Personalized copy number and segmental duplication maps using next-generation sequencing

Can Alkan, Jeffrey M Kidd, Tomas Marques-Bonet, Gozde Aksay, Francesca Antonacci, Fereydoun Hormozdiari, Jacob O Kitzman, Carl Baker, Maika Malig, Onur Mutlu, S Cenk Sahinalp, Richard A Gibbs & Evan E Eichler

Despite their importance in gene innovation and phenotypic variation, duplicated regions have remained largely intractable owing to difficulties in accurately resolving their structure, copy number and sequence content. We present an algorithm (mrFAST) to comprehensively map next-generation sequence reads, which allows for the prediction of absolute copy-number variation of duplicated segments and genes. We examine three human genomes and experimentally validate genome-wide copy number differences. We estimate that, on average, 73–87 genes vary in copy number between any two individuals and find that these genetic differences overwhelmingly correspond to segmental duplications (odds ratio = 135; P < 2.2 × 10^-16). Our method can distinguish between different copies of highly identical genes, providing a more accurate assessment of gene content and insight into functional constraint without the limitations of array-based technology.

The human genome is enriched for gene-rich segmental duplications that vary extensively in copy number. Variation in the content and copy number of these duplicated genes has been associated with recurrent genomic rearrangements as well as with a variety of diseases, including color blindness, psoriasis, HIV susceptibility, Crohn’s disease and lupus glomerulonephritis. Despite recent technological advances in copy number detection, a global assessment of genetic variation of these regions has remained elusive. Commercial SNP microarrays frequently bias against probe selection within these regions. Array comparative genomic hybridization (array CGH) approaches have limited power to discern copy number differences, especially as the underlying number of duplicated genes increases and the difference in copy number with respect to a reference genome becomes vanishingly small. Even sequence-based strategies such as paired-end mapping frequently cannot unambiguously assign end sequences in duplicated regions, making it impossible to distinguish allelic and paralogous variation. Consequently, duplicated regions have been largely refractory to standard human genetic analyses.

One promising approach for assessing copy number variation has involved measuring the depth of coverage of whole-genome shotgun (WGS) sequencing reads aligned to the human reference genome. Recent applications of this approach to next-generation sequencing (NGS) technology have provided high-resolution mapping of copy number alterations. However, most of these approaches assay only the ‘unique’ regions of the genome. For example, MAQ reports only unique alignments and arbitrarily selects one position in the case of tied map positions, not reporting any sequence variation. Although it is possible to run MAQ with an option to return all possible map locations of the sequence reads, it reports only the anchoring position and does not return any information on sequence variation. Here, we develop a read-mapping algorithm to rapidly assay copy number variation and experimentally verify its ability to accurately predict copy number in some of the most complex and duplicated regions of three human genomes.

RESULTS

Algorithm development

We developed mrFAST (micro-read fast alignment search tool) to (i) effectively map large amounts of short sequence read data to the human genome reference assembly, (ii) calculate accurate read depth and (iii) return all possible single-nucleotide differences within both unique and duplicated portions of the genome (Supplementary Figs. 1 and 2a). We have shown previously that the ability to place reads to all possible locations in the reference genome is critical to accurately predicting the absolute copy number of duplicated sequences.

mrFAST is designed for short sequence reads (>25 bp). It uses a seed-and-extend method similar to BLAST and implements a hash table to create indices (n = 300 indices of 10 Mb each) of the reference genome that can efficiently use the main memory of the system (Supplementary Fig. 1). For each read, the first, middle and last k-mers are interrogated in the hash table to place initial seeds, where k is the ungapped seed length (we set k = 12 by default). A rapid version of the edit distance computation is then performed.

Received 16 April; accepted 23 July; published online 30 August 2009; doi:10.1038/ng.437
to extend the seed to discover all possible map locations, allowing ≤2 indels. We optionally exclude most of the ‘non-extensible’ seeds, bypassing the high cost of the edit distance computation. For this analysis, we selected an edit distance threshold of two mismatches or indels to account for allelic variants and sequencing error. Moreover, querying three distinct k-mers guarantees discovery of all possible locations of reads within an edit distance of 2 if the length is ≥35 bp and k = 12. As a benchmark, mapping of one human genome (21-fold) against the repeat masked reference genome was achieved in 13.5 h using a 100-CPU cluster.

**Personal duplication maps**

We tested the utility of mrFAST to accurately construct duplication maps by obtaining WGS sequence data from three human males from the NCBI short-read archive and European Read Archive (see URLs section). These included a genome sequence generated with 454 Life Sciences GS-FLX sequence data (for an individual of European descent, JDW)\(^1\) and two genome sequences generated with Illumina WGS data (for a Yoruba African individual, NA18507, and a Han Chinese individual, YH)\(^1\)\(^8\)\(^\text{,}\)\(^9\)\(^2\)\(^2\)\(^4\)\(^,\)\(^5\) (Table 1). All loci were first masked for high-copy common repeat elements (retroposons and short high-copy repeats) using RepeatMasker\(^8\)\(^\text{,}\)\(^9\)\(^2\)\(^2\)\(^4\)\(^,\)\(^5\) and WindowMasker\(^10\). We initially assessed the dynamic range response of shotgun sequence data mapped by mrFAST by determining the read depth for a set of 32 duplicated and unique loci where copy number status had been previously confirmed using experimental methods\(^1\)\(^,\)\(^2\). Using these benchmark loci, we determined the average read depth and variance for 5-kb (unmasked) regions for autosomal and X chromosomal loci (Table 1). For each of the three libraries, we found that read depth strongly correlated with the known copy number (\(R^2 = 0.83–0.90;\) Fig. 1a). Owing to the known sequencing biases of high-throughput sequencing technologies in GC-rich and GC-poor regions\(^3\)\(^1\)\(^5\), we also applied a statistical correction to normalize the read depth based on the GC content of each window (Online Methods and Supplementary Note).

We next assessed the ability of mrFAST read depth to accurately predict the boundaries of known duplicated sequences. We selected a set of 961 autosomal duplication intervals (745 intervals ≥20 kb) that were predicted both by the analysis of the human genome assembly\(^2\)\(^9\)\(^\text{,}\)\(^10\)\(^\text{a}\)\(^,\)\(^b\) and by an independent assessment of Celera capillary WGS sequences\(^1\)\(^,\)\(^2\)\(^1\)\(^5\) where the 20-kb threshold was applied. We reasoned that duplications detected by both methods probably represented a set of true positive duplications whose boundaries would remain largely invariant in additional human genomes. We mapped each of the three WGS sequence libraries (JDW, NA18507 and YH) to the human reference genome (build35) using mrFAST and identified all intervals where at least six out of seven consecutive windows showed an excess depth of coverage (number of reads greater than or equal to the mean plus 3 s.d.). A threshold of 3 s.d. corresponds to a diploid copy number of approximately 3.5, which means that a fraction of sequences with a hemizygous duplication may be missed by this approach. We compared the predicted sizes of intervals in each genome with the duplications predicted from the assembly\(^3\)\(^4\)\(^,\)\(^2\)\(^3\)\(^4\)\(^,\)\(^5\) and determined that the boundaries of known duplications could be accurately predicted (\(R^2 = 0.92;\) Fig. 1b). Because sequence coverage directly affects the power to detect duplications by read depth, we computed the fraction of high-confidence duplication intervals that could be detected at various WGS sequence coverages (Fig. 1c).

### Table 1 Summary statistics for three human genome libraries

| Genome  | Platform | Number of reads | Number of mapped reads\(^a\) | Autosomes | Chromosome X |
|---------|----------|----------------|-------------------------------|-----------|--------------|
| JDW     | 454      | 509,667,772    | 159,293,568                  | 694.93    | 400.58       |
| NA18507 | Illumina | 1,776,928,308  | 556,713,986                  | 2,393.52  | 1,247.50     |
| YH      | Illumina | 1,315,249,404  | 375,234,167                  | 1,645.51  | 971.58       |

\(^a\)Reads mapped against the human reference genome (build35) using mrFAST. Average read length is 36 bp in the JDW and NA18507 genomes and 35 bp in the YH genome.

We inspected three different WGS libraries from three individuals. Approximately 74 million 454-based reads from the JDW genome were rendered into 36-bp reads (Supplementary Note). In total, we processed 4.9 billion reads (206 Gb), and approximately 1.4 billion reads (~28.5%) were mapped to repeat-masked human genome build35 (see duplication maps of three individual genomes and Supplementary Note).

**Figure 1** Correlation of predicted and known segmental duplications in NA18507. (a) mrFAST sequence read depth per 5-kb window along the human genome correlates well (\(R^2 = 0.87\)) with the known copy number of duplicated sequences. (b) Predicted duplication interval length versus the assembly-based length of known duplications (whole-genome assembly comparison; ≥94% sequence identity)\(^3\)\(^4\) shows that boundaries of duplications can be accurately predicted. A few intervals show discrepancy in boundary prediction; this is largely due to deletion polymorphism in the NA18507 genome within duplications (confirmed by array CGH). WSSD, whole-genome shotgun sequence detection; Venter, J. Craig Venter, whose genome has previously been sequenced with Sanger sequencing. (c) A cumulative plot of the fraction of duplication intervals detected as a function of read depth (sequence coverage). The segmental duplication size is given in cumulative intervals (≥10 kb, ≥20 kb, etc.) and represents the set of intervals identified both within the reference assembly (build35) and the Celera whole-genome shotgun sequence reads. As expected, the sensitivity of our method increases with greater genome coverage; the most marked difference in detection is observed between three- and fourfold coverage. WGAC, whole-genome assembly comparison.
Figure 2  Computational prediction and array CGH validation of segmental duplication copy number differences for three human genomes. Red indicates regions of excess read depth (mean + 3 s.d.); gray, regions of intermediate read depth (mean + 2 s.d. and - 3 s.d.); green, regions of normal read depth (mean ± 2 s.d.). The absolute copy number and array CGH results for specific individual genome comparisons are shown in the context of RefSeq-annotated genes. Oligonucleotide relative log2 ratios are depicted as red-green histograms and correspond to an increase and decrease in signal intensity when test-reference is reverse labeled. (a) A known copy number polymorphism on 17q21.31 associated with the H2 haplotype among Europeans (build35 coordinates: chr17:41000000–42300000). Compared with NA18507, the JDW genome shows one to two additional copies of a 459-kb segmental duplication mapping to 17q21.31. (b) An expansion of the complement factor H–related gene family (chr1:193350000–193700000) within JDW. (c) An increase in NA18507 copy number for the defensin gene cluster in 8p23.1 is confirmed by array CGH (aCGH).

As an independent and more sensitive test within unique regions of the genome, we compared copy number variant (CNV) genotype calls for NA18507, with calls recently assessed by another group using the Affymetrix 6.0 platform. We found that 250/282 (88.7%) of CNVs >10 kb and 120/128 (93.8%) of CNVs >20 kb were consistent between the two platforms (Supplementary Note). In two of the most extreme cases of discrepancy, we found that the Affymetrix 6.0 genotypes probably misassigned absolute copy numbers, possibly owing to an incorrect assignment of the population average genotype based on fluorescence intensities. These results highlight the potential of mrFAST read depth to provide precise estimates of copy number across all genomic regions.

We constructed duplication maps for each of the three genomes and estimated the absolute copy number of each duplication interval larger than 20 kb in length. We considered a given segment to be duplicated within an individual if the median estimated copy number for that individual was >2.5 (diploid copy number; see Supplementary Note). We compared the extent of overlap among duplicated sequences (Fig. 2 and Online Methods) and reclassified duplicated sequences as shared or specific to an individual based on the predicted copy numbers in the analysis of these three genomes (Supplementary Note). We defined a total of 725 non-overlapping duplication intervals across the three individuals that total 84.76 Mb.
Figure 4 Correlation between computational and experimental copy number for NA18507 and JDW. 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 data against the human reference assembly (build35). Based on these computational estimates of copy number, we calculated a predicted log2 copy number ratio for each autosomal duplication interval >20 kb (and with <80% of total common repeat content). These values were plotted against the experimental log2 ratios determined by oligonucleotide array CGH. The vertical lines indicate the threshold used for the validated calls (Supplementary Note).

Regardless of the NGS platform, the pattern of read depth was reproducible for 48% of the shared duplications (44,711/94,070; Supplementary Fig. 4). However, among the remaining 52% of duplications, read depth did not correlate between individuals. This suggests that shared duplications show the greatest extremes of copy number variation between individuals (Supplementary Fig. 3). Using absolute estimates of copy number, we calculated an in silico log2 ratio for each of the three genome-wide comparisons and compared it with the experimental values determined by array CGH (Fig. 4 and Supplementary Fig. 6). Overall, we found a positive correlation with copy number predictions ($R^2 = 0.52–0.63$, depending on the pairwise comparison). We note that the ability of array CGH to discriminate absolute differences diminishes as the duplication copy number increases.

We selected 11 duplicated loci that showed copy number differences between the YH and NA18507 genomes and performed FISH analysis on interphase nuclei (Fig. 5 and Supplementary Note) from immortalized cell lines from YH and NA18507. The FISH results were highly consistent with the absolute copy number predicted by mrFAST. For

Figure 5 FISH validation. (a) Sequence read depth predicts five copies of this segment of 17q21.31 in the YH genome and two ‘unique’ copies in NA18507. Array CGH shows a higher copy number in the YH genome, and FISH on interphase nuclei confirms the absolute copy number difference between the two genomes. (b) Similarly, interphase FISH confirms five copies of NPEPPS in YH versus 12 copies in NA18507. (c) YH is predicted and validated to have two more copies of the ‘defensin gene family cluster of Bp23.1. (d) Dovoto the known mosaic architecture for this high-copy locus (>30 copies), both array CGH and FISH cannot accurately estimate copy number difference between the NA18507 and YH genomes, despite the fact that sequence depth predicts approximately two more copies in NA18507. aCGH, array CGH; CN, copy number.
cases where the copy number was >15, FISH was unable to provide a precise estimate of copy number difference because of the technical limitations of this procedure (Fig. 5d and Supplementary Note). With one exception, interphase FISH analysis showed that differences in copy number involved local changes in copy number, suggesting that duplicative transpositions to new locations were exceedingly rare.

Copy number polymorphic genes
This analysis validated 68 gene families as completely or partially copy number–variable among these three individual genomes (Supplementary Table 1). This included a complete duplication of the complement factor H–related complex (consisting of four genes, CFHR1 through CFHR4) within the JDW genome (Fig. 2b). We also confirmed one additional copy of the 8p23.1 defense gene family (DEFB103B) in the YH genome compared with NA18507 and one additional copy in NA18507 compared with JDW. We predict about twice as many copies of the amylase (AMY1A) gene family in NA18507 (n = 9) and YH (n = 10) as in JDW (n = 5). As expected, the African genome (NA18507) showed the greatest number of CCL3L1 copies (n = 7) compared with either JDW (n = 3) or YH (n = 5). We also validated increases in gene segments of functional relevance. For example, we found ten fewer copies of the kringle IV domain of the lipoprotein(a) gene (LPA) in NA18507 (22 copies versus 35 in JDW and 26 in YH)—a polymorphism known to be associated with risk of coronary heart disease\(^\text{35}\).

Although many of these differences are consistent with previous studies, the analysis also confirmed differences in rapidly evolving human and great ape gene families that have been previously difficult to ascertain. For example, our results suggested a higher copy number for the TBC1D3 gene family in NA18507 (29 copies) than in the other two genomes (26 copies in JDW and 17 in YH). Similarly, we predicted absolute differences in the copy number of NPIP (a member of the morpheus gene family) between three human genomes. Unlike FISH or array CGH, sequencing data provide very high specificity to assess the presence or absence of individual paralogous genes. We examined three gene families (morpheus, opsin and CFHR) in more detail by identifying single-nucleotide variants that distinguish the different paralogs. Despite the high degree of sequence identity among the duplicated genes, we found approximately 300 distinct paralogous sequence variants per duplicated gene (one variant per 91 bp) (Supplementary Table 2). We determined which specific duplicate genes were present in each individual, providing, to our knowledge, the first accurate assessment of specific genes, as opposed to copy number differences in the aggregate (Supplementary Fig. 7). Because we tracked all single-nucleotide differences using mrFAST, we were able to assess the relative proportion of disruptive stop codons
Supplementary Table 3

Throughput fashion. The main drawback of existing technologies is the limited length, which renders them unsuitable for identifying segmental duplications. The next challenge will be further development of mapping algorithms that can accurately predict absolute copy number differences, such as those involving complex regions of the genome.

We draw three major conclusions from our computational and experimental analyses. First, NGS technology offers the ability to obtain an unprecedented amount of sequence information at a low-cost, high-throughput fashion. The next challenge will be further development of mapping algorithms that can accurately predict absolute copy number differences, such as those involving complex regions of the genome.

We thank D. Bentley for early access to the Illumina WGS dataset for NA18507; J. Wang for the YH DNA and the cell line; M. Egholm and B. Simen for the JDW DNA. We thank D. Bentley for early access to the Illumina WGS dataset for NA18507; J. Wang for the YH DNA and the cell line; M. Egholm and B. Simen for the JDW DNA.

DISCUSSION

NGS platforms are fundamentally altering genetic and genomic research. Compared to other methods, these platforms offer the ability to obtain an unprecedented amount of sequence information in a low-cost, high-throughput fashion. The main drawback of existing technologies is the comparably short sequence read lengths they produce. As a result, some regions of the human genome—particularly duplication- or repeat-rich regions—have already begun to be excluded as part of standard NGS analyses. We specifically designed our new mapping algorithm, mrFAST, to address this limitation. By considering all possible mappings for a read in an efficient manner, we have been able to apply the high potential of NGS to some of the most structurally complex and dynamic regions of the human genome. By including these regions, we provide one of the first comprehensive estimates of absolute copy number differences among three human genomes.

We draw three major conclusions from our computational and experimental analyses. First, NGS read depth can be used to accurately predict absolute copy number, such that even multicyclic duplications (5 versus 12; Fig. 5) can be reliably predicted between different individuals. Second, the duplication status of the largest segmental duplications (>20 kb in length) is largely invariant, with only 3% of the duplications being specific to an individual. Third, the most extreme copy number variation corresponds to genes embedded within segmental duplications, and most of these differences involve tandem changes in copy as opposed to duplications to new locations. We have validated 113 complete genes as copy number variable among these three individuals. Several of the validated loci are known to be biomedically relevant (for example, opsins, Supplementary Fig. 2d, psoriasis, Supplementary Note) and age-related macular degeneration (Fig. 2b). In addition, several human genes with the most variable copy number (Table 2 and Supplementary Fig. 2f) correspond to rapidly evolving gene families that emerged within the common ancestor of human and African great apes (for example, TBC1D, LRR37, GOLGA, and NBP). Members of these gene families correspond to the core duplications that have been implicated in the expansion of intrachromosomal segmental duplications during hominid evolution. Although the function of these genes is largely unknown, the ability to use NGS to accurately predict their copy number provides the ability to make genotype and phenotype correlations in these complex areas of the genome.

Copy number differences, including variable duplications of entire genes, are now recognized as making substantial contributions to variation in human phenotypes. The ability to accurately and systematically determine the absolute copy number for any genomic segment is a notable first step toward a true and complete picture of individual genomes and phenotypes. In light of the sensitivity and specificity of read depth approaches, we anticipate that this strategy will eventually replace array CGH–based methods. The next challenge will be further definition of the sequence content and structural organization of these dynamic and important regions of the human genome.

URLs. NCBI short-read archive: http://www.ncbi.nlm.nih.gov/Traces/sra/sra.cgi; European Read Archive: ftp://ftp.era.ebi.ac.uk/1000 Genomes Project: http://www.1000genomes.org; mrFAST: http://mrfast.sourceforge.net.

METHODS

Methods and any associated references are available in the online version of the paper at http://www.nature.com/naturegenetics/.

Note: Supplementary information is available on the Nature Genetics website.

ACKNOWLEDGMENTS

We thank D. Bentley for early access to the Illumina WGS dataset for NA18507; J. Wang for the YH DNA and the cell line; M. Egholm and B. Simen for the JDW DNA.

| Number of genes | 0 | 5 | 10 | 15 | 20 | 25 | 30 | 35 | 40 | 45 | 50 |
|-----------------|---|---|----|----|----|----|----|----|----|----|----|
| Difference in copy number between JDW, NA18507 and YH | ≤ 50% overlap with segmental duplications | > 50% overlap with segmental duplications |

Figure 6 Copy number differences between unique and duplicated regions. The 113 genes that vary in copy number were partitioned based on copy number difference and on their intersection with annotated segmental duplications (WGAC*). Duplicated genes show a greater extent of copy number variation than genes mapping to unique regions of the genome.
DNA and J.D. Watson for permission to analyze his genome. We also thank M. Shumway, P. Flicek and R. Leinonen for technical assistance in transferring large sequence datasets; E. Tüzün for help in parallelizing mrFAST for computation clusters through message passing interface; S. Girirajan for assistance with experiments and T. Brown for her help in manuscript preparation. J.M.K. is supported by a US National Science Foundation Graduate Research Fellowship. T.M.-B. is supported by a Marie Curie fellowship (FP7). This work was supported, in part, by U.S. National Institutes of Health grant HG004120 to E.E.E. E.E.E. is an investigator of the Howard Hughes Medical Institute.

AUTHOR CONTRIBUTIONS
C.A., J.M.K., T.M.-B. and E.E.E. designed the study, performed analytical work and wrote the manuscript. C.A., E.H. and O.M. designed and implemented the mrFAST algorithm. C.A., J.M.K., G.A. and J.O.K. performed computational analysis. T.M.-B., F.A., C.B. and M.M. performed validation experiments. R.A.G. advised on handling algorithm. C.A., J.M.K., G.A. and J.O.K. performed computational analysis. T.M.-B., F.A. and M.M. performed validation experiments. R.A.G. advised on handling.

Published online at http://www.nature.com/naturegenetics/
Reprints and permissions information is available online at http://npg.nature.com/

1. Bailey, J.A. et al. Recent segmental duplications in the human genome. Science 297, 1003–1007 (2002).
2. Iafrate, A.J. et al. Detection of large-scale variation in the human genome. Nat. Genet. 36, 949–951 (2004).
3. Redon, R. et al. Global variation in copy number in the human genome. Nature 444, 444–454 (2006).
4. Kidd, K.M. et al. Mapping and sequencing of structural variation from eight human genomes. Nature 453, 56–64 (2008).
5. Fanciulli, M. et al. FGR3B copy number variation is associated with susceptibility to systemic, but not organ-specific, autoimmunity. Nat. Genet. 39, 721–723 (2007).
6. Atkinson, T.J. et al. Copy number polymorphism in Fgr3 predisposes to glomerulonephritis in rats and humans. Nature 439, 851–855 (2006).
7. Gonzalez, E. et al. The influence of CCL3L1 gene-containing segmental duplications on HIV-1/AIDS susceptibility. Science 307, 1434–1440 (2005).
8. Feilermann, K. et al. A chromosome 8 gene-cluster polymorphism with low human beta-defensin 2 gene copy number predisposes to Crohn disease of the colon. Am. J. Hum. Genet. 79, 439–448 (2006).
9. Yang, Y. et al. Gene copy-number variation and associated polymorphisms of complement component C4 in human systemic lupus erythematosus (SLE): low copy number is a risk factor for and high copy number is a protective factor against SLE susceptibility in European Americans. Am. J. Hum. Genet. 80, 1037–1054 (2007).
10. Hollos, E.J. et al. Psoriasis is associated with increased beta-defensin genomic copy number. Nat. Genet. 40, 23–25 (2008).
11. Estivill, X. et al. Chromosomal regions containing high-density and ambiguously mapped putative single nucleotide polymorphisms (SNPs) correlate with segmental duplications in the human genome. Hum. Mol. Genet. 11, 1987–1995 (2002).
12. Locke, D.P. et al. Linkage disequilibrium and heritability of copy-number polymorphisms within duplicated regions of the human genome. Am. J. Hum. Genet. 79, 275–290 (2006).
13. 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, 1199–1203 (2008).
14. Locke, D.P. et al. Large-scale variation among human and great ape genomes determined by array comparative genomic hybridization. Genome Res. 13, 347–357 (2003).
15. Sharp, A.J. et al. Segmental duplications and copy-number variation in the human genome. Am. J. Hum. Genet. 77, 78–88 (2005).
16. Tuzun, E. et al. Fine-scale structural variation of the human genome. Nat. Genet. 37, 727–732 (2005).
17. Korb, J.O. et al. Pair-end mapping reveals extensive structural variation in the human genome. Science 318, 420–426 (2007).
18. Wang, J. et al. The diploid genome sequence of an Asian individual. Nature 456, 60–68 (2008).
19. Campbell, P.J. et al. Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel pair-end sequencing. Nat. Genet. 40, 722–729 (2008).
20. Wheeler, D.A. et al. The complete genome of an individual by massively parallel DNA sequencing. Nature 452, 872–876 (2008).
21. Chiang, D.T. et al. High-resolution mapping of copy-number alterations with massively parallel sequencing. Nat. Methods 6, 99–103 (2009).
22. Bentley, D.R. et al. Accurate whole human genome sequencing using reversible terminator chemistry. Nature 456, 53–59 (2008).
23. Li, H., Ruan, J. & Durbin, R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome Res. 18, 1851–1858 (2008).
24. Hillier, L.W. et al. Whole-genome sequencing and variant discovery in C. elegans. Nat. Methods 5, 183–188 (2008).
25. Altshul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. Basic local alignment search tool. J. Mol. Biol. 215, 403–410 (1990).
26. Levenshtein, V.I. Binary codes capable of correcting deletions, insertions and reversals. Sov. Phys. Dokl. 10, 707–710 (1966).
27. Ukkonen, E. On approximate string matching. in Fundamentals of Computation Theory, Proceedings of the 1983 International FCT Conference 487–495 (Springer-Verlag, London, 1983).
28. Smit, A.F.A., Hubley, R. & Green, P. RepeatMasker Open-3.0. <http://www.repeatmasker.org> (1996–2004).
29. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res. 27, 573–580 (1999).
30. Morgulis, A., Gertz, E.M., Schaffer, A.A. & Agarwala, R. WindowMasker: window-based masker for sequenced genomes. Bioinformatics 22, 134–141 (2006).
31. Smith, D.R. et al. Rapid whole-genome mutational profiling using next-generation sequencing technologies. Genome Res. 18, 1638–1642 (2008).
32. She, X. et al. Shotgun sequence assembly and recent segmental duplications within the human genome. Nature 431, 927–930 (2004).
33. Istrail, S. et al. Whole-genome shotgun assembly and comparison of human genome assemblies. Proc. Natl. Acad. Sci. USA 101, 1916–1921 (2004).
34. 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, 1005–1017 (2001).
35. McCarroll, S.A. et al. Integrated detection and population-genetic analysis of SNPs and copy number variation. Nat. Genet. 40, 1166–1174 (2008).
36. Lackner, C., Cohen, J.C. & Hobbs, H.H. Molecular definition of the extreme size polymorphism in apolipoprotein(a). Hum. Mol. Genet. 2, 933–940 (1993).
37. Prutt, K.D., Tatusova, T. & Maglott, D.R. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res. 35, D61–D65 (2007).
38. Jiang, Z. et al. Ancestral reconstruction of segmental duplications reveals punctuated cores of human genome evolution. Nat. Genet. 39, 1361–1368 (2007).
ONLINE METHODS

Computational analyses. Details regarding the mrFAST algorithm are described at length in the Supplementary Note. mrFAST is freely available to not-for-profit institutions (see URLs section). Segmental duplication maps were constructed from approximately sixfold 454 sequence coverage of the JDW genome, 21-fold Illumina sequence coverage of NA18507 and 16-fold Illumina sequence coverage of YH. 454-based WGS sequence reads of JDW (average length, 266 bp) were broken into 36-bp sequences to make the read-length properties comparable among the three sequence libraries (Supplementary Note). Sequence reads were mapped using mrFAST against human genome reference build35 (Supplementary Note) to define duplication intervals and calculate absolute copy numbers. Read depth was normalized with respect to GC content via a LOESS-based smoothing technique (Supplementary Note). For cross-sample comparisons, the duplication status of each individual over each interval was reassessed based on the estimated absolute copy number (Supplementary Note).

Array CGH validation. We performed array CGH to confirm duplications specific to individual genomes and to confirm copy number differences for shared duplications. A total of six experiments were performed in duplicate with dye-reversals performed between test and reference: NA18507 versus JDW, NA18507 versus YH and JDW versus YH. The log$_2$ relative hybridization intensity was calculated for each probe. In this analysis, we restricted our analysis to regions >20 kb that contained at least 20 probes. We used a heuristic approach to calculate log$_2$ thresholds of significance for each comparison, dynamically adjusting the thresholds for each hybridization to result in a false discovery rate of <1% in the control regions $^{39}$.

FISH analysis. Metaphase spreads were obtained from lymphoblast cell lines from NA18507 (Coriell Cell Repository) and YH (Han Chinese) $^{18}$. FISH experiments were performed using fosmid clones $^{4}$ (Supplementary Note) directly labeled by nick-translation with Cy3-dUTP (Perkin-Elmer) as described previously $^{40}$ with minor modifications (Supplementary Note). Digital images were captured using a Leica DMRXA2 epifluorescence microscope equipped with a cooled CCD camera (Princeton Instruments). 4,6-diamidino-2-phenylindole (DAPI) and Cy3 fluorescence signals, detected with specific filters, were recorded separately as grayscale images. Pseudocoloring and merging of images was performed using Adobe Photoshop software. A minimum of 50 interphase cells were scored for each probe.

$^{39}$ Marques-Bonet, T. et al. A burst of segmental duplications in the genome of the African great ape ancestor. Nature 457, 877–881 (2009).

$^{40}$ Lichter, P. et al. High-resolution mapping of human chromosome 11 by in situ hybridization with cosmids clones. Science 247, 64–69 (1990).