs. YH. All the

Nature Genetics: doi:10.1038/ng.437

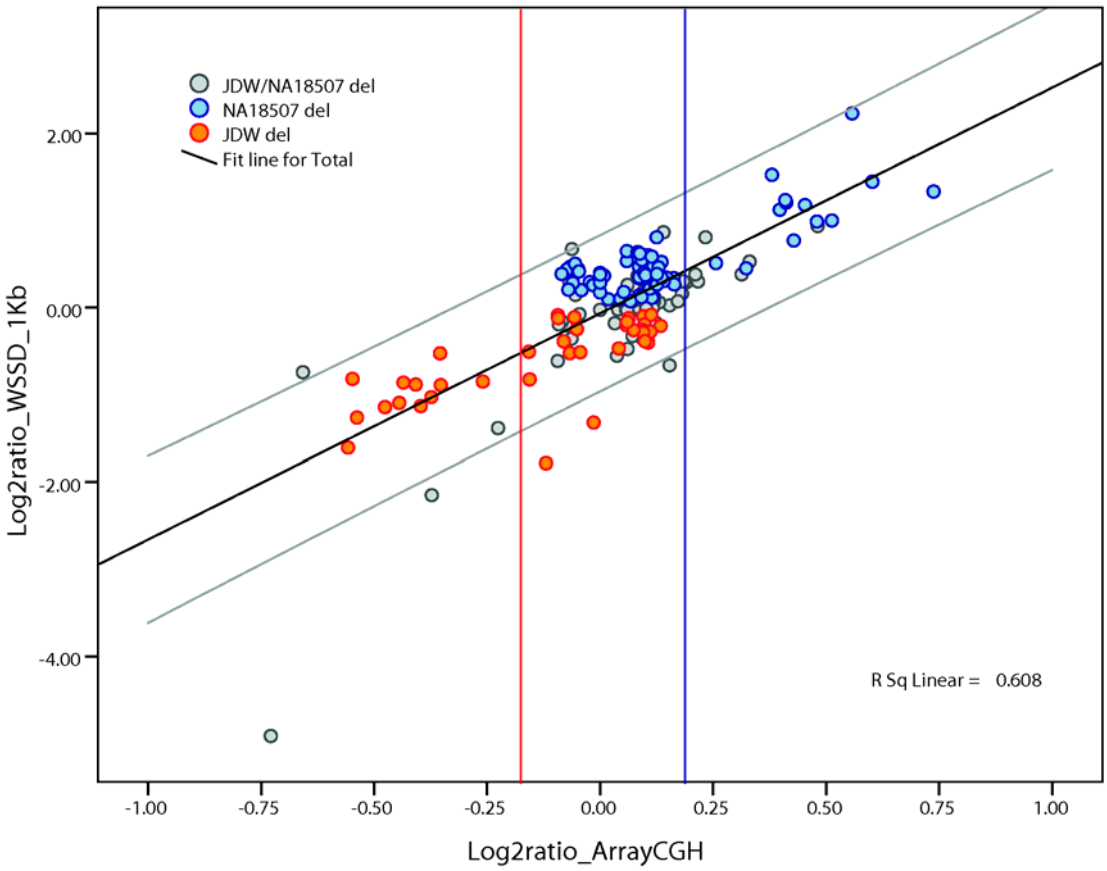
experiments were performed with a standard replicate dye-swap experimental design (reverse labeling of test and reference samples).

To analyze the results of the hybridizations and to validate our predictions, we considered only those probes that showed a consistent result in replicate dye-swap experiments (~55% of probes). We further restricted our analysis to those regions that were greater than 20 kbp in length and had less than 80% of common repeats. We first correlated experimental and computational copy-number estimations. We computed the copy number for individuals based on the depth-of-coverage of aligned WGS (WSSD duplication) against the human reference assembly (hg17). Based on this computational estimate of copy number, we calculated a predicted $\log_2$ copy-number ratio for each autosomal duplication interval >20 kbp in length (and with less than 80% of total common repeat content). These values were plotted against the experimental $\log_2$ ratios determined by array comparative genomic hybridization specific for each individual. The correlations (R^2) ranged from 0.9 for specific duplications to 0.5 for some shared duplications (Figure 4 and Supplementary Figure 4). Specific duplications are more accurately correlated with predicted copy number because in shared duplications with similar copy number there is an associated noise in both the computational predictions and experimental $\log_2$ (although usually centered around zero).

Secondly, we used a heuristic approach to calculate specific $\log_2$ thresholds of significance for each comparison and experimental array [29]. In short, we dynamically adjusted the thresholds for each hybridization to result in a false discovery rate of <1% in the control regions.

With this method, we statistically validated 17 duplication intervals (68%) of the initial 25 predicted intervals greater than 20 kb that were found not shared by all three individuals. To make calls on validated sites, we required the interval to be statistically significant in both complementary hybridizations (so for a JDW-specific SD it had to be validated in both JDW/NA18507 and JDW/YH arrayCGH experiments). Those validated intervals encompassed 1.1 Mb of sequence not duplicated in at least one of the three individuals. Similarly, we found 28 validated deletions (12% of the 134 predicted not shared in the three individuals). This accounted for 1.4 Mb (Supplementary Note Figures 8, 9, 10, Supplementary Note Table 3). See also Supplementary Note Figure 11 for *in silico* vs. experimental $\log_2$ ratios of copy numbers in deletion regions.

Nature Genetics: doi:10.1038/ng.437



Supplementary Note Figure 11. Correlation between computational and experimental copy number in deletion regions. Based on our computational estimates of copy number, we calculated a predicted $\log_2$ copy-number ratio for each autosomal duplication interval >20 kbp in length (and with less than 80% of total common repeat content). These values were plotted against the experimental $\log_2$ ratios determined by oligonucleotide intraspecific array comparative genomic hybridization. The vertical bars represent an approximation of the threshold used for the validated calls of the specific experiment (see Methods).

Validated gene list

The RefSeq gene list was retrieved from the RefSeq May 08 dataset (UCSC Genome browser http://genome.ucsc.edu/). Validated segmental duplication and deletion intervals were classified by type and cross-referenced within RefSeq transcript assignments. Complete genes were classified as only those genes where the full transcript mapped within the segmental duplication interval.

We detected 68 non-redundant (gene-families collapsed) genes overlapping validated segmental duplications. Several genes are found to be distinctly *duplicated* in the three individuals. For instance, we found that JDW had an excess of copies of Complement

Nature Genetics: doi:10.1038/ng.437

Factor H-related 1 and 4 (*CFHR1* and *CFHR4*). There is a well-established inverse relationship between the highly variable Kringle IV size polymorphism and Lp(a) levels in humans, and it is know that African-descent populations have on average 2–3 times higher levels of Lp(a) compared with European-descent populations. This will be congruent with our result in which NA18507 has less copies than the other individuals and JDW has more copies than any of the other two. The *defensin* cluster (8p23.1, region of clinical relevance for innate immunity, inflammation and cancer) is found to be highly variable among individuals and, in our case, JDW has fewer copies than NA18507 which in turn has fewer copies than the Asian individual. Another interesting example is Amylase. In our study, we found that JDW had less copies (5.5) than both NA18507 (9.5) and YH (9.7). Here we found JDW to have fewer copies of *CCL3L* (3.2) than YH (4.5) or NA18507 (6.5). Finally, two genes (*LRRC37B* and *TBC1D3*) are among the core duplicon regions reported previously [29]. Those regions are highly dynamic and, hence, it is not surprising to have them variable among individuals.

VI. FISH Validation

We experimentally validated by fluorescence in situ hybridization (FISH) 11 predicted copy-number differences of the genomes of the Yoruba (NA18507, Illumina) and Chinese individual (YH, Illumina) using cell lines from the same individuals from which the computational predictions were generated. Using FISH we found that 7 out of 11 of copy-number differences were concordant with computational predictions (Supplementary Note Table 5). FISH experiments confirmed two YH-specific duplications on chromosome 17q21.31. Five copies in YH and two copies (unique) in the NA18507 genome were predicted and experimentally confirmed (Figure 5a). We validated predicted copy-number differences between NA18507 and YH in the 15q11.2 region in which we found one and two copies, respectively, and in the 16q22.1 region where YH was found to have four copies and NA18507 five copies. We validated the predicted copy-number difference between the Yoruba (12 copies) and Chinese individuals (5 copies) in a known polymorphic region in 17q21.32 (Figure 5b) and in the 7q35 region in which the Chinese was found to have more copies (7 copies) than the Yoruba individual (4 copies).

FISH experiments were concordant with the computational prediction also in the *defensin* cluster at 8p23.1, where the Chinese individual was found to have more copies (5 copies) than the Yoruba individual (3 copies) (Figure 5c).

We note that our read-depth approach returns a real-value estimate of absolute copy number. For comparisons, we simply round to the nearest integer value.

Nature Genetics: doi:10.1038/ng.437

Prediction		FISH validation				
NA18507	YH	NA18507	YH	Chromosome	Clone	Position (hg17)
4.0	6.1	4	7	7q35	WIBR2-1511F11_G248P82413C6	chr7:143,471,773-143,513,118
3.04	4.49	3	5	8p23.1	WIBR2-2553B23_G248P83461A12	chr8:7,697,239-7,734,464
0.8	1.9	1	2	15q11.2	WIBR2-2992K17_G248P89716F9	chr15:20,384,910-20,425,229
4.12 - 40.75	5.77 - 34.64	>25	>25	16p13.11	WIBR2-1564J16_G248P83059E8	chr16:14,921,936-14,963,711
4.35 - 49.77	4.29 - 44.75	>30	>30	16p12.2	WIBR2-0894D06_G248P8387B3	chr16:21,300,233-21,343,884
11.77 - 54.03	14.92 - 48.47	>50	>50	16p11.2	WIBR2-1354F18_G248P84059C9	chr16:33,235,396-33,276,997
5.1	4.0	5	4	16q22.1	WIBR2-3400B16_G248P802583A8	chr16:68,708,494-68,749,105
10.004 - 38.55	7.92 - 21.32	16	12	17q12	WIBR2-2247O08_G248P88106H4	chr17:33,353,398-33,393,623
1.77	5.04	2	5	17q21.31	WIBR2-1797D06_G248P85943B3	chr17:41,870,497-41,906,290
2.0	4.6	2	5	17q21.31	WIBR2-1854B21_G248P85429A11	chr17:42,055,754-42,093,334
12.98	4.97	12	5	17q21.32	WIBR2-0946N09_G248P801829G5	chr17:42,986,667-43,025,934

Supplementary Note Table 5. Summary of FISH Validation.

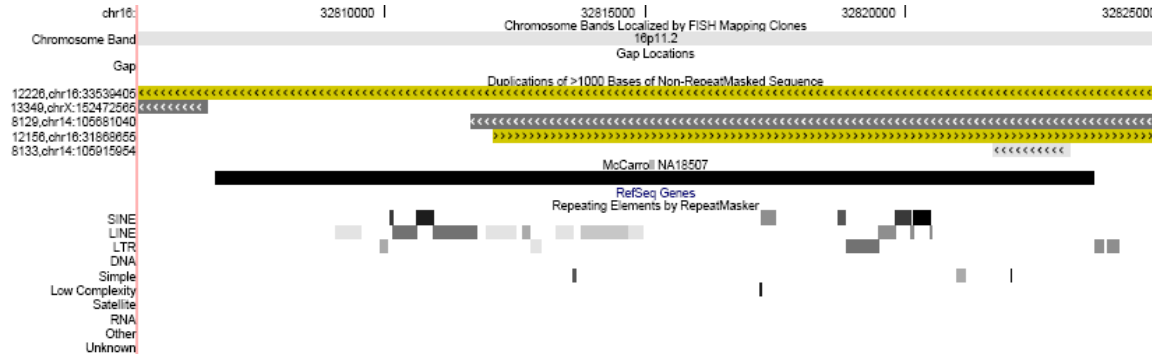
In three cases, FISH failed to accurately estimate copy-number difference between NA18507 and YH genomes. The 16p12.2 locus was predicted to have 4.35–49.77 copies in NA18507 and 4.29–44.75 in the YH. We report a range of copy numbers instead of the median due to the mosaic architecture of the region. Using FISH we were able to estimate a copy number >30 in both the genomes, but the signals visualized on interphase nuclei were too numerous to provide an exact number of copies (Figure 5d). Similar results were obtained for the 16p13.11 and 16p11.2 loci (Supplementary Note Table 5). It is worth noting that all of these regions have a mosaic architecture [30], therefore it can be possible that a small portion of the probes is copy-number polymorphic in the two genomes but is too small to be visualized by FISH. In the 17q12 region, FISH detected 16 copies in NA18507 and 12 in YH, confirming a higher number of copies predicted in the Yoruba with respect to the Chinese individual (10.004–38.55 in NA18507 and 7.92–21.32 in YH). Also in this case we report a range of copy numbers instead of the median due to the mosaic architecture of the region.

In total we validated 7/11 predicted copy-number differences of the genomes of the Yoruba and Chinese individuals. In all the cases the additional copies map on the same chromosome, except for the 16q22.1 region, in which some signals are also detected on the pericentromeric regions of chromosome 2, 9, 10, 15, and 22.

VII. SNP Microarray Comparison

As an independent test, we compared absolute copy numbers based on *mrFAST* depth-of-coverage with the copy numbers reported in McCarroll *et al.* for sample NA18507 [31]. In the McCarroll study, samples were assigned to integer copy-number states based on the clustering of fluorescence intensities from Affymetrix arrays for 270 samples. Since integer copy-number assignments (i.e., 0,1,2 or 2,3,4) were determined based on the ratio of intensity values between clusters, it is possible that the copy-number calls for some of the McCarroll loci may be correct in a relative but not in an absolute sense. This is particularly likely to occur for events involving duplicated sequences that have a normal diploid copy number greater than 2. One example of this, McCarroll CNP 12431 (chr16:32806748-32823606), is shown below.

Nature Genetics: doi:10.1038/ng.437



Supplementary Note Figure 12. Genome browser image of McCarroll CNP 12431. The yellow and gray duplication bars indicate that this sequence is represented elsewhere in the reference assembly.

Based on the array data McCarroll *et al.* report four clusters (assigned copy numbers of 1,2,3, and 4), with sample NA18507 assigned to the "copy number = 2" cluster. However, as shown in Supplementary Note Figure 12, this region is duplicated elsewhere in the genome. In addition to the coordinates given for CNP 12341 this sequence is present at two other locations on chr16, (each having >98% sequence identity) as well as at a single location on chr14 (>95% sequence identity). The region is represented in four locations in the genome assembly, resulting in an expected diploid copy number of 8. This agrees with the copy number estimated for NA18507 by *mrFAST* using the depth of Illumina reads (diploid copy-number estimate = 8.02). The copy-number state of "2" is the most common cluster reported in the McCarroll data, suggesting that in this case the reported numbers correspond to relative copy-number differences rather than absolute copy numbers for this sequence. Alternatively, it is possible that it is this specific paralog that is variable in copy, and that this is accurately reflected by the McCarroll cluster definitions. In either case, it is clear that the absolute copy number estimated for duplicated loci do not correspond to genotypes reported from the array data.

Copy Number State	Number of Individuals
1	16
2	201
3	16
4	2
NA	35

Supplementary Note Table 6: Distribution of assigned copy numbers at CNP 12341.

Given this potential effect we focused our analysis on the 992 loci reported by McCarroll *et al.* for NA18507 that do not intersect with regions annotated as segmental duplications.

Second, in order to get comparable copy-number estimates we corrected the Illumina-based copy numbers for micro-duplications, identified as short stretches of sequence that are not annotated as duplications or common repeats but are overrepresented in the genome assembly (see Supplementary Note section IIId: Removal of short-read mapping

Nature Genetics: doi:10.1038/ng.437

artifacts). Using this cleaned set of copy numbers, we then compared the copy-number estimates based on Illumina read-depth (calculated in non-overlapping windows containing 1 kbp of unmasked sequence, with the first and last window removed to reduce edge effects and only considering intervals containing at least 1 artifact-free window) with those defined by McCarroll *et al.* Not surprisingly, the correlation is stronger for longer intervals.

Locus Size	Number of Loci	R^2
>= 10 kbp	282	0.5188
>=20 kbp	128	0.4988
>=40 kbp	57	0.8343
>=100 kbp	11	0.9189

Supplementary Note Table 7. Correlation between copy-number estimates based on Illumina depth-of-coverage and those reported in McCarroll et al. for sample NA18507.

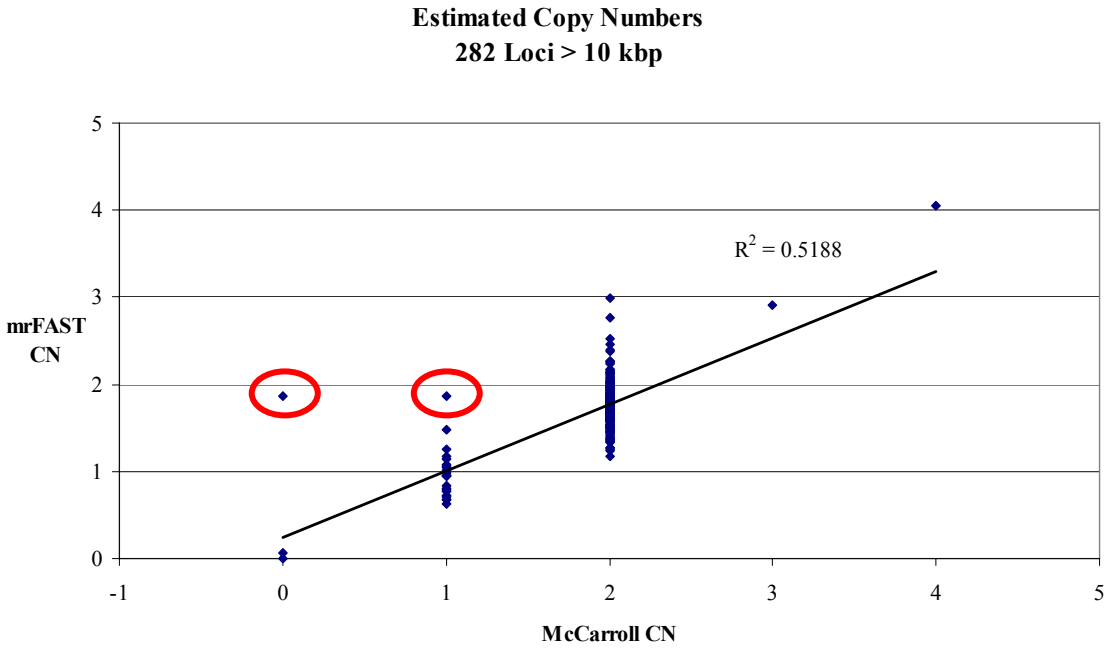
We performed a further comparison of our copy-number predictions for sample NA18507. We found a total of 88.7% to 100% agreement, with the copy number=2 class showing the lowest level of agreement. The data show that the larger a copy-number variant event the better the correspondence. Thus, this method can be used for accurate genotyping of copy-number variation.

Nature Genetics: doi:10.1038/ng.437

Locus Size	Copy Number (McCarroll)	Number of Loci (McCarroll)	Number With Concordant Read Depth Classificaiton	Percent Agreement
>=1 kbp	0	18	10	55.6%
	1	97	81	83.5%
	2	829	669	80.7%
	3	9	4	44.4%
	4	2	2	100.0%
	Total	955	766	80.2%
>=5 kbp	0	8	5	62.5%
	1	48	42	87.5%
	2	478	391	81.8%
	3	3	2	66.7%
	4	1	1	100.0%
	Total	538	441	82.0%
>=10 kbp	0	3	2	66.7%
	1	22	21	95.5%
	2	255	225	88.2%
	3	1	1	100.0%
	4	1	1	100.0%
	Total	282	250	88.7%
>=20 kbp	0	1	0	0.0%
	1	8	8	100.0%
	2	117	110	94.0%
	3	1	1	100.0%
	4	1	1	100.0%
	Total	128	120	93.8%
>=40 kbp	1	3	3	100.0%
	2	52	50	96.2%
	3	1	1	100.0%
	4	1	1	100.0%
	Total	57	55	96.5%
>=100 kbp	1	1	1	100.0%
	2	9	9	100.0%
	3	1	1	100.0%
	Total	11	11	100.0%

Supplementary Note Table 8. Copy-number estimates based on Illumina data were rounded to the nearest integer and compared with the results reported by McCarroll et al. The resulting assignments were then compared. This is the uncorrected comparison including the two CNVs that were likely misclassified by McCarroll *et al.* (see below).

Nature Genetics: doi:10.1038/ng.437



Supplementary Note Figure 13. Comparison of copy numbers reported in McCarroll et al. with those estimated using the depth of Illumina reads in sample NA18507. Some clear outliers, such as CNP 2434 and CNP 1203 (circled in red), may represent copy-number state misclassifications.

Among events >10 kbp in length, we noted two striking outliers among the 282 loci and investigated these in more detail. The most striking is McCarroll CNP 2434 (chr19:58210563-58244245). At this locus, McCarroll *et al.* assigns NA18507 a copy number of '0'. Based on the depth of Illumina reads a copy number of 1.86 (which rounds to 2) is estimated. This is surprising, since a homozygous deletion event should result in nearly zero mapped reads. Neither fosmid end pair mapping[32] nor intensity data from an Illumina 1M genotyping array, support a deletion for NA18507 at this locus. Also, this interval contains five heterozygous SNPs according to HapMap Phase II genotypes. Together, these additional data sets indicate that the sample has likely been misclassified. Examination of the McCarroll genotypes for this locus indicates that majority of the samples were assigned to the copy number '0' state. This suggests that the intensity values may be been incorrectly normalized and that, rather than representing a common deletion event, CNP 2434 actually corresponds to a more rare duplication. Reassigning this value to a predicted copy-number state of '2' increases the correlation to 0.5893 and improves the percent agreement reported in Supplementary Note Table 7.

Nature Genetics: doi:10.1038/ng.437

Copy Number State	Number of Individuals
0	241
1	21
2	6
NA	2

Supplementary Note Table 9: Distribution of assigned copy numbers at CNP 2434 as reported in McCarroll et al.

A second outlier is CNP 1203 (chr7:156837843-156850316), a predicted 12.4-kbp deletion having an estimated copy number of 1 from McCarroll *et al.* and of 1.86 based on Illumina read-depth. There is no fosmid end-sequence support for a 12.4-kbp deletion at this position, and four heterozygous SNPs are reported in the HapMap for this sample. As before, these external findings are consistent with a copy-number state of '2' at this locus. Reassigning CNP1203 along with CNP 2434 to a copy-number status of 2 increases the correlation with Illumina read-depth to 0.6115. Following the correction for these two sites, we correctly assign all (25/25) sites larger than 10 kbp that have a copy number other than 2. Overall, we conclude that there is excellent concordance with the genotypes for NA18507 reported by McCarroll *et al.* but, in some cases, the copy number estimated from Illumina read-depth may be a more accurate representation of the absolute copy number.

We found a stronger correlation of estimated copy number for larger sites and identified regions that may have the wrong absolute copy number reported by McCarroll *et al.* After correcting two of these loci, we find a strong correlation (R^2=0.62) with McCarroll copy numbers for regions >10 kbp. If we round our absolute copy-number estimates to the nearest integer, we agree with 89% of the McCarroll assignments.

VIII. Quantitative PCR Comparison

Neither FISH nor array data offer accurate estimate for certain regions (usually regions of high copy number). Quantitative PCR is at best qualitative for these types of sequences. There is no ideal experimental technology for these high-copy duplications. Nevertheless, we designed quantitative PCR assays targeted to 7 genes with high copy number differences not validated by arrayCGH. We reasoned that in most of these cases, arrayCGH did not have enough power to discriminate because of saturation in copy number. We performed the experiments comparing copy-number predictions for two genomes, NA18507 and YH. The results of the experiment show that in 5 out of the 7 experiments performed, we obtained a "good correlation" between the predicted ratio of mrFAST and qPCR (Supplementary Note Table 10 and Supplementary Note Figure 14) that was superior to arrayCGH. The 2 cases that are most discrepant contained microduplications and may have unstable copy number (although PCDHB2 also contained microduplications and had a good correlation).

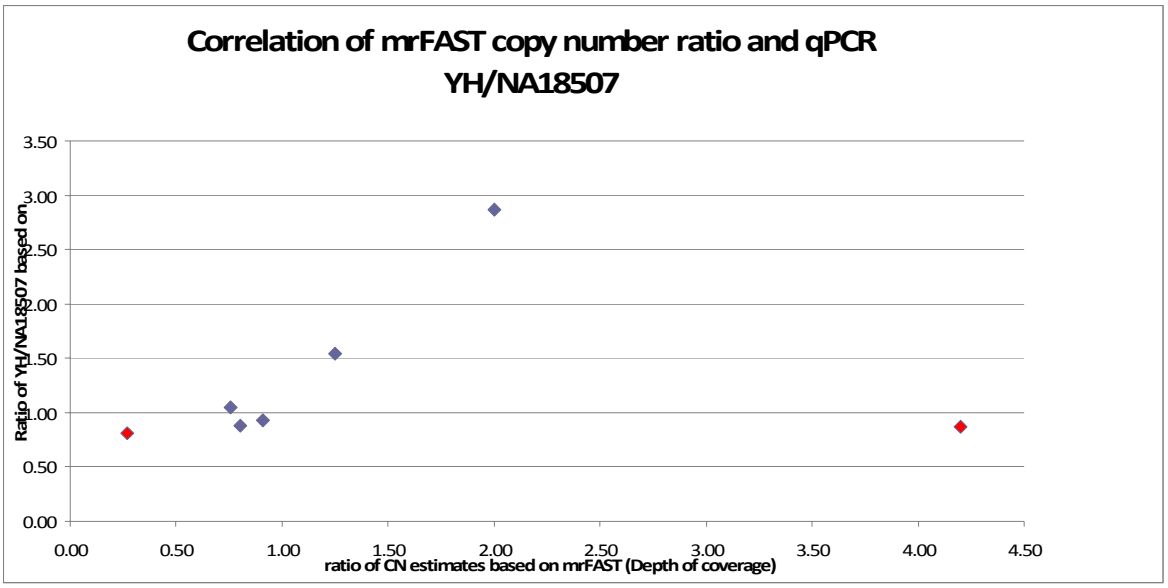
Nature Genetics: doi:10.1038/ng.437

In summary, mrFAST might be a potential replacement for arrayCGH, even for those genes with high copy number in which arrayCGH can not accurately detect variation among individuals.

GeneName	NA118507 CN	YH CN	mrFAST ratio YH/NA18507	qPCR (2 replicates)	arrayCGH log2
ANKRD20A1	33	25	**0.76**	**1.05**	-0.18
GALP*	5	21	**4.20**	**0.875**	0.16
NBPF14	282	227	**0.80**	**0.88**	0.18
NBPF20	56	51	**0.91**	**0.925**	0.1
PCDHB2*	6	12	**2.00**	**2.87**	0.04
RPS3A*	15	4	**0.27**	**0.81**	0.05
WASH1	16	20	**1.25**	**1.54**	0.21

*= regions which intersect with microduplications and may have unstable copy number estimates.

Supplementary Note Table 10. qPCR comparison.



Supplementary Note Figure 14. Scatterplot of the ratio of copy number in NA18507 and YH individuals estimated computationally by mrFAST and experimentally by qPCR. Notice that there is a good correlation in the blue dots, but they are the two genes that failed to confirm the power of our approach. Both of them contained microduplications which complicates the estimation of copy number.

IX. Simple Gene Table Analysis

Our gene analysis began with RefSeq transcripts (n=29,129 transcript structures, n=20,301 gene names, n=29,870 genomic segments). We limited our analysis to coding transcripts located on autosomes and created a non-redundant set of genes as follows: for each alternative transcript, we chose the largest gene model and a single chromosome

Nature Genetics: doi:10.1038/ng.437

location; and for multiple distinct entries having the same transcription coordinates we chose a single entry. An additional bias may occur since a deletion or duplication of a single member of highly homologous gene family may be manifested as a variation in the copy number of all members of the family. In order to correct for this bias we created a non-redundant set of genes by collapsing RefSeq entries related by WGAC segmental duplication alignments of 95% sequence identity or greater. Briefly, using the pair-wise segmental duplication definitions (UCSC browser), we created connected sets of RefSeq genes that were entirely contained within related duplication blocks. We then chose a single gene from each set of connected RefSeqs. In cases where unrelated genes were grouped together (because of larger duplications that spanned multiple genes), we manually split the connections and chose one representative from each distinct gene. For example, we chose *TMPRSS11E* and *UGT2B15* as representatives from a set of duplications on chr4 that encompasses four genes: *TMPRSS11E* and *TMPRSS11E2* (two protease genes) as well as *UGT2B15* and *UGT2B17* (two glucuronosyltransferases).

In total, this procedure resulted in 17,601 nonredundant genes. 3.8% of these genes (662/17,601) have a range of estimated copy number among the three individuals that is at least 1.0. Half of the copy-number variable genes (51.5%; 341/662) intersect with duplicated segments represented multiple times in the assembly (WGAC). This represents a significant enrichment (Fisher's exact test, p-value <2.2e-16, odds ratio 7.053548), indicating that duplicated genes are more likely to show copy-number differences.

This analysis utilized the absolute copy number estimated over the entire genomic position of each gene. Cases where specific exons or functional domains have been expanded or contracted may have been missed. We searched for such situations by estimating absolute copy number for 192,121 autosomal exons. The calculated copy numbers were based on 1-kbp non-overlapping windows located within 5 kbp of each exon. Exons from 845 genes were copy-number variable. Of these, 504 genes were not identified as variable at the whole gene level. This set includes variants such as *LPA* (JDW 40.4 copies) and *MUC4* (NA18507 4.36 copies), which result in altered numbers of functional domains.

Unlike the predictions of segmental duplications, this analysis was performed for all genes regardless of their length. In order to validate the predictions we performed a set of arrayCGH experiments targeting all 662 genes that show at least 1 copy-number difference in any pairwise comparison (24.2 Mb of coverage, 1 probe every/63 bps). We performed the same pairwise experiments (JDW vs. NA18507, NA18507 vs. YH) in a typical dye-swap experiment as described above with one exception: the hybridization of JDW vs. YH could not be performed because of a limited quantity of JDW DNA sample. We found that the correlations between predicted copy number and experimental hybridizations are affected by both the size of the gene and the presence of a variable copy number of shorter segmental duplications (<1 kbp) (hereafter called microSDs, See Supplementary Notes section IIId). The length of the gene appears to be the most critical parameter since we need sufficient repeat-free 1-kb windows to avoid local fluctuations in copy-number estimations and sufficient probe coverage in order to experimentally

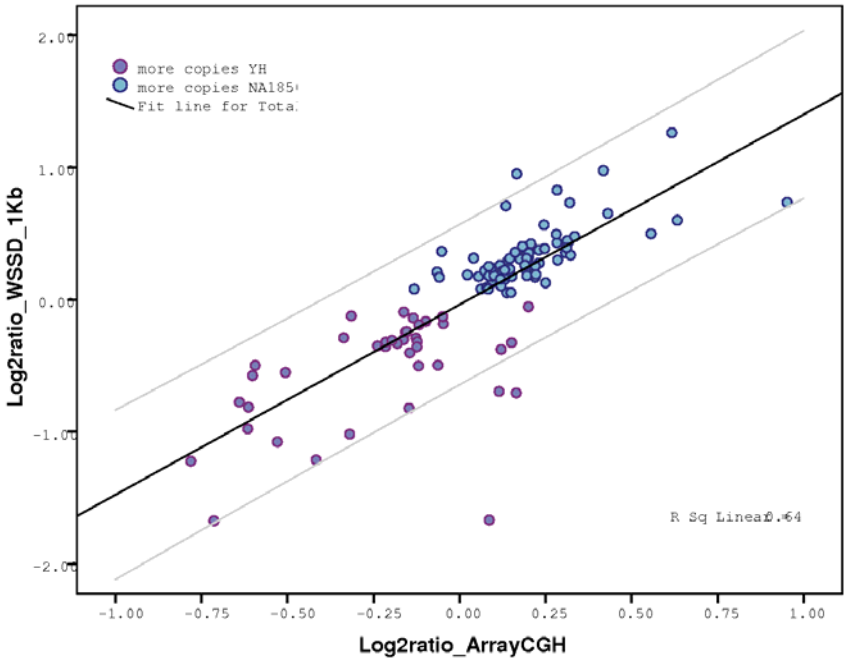
Nature Genetics: doi:10.1038/ng.437

validate a given prediction. We applied two different copy-number estimations: a) the standard approach, described previously in the paper and b) copy-number recalculation by removing small regions with highly expanded copy number (but not considered common repeats).

We identified a total of 113 genes (Supplementary Table 6) with computational read-depth and experimental evidence of copy differences among the three individuals. Due to the inability of arrayCGH to validate slight differences in copy number (e.g. 12 vs. 13 copies) and limitations of detecting smaller variants by this method, this 113 gene set represents a conservative estimate. For example, we find that if the differential in copy number is less than 0.25, we have almost no power to detect the difference by arrayCGH (Supplementary Note Figure 18).

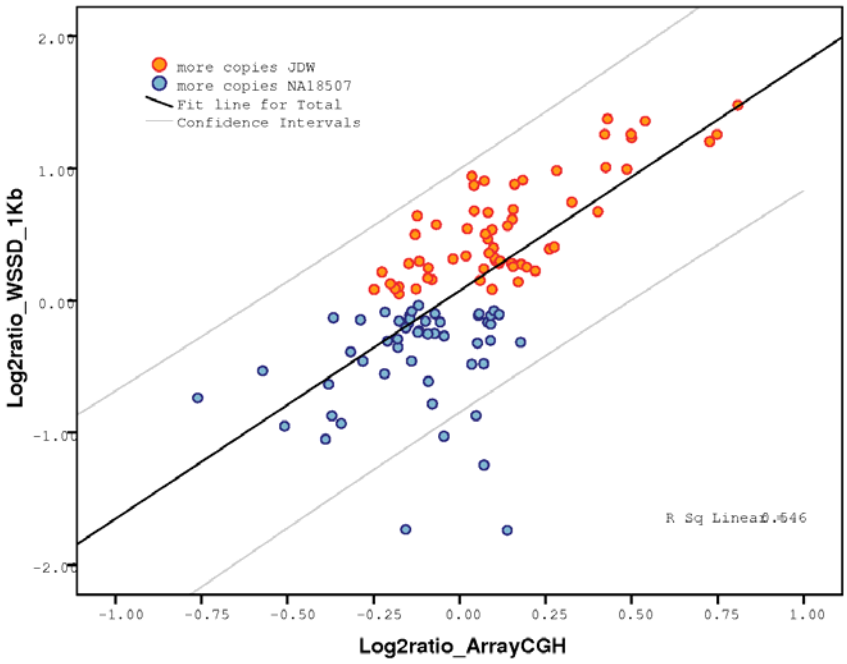
		Correlation of aCGH vs DOC		
		JDW/NA18507	JDW/YH (1hyb)	NA18507/YH
Standard copy number	R sq ALL	0.247	0.13	0.193
	R sq | ABS CN > 1 | ALL	0.286	0.159	0.246
Copy number (no microSD)	R sq | ABS CN > 1 | ALL	0.312	0.179	0.316
	R sq | ABS CN > 1 | Genes > 5kb	0.546	0.43	0.64
	R sq | ABS CN > 1 | Genes > 10kb	0.701	0.641	0.812
	R sq | ABS CN > 1 | Genes > 20kb	0.689	0.6	0.829

Supplementary Note Table 11. Summary of the correlations between experimental validation and the predicted copy number of copy-number variant genes among humans.



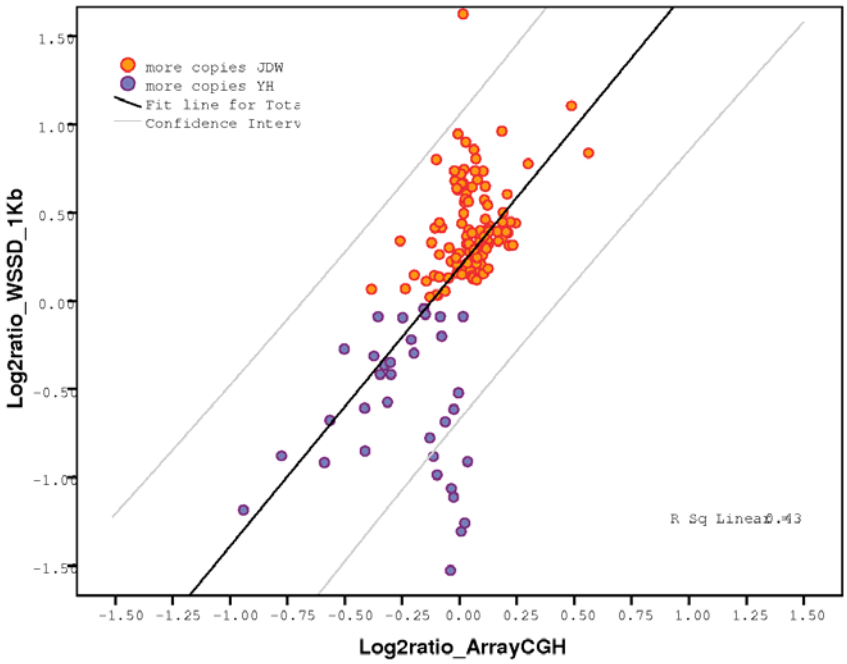
Nature Genetics: doi:10.1038/ng.437

Supplementary Note Figure 15. Predicted copy number log_2 ratio (no microSDs) vs. arrayCGH log_2 ratio of genes larger than 5 kb in NA18507 and YH genomes.

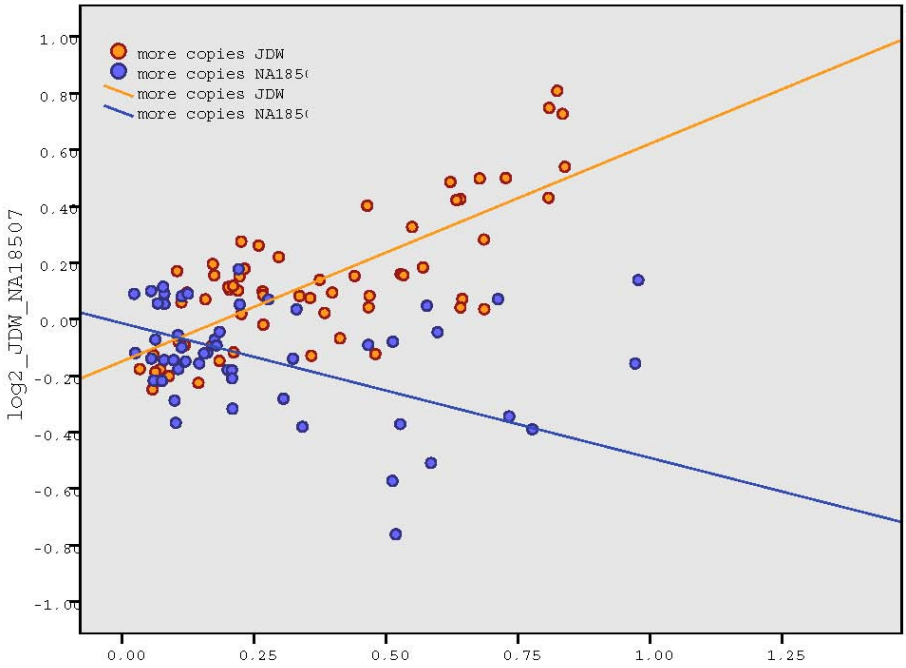


Supplementary Note Figure 16. Predicted copy number log_2 ratio (no microSDs) vs. arrayCGH log_2 ratio of genes larger than 5 kb in JDW and NA18507 genomes.

Nature Genetics: doi:10.1038/ng.437



Supplementary Note Figure 17. Predicted copy number $\log_2$ ratio (no microSDs) vs. arrayCGH $\log_2$ ratio of genes larger than 5 kb in JDW and YH genomes.



Supplementary Note Figure 18. Limitation in detection of copy-number differences by arrayCGH. The proportion of differences, calculated as the difference in copy

Nature Genetics: doi:10.1038/ng.437

number between JDW and NA18507 over the average of copy number, is plotted against the results from the experimental $\log_2$ for genes longer than 5 kb. When the differential in copy number is small (<0.20), the experimental $\log_2$ values tend to cluster within the background noise level (around 0.25). Only when the proportion of differences is substantial, the difference supported by experimental results. This means that validation by arrayCGH will penalize against higher copy-number duplications—a phenomenon referred to as duplication sensitivity[33] and, thus, the 113 genes that we have confirmed as copy-number variable should be considered a conservative set.

Limiting analysis to the 113 genes confirmed to vary in copy number, we can estimate the number of gene-copy differences expected between any two individuals. We estimate that, on average, each pair of individuals differs at 73–87 of these loci.

Number of genes variable among two humans

This analysis of three individuals permits us to estimate how many genes are variable in copy number between any two individuals.

	Genes with ΔCN > 1	Genes with ΔCN > 3	Genes with ΔCN > 5
JDW vs NA18507	73	31	14
JDW vs YH	80	24	11
NA18507 vs YH	87	26	9
Mean	80	27	11.3

Supplementary Note Table 12. The number of genes predicted to differ by at least 1, 3, and 5 copies between any two individuals. This analysis is limited to the 113 genes, which are validated by arrayCGH to vary among these three individuals.

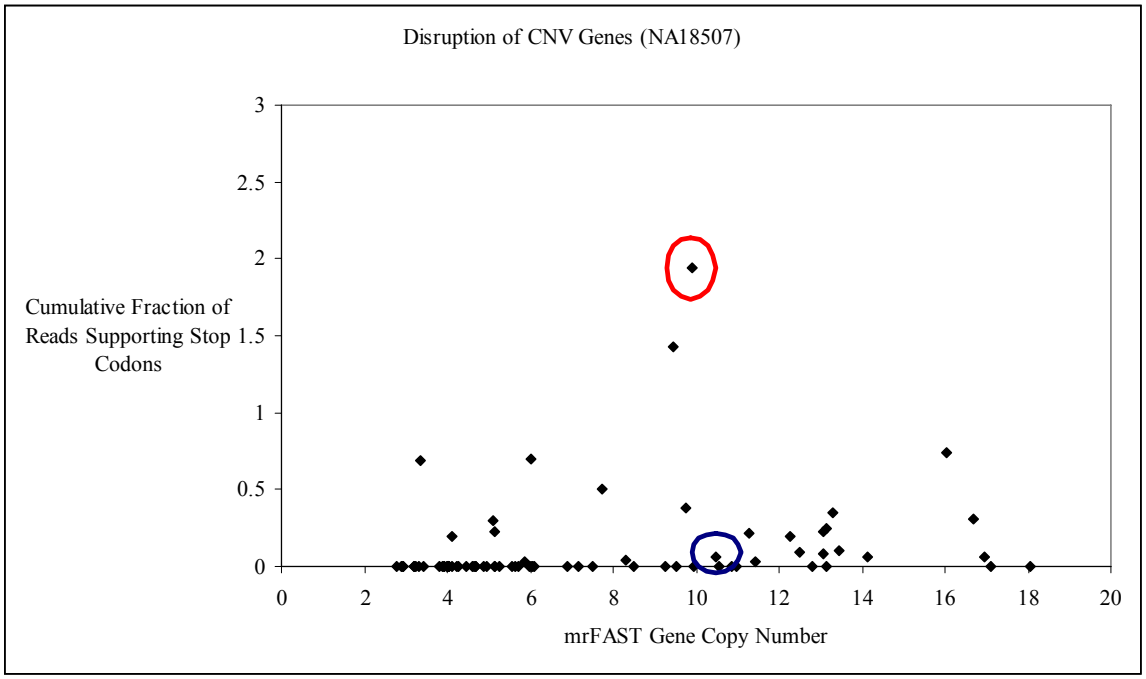
Disrupted Gene Analysis

Owing to the short nature of the sequence reads, it is currently impossible to completely establish phase across duplicated segments in order to assess the proportion of real genes (i.e. with complete ORF) versus pseudogenes (stop codons). Such analyses generally require high-quality sequence data typically from large insert clones to determine. Since *mrFAST* however does track all sequence variants, we did analyze the proportion of disruptive stop codon mutations in unique versus duplicated genes. We analyzed read data from sample NA18507. We limited our analysis to the genes for which we have experimental validation for copy-number variation and only considered the 92 validated CNV genes that are predicted to be duplicated in sample NA18507. Analysis was limited to those changes supported by three or more Q30 reads. For each gene, we identified single nucleotide changes that result in the formation of stop codons and recorded the fraction of reads at the position supporting the change (Supplementary Table 3). As expected, some duplicated genes appear to have been pseudogenized. Interestingly, this property is variable among different genes. For example, the *FAM157A* gene has a

Nature Genetics: doi:10.1038/ng.437

predicted copy number of 10. For this gene, six stop-codon forming, single-nucleotide changes were detected. On average, each of these changes is supported by 33% of the reads mapped at that position, yielding a cumulative fraction of 1.98 (circled in red). In contrast, the *AMY2A* gene also has a predicted copy number of 10. However, for this gene there is only a single stop codon detected which is supported by 6% of the mapped reads (circled in blue).

We stress that this analysis is limited by the short sequence reads and the lack of resolved haplotypes. For example, this approach cannot distinguish between a duplicated gene that is present in four copies with each copy having a single (different) stop codon and the alternative case where one copy has acquired different stop codons while the other three copies remain intact.



Supplementary Note Figure 19. The cumulative fraction of reads from sample NA18507 supporting stop codons in validated as being CNV and predicted to be duplicated in NA18507. The *FAM157A* gene is circled in red and the *AMY2A* gene is circled in blue. Only genes having a predicted copy number less than 20 are depicted. See Supplementary Table 3 for a full description of the genes.

Nature Genetics: doi:10.1038/ng.437

X. References

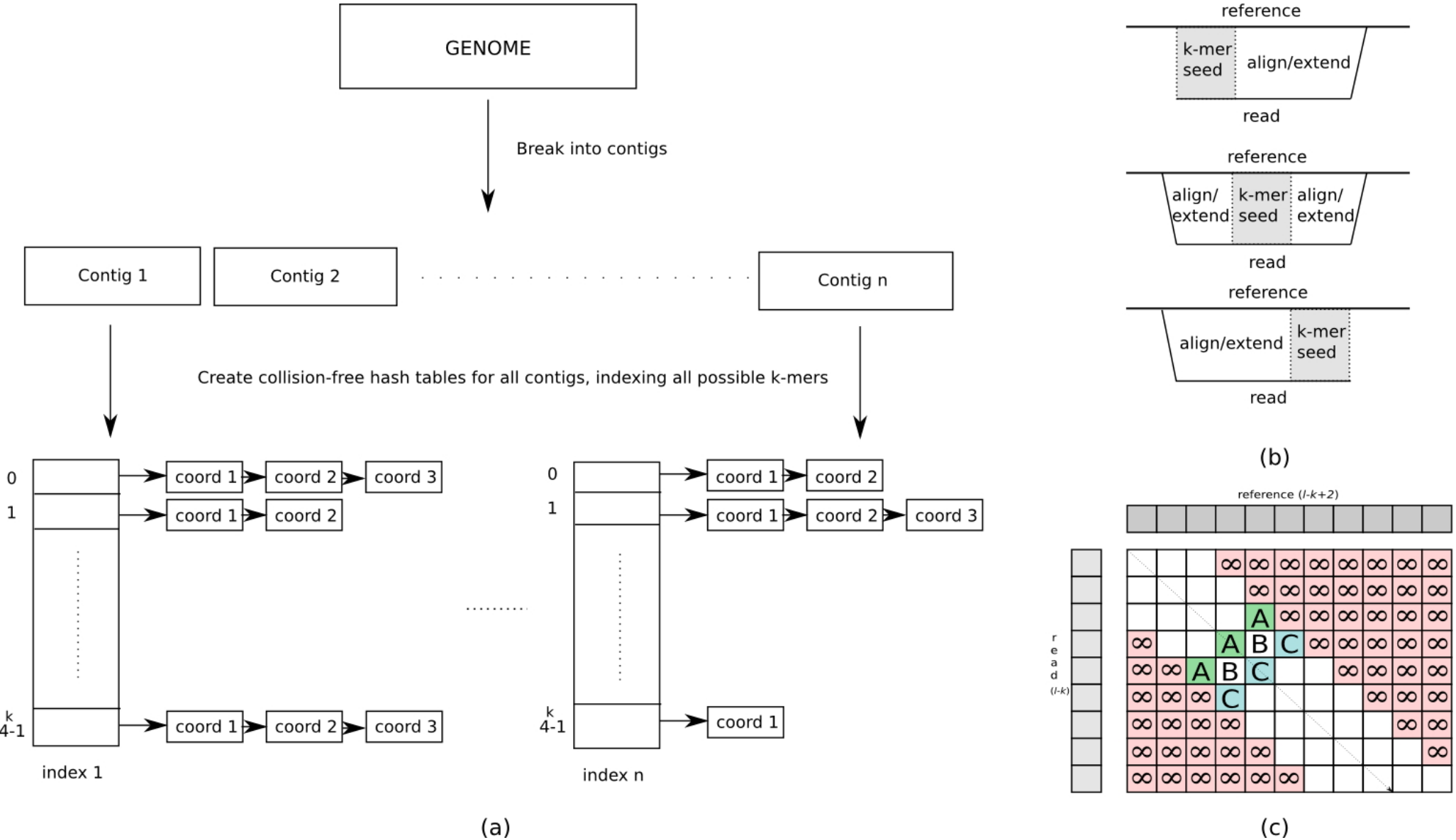
1. Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. Basic local alignment search tool. *J. Molec. Biol.* **215**, 403-410 (1990).
2. Kent, W.J. BLAT--the BLAST-like alignment tool. *Genome Res* **12**, 656-64 (2002).
3. Yanovsky V, Rumble SR & M, B. Read Mapping Algorithms for Single Molecule Sequencing Data. in *Workshop on Algorithms in Bioinformatics (WABI)* (Springer-Verlag, Karlsruhe, Germany 2008).
4. Ma, B., Tromp, J. & Li, M. PatternHunter: faster and more sensitive homology search. *Bioinformatics* **18**, 440-5 (2002).
5. Ning, Z., Cox, A.J. & Mullikin, J.C. SSAHA: a fast search method for large DNA databases. *Genome Res* **11**, 1725-9 (2001).
6. Pearson, W.R. & Lipman, D.J. Improved tools for biological sequence comparison. *Proc Natl Acad Sci U S A* **85**, 2444-8 (1988).
7. Schwartz, S. et al. Human-mouse alignments with BLASTZ. *Genome Res* **13**, 103-7 (2003).
8. Smith, T.F. & Waterman, M.S. Identification of common molecular subsequences. *J Mol Biol* **147**, 195-7 (1981).
9. Levenshtein, V.I. Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physics Doklady* **10**, 707-710 (1966).
10. Ukkonen, E. On approximate string matching. in *International FCT-Conference on Fundamentals of Computation Theory* 487-495 (Springer-Verlag, London, UK, 1983).
11. Cozen, G. Simon Fraser University (2007).
12. Higgins, D.G. CLUSTAL V: multiple alignment of DNA and protein sequences. *Methods Mol Biol* **25**, 307-18 (1994).
13. Wozniak, A. Using video-oriented instructions to speed up sequence comparison. *Computer Applications in the Biosciences (CABIOS)* **13**, 145-150 (1997).
14. Liu, Y., Huang, W., Johnson, J. & Vaidya, S. GPU accelerated Smith-Waterman. in *International Conference on Computational Science (ICCS 2006)* Vol. 3994 188-195 (Springer-Verlag, 2006).
15. Farrar, M. Striped Smith-Waterman speeds database searches six times over other SIMD implementations. *Bioinformatics* **23**, 156-61 (2007).
16. Rognes, T. & Seeberg, E. Six-fold speed-up of Smith-Waterman sequence database searches using parallel processing on common microprocessors. *Bioinformatics* **16**, 699-706 (2000).
17. Li, H., Ruan, J. & Durbin, R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res* **18**, 1851-8 (2008).
18. Hillier, L.W. et al. Whole-genome sequencing and variant discovery in C. elegans. *Nat Methods* **5**, 183-8 (2008).
19. Hamming, R.W. Error-detecting and error-correcting codes. *Bell System Technical Journal* **29**, 147-160 (1950).
20. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754-60 (2009).
21. Li, R., Li, Y., Kristiansen, K. & Wang, J. SOAP: short oligonucleotide alignment program. *Bioinformatics* **24**, 713-4 (2008).

Nature Genetics: doi:10.1038/ng.437

22. Smit, A.F.A., Hubley, R. and Green, P. RepeatMasker Open-3.0. (1996-2004).
23. Morgulis, A., Gertz, E.M., Schaffer, A.A. & Agarwala, R. WindowMasker: window-based masker for sequenced genomes. *Bioinformatics* **22**, 134-41 (2006).
24. Smith, D.R. et al. Rapid whole-genome mutational profiling using next-generation sequencing technologies. *Genome Res* **18**, 1638-42 (2008).
25. Bailey, J.A. et al. Recent segmental duplications in the human genome. *Science* **297**, 1003-7 (2002).
26. Bailey, J.A., Yavor, A.M., Massa, H.F., Trask, B.J. & Eichler, E.E. Segmental duplications: organization and impact within the current human genome project assembly. *Genome Res* **11**, 1005-17. (2001).
27. She, X. et al. Shotgun sequence assembly and recent segmental duplications within the human genome. *Nature* **431**, 927-30 (2004).
28. Marques-Bonet, T. et al. A burst of segmental duplications in the genome of the African great ape ancestor. *Nature* **457**, 877-81 (2009).
29. Jiang, Z. et al. Ancestral reconstruction of segmental duplications reveals punctuated cores of human genome evolution. *Nat Genet* **39**, 1361-8 (2007).
30. McCarroll, S.A. et al. Integrated detection and population-genetic analysis of SNPs and copy number variation. *Nat Genet* **40**, 1166-74 (2008).
31. Kidd, J.M. et al. Mapping and sequencing of structural variation from eight human genomes. *Nature* **453**, 56-64 (2008).
32. Cooper, G.M., Zerr, T., Kidd, J.M., Eichler, E.E. & Nickerson, D.A. Systematic assessment of copy number variant detection via genome-wide SNP genotyping. *Nat Genet* **40**, 1199-203 (2008).
33. Locke, D.P. et al. Large-scale variation among human and great ape genomes determined by array comparative genomic hybridization. *Genome Res* **13**, 347-57 (2003).

Nature Genetics: doi:10.1038/ng.437

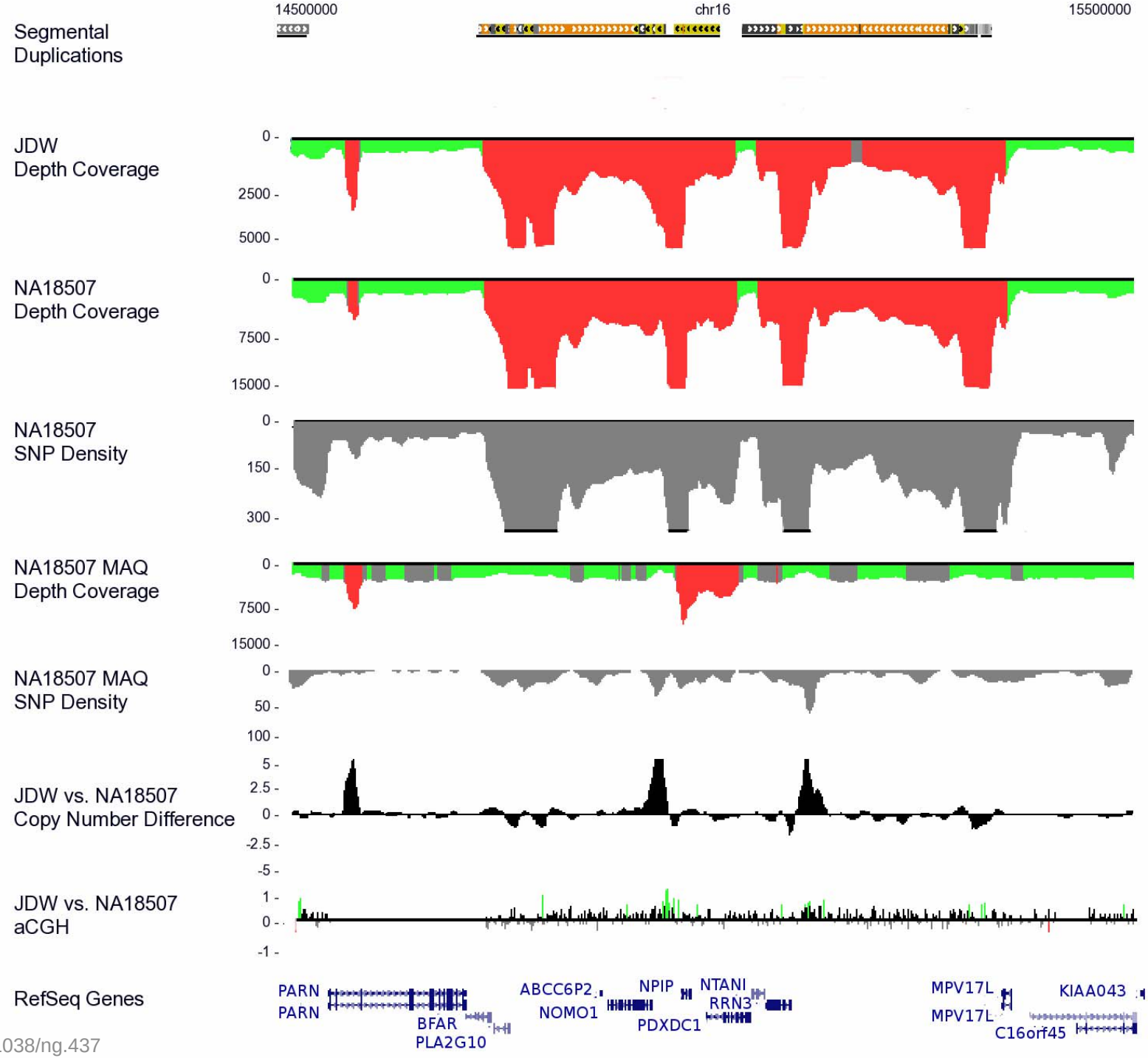
Supplementary Figure 1



Supplementary Figure 1. mrFAST sequence search algorithm. a) The reference genome is first partitioned into contigs to limit the main memory usage to <~1GB. Each contig is separately indexed with collision-free hash tables of size O(4k), where *k* is the ungapped seed length (we set *k*=12 by default). b) When searching for a read, the first *k*-mer of the query is placed using the previously created indices, then the remainder of the read is aligned to the reference genome to extend the initial seed match. This procedure is repeated for the second and the last *k*-mers and the reverse complement of the read. c) Dynamic programming (DP) matrix for Levenshtein distance computation with Ukkonen's improvements. When the maximum allowed edit distance is low, it is sufficient to compute only a narrow band along the main diagonal of the DP matrix. The rest of the matrix entries are set to infinity.

Supplementary Figure 2

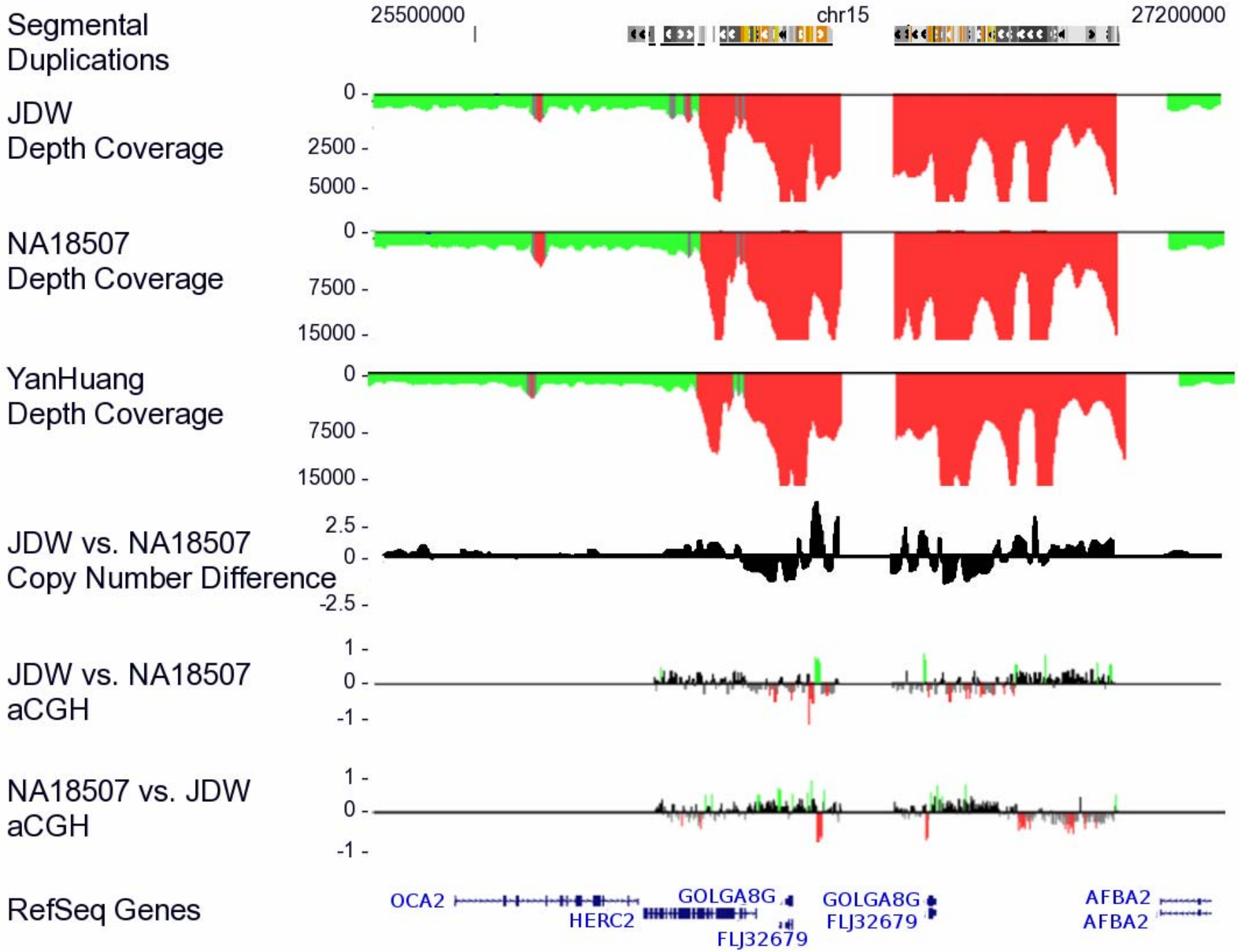
A



Nature Genetics: doi:10.1038/ng.437

Supplementary Figure 2

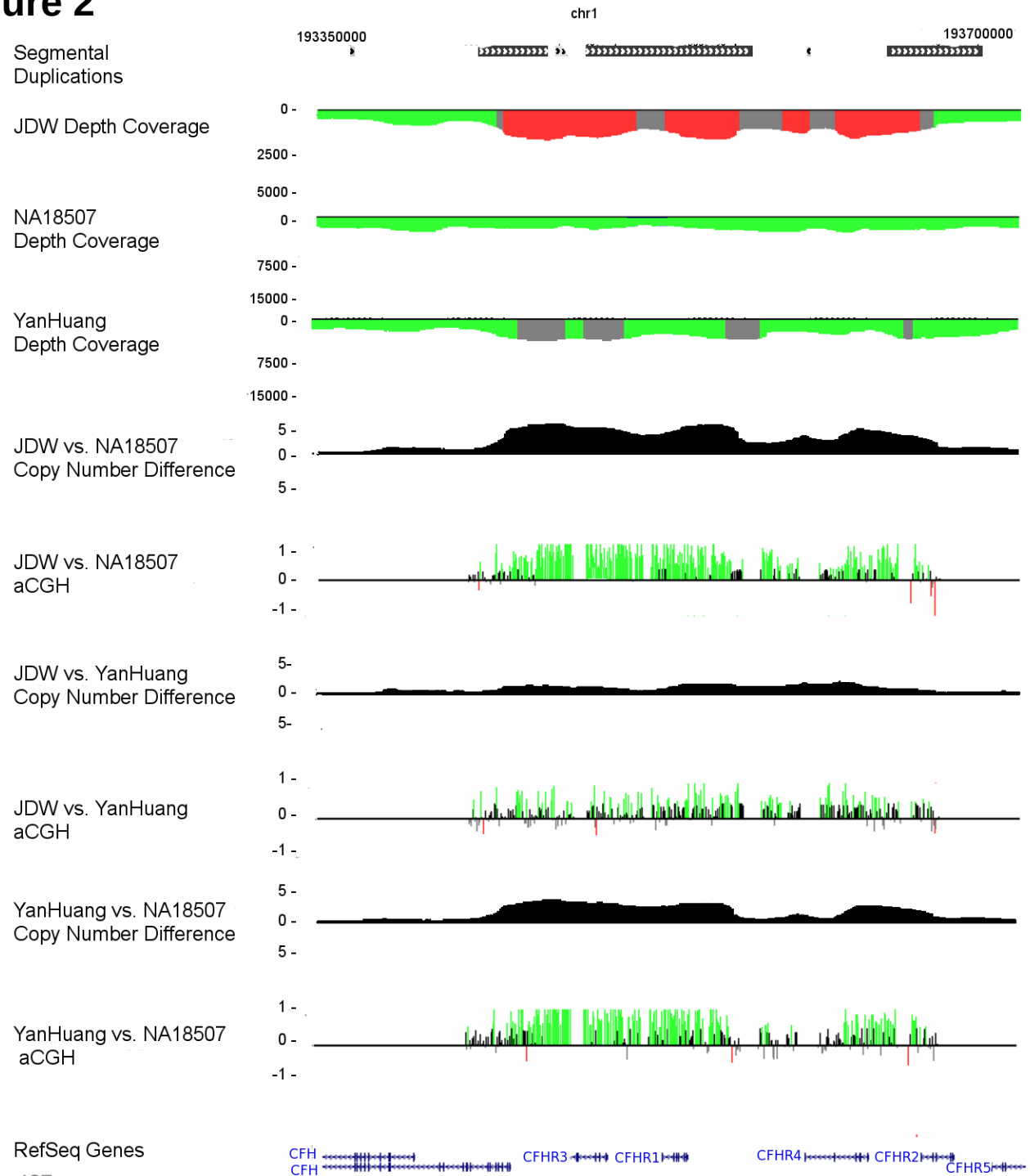
B



Nature Genetics: doi:10.1038/ng.437

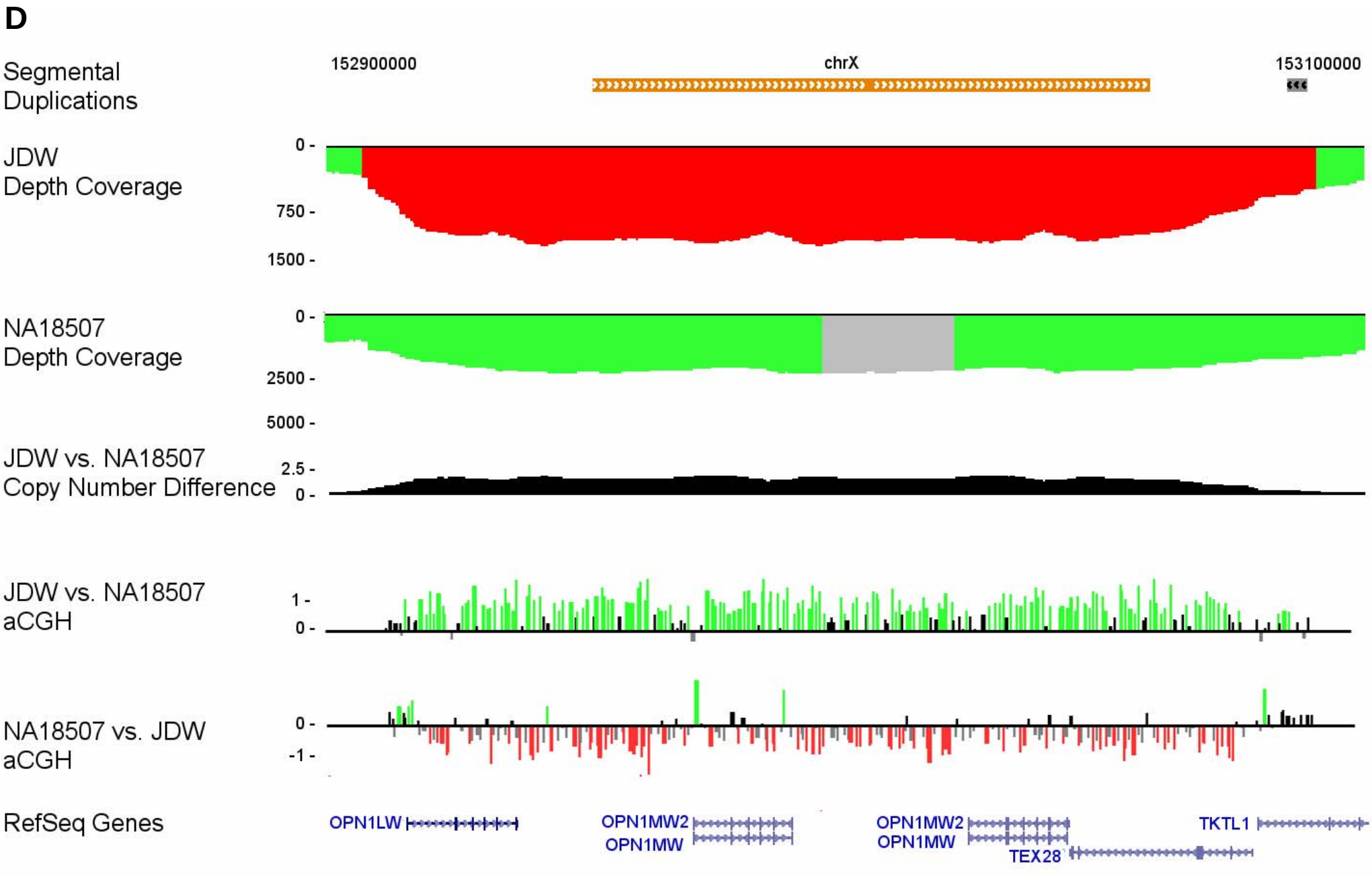
Supplementary Figure 2

C



Nature Genetics: doi:10.1038/ng.437

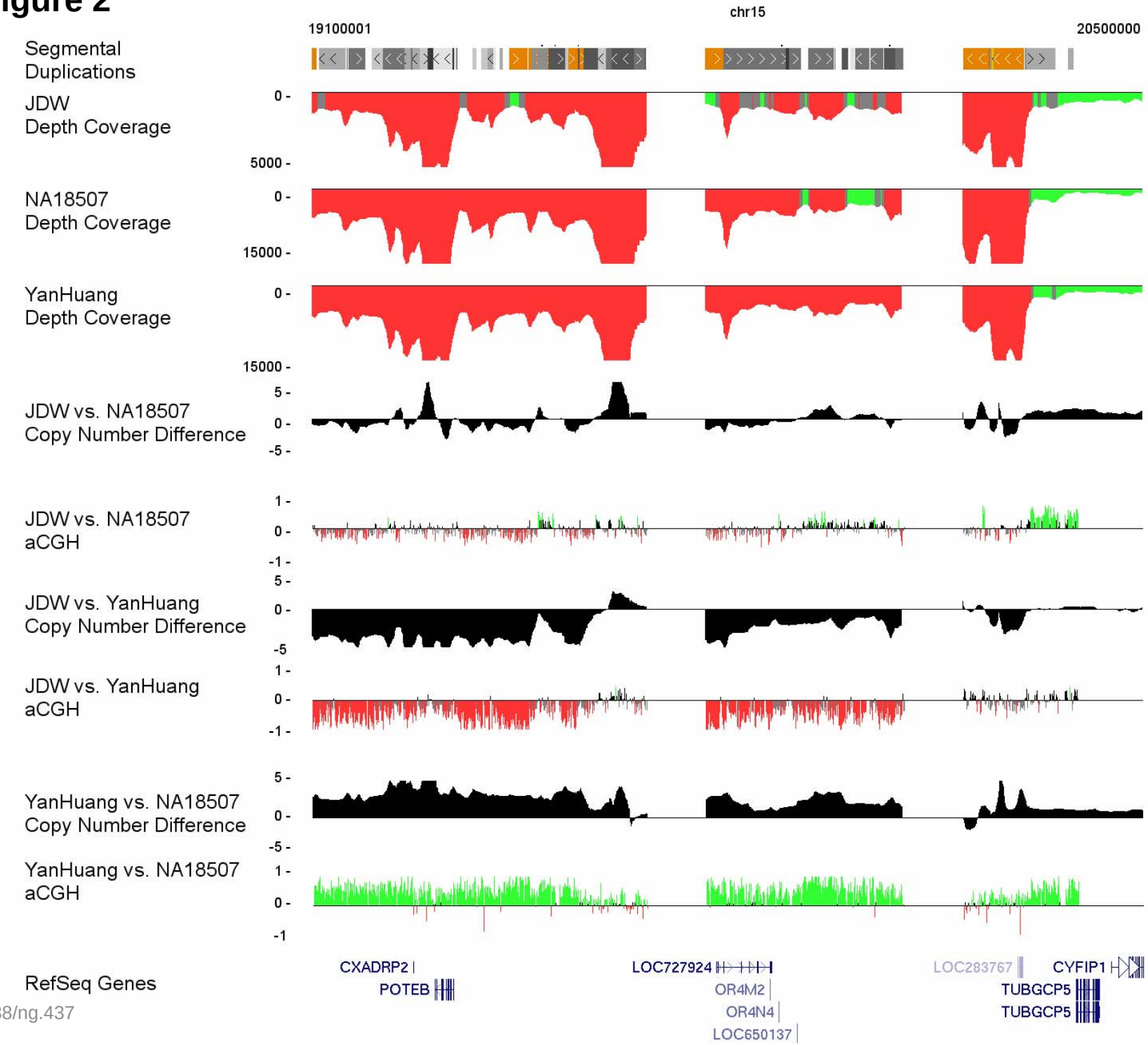
Supplementary Figure 2



Nature Genetics: doi:10.1038/ng.437

Supplementary Figure 2

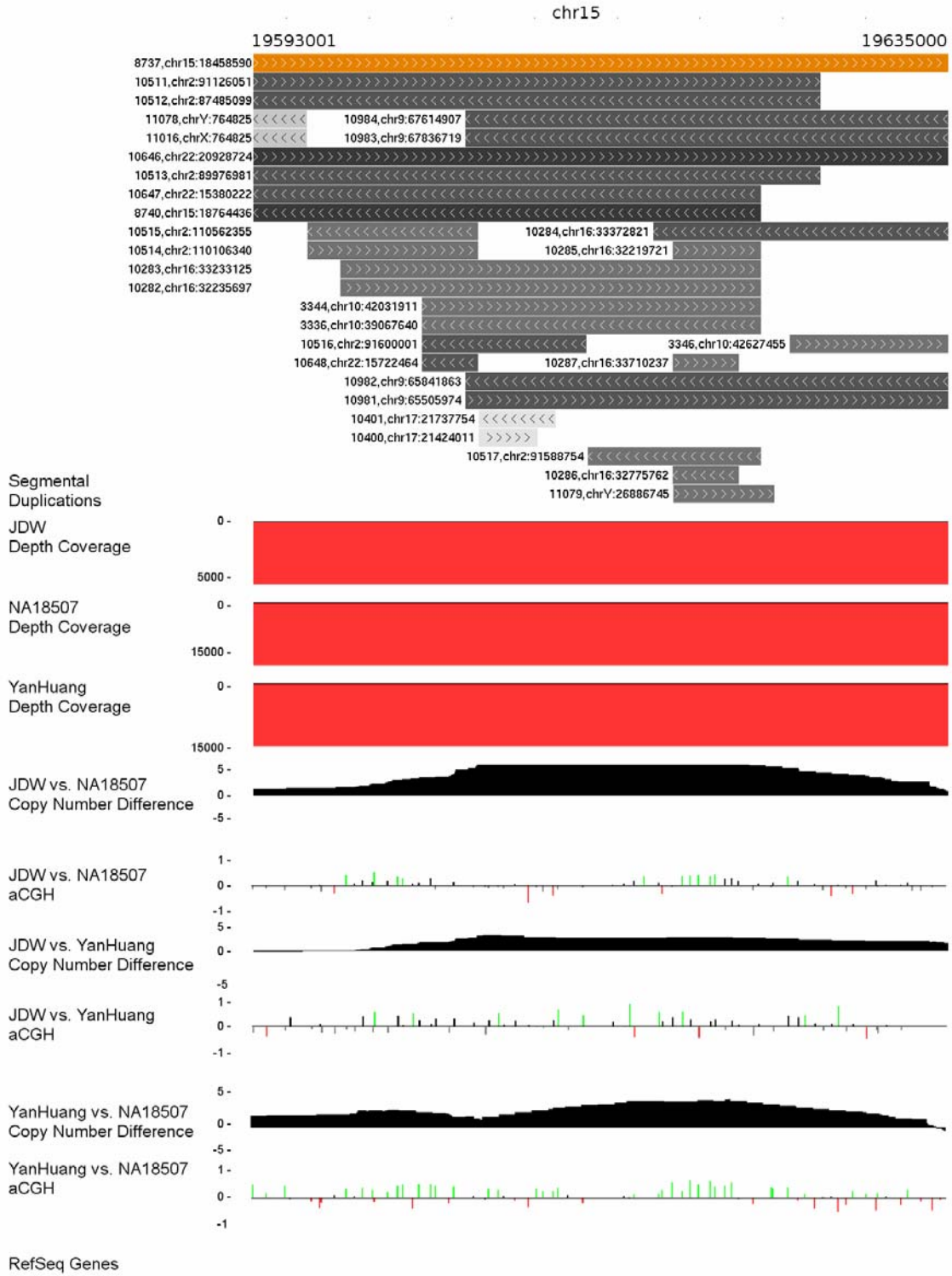
E



Nature Genetics: doi:10.1038/ng.437

Supplementary Figure 2

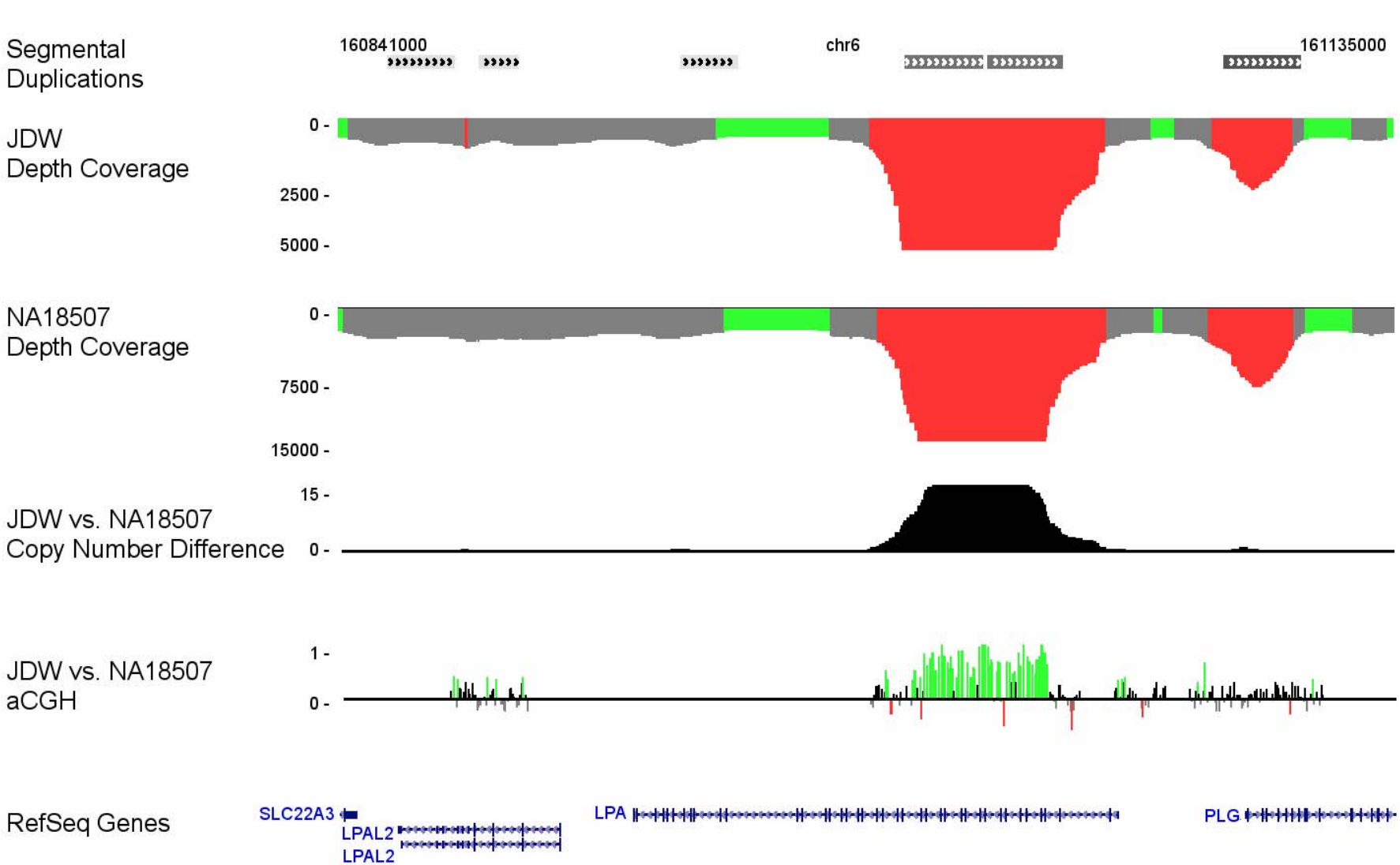
F



Nature Genetics: doi:10.1038/ng.437

Supplementary Figure 2

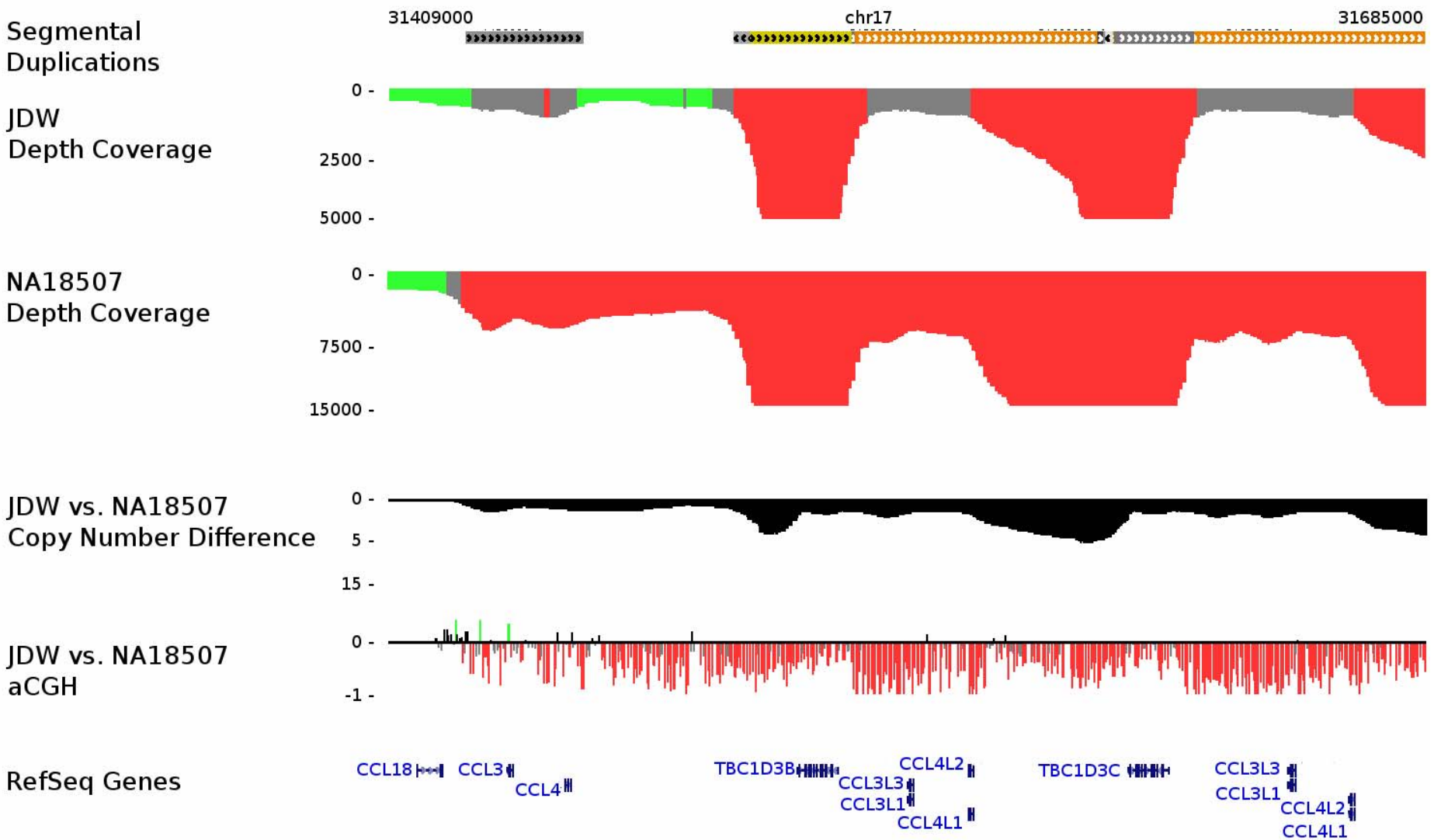
G



Nature Genetics: doi:10.1038/ng.437

Supplementary Figure 2

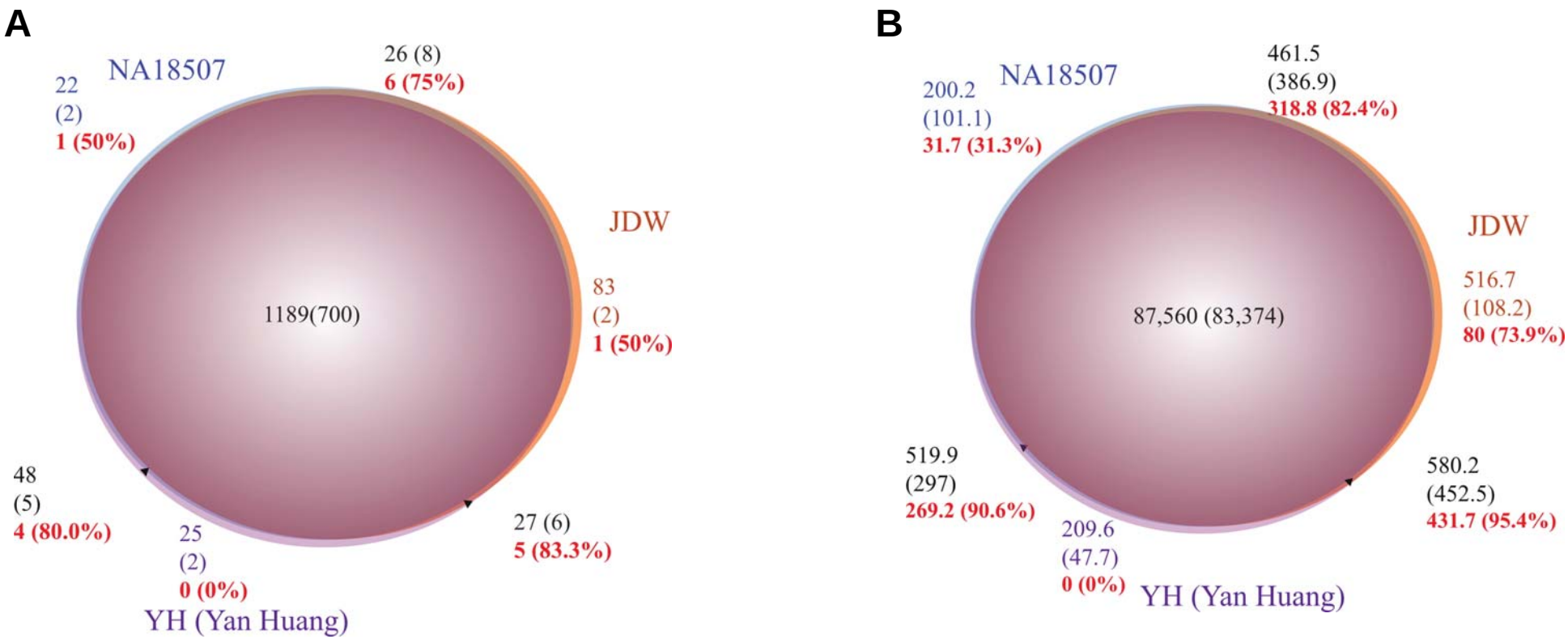
H



Nature Genetics: doi:10.1038/ng.437

Supplementary Figure 2. Visualization of depth-of-coverage and arrayCGH validation in the UCSC Genome Browser. Regions of excess read-depth (average+3std) are shown in red in contrast to regions of intermediate read-depth (gray; average + 2std-3std) or normal read-depth (green, average +/- 2std). The absolute copy number and arrayCGH results for specific individual genome comparisons are shown in the context of RefSeq annotated genes. Significant departures in the relative log2 ratios are depicted as red/green histograms, corresponding to an increase and decrease, respectively, in signal intensity when test/reference are reverse labeled. a) Depth-of-coverage and single nucleotide variant density (SNPs and paralogous sequence variants) in 16p13.11 as detected by *mrFAST* and *MAQ* 1. This region is previously characterized to have more than 24 copies and be copy-number polymorphic 2,3. *mrFAST* mapping clearly shows increased read-depth and single nucleotide variant density that correlates with read-depth: however *MAQ* fails to report accurate read-depth by placing reads to a single location and calling few if any single nucleotide variant in **duplicated** sequence. *MAQ* was run on the human genome build35 with the same repeat masking parameters and same 5-kbp windows used with *mrFAST*. b) Copy-number polymorphic region 15.q13.1 (build35 coordinates: chr15:25,500,000-27,200,000). Predicted copy-number differences between JDW and NA18507 genomes correlates with the log2 ratios detected by arrayCGH experiment. c) Copy-number polymorphism in 1q31.3 in three individuals as predicted by read-depth analysis and confirmed by arrayCGH: JDW has the most number of copies, and NA18507 has the least (JDW=5, YH=3, NA18507=2). d) *Opsin* gene cluster in Xq28, where JDW is predicted and confirmed to have more copies than NA18507. Variation of this locus is associated with color-blindness. e) Pericentromeric locus in chr15:19,100,000-20.500,000. Additional copies are predicted in YH genome, and validated by arrayCGH. JDW/NA18507 show similar sequence properties. f) Closer view of chr15:19,593,001-19,635,000; where WGAC 3 analysis predicts multiple copies. In this more complex region, arrayCGH loses sensitivity to predict copy-number difference between individual genomes. g) *LPA* gene is predicted and validated to be partially duplicated in JDW genome with respect to NA18507 (chr6:160,841,000-161,135,000). Decrease in copy number is associated with increased lipoprotein A serum levels and risk for coronary heart disease. h) Increased copy number of the *CCL3* gene family in the NA18507 genome when compared to JDW (chr17:31,409,000-31,685,000). Individuals of African ancestry are known to carry additional copies of this variant which has been shown to be protective against AIDS progression.

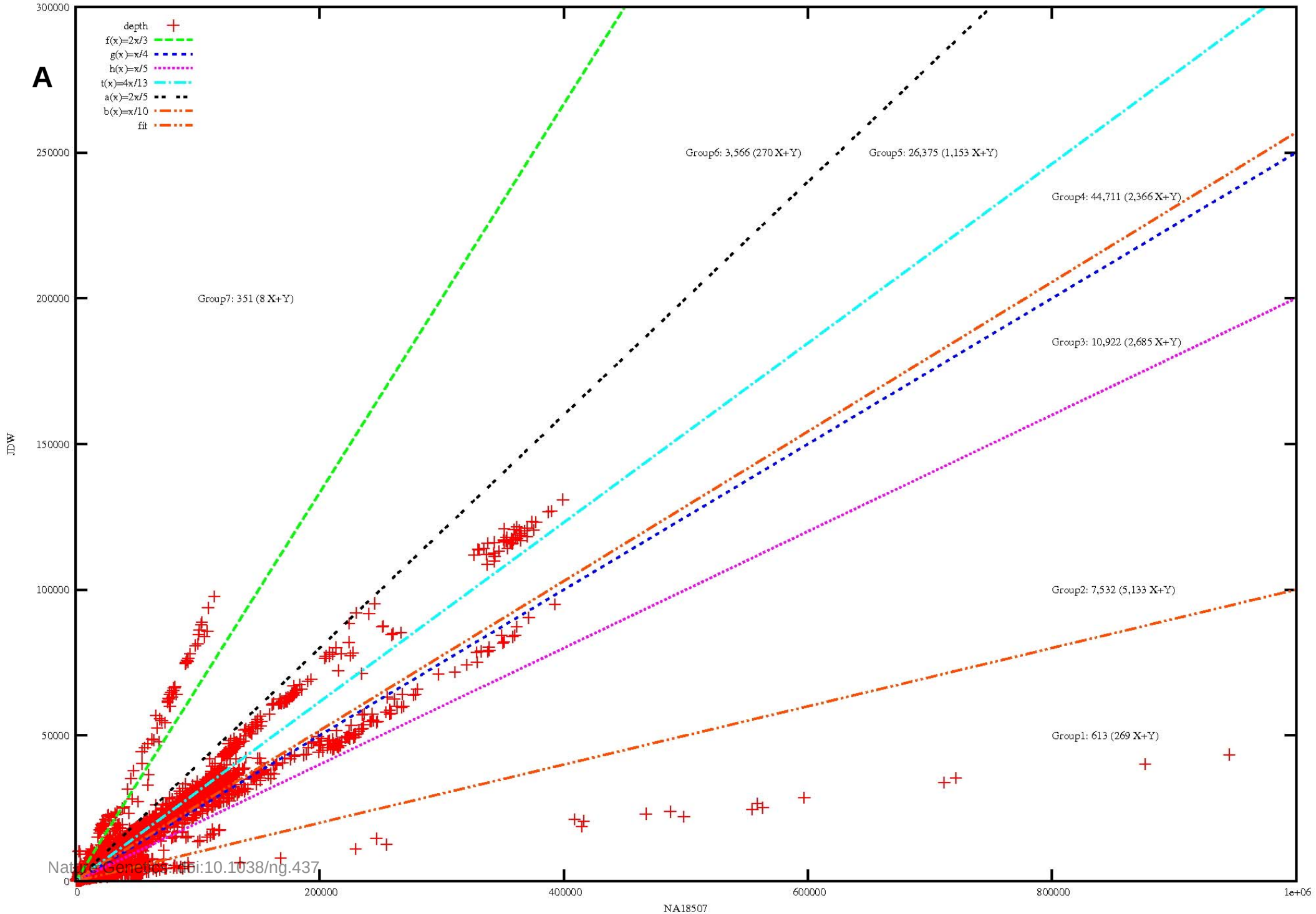
Supplementary Figure 3



Supplementary Figure 3. Venn diagrams of shared and individual specific segmental duplications. The numbers correspond to duplication intervals (a) or number of kilobases (b) that are either shared or predicted to be specific to an individual. Parentheses denote the subset that is greater than 20 kbp in length while numbers in red indicate the fraction that was experimentally validated by arrayCGH.

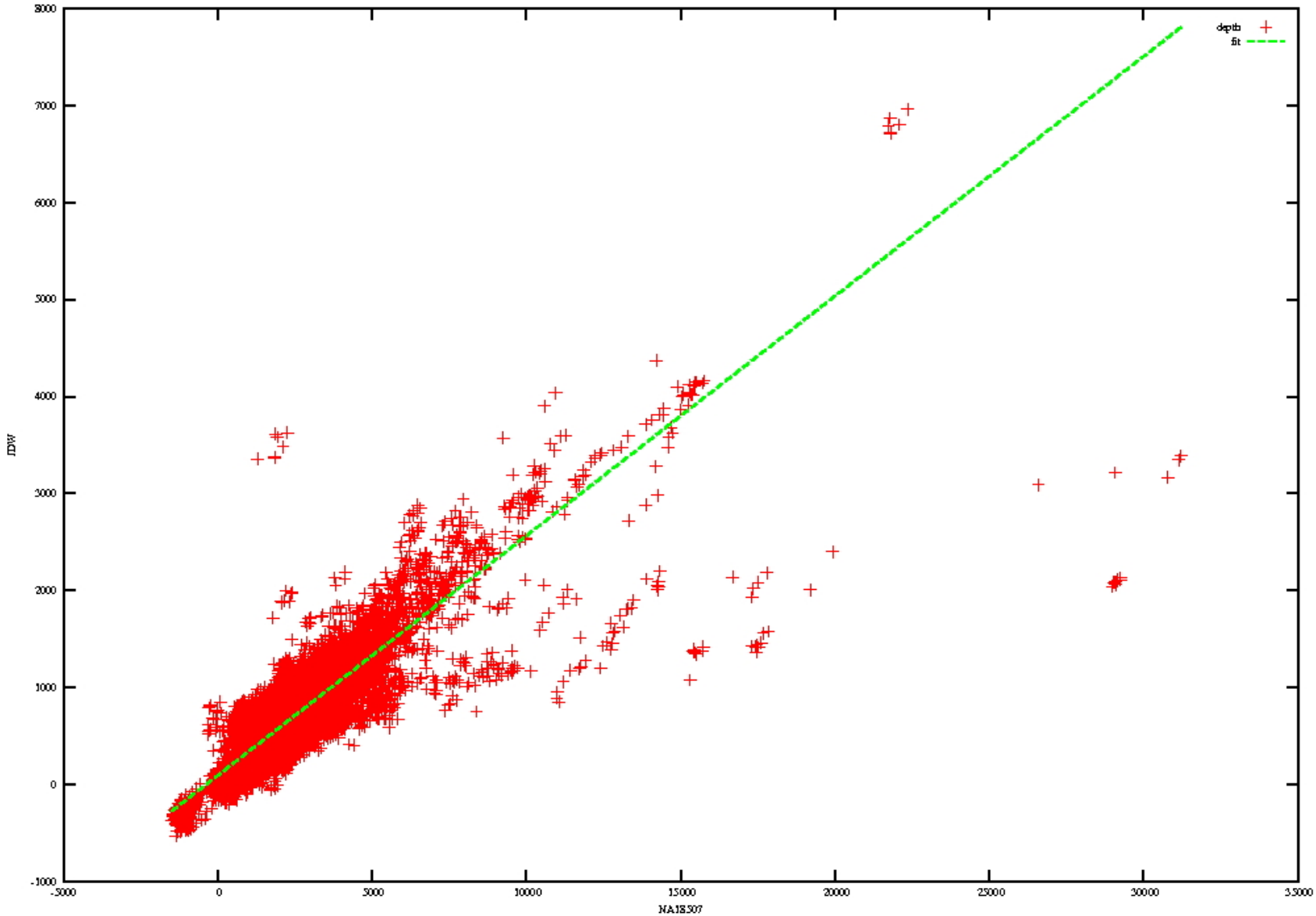
Nature Genetics: doi:10.1038/ng.437

Supplementary Figure 4



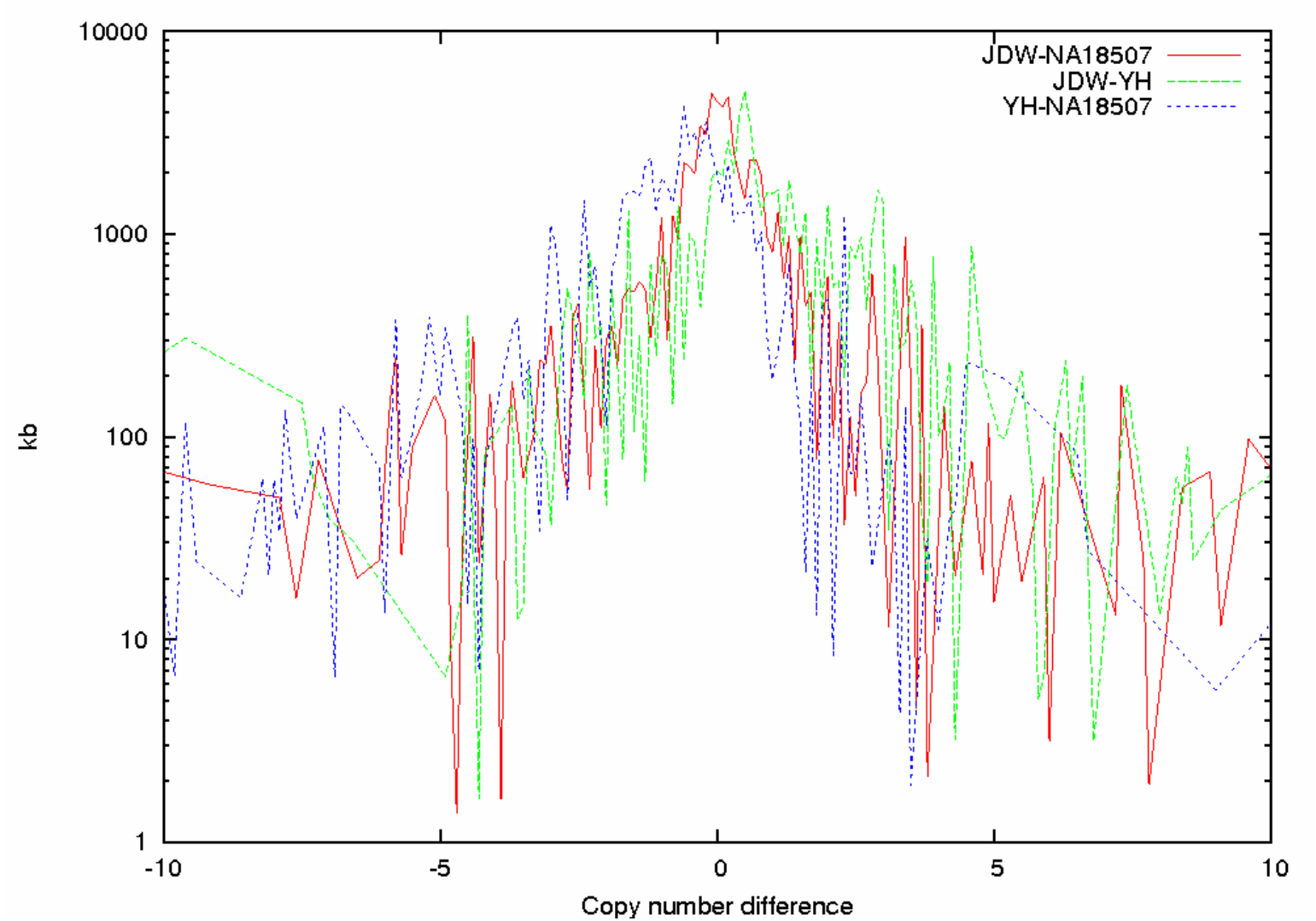
Supplementary Figure 4

B



Supplementary Figure 4. Read-depth correlation for two individuals. Absolute read-depth for an Illumina/Solexa sequenced and 454-sequenced genome is compared for regions annotated as a) segmental duplications or b) unique. Depth-of-coverage was computed over 5-kbp windows for the NA18507 and JDW genomes. Segmental duplications were defined based on the union of WSSD 4 and WGAC 3. Regions with more than 80% common repeat content and more than 30% mapping artifacts (see Supplementary Note) are removed. Note only 48% of the read-depth within duplicated regions correlates, while ~52% of the regions read-depth varies as discrete steps suggesting extensive copy-number variation.

Supplementary Figure 5

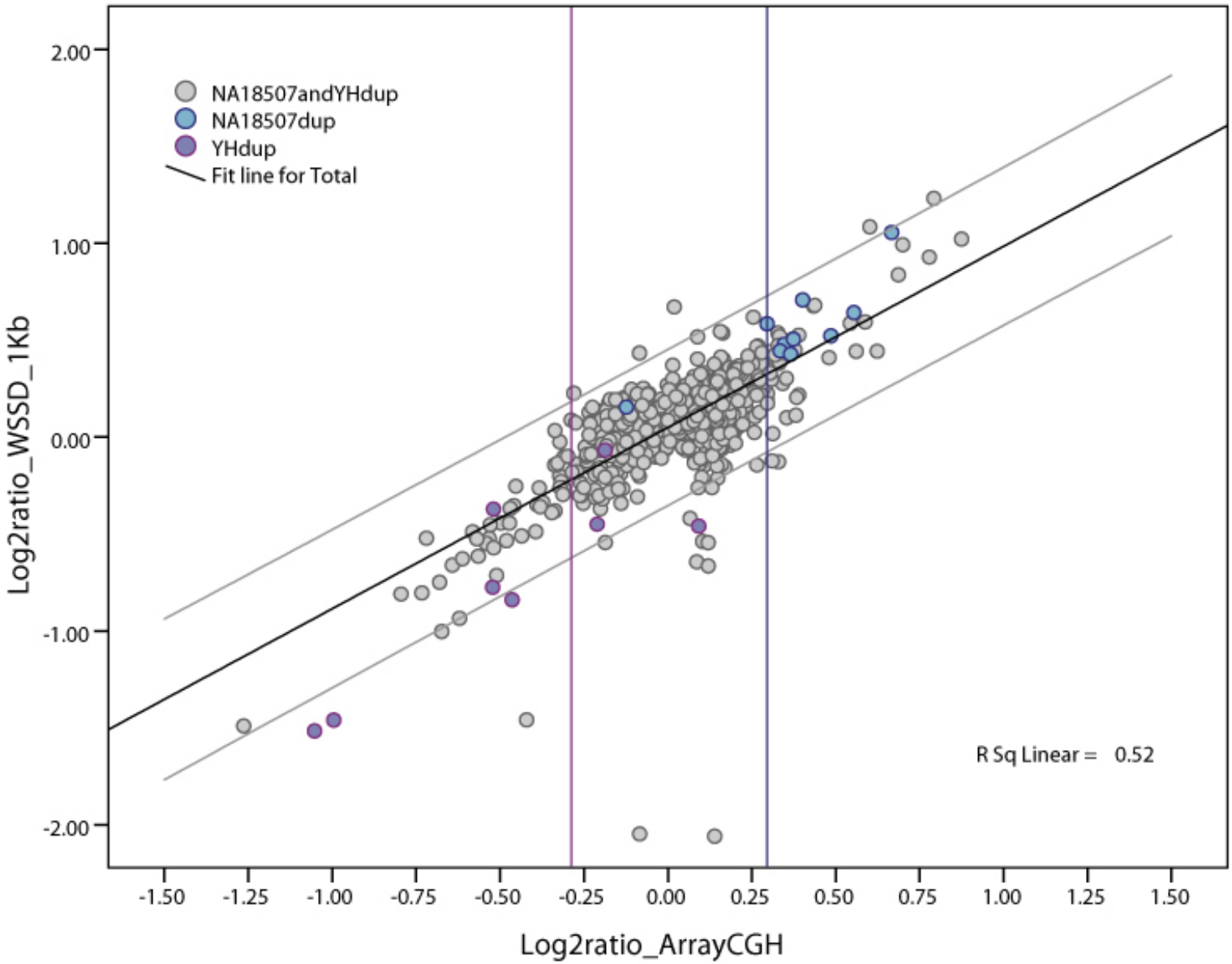


Supplementary Figure 5. Copy-number difference histogram. Predicted copy-number differences among three human individuals within the segmental duplication regions are shown. Absolute copy-number estimates were estimated by mapping to the reference genome.

Nature Genetics: doi:10.1038/ng.437

Supplementary Figure 6

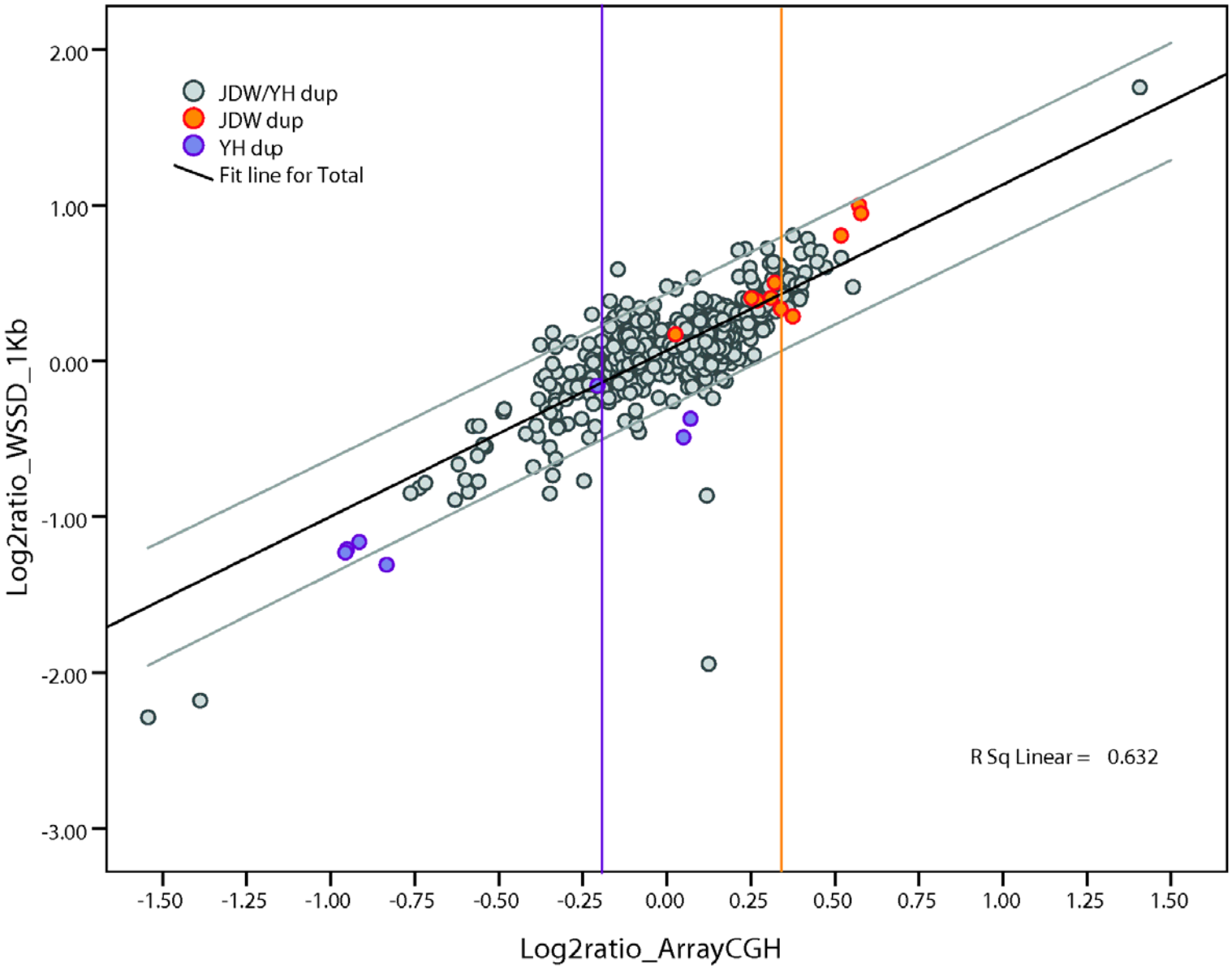
A



Nature Genetics: doi:10.1038/ng.437

Supplementary Figure 6

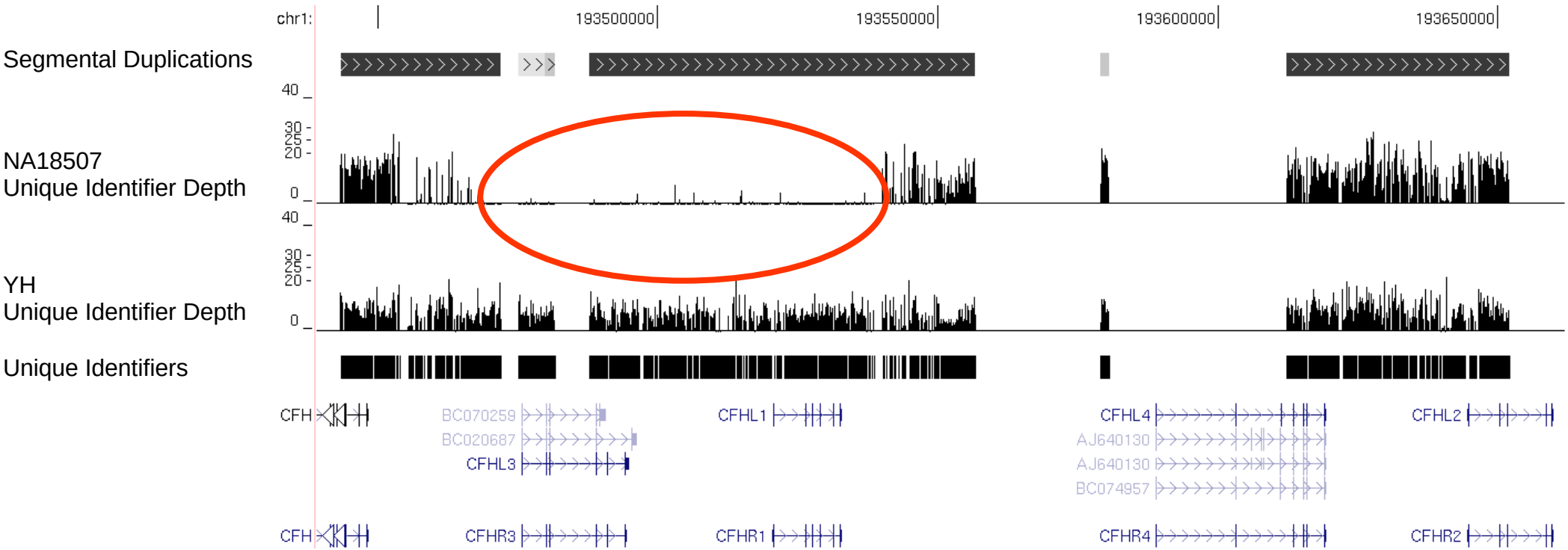
B



Supplementary Figure 6. Copy-number comparison. Correlation between computational and experimental copy number. a) NA18507 vs. JDW and b) JDW vs. YH. We computed the copy number for each shared (gray) and individual specific duplication interval (blue or orange) based on the depth-of-coverage of aligned WGS against the human reference assembly (build35). Based on this computational estimates of copy number, we calculated a predicted log2 copy-number ratio for each autosomal duplication interval >20 kbp in length (and with less than 80% of total common repeat content). These values were plotted against the experimental log2 ratios determined by oligonucleotide array comparative genomic hybridization. The vertical red lines indicating the threshold used for the validated calls (see Supplementary Note).

Supplementary Figure 7

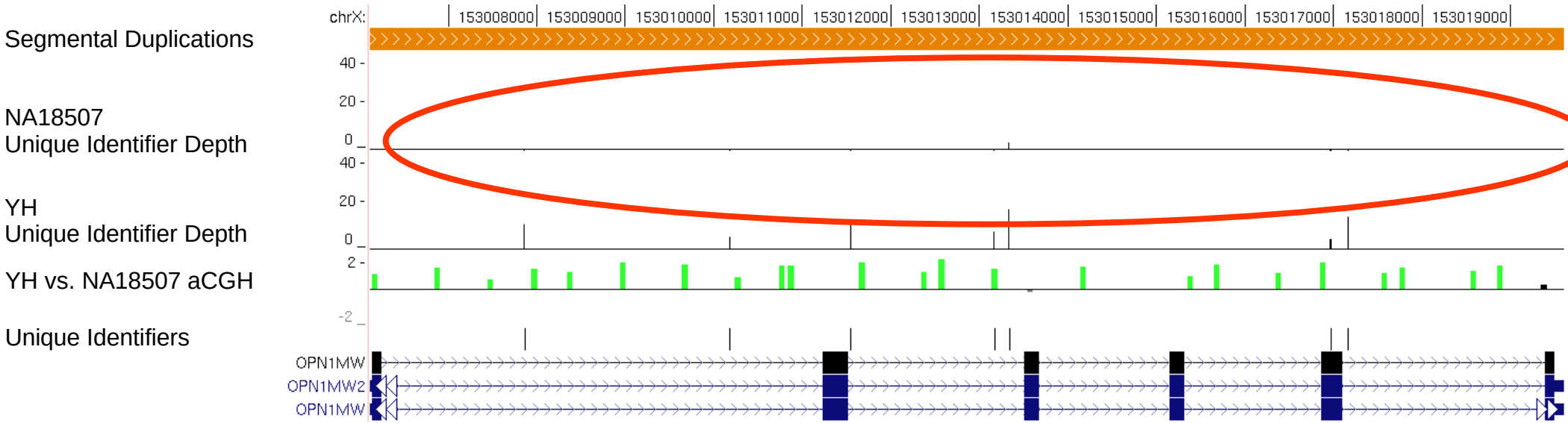
A



Nature Genetics: doi:10.1038/ng.437

Supplementary Figure 7

B



Supplementary Figure 7

C

Supplementary Figure 7. Characterization of duplicated gene paralogs. Unique identifier (UI) loci are shown along with the UI depth in the NA18507 and YH genomes. a) *CFHR3* and *CFHR1* genes are predicted to be deleted in NA18507 and present in YH. b) No UIs were observed in the NA18507 genome or in the *OPN1MW* locus, as confirmed by arrayCGH. c) View of the entire *opsin* locus indicates a deletion of the distal copy of *OPN1MW* in NA18507 and the presence of the proximal copy.

Nature Genetics: doi:10.1038/ng.437

Supplementary Table 1. RefSeq genes assigned to validated segmental duplications and deletions.

Gene name	Gene ID	Complete/ Partial	DUP/ DEL	Chr (B35)	start	end	WGS prediction	JDW CN	NA18507 CN	YH CN	# probes	Log2 ArrayCGH JDW/NA18507	Log2 ArrayCGH JDW/YH	Log2 ArrayCGH NA18507/YH
AMY1A	NM_001008221	Complete	dup	chr1	103,871,301	103,922,522	JDW_NA18507_YH	5.6	9.6	9.8	84	-0.46861	-0.73642	-0.321885
AMY1B	NM_001008218	Complete	dup	chr1	103,924,688	104,019,040	JDW_NA18507_YH	5.1	9.5	9.3	154	-0.47786	-0.76242	-0.336635
AMY2A	NM_000699	Complete	dup	chr1	103,871,301	103,922,522	JDW_NA18507_YH	5.6	9.6	9.8	84	-0.46861	-0.73642	-0.321885
ARHGEF5	NM_005435	Partial	dup	chr7	143,313,333	143,512,027	JDW_NA18507_YH	4.6	3.9	6.2	297	0.27214	-0.38842	-0.642135
ARL17	NM_001103154	Partial	dup	chr17	41,713,832	41,781,181	JDW_NA18507_YH	11.6	8.8	11.3	89	0.30764	-0.16492	-0.458635
ARL17P1	NM_016632	Partial	dup	chr17	41,915,001	41,998,660	JDW_NA18507_YH	11.5	8.4	10.9	118	0.31364	-0.21367	-0.469635
BMPR2	NM_001204	Partial	del	chr2	203,109,001	203,137,590	JDW_NA18507	1.3	1.1	1.7	52	0.21632	0.060065	-0.23952
C2orf78	NM_001080474	Partial	dup	chr2	73,921,368	73,948,837	JDW_NA18507_YH	10.1	8.1	15.6	101	0.29757	-0.330435	-0.62077
C4A	NM_007293	Complete	dup	chr6	32,066,001	32,121,881	JDW_NA18507	3.5	3.8	2.5	66	-0.25186	0.320585	0.553865
C4B	NM_001002029	Partial	dup	chr6	32,066,001	32,121,881	JDW_NA18507	3.5	3.8	2.5	66	-0.25186	0.320585	0.553865
CCDC146	NM_020879	Partial	dup	chr7	75,792,001	76,453,431	JDW_NA18507_YH	3.8	3.9	2.8	688	-0.04136	0.301085	0.353865
CCL3L3	NM_001001437	Complete	dup	chr17	31,505,927	31,698,947	JDW_NA18507_YH	4.8	8.3	6.1	302	-0.59836	-0.33292	0.623365
CCL4L2	NM_207007	Complete	dup	chr17	31,505,927	31,698,947	JDW_NA18507_YH	4.8	8.3	6.1	302	-0.59836	-0.33292	0.623365
CDH12	NM_004061	Partial	dup	chr5	21,971,361	22,051,302	JDW_NA18507_YH	5.7	5.0	3.7	114	0.27889	0.446835	0.279865
CES1	NM_001025195	Partial	dup	chr16	54,379,001	54,421,563	JDW_NA18507_YH	4.1	3.7	4.1	46	0.12814	-0.33992	-0.337635
CFH	NM_000186	Partial	dup	chr1	193,445,001	193,471,023	JDW_YH	5.7	2.0	3.5	42	0.65989	0.234085	-0.521885
CFHR1	NM_002113	Complete	dup	chr1	193,513,754	193,622,002	JDW_YH	4.4	2.3	3.0	133	0.57814	0.338585	-0.519635
CFHR4	NM_006684	Complete	dup	chr1	193,513,754	193,622,002	JDW_YH	4.4	2.3	3.0	133	0.57814	0.338585	-0.519635
CR1	NM_000573	Partial	dup	chr1	204,085,341	204,140,002	JDW_NA18507_YH	5.4	4.1	4.4	70	0.42889	0.301335	-0.309885
CROCC	NM_014675	Partial	dup	chr1	17,000,001	17,026,544	JDW_NA18507_YH	4.2	4.9	5.6	25	-0.22086	-0.21292	0.148365
CYFIP1	NM_014608	Partial	del	chr15	20,357,001	20,464,046	NA18507	2.0	0.9	1.9	321	0.39732	0.094065	-0.43527
CYP21A2	NM_000500	Complete	dup	chr6	32,066,001	32,121,881	JDW_NA18507	3.5	3.8	2.5	66	-0.25186	0.320585	0.553865
DAB2IP	NM_032552	Partial	del	chr9	121,406,001	121,440,540	JDW	0.7	1.6	1.7	141	-0.39693	-0.415935	-0.04302
DEFB103B	NM_001081551	Partial	dup	chr8	7,693,579	7,777,000	NA18507_YH	1.9	2.9	4.5	148	-0.42786	-0.95642	-0.564885
DEFB104A	NM_080389	Complete	dup	chr8	7,693,579	7,777,000	NA18507_YH	1.9	2.9	4.5	148	-0.42786	-0.95642	-0.564885
DEFB105B	NM_001040703	Complete	dup	chr8	7,275,228	7,390,000	NA18507_YH	2.0	3.0	4.6	190	-0.39236	-0.95167	-0.611885
DEFB106B	NM_001040704	Complete	dup	chr8	7,275,228	7,390,000	NA18507_YH	2.0	3.0	4.6	190	-0.39236	-0.95167	-0.611885
DEFB107A	NM_001037668	Complete	dup	chr8	7,693,579	7,777,000	NA18507_YH	1.9	2.9	4.5	148	-0.42786	-0.95642	-0.564885
DEFB107B	NM_001040705	Complete	dup	chr8	7,275,228	7,390,000	NA18507_YH	2.0	3.0	4.6	190	-0.39236	-0.95167	-0.611885
DEFB130	NM_001037804	Complete	dup	chr8	11,898,231	12,097,364	JDW_NA18507_YH	9.8	13.0	13.2	234	-0.22336	-0.32467	-0.173635
DEFB4	NM_004942	Complete	dup	chr8	7,789,060	7,825,000	NA18507_YH	1.8	2.8	4.5	57	-0.38336	-0.83342	-0.510135
DGCR6L	NM_033257	Partial	dup	chr22	18,680,001	18,878,620	JDW_NA18507_YH	3.7	4.3	3.2	242	-0.31061	0.284335	0.560865
DTX2	NM_001102594	Partial	dup	chr7	75,707,830	75,779,002	JDW_NA18507_YH	3.8	3.5	2.9	79	0.08989	0.395585	0.352865
DUB3	NM_201402	Complete	dup	chr8	11,898,231	12,097,364	JDW_NA18507_YH	9.8	13.0	13.2	234	-0.22336	-0.32467	-0.173635
DUX4	NM_033178	Complete	dup	chr10	135,323,001	135,408,252	JDW_NA18507_YH	27.6	16.3	22.8	109	0.36314	0.178085	-0.393635
FAM86B1	NM_001083537	Complete	dup	chr8	11,898,231	12,097,364	JDW_NA18507_YH	9.8	13.0	13.2	234	-0.22336	-0.32467	-0.173635
FBXO25	NM_012173	Partial	dup	chr8	329,001	389,952	JDW_NA18507_YH	3.3	4.4	4.0	57	-0.20136	-0.21867	-0.110135
FCGBP	NM_003890	Partial	dup	chr19	45,065,108	45,096,549	JDW_NA18507_YH	5.0	4.7	5.1	29	-0.23886	-0.12592	0.310365
FLJ43692	NM_001003702	Complete	dup	chr7	143,313,333	143,512,027	JDW_NA18507_YH	4.6	3.9	6.2	297	0.27214	-0.38842	-0.642135
FRG2	NM_001005217	Complete	dup	chr4	191,300,364	191,397,704	JDW_NA18507_YH	29.6	19.4	24.5	110	0.39064	0.1305825	-0.372885
FRG2B	NM_001080998	Complete	dup	chr10	135,323,001	135,408,252	JDW_NA18507_YH	27.6	16.3	22.8	109	0.36314	0.178085	-0.393635
GTF2H2B	NM_001098729	Complete	dup	chr5	69,440,001	70,231,831	JDW_NA18507_YH	11.0	9.5	7.8	921	0.27364	0.383585	0.189615

HIC2	NM_015094	Partial	dup	chr22	19,790,229	20,120,277	JDW_NA18507_YH	11.1	10.2	6.8	414	-0.16936	0.457335	0.542865
HLA	NM_002125	Partial	del	chr6	32,558,001	32,598,262	NA18507	1.8	1.3	2.0	113	-0.05543	-0.471935	-0.57627
KIAA1267	NM_015443	Partial	dup	chr17	41,569,001	41,649,000	JDW	3.8	2.0	1.9	228	0.53882	0.571815	0.14473
LCE3B	NM_178433	Complete	del	chr1	149,366,001	149,403,347	JDW_NA18507_YH	0.0	1.2	0.2	120	-0.72843	-0.091435	0.88873
LCE3D	NM_032563	Partial	del	chr1	149,366,001	149,403,347	JDW_NA18507_YH	0.0	1.2	0.2	120	-0.72843	-0.091435	0.88873
LGALS9B	NM_001042685	Complete	dup	chr17	20,162,001	20,404,817	JDW_NA18507_YH	6.4	6.9	5.9	243	-0.22486	-0.12142	0.277865
LOC650137	NM_001080841	Complete	dup	chr15	19,764,002	19,924,652	JDW_NA18507_YH	3.6	4.5	6.2	222	-0.20261	-0.56042	-0.529385
LOC728340	NM_001098728	Complete	dup	chr5	68,887,001	69,356,689	JDW_NA18507_YH	10.5	9.0	7.8	555	0.25764	0.399835	0.172365
LOC728358	NM_001042500	Complete	dup	chr8	6,816,001	6,853,605	JDW_NA18507_YH	8.2	7.1	5.8	46	0.27364	0.401335	0.205365
LOC729355	NM_001099687	Complete	dup	chr16	32,561,838	33,432,843	JDW_NA18507_YH	14.0	10.7	9.5	987	0.29864	0.411585	0.173115
LPA	NM_005577	Partial	dup	chr6	160,996,001	161,039,550	JDW_NA18507_YH	35.4	22.0	26.4	64	0.69564	0.346085	-0.383385
LRRC37A	NM_014834	Complete	dup	chr17	41,713,832	41,781,181	JDW_NA18507_YH	11.6	8.8	11.3	89	0.30764	-0.16492	-0.458635
LRRC37A2	NM_001006607	Complete	dup	chr17	41,915,001	41,998,660	JDW_NA18507_YH	11.5	8.4	10.9	118	0.31364	-0.21367	-0.469635
LRRC37A3	NM_199340	Complete	dup	chr17	60,211,001	60,346,727	JDW_NA18507_YH	6.4	5.4	6.5	143	0.23489	-0.17017	-0.321635
MGC119295	NM_001031618	Complete	dup	chr7	101,708,106	102,012,000	JDW_NA18507_YH	5.9	6.5	7.9	287	-0.21936	-0.32892	-0.147385
MNS1	NM_018365	Complete	del	chr15	54,447,001	54,683,677	NA18507	1.9	0.8	1.8	666	0.41157	-0.042935	-0.48377
MRPL45	NM_032351	Partial	dup	chr17	33,530,002	33,716,058	JDW_NA18507_YH	8.5	13.8	5.9	229	-0.40436	0.314085	0.792115
NAIP	NM_022892	Partial	dup	chr5	70,231,832	70,316,000	JDW_NA18507	3.8	3.0	2.1	106	0.20414	0.518085	0.373865
NEB	NM_004543	Partial	dup	chr2	152,259,862	152,291,645	JDW_NA18507_YH	3.8	5.1	4.9	34	-0.24536	-0.25367	0.073615
NIPA1	NM_144599	Partial	del	chr15	20,546,001	20,636,690	NA18507	1.9	0.8	1.9	298	0.45307	0.109565	-0.46302
NIPA2	NM_001008860	Complete	del	chr15	20,546,001	20,636,690	NA18507	1.9	0.8	1.9	298	0.45307	0.109565	-0.46302
NPEPPS	NM_006310	Partial	dup	chr17	42,907,079	43,026,438	JDW_NA18507_YH	4.4	7.1	3.6	129	-0.37836	0.325085	0.699365
NSF	NM_006178	Partial	dup	chr17	41,998,661	42,133,889	JDW_YH	2.8	1.8	5.0	226	0.35814	-0.59017	-0.995135
OR2A42	NM_001001802	Complete	dup	chr7	143,313,333	143,512,027	JDW_NA18507_YH	4.6	3.9	6.2	297	0.27214	-0.38842	-0.642135
OR2A7	NM_001005328	Complete	dup	chr7	143,313,333	143,512,027	JDW_NA18507_YH	4.6	3.9	6.2	297	0.27214	-0.38842	-0.642135
OR2T10	NM_001004693	Complete	del	chr1	245,072,001	245,121,144	JDW_YH	1.0	1.8	0.9	146	-0.54793	-0.065435	0.56923
OR4F16	NM_001005277	Complete	dup	chr1	358,002	510,787	JDW_NA18507_YH	18.2	20.1	14.8	189	-0.42311	-0.22267	0.303615
OR4F17	NM_001005240	Complete	dup	chr19	11,002	210,400	JDW_NA18507_YH	21.7	18.4	15.0	205	0.21289	0.363585	0.193865
OR4F21	NM_001005504	Complete	dup	chr8	2	160,543	JDW_NA18507_YH	14.0	19.1	14.1	194	-0.45436	-0.21842	0.345365
OR4F29	NM_001005221	Complete	dup	chr1	562,002	796,646	JDW_NA18507_YH	18.3	20.3	15.1	253	-0.32236	-0.15567	0.264365
OR4F4	NM_001004195	Complete	dup	chr15	100,218,756	100,336,307	JDW_NA18507_YH	17.6	12.7	10.2	137	0.40164	0.419085	0.172615
OR4M1	NM_001005500	Complete	dup	chr14	19,093,852	19,393,002	JDW_NA18507_YH	8.6	9.0	10.1	350	-0.20386	-0.38192	-0.270885
OR4M2	NM_001004719	Complete	dup	chr15	19,764,002	19,924,652	JDW_NA18507_YH	3.6	4.5	6.2	222	-0.20261	-0.56042	-0.529385
OR4N2	NM_001004723	Complete	dup	chr14	19,093,852	19,393,002	JDW_NA18507_YH	8.6	9.0	10.1	350	-0.20386	-0.38192	-0.270885
OR4N4	NM_001005241	Complete	dup	chr15	19,764,002	19,924,652	JDW_NA18507_YH	3.6	4.5	6.2	222	-0.20261	-0.56042	-0.529385
OR4Q3	NM_172194	Complete	dup	chr14	19,093,852	19,393,002	JDW_NA18507_YH	8.6	9.0	10.1	350	-0.20386	-0.38192	-0.270885
OR52N1	NM_001001913	Partial	del	chr11	5,740,001	5,765,881	JDW_NA18507	0.9	1.0	2.0	71	-0.07168	-0.407435	-0.52877
OR52N5	NM_001001922	Complete	del	chr11	5,740,001	5,765,881	JDW_NA18507	0.9	1.0	2.0	71	-0.07168	-0.407435	-0.52877
PDXDC1	NM_015027	Partial	dup	chr16	14,966,820	15,035,106	JDW_NA18507_YH	5.1	5.3	4.0	72	0.03439	0.320835	0.314865
PI4KA	NM_002650	Partial	dup	chr22	19,357,191	19,397,302	JDW_NA18507_YH	3.5	4.2	3.5	35	-0.21036	0.169585	0.341615
POLR2J	NM_006234	Partial	dup	chr7	101,708,106	102,012,000	JDW_NA18507_YH	5.9	6.5	7.9	287	-0.21936	-0.32892	-0.147385
POLR2J3	NM_001097615	Complete	dup	chr7	101,708,106	102,012,000	JDW_NA18507_YH	5.9	6.5	7.9	287	-0.21936	-0.32892	-0.147385
POMZP3	NM_012230	Complete	dup	chr7	75,792,001	76,453,431	JDW_NA18507_YH	3.8	3.9	2.8	688	-0.04136	0.301085	0.353865
POTEB	NM_207355	Complete	dup	chr15	19,220,342	19,344,722	JDW_NA18507_YH	12.4	13.1	15.6	176	-0.12211	-0.48642	-0.452885
PPP1R12B	NM_002481	Partial	dup	chr1	199,164,001	199,260,536	JDW_NA18507_YH	2.7	3.9	4.4	112	-0.23836	-0.39742	-0.203635
PRAMEF1	NM_023013	Complete	dup	chr1	12,774,525	12,799,637	JDW_NA18507_YH	7.9	7.6	8.3	21	0.06339	-0.32842	-0.274385
RASA4	NM_001079877	Complete	dup	chr7	101,708,106	102,012,000	JDW_NA18507_YH	5.9	6.5	7.9	287	-0.21936	-0.32892	-0.147385

REXO1L1	NM_172239	Partial	dup	chr8	86,739,822	86,761,576	JDW_NA18507_YH	176.6	133.8	136.8	12	0.21264	0.295585	0.201115
RIMBP3	NM_015672	Complete	dup	chr22	18,680,001	18,878,620	JDW_NA18507_YH	3.7	4.3	3.2	242	-0.31061	0.284335	0.560865
RNFT1	NM_016125	Partial	dup	chr17	55,354,088	55,388,002	JDW_NA18507_YH	4.9	4.9	4.1	76	-0.09143	0.149315	0.16323
SERF1B	NM_022978	Complete	dup	chr5	70,231,832	70,316,000	JDW_NA18507	3.8	3.0	2.1	106	0.20414	0.518085	0.373865
SMN1	NM_000344	Complete	dup	chr5	70,231,832	70,316,000	JDW_NA18507	3.8	3.0	2.1	106	0.20414	0.518085	0.373865
SMN2	NM_022875	Complete	dup	chr5	69,356,690	69,440,000	JDW_NA18507	3.7	3.1	1.9	103	0.24839	0.577835	0.401365
SPAG11A	NM_001081552	Complete	dup	chr8	7,693,579	7,777,000	NA18507_YH	1.9	2.9	4.5	148	-0.42786	-0.95642	-0.564885
SPAG11B	NM_058200	Complete	dup	chr8	7,275,228	7,390,000	NA18507_YH	2.0	3.0	4.6	190	-0.39236	-0.95167	-0.611885
SPAG11B	NM_058206	Complete	dup	chr8	7,275,228	7,390,000	NA18507_YH	2.0	3.0	4.6	190	-0.39236	-0.95167	-0.611885
TBC1D3B	NM_001001417	Complete	dup	chr17	31,505,927	31,698,947	JDW_NA18507_YH	4.8	8.3	6.1	302	-0.59836	-0.33292	0.623365
TBC1D3C	NM_001001418	Complete	dup	chr17	31,800,002	31,889,187	JDW_NA18507_YH	23.1	28.7	15.1	117	-0.42111	0.336085	0.778615
TBC1D3E	NM_001123392	Complete	dup	chr17	33,323,961	33,428,689	JDW_NA18507_YH	20.9	26.3	12.9	146	-0.44986	0.399585	0.874365
TEX9	NM_198524	Partial	del	chr15	54,447,001	54,683,677	NA18507	1.9	0.8	1.8	666	0.41157	-0.042935	-0.48377
TNXB	NM_032470	Partial	dup	chr6	32,066,001	32,121,881	JDW_NA18507	3.5	3.8	2.5	66	-0.25186	0.320585	0.553865
TPPP	NM_007030	Partial	dup	chr5	741,001	868,717	JDW_NA18507_YH	3.3	4.6	5.6	181	-0.41836	-0.59992	-0.312885
TUBGCP5	NM_001102610	Complete	del	chr15	20,357,001	20,464,046	NA18507	2.0	0.9	1.9	321	0.39732	0.094065	-0.43527
UGT2B11	NM_001073	Complete	dup	chr4	70,210,001	70,362,484	JDW_NA18507_YH	5.6	7.1	5.4	148	-0.22586	-0.02292	0.280865
UGT2B28	NM_053039	Complete	dup	chr4	70,210,001	70,362,484	JDW_NA18507_YH	5.6	7.1	5.4	148	-0.22586	-0.02292	0.280865
UPK3B	NM_030570	Partial	dup	chr7	75,792,001	76,453,431	JDW_NA18507_YH	3.8	3.9	2.8	688	-0.04136	0.301085	0.353865
UPLP	NM_001114403	Complete	dup	chr7	101,708,106	102,012,000	JDW_NA18507_YH	5.9	6.5	7.9	287	-0.21936	-0.32892	-0.147385
USP18	NM_017414	Partial	dup	chr22	17,018,001	17,265,089	JDW_NA18507_YH	13.5	14.7	9.7	312	-0.06561	0.553585	0.587615
ZDHHC11	NM_024786	Partial	dup	chr5	741,001	868,717	JDW_NA18507_YH	3.3	4.6	5.6	181	-0.41836	-0.59992	-0.312885
ZNF705D	NM_001039615	Complete	dup	chr8	11,898,231	12,097,364	JDW_NA18507_YH	9.8	13.0	13.2	234	-0.22336	-0.32467	-0.173635
ZP3	NM_001110354	Partial	dup	chr7	75,707,830	75,779,002	JDW_NA18507_YH	3.8	3.5	2.9	79	0.08989	0.395585	0.352865

RefSeq genes assigned to ArrayCGH validated human segmental duplications (or deletions) based on human genome annotation (build35). Intervals were classified by type (duplication or deletion and in which individual) and cross-referenced within Refseq transcript assignments. Complete genes were classified as only those genes where the full transcript mapped within the segmental duplication interval. The estimated copy number for the 3 individuals based on depth of coverage is also reported as well as the Intra-species arrayCGH results. The unique identifiers correspond to members of gene-family collapsed genes.

Supplementary Table 2. Gene family member census

chrom	start	stop	gene	gene family	UI expected	NA18507 UI observed	YH UI observed	NA18507 diploid copy	YH diploid copy
chr1	193475586	193494529	CFHR3	CFHR	3619	142 (4%)	3397 (93.86%)	0	2
chr1	193520517	193532973	CFHR1	CFHR	3022	78 (2.5%)	2765 (91.49%)	0	2
chr1	193588868	193619419	CFHR4	CFHR	1736	1640 (94.4%)	1603 (92.24%)	2	2
chr1	193644590	193660013	CFHR2	CFHR	1796	1736 (96.7%)	1702 (94.77%)	2	2
chr16	11926089	11945037		Morpheus	626	511 (81.6%)	482 (76.99%)	2	2
chr16	14711393	14730015	AF132984	Morpheus	46	25 (54.3%)	10 (21.74%)	2	1
chr16	14750515	14769609	AF132984	Morpheus	27	27 (100%)	22 (81.48%)	2	2
chr16	14917279	14956148	NPIP	Morpheus	455	386 (84.8%)	349 (76.70%)	2	2
chr16	15102862	15119044	BC053946	Morpheus	248	225 (90.7%)	226 (91.13%)	2	2
chr16	15118884	15136222	BC053946	Morpheus	383	276 (72.06%)	248 (64.75%)	1	1
chr16	15362179	15381286		Morpheus	406	344 (84.72%)	324 (79.8%)	2	2
chr16	16216473	16354758	AF132984	Morpheus	478	404 (84.5%)	175 (36.6%)	1	1
chr16	16355958	16398119	AF132984	Morpheus	252	176 (69.84%)	138 (54.76%)	1	1
chr16	18316491	18355459	AF132984	Morpheus	68	24 (35.3%)	49 (72%)	0	1
chr16	18356640	18498367	AF132984	Morpheus	576	376 (65.28%)	421 (73%)	2	2
chr16	21323738	21340255	LOC23117	Morpheus	32	18 (56.25%)	17 (53.12%)	0	0
chr16	21756051	21772533	BC036263	Morpheus	113	85 (75.22%)	59 (52.2%)	3	1
chr16	22436298	22452964	LOC100132247	Morpheus	13	5 (38.46%)	9 (69.23%)	1	4
chr16	28258506	28277984		Morpheus	156	131 (83.97%)	59 (37.82%)	2	0
chr16	28372312	28390714		Morpheus	173	118 (68.2%)	114 (65.89%)	2	2
chr16	28560874	28580375		Morpheus	172	86 (50%)	78 (45.34%)	1	1
chr16	28674964	28694500	AF034373	Morpheus	113	75 (66.38%)	78 (69.02%)	1	1
chr16	28954934	28974417		Morpheus	201	110 (54.73%)	94 (46.77%)	1	1
chr16	29302710	29318970	AK023827	Morpheus	299	247 (82.61%)	244 (81.6%)	2	2

chr16	29404278	29420760	LOC440353	Morpheus	50	22 (44%)	14 (28%)	1	2
chr16	30144110	30160729	LOC613037	Morpheus	31	25 (80.65%)	22 (70.97%)	2	3
chr16	68564870	68583940	PDXCD2	Morpheus	809	598 (73.92%)	526 (65%)	2	2
chr16	72966815	72986348		Morpheus	452	396 (87.61%)	379 (83.85%)	3	3
chr18	11606449	11625790		Morpheus	748	625 (83.56%)	549 (73.4%)	1	2
chrX	152930571	152945354	OPN1LW	Opsin	0*	0	0	0	0
chrX	152968995	152982485	OPN1MW2	Opsin	223	137 (61.43%)	127 (56.7%)	1	1
chrX	153006113	153019603	OPN1MW2	Opsin	129	19 (14.72%)	119 (92.25%)	0	2

* OPN1LW gene does not overlap segmental duplications, no unique identifiers were determined. Diploid copy numbers in the NA18507 and YH genomes were calculated by normalizing the unique identifier depth seen in these genomes.

Supplementary Table 3. Gene disruption analysis

Gene Name	mrFAST Copy Number	Number Stop Codons (>= 3 Reads)	Mean Fraction of Reads Supporting Stop	Cumulative Stop Fraction
DUX4	96.6	289	0.0	10.2
ZNF717	26.8	54	0.2	9.5
NPIP	33.6	12	0.2	2.2
FAM157A	9.9	6	0.3	1.9
NBPF1	48.0	17	0.1	1.5
FLG	9.4	8	0.2	1.4
LOC100132832	22.6	1	0.8	0.8
WASH1	16.0	2	0.4	0.7
MST1	6.0	2	0.3	0.7
PKD1	3.4	2	0.3	0.7
HRNR	7.7	1	0.5	0.5
C2orf27	9.8	2	0.2	0.4
PSG8	13.3	4	0.1	0.3
FAM86B2	16.7	3	0.1	0.3
OR2A1	5.1	1	0.3	0.3
PSG6	13.1	3	0.1	0.3
PSG9	13.1	3	0.1	0.2
PDPR	5.1	1	0.2	0.2
NBPF10	51.7	11	0.0	0.2
PSG5	11.3	3	0.1	0.2
UGT2B15	4.1	2	0.1	0.2
PSG7	12.3	2	0.1	0.2
TBC1D3	28.7	6	0.0	0.2
PSG1	13.4	1	0.1	0.1
ZNF705A	12.5	2	0.0	0.1
PSG11	13.1	1	0.1	0.1
DUB3	186.1	7	0.0	0.1
BAGE	17.0	2	0.0	0.1
OR11H1	14.1	1	0.1	0.1
AMY2A	10.5	1	0.1	0.1
NPEPPS	8.3	1	0.0	0.0
PCDHB8	5.9	1	0.0	0.0
ZNF705D	11.4	1	0.0	0.0
DND1	2.8	0	0.0	0.0
SPAG11A	2.9	0	0.0	0.0
OR2A25	2.9	0	0.0	0.0
SERF1A	2.9	0	0.0	0.0
C17orf58	3.2	0	0.0	0.0
DEFB103A	3.2	0	0.0	0.0
OR4K5	3.2	0	0.0	0.0
OR4K2	3.3	0	0.0	0.0
DTX2	3.4	0	0.0	0.0
OR51A4	3.8	0	0.0	0.0
LHB	3.9	0	0.0	0.0
PCDHB7	3.9	0	0.0	0.0
POMZP3	3.9	0	0.0	0.0
SULT1A3	4.0	0	0.0	0.0

ARHGEF5	4.0	0	0.0	0.0
OR51A2	4.0	0	0.0	0.0
C4A	4.1	0	0.0	0.0
KRTAP9-3	4.1	0	0.0	0.0
BOLA2	4.1	0	0.0	0.0
GIYD1	4.2	0	0.0	0.0
RAB6C	4.3	0	0.0	0.0
OR2A4	4.5	0	0.0	0.0
MUC20	4.6	0	0.0	0.0
NAIP	4.6	0	0.0	0.0
OR4M1	4.7	0	0.0	0.0
ZDHHC11	4.7	0	0.0	0.0
CCL3	4.9	0	0.0	0.0
CNTNAP3	4.9	0	0.0	0.0
FAM21A	5.1	0	0.0	0.0
CCL4	5.2	0	0.0	0.0
PDXDC1	5.3	0	0.0	0.0
NOMO1	5.6	0	0.0	0.0
LOC646227	5.6	0	0.0	0.0
CCL3L1	5.7	0	0.0	0.0
LGALS9B	6.0	0	0.0	0.0
LOC650137	6.0	0	0.0	0.0
FLJ43692	6.0	0	0.0	0.0
RASA4	6.1	0	0.0	0.0
TP53TG3	6.9	0	0.0	0.0
C2orf78	7.2	0	0.0	0.0
TAS2R43	7.5	0	0.0	0.0
POLR2J	8.5	0	0.0	0.0
KIR2DS4	9.3	0	0.0	0.0
LOC136157	9.5	0	0.0	0.0
PMS2L3	9.9	0	0.0	0.0
TBC1D29	10.6	0	0.0	0.0
CBWD1	10.8	0	0.0	0.0
AMY1A	11.0	0	0.0	0.0
OR4F17	12.8	0	0.0	0.0
GOLGA6	13.2	0	0.0	0.0
OR4F16	17.1	0	0.0	0.0
TCEB3C	18.1	0	0.0	0.0
POTEB	20.6	0	0.0	0.0
PRR20	22.4	0	0.0	0.0
POTEG	22.8	0	0.0	0.0
LOC100132247	30.5	0	0.0	0.0
GOLGA8E	34.5	0	0.0	0.0
FOXD4L4	43.3	0	0.0	0.0
FAM90A7	44.3	0	0.0	0.0

The fraction of new stop codons based on sequencing reads from sample NA18507 is shown. Analysis was limited to the 92 valided CNV genes which are predicted to be duplicated in sample NA18507. Analysis was limited to those changes supported by 3 or more Q30 reads.

Supplementary Table 6. Validated copy-number polymorphic genes among three individuals.

Gene Name	Transcript ID	Chr (B35)	Start	End	Gene Size	Duplicated Bp	Duplicated Bp %	JDW Copy Number	NA18507 Copy Number	YH Copy Number	#Probes	Log2 ArrayCGH NA18507_YH	Log2 ArrayCGH JDW_YH	Log2 ArrayCGH JDW_NA18507
NBPF10	NM_001039703	chr1	16,635,718	16,663,995	28278	28278	1.00	48.7	51.7	48.9	239	0.062105	-0.24803	-0.21661
NBPF1	NM_017940	chr1	16,635,718	16,685,288	49571	49533	1.00	43.4	48.0	46.3	577	0.137605	-0.24403	-0.28786
AMY2A	NM_000699	chr1	103,872,020	103,880,414	8395	8395	1.00	5.7	10.5	10.8	139	-0.212395	-0.58903	-0.37111
AMY1A	NM_004038	chr1	104,004,461	104,013,331	8871	8871	1.00	5.7	11.0	10.4	164	-0.250895	-0.77603	-0.50861
HRNR	NM_001009931	chr1	148,997,631	149,009,742	12112	7721	0.64	19.5	7.7	15.4	226	-0.529395	0.03997	0.80864
FLG	NM_002016	chr1	149,087,724	149,110,752	23029	10606	0.46	12.6	9.4	13.3	347	-0.196895	-0.11903	-0.04536
LCE3C	NM_178434	chr1	149,386,281	149,386,564	284	0	0.00	0.0	1.1	0.3	5	0.492105	0.09647	-0.41536
LCE1D	NM_178352	chr1	149,582,300	149,583,730	1431	477	0.33	1.1	0.4	2.2	27	-0.716395	-0.48403	0.51589
CFHR3	NM_021023	chr1	193,475,587	193,494,529	18943	12709	0.67	5.0	2.0	3.8	234	-0.639895	0.12897	0.72689
CFHR1	NM_002113	chr1	193,520,518	193,532,973	12456	12455	1.00	3.8	1.6	2.8	212	-0.613395	0.21947	0.74789
CFHR4	NM_006684	chr1	193,588,869	193,619,419	30551	7135	0.23	4.3	2.3	2.8	463	-0.155895	0.29897	0.42489
OR2T11	NM_001001964	chr1	245,115,520	245,116,470	951	0	0.00	1.1	2.3	0.8	21	0.594105	0.03497	-0.86086
FAM21A	NM_001005751	chr10	51,497,690	51,563,274	65585	65585	1.00	4.4	5.1	3.8	840	0.207105	0.03297	-0.14411
DUX4	NM_033178	chr10	135,372,560	135,380,764	8205	8205	1.00	247.8	96.6	195.7	13	-0.319395	-0.25903	0.53939
OR51A4	NM_001005329	chr11	4,923,967	4,924,906	940	940	1.00	2.3	3.8	2.5	18	0.350855	-0.09903	-0.27661
OR51A2	NM_001004748	chr11	4,932,580	4,933,519	940	940	1.00	2.3	4.0	5.1	19	0.363105	-0.05553	-0.29786
OR52N1	NM_001001913	chr11	5,765,662	5,766,622	961	0	0.00	0.9	1.6	2.1	15	-0.322895	-0.42803	-0.22286
OR52E4	NM_001005165	chr11	5,862,099	5,863,035	937	0	0.00	2.2	2.1	3.2	15	-0.359645	-0.49203	-0.17811
OR4P4	NM_001004124	chr11	55,162,410	55,163,347	938	0	0.00	1.0	2.1	2.2	18	-0.177395	-0.76603	-0.68036
OR4S2	NM_001004059	chr11	55,174,956	55,175,891	936	0	0.00	1.0	2.1	2.5	16	-0.234895	-0.33003	-0.28361
ZNF705A	NM_001004328	chr12	8,216,417	8,223,909	7493	7493	1.00	10.1	12.5	11.8	121	-0.074895	-0.21103	-0.20936
TAS2R43	NM_176884	chr12	11,135,153	11,136,179	1027	1027	1.00	4.7	7.5	7.6	19	-0.133395	-0.82853	-0.61161
PRR20	NM_198441	chr13	56,639,332	56,642,353	3022	3022	1.00	27.9	22.4	10.8	59	1.073605	0.87997	-0.22661
POTEG	NM_001005356	chr14	18,623,365	18,654,942	31578	31578	1.00	19.4	22.8	22.9	386	-0.226145	-0.20303	-0.06436
OR4M1	NM_001005500	chr14	19,318,322	19,319,262	941	941	1.00	3.8	4.7	6.0	18	-0.303895	-0.74103	-0.40361
OR4K2	NM_001005501	chr14	19,414,267	19,415,211	945	0	0.00	2.3	3.3	3.6	18	-0.193145	-0.36853	-0.46861
OR4K5	NM_001005483	chr14	19,458,606	19,459,575	970	0	0.00	2.1	3.2	2.4	17	0.057355	-0.26753	-0.47411
POTEB	NM_207355	chr15	19,305,253	19,336,667	31415	31415	1.00	17.4	20.6	22.4	457	-0.314895	-0.32803	-0.12111
LOC650137	NM_001080841	chr15	19,915,066	19,915,749	684	684	1.00	6.1	6.0	7.9	13	-0.494895	-0.61703	-0.18161
TUBGCP5	NM_052903	chr15	20,384,836	20,425,332	40497	0	0.00	2.1	0.8	1.9	538	-0.416895	0.03847	0.42989
CYFIP1	NM_001033028	chr15	20,507,196	20,555,043	47848	0	0.00	2.3	1.2	2.1	711	-0.461645	-0.01203	0.48589
NIPA2	NM_030922	chr15	20,556,790	20,585,849	29060	0	0.00	1.8	0.8	1.7	374	-0.495145	0.06997	0.49989
NIPA1	NM_001142275	chr15	20,594,722	20,638,284	43563	0	0.00	2.0	0.9	1.8	528	-0.479395	0.05097	0.49814
GOLGA8E	NM_001012423	chr15	20,986,537	20,999,864	13328	13328	1.00	35.7	34.5	37.9	118	-0.134395	-0.35503	-0.17611
GOLGA6	NM_001038640	chr15	72,149,251	72,161,944	12694	12694	1.00	17.2	13.2	16.8	173	-0.238395	-0.11053	0.26089
PKD1	NM_000296	chr16	2,078,712	2,125,900	47189	38481	0.82	6.6	3.4	4.4	593	0.120605	0.02947	0.28214
NOMO1	NM_014287	chr16	14,835,144	14,897,514	62371	62371	1.00	5.6	5.6	4.3	814	0.246105	0.20547	-0.07536
NPIP	NM_006985	chr16	14,938,801	14,953,432	14632	14632	1.00	30.0	33.6	30.7	20	0.249855	0.02297	0.08139
PDXDC1	NM_015027	chr16	14,976,334	15,039,053	62720	55935	0.89	5.0	5.3	3.9	706	0.283105	0.18547	0.05039
LOC100132247	NM_001135865	chr16	22,432,345	22,455,329	22985	22985	1.00	27.0	30.5	27.2	41	0.219105	-0.00703	0.09039
BOLA2	NM_001031827	chr16	30,111,758	30,113,128	1371	1371	1.00	7.2	4.1	4.6	23	-0.182395	0.22347	0.44989
GIYD1	NM_001015000	chr16	30,112,718	30,116,381	3664	3664	1.00	5.9	4.2	5.4	40	0.150605	0.09347	0.37489
SULT1A3	NM_003166	chr16	30,113,970	30,123,150	9181	9181	1.00	6.7	4.0	4.0	96	0.153105	0.10297	0.32639

TP53TG3	NM_016212	chr16	33,112,481	33,115,680	3200	3200	1.00	15.9	6.9	6.1	67	0.331605	0.75297	0.54689
PDPR	NM_017990	chr16	68,705,030	68,752,685	47656	35926	0.75	4.0	5.1	4.0	537	0.307855	-0.00453	-0.25811
LGALS9B	NM_001042685	chr17	20,293,768	20,311,440	17673	17673	1.00	3.6	6.0	4.2	330	0.556105	-0.38253	-0.76111
TBC1D29	NM_015594	chr17	25,910,710	25,914,633	3924	3924	1.00	10.9	10.6	9.0	31	0.337605	-0.11903	0.13039
CCL3	NM_002983	chr17	31,439,716	31,441,619	1904	1904	1.00	2.9	4.9	3.9	34	0.224605	-0.28403	-0.29261
CCL4	NM_002984	chr17	31,455,333	31,457,127	1795	1795	1.00	2.8	5.2	3.9	33	0.296855	-0.23103	-0.36736
CCL3L1	NM_021006	chr17	31,647,958	31,649,843	1886	1886	1.00	2.2	5.7	4.2	39	0.373105	-0.67003	-0.80161
TBC1D3	NM_032258	chr17	33,541,292	33,552,188	10897	10897	1.00	26.2	28.7	17.2	12	0.951855	0.20747	-0.36661
KRTAP9-3	NM_031962	chr17	36,642,241	36,643,231	991	991	1.00	4.7	4.1	5.4	18	-0.297645	-0.20603	0.28189
KRT14	NM_000526	chr17	36,992,059	36,996,673	4615	4615	1.00	5.4	2.3	4.5	73	-0.269395	-0.19303	0.32639
ARL17	NM_001039083	chr17	41,989,936	42,012,375	22440	22440	1.00	3.9	1.6	5.2	283	-0.713895	-0.34603	0.42164
NSF	NM_006178	chr17	42,023,354	42,190,000	166647	116319	0.70	2.5	2.0	4.6	2399	-0.780895	-0.41203	0.27139
NPEPPS	NM_006310	chr17	42,963,443	43,055,641	92199	62993	0.68	4.4	8.3	3.5	974	0.616855	0.19897	-0.34411
C17orf58	NM_181656	chr17	63,417,679	63,420,227	2549	2549	1.00	4.3	3.2	2.8	40	0.283105	0.36097	0.36189
TCEB3C	NM_145653	chr18	42,808,571	42,810,447	1877	1877	1.00	23.1	18.1	16.8	30	0.670355	-0.00753	-0.20561
OR4F17	NM_001005240	chr19	61,679	62,596	918	918	1.00	18.1	12.8	9.4	17	0.277105	0.76097	0.55889
PSG8	NM_182707	chr19	47,950,225	47,961,671	11447	11447	1.00	12.6	13.3	11.2	164	0.213355	0.04747	-0.11661
PSG1	NM_006905	chr19	48,063,198	48,075,711	12514	12514	1.00	13.0	13.4	9.9	196	0.316105	0.21097	-0.14011
PSG6	NM_002782	chr19	48,099,608	48,113,829	14222	14222	1.00	13.1	13.1	9.6	218	0.313105	0.24397	-0.11311
PSG7	NM_002783	chr19	48,120,124	48,133,170	13047	13047	1.00	12.1	12.3	9.7	207	0.322605	0.21597	-0.10236
PSG11	NM_203287	chr19	48,203,649	48,222,471	18823	18823	1.00	12.4	13.1	9.9	278	0.311105	0.23197	-0.10061
PSG5	NM_002781	chr19	48,363,735	48,382,528	18794	18794	1.00	10.9	11.3	9.4	267	0.212605	0.09297	-0.10361
PSG9	NM_002784	chr19	48,449,275	48,465,522	16248	16248	1.00	13.3	13.1	10.1	229	0.230105	0.09097	-0.09361
LHB	NM_000894	chr19	54,211,050	54,212,159	1110	1110	1.00	5.1	3.9	4.7	14	0.204355	-0.14103	0.33614
KIR2DS4	NM_012314	chr19	60,035,986	60,051,835	15850	15850	1.00	9.6	9.3	7.7	196	0.231105	0.12397	-0.07936
C2orf78	NM_001080474	chr2	73,922,971	73,955,929	32959	26245	0.80	9.5	7.2	14.1	419	-0.615395	-0.31403	0.27489
RAB6C	NM_032144	chr2	130,453,465	130,456,541	3077	3077	1.00	3.1	4.3	5.6	44	-0.271395	-0.06603	-0.05361
C2orf27	NM_013310	chr2	132,313,796	132,358,709	44914	44914	1.00	8.7	9.8	10.7	620	-0.174895	-0.19903	-0.05636
BAGE	NM_001187	chr21	10,079,667	10,120,808	41142	40371	0.98	17.6	17.0	13.7	471	0.201605	0.12147	-0.05611
OR11H1	NM_001005239	chr22	14,823,380	14,824,358	979	979	1.00	14.7	14.1	16.5	17	-0.282395	-0.29903	-0.06461
DDT	NM_001355	chr22	22,638,108	22,646,573	8466	8056	0.95	3.3	1.8	3.6	137	-0.518145	-0.29053	0.31064
GSTT2	NM_000854	chr22	22,646,868	22,650,660	3793	3793	1.00	2.8	2.0	3.6	62	-0.489895	-0.32403	0.42839
MST1	NM_020998	chr3	49,696,385	49,701,200	4816	4776	0.99	11.5	6.0	10.7	73	-0.472395	-0.59203	0.33489
ZNF717	NM_001128223	chr3	75,868,719	75,916,945	48227	24791	0.51	36.2	26.8	31.6	560	-0.058895	0.16497	0.21989
MUC20	NM_001098516	chr3	196,940,682	196,954,124	13443	2855	0.21	2.8	4.6	2.9	197	0.632605	0.01797	-0.57211
FAM157A	NM_001145248	chr3	199,367,547	199,396,038	28492	28492	1.00	11.0	9.9	8.3	295	0.219855	0.12697	0.05964
UGT2B15	NM_001076	chr4	69,693,104	69,717,150	24047	24047	1.00	4.5	4.1	2.6	284	0.283105	0.18447	0.05114
ZDHHC11	NM_024786	chr5	848,720	904,101	55382	52088	0.94	3.7	4.7	6.1	887	-0.336645	-0.56503	-0.31661
SERF1A	NM_021967	chr5	70,232,246	70,250,112	17867	17867	1.00	3.8	2.9	1.8	175	0.320105	0.48797	0.30239
SMN1	NM_000344	chr5	70,256,524	70,284,592	28069	28069	1.00	3.2	2.3	1.8	309	0.340605	0.56197	0.28489
NAIP	NM_022892	chr5	70,300,066	70,356,697	56632	50992	0.90	4.5	4.6	3.6	590	0.208605	0.12397	0.10589
DND1	NM_194249	chr5	140,030,566	140,033,355	2790	2594	0.93	5.2	2.8	3.5	42	0.181605	0.08697	0.38539
PCDHB7	NM_018940	chr5	140,532,427	140,536,140	3714	2333	0.63	10.0	3.9	6.6	53	-0.102895	0.10497	0.40539
PCDHB8	NM_019120	chr5	140,537,614	140,540,203	2590	2508	0.97	12.3	5.9	7.9	39	0.064855	0.27997	0.44164
LOC646227	NM_001101396	chr5	180,341,824	180,345,858	4035	3961	0.98	4.3	5.6	4.0	74	0.445355	0.00797	-0.39436
BTNL3	NM_197975	chr5	180,348,507	180,366,332	17826	1262	0.07	1.0	2.0	1.0	274	0.417855	-0.01603	-0.38961
OR4F16	NM_001005277	chr5	180,726,894	180,727,830	937	937	1.00	17.9	17.1	12.1	19	0.316855	0.23297	-0.12711
C4A	NM_007293	chr6	32,090,550	32,111,173	20624	20624	1.00	3.6	4.1	2.6	361	0.430855	0.11597	-0.17061

OR2A4	NM_030908	chr6	132,063,302	132,064,234	933	933	1.00	5.8	4.5	7.6	18	-0.555895	-0.49453	0.34889
PMS2L3	NM_005395	chr7	74,781,721	74,802,045	20325	20325	1.00	9.3	9.9	7.1	183	0.281105	0.20297	0.00539
DTX2	NM_020892	chr7	75,735,623	75,779,963	44341	44341	1.00	3.9	3.4	2.7	519	0.299855	0.18997	0.09764
POMZP3	NM_012230	chr7	75,883,955	75,901,271	17317	17317	1.00	3.9	3.9	2.8	149	0.334355	0.22197	0.05639
LOC100132832	NM_001129851	chr7	76,313,448	76,327,006	13559	13559	1.00	21.7	22.6	19.8	60	0.221605	-0.04803	-0.05011
POLR2J	NM_006234	chr7	101,707,270	101,713,101	5832	4996	0.86	7.8	8.5	10.4	72	-0.068145	-0.41353	-0.21811
RASA4	NM_001079877	chr7	101,813,883	101,851,140	37258	37258	1.00	6.5	6.1	7.8	395	-0.215395	-0.50253	-0.25111
LOC136157	NM_001085395	chr7	123,710,376	123,710,960	585	585	1.00	7.9	9.5	6.9	9	0.444105	-0.12503	-0.36161
OR2A25	NM_001004488	chr7	143,208,961	143,209,892	932	0	0.00	3.3	2.9	3.9	16	-0.309145	-0.13303	0.23564
FLJ43692	NM_001003702	chr7	143,321,325	143,330,384	9060	9051	1.00	6.4	6.0	8.5	157	-0.593895	-0.29903	0.22789
OR2A1	NM_001005287	chr7	143,452,866	143,453,796	931	931	1.00	6.3	5.1	9.3	17	-0.477895	-0.44503	0.33714
ARHGEF5	NM_005435	chr7	143,490,137	143,515,372	25236	21888	0.87	5.0	4.0	6.1	439	-0.601895	-0.37353	0.23389
FAM90A7	NM_001136572	chr8	7,408,721	7,427,585	18865	18865	1.00	6.8	44.3	36.0	29	0.284855	-1.51403	-1.50786
SPAG11A	NM_001081552	chr8	7,742,812	7,758,729	15918	15918	1.00	1.9	2.9	4.2	294	-0.505895	-0.94203	-0.38061
DEFB103A	NM_018661	chr8	7,776,324	7,777,590	1267	1267	1.00	2.8	3.2	5.5	19	-0.442395	-0.75953	-0.35311
ZNF705D	NM_001039615	chr8	11,984,256	12,010,434	26179	26179	1.00	8.3	11.4	10.6	384	-0.074395	-0.30203	-0.28111
DUB3	NM_201402	chr8	12,032,086	12,033,678	1593	1593	1.00	139.4	186.1	121.8	18	0.677605	-0.15853	-0.59161
FAM86B2	NM_001137610	chr8	12,327,497	12,338,223	10727	10727	1.00	20.2	16.7	20.8	134	-0.214395	-0.37953	-0.14761
WASH1	NM_182905	chr9	4,511	19,739	15229	15229	1.00	25.6	16.0	19.8	268	-0.196645	0.10647	0.40189
CBWD1	NM_018491	chr9	111,038	169,075	58038	58038	1.00	11.3	10.8	8.3	660	0.247105	0.18747	0.10939
CNTNAP3	NM_033655	chr9	39,062,766	39,278,300	215535	215535	1.00	5.9	4.9	4.6	2690	0.124605	0.16897	0.18189
FOXD4L4	NM_199244	chr9	67,935,112	67,938,218	3107	3106	1.00	40.5	43.3	37.7	41	0.268605	0.05597	0.20314

Validated copy-number polymorphic genes among three individuals. 113 genes where copy-number difference was at least 1 among three individuals are shown.