nature genetics

Fine-scale structural variation of the human genome

Eray Tuzun[1,5], Andrew J Sharp[1,5], Jeffrey A Bailey[2,5], Rajinder Kaul[3], V Anne Morrison[1], Lisa M Pertz[2], Eric Haugen[3], Hillary Hayden[3], Donna Albertson[4], Daniel Pinkel[4], Maynard V Olson[3] & Evan E Eichler[1]

Inversions, deletions and insertions are important mediators of disease and disease susceptibility[1]. We systematically compared the human genome reference sequence with a second genome (represented by fosmid paired-end sequences) to detect intermediate-sized structural variants >8 kb in length. We identified 297 sites of structural variation: 139 insertions, 102 deletions and 56 inversion breakpoints. Using combined literature, sequence and experimental analyses, we validated 112 of the structural variants, including several that are of biomedical relevance. These data provide a fine-scale structural variation map of the human genome and the requisite sequence precision for subsequent genetic studies of human disease.

Two recent studies presented surveys of copy-number variation in the human genome using comparative microarray technology (oligonucleotide and BAC-based)[2,3]. These papers suggested that large-scale variation, up to hundreds of kilobases in size, occurs in the human genome and that gene-rich regions are common sites of copy-number polymorphism. Both studies, however, surveyed only a small fraction of the sequence, and the results lacked the precision necessary to demarcate the boundaries of specific copy-number changes. In addition, neither method could reliably detect more subtle variation such as inversions or small insertions and deletions. Detection and characterization of such variation is an important aspect of both selection and susceptibility to disease[4,5].

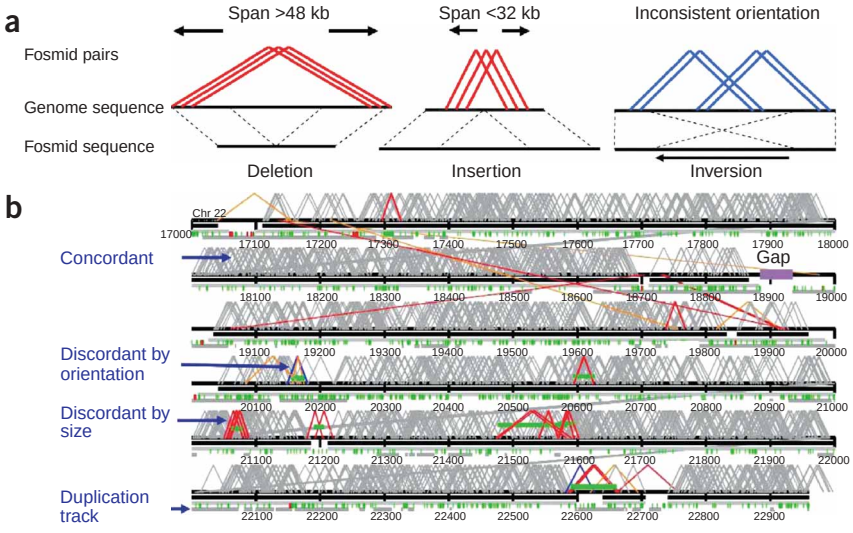

Figure 1 Detection of structural variation. (**a**) The underlying paired-end sequence methodology used to detect structural variation. Deletions in the fosmid source are defined as sites where two or more fosmid end-sequence pairs span >48 kb. Insertions are defined as sites where two or more fosmids span <32 kb (red). These length thresholds are 3 s.d. from the mean insert size (**Supplementary Fig. 1** online). Inversions in DNA show two or more fosmids (blue) with an inconsistent orientation of the end sequences with respect to the human genome for each breakpoint. (**b**) Sites of discordance. Two or more fosmids showing a consistent size discordance (red), discrepancy by orientation (blue) or both size and orientation differences (yellow) were mapped with respect to recent retrotransposon insertions (green), segmental duplications (dark gray bars) and sequence gaps (purple). Concordant best placement fosmids are shown as gray triangles. Below the sequence (horizontal black line), a track depicting the span of best hits (black bar) and tied (gray bar) fosmid pairs is shown. A 7-Mb region of chromosome 22 (May 2004 assembly; 17,000–24,000 kb) is depicted, with violet arrows highlighting various features. Overall, 99% of the pairs were concordant by size and orientation. Sites ($n = 7$) that met our criteria are indicated by a green bar above the line. Regions where concordant and discordant pairs overlap suggest heterozygosity in the fosmid DNA source, as opposed to regions where only discordant pairs are noted, which may represent sequence error or homozygosity.

[1]Department of Genome Sciences, University of Washington School of Medicine, 1705 NE Pacific Street, Seattle, Washington 98195, USA. [2]Department of Genetics, Case Western Reserve University, Cleveland, Ohio 44106, USA. [3]Department of Medicine and the University of Washington Genome Sequencing Center, University of Washington, Seattle, Washington 98195, USA. [4]Comprehensive Cancer Center, University of California San Francisco, San Francisco, California 94143, USA. [5]These authors contributed equally to this work. Correspondence should be addressed to E.E.E. (eee@gs.washington.edu).

Published online 15 May 2005; doi:10.1038/ng1562

We developed a computational strategy to detect systematically such fine-scale structural variation, which includes inversions and more subtle variation, by mapping paired-end sequence data[6] from a human fosmid DNA genomic library (**Supplementary Note** online). The fosmid DNA represented a different source individual than had been used to generate the bulk of the genome reference sequence and therefore offers, in theory, exquisite power to detect structural variations between these two sources. Consistent discrepancies in size and orientation of multiple paired-end sequences have the potential to detect insertions, deletions and inversions between the reference human genome and this individual (**Fig. 1a**). We optimally mapped 589,275 pairs of fosmid end sequences (from an initial total of 1,113,518 distinct pairs) to their best locations in the human genome assembly (build 35, May 2004). This 581 Mb of mapped sequence (0.19× sequence coverage) represents approximately eightfold physical coverage of the human euchromatic regions of the human genome (average insert size of fosmid = 40 kb).

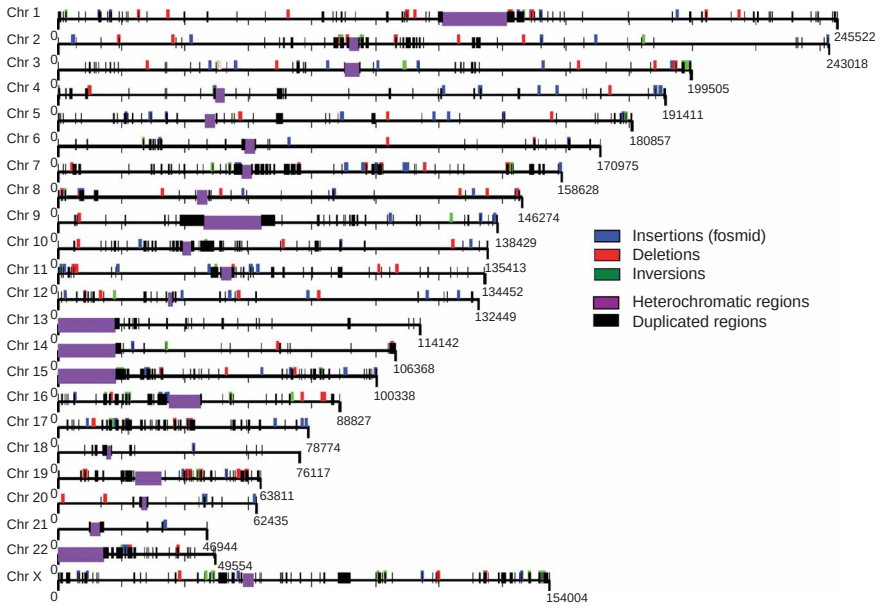

Figure 2 Structural variation map. The schematic summarizes the distribution of insertions, deletions and inversions on each human chromosome. A total of 297 clusters were identified: 139 insertions, 102 deletions and 56 inversions breakpoints (**Supplementary Table 1** online). Across the genome, 163 of the structural variants map to regions of segmental duplication. Detailed views of each chromosome are given in **Supplementary Figs. 2–25** online.

Next, we searched for regions where multiple independent fosmids showed discrepancy by their predicted size. We categorized fosmids as discordant if the *in silico* size was in excess of three standard deviations from the mean (<32 kb or >48 kb; **Fig. 1a** and **Supplementary Fig. 1** online). Using this threshold, 99% (583550 of 589275) of the pairs are concordant with the genome by size and by orientation. Discordant pairs (n = 3,169) are classified as those whose insert size is predicted to be too large (n = 1,531) or too small (n = 1,638). Some of these (n = 698) also had ends that were incorrectly oriented with respect to the human reference genome and are indicative of potential inversions.

We graphically mapped all sites with two or more discordant fosmids in the context of repeat, gap and duplication properties of each human chromosome (**Fig. 1b** and **Supplementary Figs. 2–25** and **Supplementary Table 1** online). After elimination of clonal propagation and other assembly artifacts (**Supplementary Note** online), we identified 297 sites of putative structural variation, corresponding to 139 insertions, 102 deletions and 56 inversion breakpoints (**Fig. 2** and **Table 1** and **Supplementary Table 1** online). Seventy-five percent (228 of 297) of these sites of putative structural variation also showed spanning fosmids consistent with the human genome reference sequence, in addition to two or more discordant fosmid pairs. This indicates that the diploid individual from whom the fosmid library was constructed is probably heterozygous with respect to these structural variants, and they probably do not represent genome assembly errors. We estimate that most of these putative structural rearrangements ranged in size from 8 to 40 kb. Deletions and inversions as large as 329 kb and 1.9 Mb, respectively, were also predicted (**Supplementary Note** online).

Structural variation in the human genome may arise by a variety of mechanisms, including retrotransposition of mobile elements (*i.e.*, LINE elements and retroviruses, nonhomologous recombination, etc.). We examined the sequence underlying the sites of structural variation and found that although variant regions are enriched near or in repetitive DNA (**Table 1** and **Supplementary Table 1** online), only 18 regions show a full-length L1 (L1HS, PA2, PA3) or retrovirus element consistent with a variable retrotransposition insertion with the human reference genome. A much more conspicuous association is noted when sites of recent duplication (>90% and >1 kb in length) are compared[7]. More than half of the sites (163 of 297) completely

Table 1 Summary of structural variants determined by paired-end sequence analysis

	Total deletions (S:N)	Total insertions (S:N)	Total inversions (S:N)	Total
Seg. dup.	52 (35:17)	68 (55:13)	43 (32:11)	163
HS repeat	28 (24:4)	12 (11:1)	1 (0:1)	42
Unique	22 (15:7)	59 (46:13)	12 (10:2)	95
Total	102 (74:28)	139 (112:27)	56 (42:14)	297

Deletions are defined as regions that contain two or more fosmids where the distance between paired-end sequences is significantly (>3 s.d.) greater than >48 kb when mapped against the human genome reference sequence (build 35). Insertions are regions containing two or more fosmids with paired ends showing a significantly (>3 s.d.) smaller span (<32 kb) on the human reference sequence. Inversions refer to breakpoint regions where two or more fosmids show a consistent misorientation of end sequence, irrespective of size. S refers to the number of variant sites with support for the human genome reference as well as the discordance, suggestive of a heterozygous fosmid donor genotype. N is the number of variant sites lacking support for the human genome reference, suggestive of a homozygous fosmid donor genotype or reference genome error. Seg. dup., sites containing or flanked by segmental duplications; HS, human specific retroposons; Unique, neither. A complete list of the 297 sites is shown in **Supplementary Table 1** online.

Table 2 Validated structural polymorphisms

Gene	Type	Frequency[a]	Locus	Expected	Detected	Phenotype	Reference
GSTT1	Deletion	20% −/−	22q11.2	54.3 kb	ND	Halothane/epoxide sensitivity	23
EMD-FLNA	Inversion	33% −/+	Xq28	219 kb	~100 kb	None	9
GSTM1	Deletion	50% −/−	1p13.3	18 kb	17.8 kb	Toxin resistance, cancer susceptibility	24
CYP2D6	Duplication	1−29% +++	22q13.1	5 kb[b]	10 kb	Antidepressant drug sensitivity	25
CYP21A2	Duplication	1.6% +/−	6p21.3	35 kb	22.1 kb	Congenital adrenal hyperplasia	26
LPA	VNTR	94% H	6q27	5.5 kb[b]	14 kb	Coronary heart disease risk	11
RHD	Deletion	15−20% −/−	1p36.11	~60 kb	67.8 kb	Rhesus blood group sensitivity	10

Published structural polymorphisms (expected) that were validated by the fosmid paired-end sequence approach (**Supplementary Fig. 28** online). The detected size is based on average discordance. H refers to heterozygosity[11]. ND, not detected by two or more fosmids but by a single pair of fosmid ends; VNTR, variable number of tandem repeats.
[a]The frequency of homozygotes (−/−) or heterozygotes (+/−) is shown. The frequency of homozygous deletion of *GSTT1* among Asians is 57% (ref. 27). Haplotypes with multiple copies of *CYP2D6* (+++) have been identified in 21−29% of Ethiopians[25]. [b]Multiple copies of the underlying repeat unit (multiallelic) have been identified. For example, the 5.5-kb Kringle IV domain of the apolipoprotein a gene (*LPA*) ranges in copy number from 2 to 40.

contain or map one of their boundaries in a segmental duplication, with the most pronounced association seen for intrachromosomal segmental duplications where the degree of sequence identity exceeds 98% (88 of 163 sites; **Supplementary Fig. 26** online). Overall, ~41% (123 of 297) of the structural variants mapped to the introns or exons of RefSeq genes (**Supplementary Table 1** online).

We carried out a series of analyses to assess the biological relevance of these putative variants. First, we compared our sites to 16 previously characterized polymorphisms[8–11]. We detected 7 of 16 common polymorphisms that ranged in size from 5 kb to 219 kb (**Table 2**). Many of these are biomedically relevant, including deletion of the gene *RHD* (Rhesus blood group susceptibility factor), variable copy-number of the Kringle IV domain of the gene *LPA* (coronary heart disease risk factor) and deletion of the gene *GSTM1* (cancer susceptibility or drug detoxification). In most cases, the boundaries of the deletion or duplication map within 2–3 kb of published breakpoints (**Supplementary Table 1** online). Five of seven of these confirmed structural polymorphisms show minor allele frequencies greater than 20% (**Table 2**).

In this study, we compared the fosmid source genome (two haplotypes) with a composite reference genome of which >70% was derived from a single individual (RPCI-11). Theoretical calculations predict that complete sequence information from two to three individuals should identify 40–60% of all common variation where minor allele frequencies exceed 20% (ref. 12). In addition to these germline events, we also detected the expected somatic variation associated with the kappa, lambda and heavy chain immunoglobulin loci at 2q11, 22q11.2 and 14q32.3, respectively.

Next, we selected 57 BACs corresponding to 91 regions of putative structural variation and constructed a specialized BAC microarray to assess copy-number differences by array

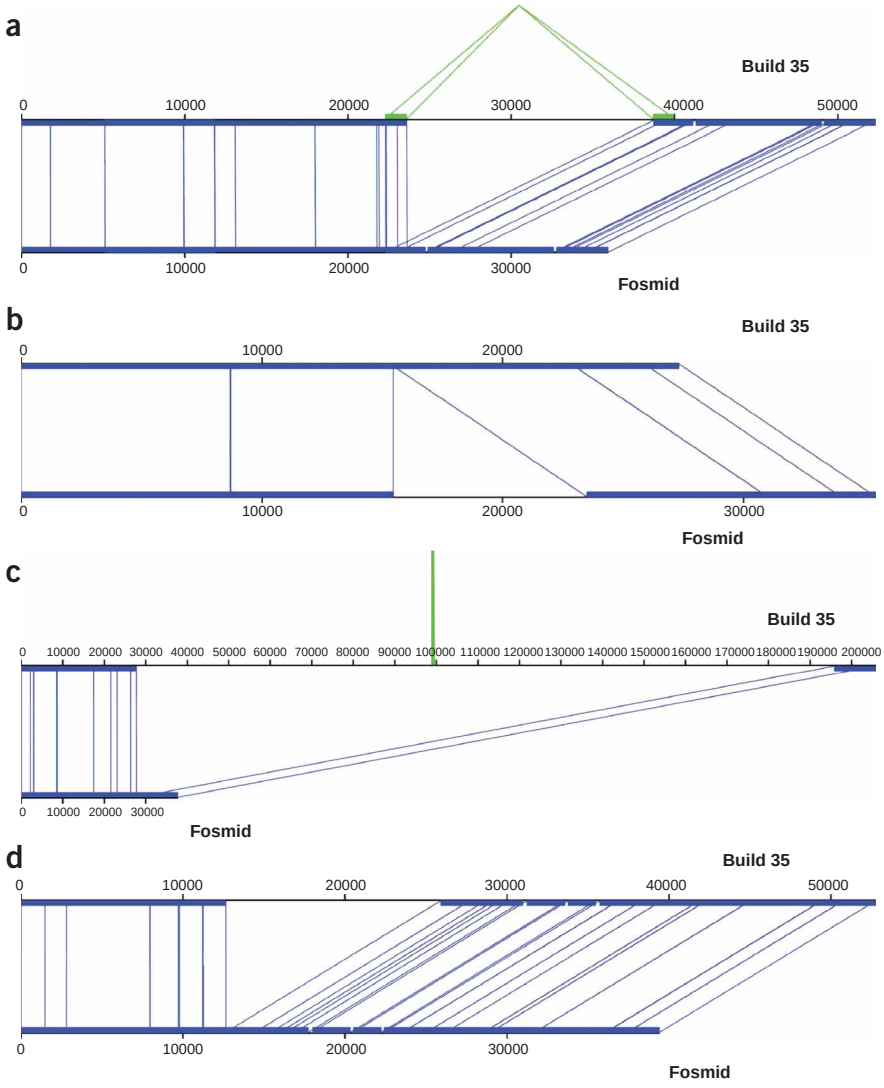

Figure 3 Sequence analysis of structural variants. Representative fosmids from 40 sites of structural variation were sequenced corresponding to ~1.5 Mb of sequence. Fosmid genome sequence was compared with the human genome reference sequence confirming (**a**) deletion of 16,424 bp (AC158317), (**b**) an insertion of 8,043 bp (AC158319), (**c**) a large deletion of 169,357 bp (AC158332) and (**d**) a deletion of 13,340 bp (AC153483).

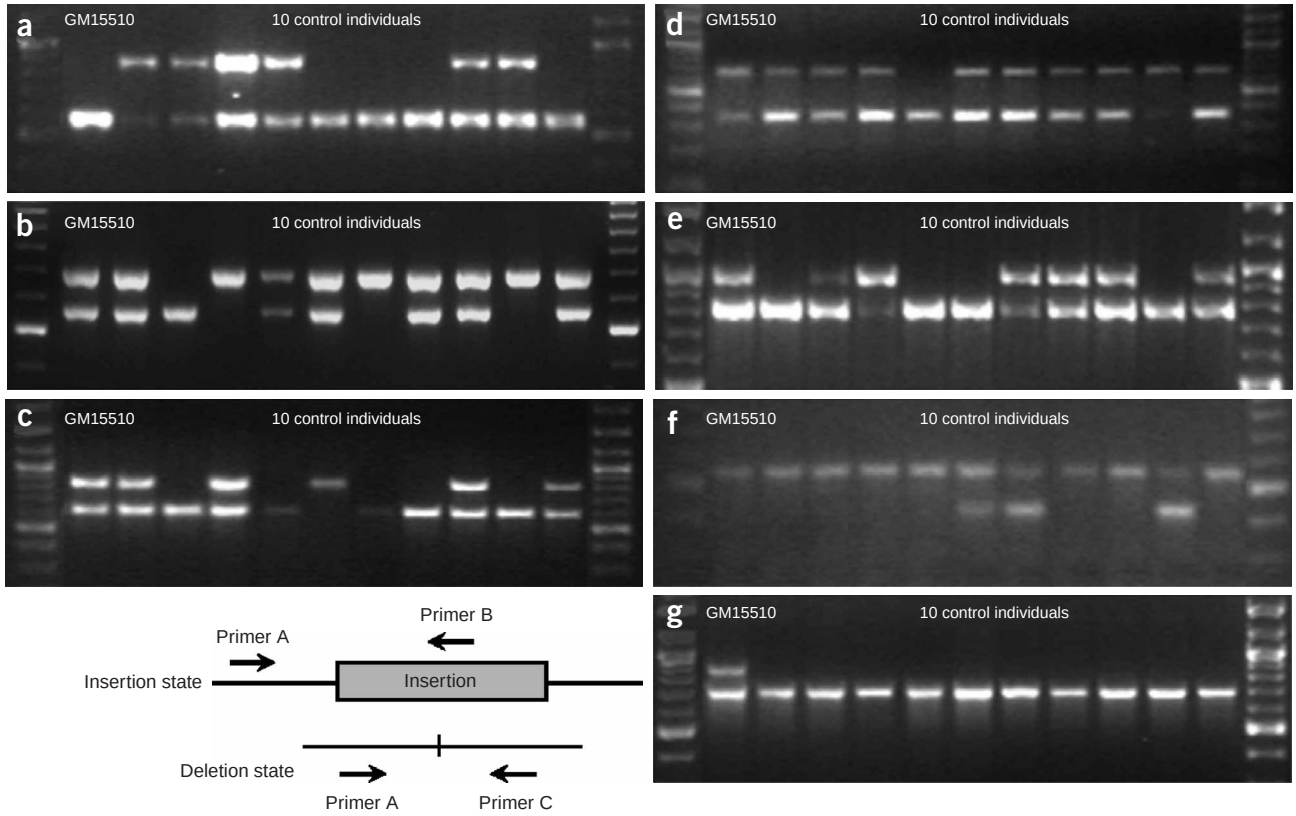

Figure 4 Genotyping analysis of structural variants. PCR validation and genotyping for seven insertion-deletion sites identified in fully sequenced fosmids, confirming (**a**) an 8.2-kb deletion in fosmid 3777M04, (**b**) a 13.3-kb deletion in fosmid 2588B13, (**c**) a 9.1-kb insertion in fosmid 913E19, (**d**) an 8.6-kb deletion in fosmid 3075L14, (**e**) a 6.4-kb deletion in fosmid 3762I17, (**f**) a 12.1-kb insertion in fosmid 647I01 and (**g**) a 10.2-kb deletion in fosmid 2840F04. PCR primers (**Supplementary Table 4** online) were designed that specifically amplify in the presence (primer A+B, upper band) or absence (primer A+C, lower band) of the insert. In each case, the genotype of the source of the fosmid library (GM15510) was consistent with the *in silico* prediction. For six of the seven structural variants, analysis of ten ethnically diverse control individuals confirms these as common polymorphisms in the normal population (**Supplementary Table 5** online).

comparative genomic hybridization (CGH). We screened a diversity panel of 46 humans and found that 28% (16 of 57) of the selected sites (**Supplementary Table 2** online) showed evidence of copy-number difference by array CGH, suggesting that these sites correspond to *bona fide* polymorphisms in the population. This represents an approximately fivefold enrichment relative to a control set of 554 randomly selected BACs, of which only 5.6% showed comparable levels of copy-number variation ($P = 7.1 \times 10^{-7}$; Fisher's exact test). To eliminate potential methodological artifacts, we repeated the experiment with a second sample set of 51 individuals and a second reference DNA and obtained nearly identical results (D. Locke, unpublished data).

We compared our complete data set with two previous surveys of large-scale variation in the human genome[2,3]. Eighty-five percent (252 of 297) of the sites that we detected were not detected by either of the screens and, therefore, represent potential new sites of structural variation. We called these insertions, deletions and inversions intermediate-sized structural variants (ISVs) to distinguish them from larger copy-number polymorphisms, although there seems to be some overlap between these two groups, especially for larger ISVs. As expected, the median size of sites that overlapped those detected by array CGH was nearly double that of structural variations overall (31.3 kb versus 15.2 kb, respectively). Many of the variants that we identified probably lie below the detection limit of BAC-based array CGH.

Because all the structural variants that we identified by this approach have been effectively subcloned, the availability of the fosmid genomic library now allows for the nature of the structural variants to be defined precisely at the sequence level. As a more direct means of validation, we randomly selected 40 discordant sites for complete fosmid-insert sequencing. Fingerprint analysis of the representative fosmids showed that the insert size (39.4 kb $\pm$ 3.7 kb) was consistent with our *in silico* estimate and that the fosmids were not grossly rearranged (**Supplementary Table 3** online). From these 40 fosmids, we generated $\sim$1.5 Mb of *de novo* assembled sequence and compared it with the reference genome. We confirmed large-scale structural variation ranging in size from $\sim$5 kb to 169.4 kb for 33 of the 40 selected clones (**Fig. 3**). This included 17 deletions, 13 insertions and 3 inversions. Two of the seven that were not verified correspond to atypically short fosmids ($<$30 kb); another two showed evidence of sequence assembly collapse inconsistent with the fosmid length by fingerprint analysis (**Supplementary Note** online).

The availability of underlying sequence structure provides the necessary framework for genotyping assays to be developed (**Fig. 4**). We selected 11 of the 33 sequenced regions for further experimental validation and designed a total of 22 PCR-based assays to distinguish both the human reference genome and fosmid variants in a human diversity panel (**Supplementary Table 4** online). In total, we estimated

a minor allele frequency >20% for 5 of the 11 sites that we selected. Differences in allele frequency were noted for two of these; we observed an altered frequency of heterozygotes among Amerindians and Asians. Genotyping assays such as these will be valuable for future association studies of this type of fine-scale structural variation with disease and disease susceptibility.

Two important biases in the distribution of these structural variants emerge from this study. First, 55% (163 of 297) of the structural variation maps to segmentally duplicated regions of the genome, which represent only 5.3% of the total genome sequence[7,13]. This suggests a tenfold enrichment relative to unique regions. Similar results have been obtained from surveys of larger copy-number variation[2,3,14]. But 86% of the variant sites detected in this study did not overlap those from previous studies, possibly owing to differences in methods of ascertainment. The fine-scale structural variation map of the human genome, therefore, offers a different level of resolution than do these other studies. The second bias is that insertions outnumber deletions (139 to 102). This effect does not seem to be due to an excess of retrotransposition events (at least in the case of L1 elements). Once again, the frequency of insertions is almost certainly an underestimate. Physical constraints of fosmid-insert size suggests that we would probably not have detected insertions >40 kb. These results might extend the long-standing observation that haploinsufficiency is less tolerated than trisomy in evolving populations[15,16] or, alternatively, might represent some systematic bias in the assembly of the human genome reference sequence.

Many of the sites of structural rearrangement that encompass genes correspond to tandem clusters of multigene families. Nonallelic recombination between homologous sequences is probably the mechanism underlying some of this variation. There is a general theme of molecular-environmental interaction (**Supplementary Table 1** online). As noted previously[2,13], many of the genes are associated with drug detoxification (glutathione S-transferase, cytochrome P450 genes and carboxlyesterase gene families), innate immune response and inflammation (leukocyte immunoglobulin-like receptor, defensin and *APOBEC* gene families), surface integrity (late epidermal cornified envelope and mucin gene families) and surface antigens (galectin, melanoma antigene gene and rhesus blood group gene families). Although many of these 'environmental sensor' genes may not be essential for viability, they may be an important component of adaptability. Gains and losses of several of these genes are known risk factors for disease[8]. It will be crucial to determine whether rearrangement of these genes reoccurs on different genetic backgrounds. If mutational events reoccur too frequently, association studies based on linkage disequilibrium of closely mapped SNP markers may not uncover an association with disease. The fine-scale structural variation map we generated provides a rationale for prioritizing regions for further study.

METHODS

Paired-end sequence analysis. We optimally aligned fosmid end sequences against the assembled human genome sequence as part of a three-step process: recruitment, quality rescoring and pairing. During the recruitment phase, we aligned 1,113,518 pairs of sequence, representing 2,298,774 sequence ends, using MEGABLAST (–p 80 –s 90 –v 7 –b 7 –w 12 –t 21) against the nearly finished human genome assembly (build 35, May 2004). We retained the seven highest-scoring alignments (≥80% sequence identity) in the genome for each end sequence and optimally realigned these with the Needleman-Wunsch algorithm (match = +10, mismatch = −8, gap opening = −20, gap extension = −1, no penalty for terminal gaps). We recalculated percent identity accounting for the underlying sequence quality data using only base pairs with a minimum of phred quality 30. We then rescored alignments on the basis of length and percent identity. We removed alignments that scored significantly lower and could not represent an allelic sequence relative to the other end alignments. For each fosmid, we scored all paired-end combinations of the top alignments for placement. This scoring scheme weighted heavily for allelic levels of variation (identity >99.5%), essentially penalizing duplicated regions. Concordant pairs were favored over discordant ones when the sequence and length scores were similar (**Supplementary Note** online). Ninety-nine percent (583,550 of 589,275) of the pairs were concordant and had end orientation and an insert size of 32–48 kb (<3 s.d. from the mean, 39.9 ± 2.76 kb), consistent with the reference genome. To eliminate false positives, we treated discordant alignments more stringently, requiring end alignments to have ≥99.5% identity and ≥400 bp, of which ≥150 bp was unique. Consequently, the effective quality of discordant sequences was high (at phred quality ≥27, the average and median read lengths were 510 bp and 528 bp, respectively). These high-quality discordant pairs (n = 3,189) were classified as those where the insert size was predicted to be too large (n = 1,531) or too small (n = 1,638). Some of these discordant pairs (n = 698) also showed an incorrect orientation of ends with respect to the human genome sequence. Finally, there were 220 pairs whose insert sizes were within estimated limits but showed improper orientation. Using parasight visualization software, we displayed sequence annotation, discrepant clones and putative regions of rearrangement, based on two or more similarly discordant clones, together for each chromosome. We then manually curated rearrangements in parasight to confirm and further characterize them as likely insertions, deletions or inversions. The final data set consisted of 297 sites less than 2 Mb and 291 sites less than 1 Mb. Sites of structural variation were mapped against annotated human genes (RefSeq) and the location of variants was classified as exonic, intronic, intragenic (deletion or inversion of entire gene) or intergenic (**Supplementary Table 1** online) with respect to each gene.

Array CGH. We isolated genomic DNA from a panel of 46 lymphoblastoid cell lines (Coriell Cell Repository) of diverse ethnic origin (8 Chinese, 4 Japanese, 10 Czechoslovakian, 1 Druze, 7 Biaka, 9 Mbuti and 7 Amerindian samples; **Supplementary Table 2** online). We labeled test and reference DNAs (from an anonymous male donor) with Cy3 or Cy5 by random priming and hybridized them to a BAC microarray[17] consisting of 57 clones overlapping sites of putative variation as detected by fosmid paired-end sequence analysis and 554 control clones. Each hybridization was done in duplicate, incorporating a reverse-labeling, and clones that yielded $\log_2$ hybridization ratios that were >0.3 s.d. from the mean in both replicates were scored as variant (**Supplementary Table 2** online). Analysis of self versus self hybridizations indicated that using these criteria yielded a false positive rate <1 in 4,000 clones.

Sequencing. We isolated a subset of discordant fosmids (n = 495), corresponding to the 297 putative variant sites, and confirmed their identity by sequence analysis of the insert (T7 and SP6) using a previously described protocol[18]. Because polymorphic L1 insertions are within our detection range, we screened all representative fosmids for known human-specific L1 elements by two methods (**Supplementary Note** online). Fosmid DNA was fluorescently sequenced (ABI3100) and hybridized[19] using γ-32P-ATP end-labeled L1 HS Ta oligonucleotides[20]. An L1 retrotransposition insertion event was scored if fosmids were positive by both assays and the average discordance was consistent with an L1 insertion (**Supplementary Fig. 27** online). We targeted a subset (n = 40) of sites that did not contain L1 retrotransposons for complete insert sequencing. For each fosmid, we generated a shotgun sequence library and a multiple complete digest fingerprint map with four independent restriction enzymes (*Eco*RI, *Hin*dIII, *Bgl*II and *Nsi*I)[21]. A total of 192 independent shotgun library clones were sequenced for each project, providing approximately three- to fourfold sequence coverage. Sequences were assembled and viewed using phred/phrap/consed software tools. Sequence contigs >2 kb in size were ordered and oriented, and FASTA files with underlying quality scores were generated for sequence analysis. We compared fosmid and human genome (build 34 and build 35) sequence using ClustalW[22] and graphical visualization scripts (two-way_mirror.pl and miropeats) to identify the extent of each rearrangement (**Supplementary Table 3** online).

Genotyping. We designed PCR assays to distinguish insertion versus deletion events (**Fig. 4**). Oligonucleotide sequences and PCR amplification conditions are given in **Supplementary Table 4** online.

URL. Parasight visualization software is available at http://humanparalogy.gs.washington.edu/parasight/.

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

ACKNOWLEDGMENTS

We thank C. Alkan, J. Sprague, C. Gulden, D. Locke, S. McGrath and Z. Cheng for technical assistance and A. Chakravarti and B. Waterston for comments. This work was supported, in part, by a grant from the US National Institutes of Health to E.E.E.

COMPETING INTERESTS STATEMENT

The authors declare that they have no competing financial interests.

Received 10 February; accepted 1 April 2005
Published online at http://www.nature.com/naturegenetics/

1. Botstein, D. & Risch, N. Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. *Nat. Genet.* **33** Suppl, 228–237 (2003).
2. Sebat, J. *et al.* Large-scale copy number polymorphism in the human genome. *Science* **305**, 525–528 (2004).
3. Iafrate, A.J. *et al.* Detection of large-scale variation in the human genome. *Nat. Genet.* **36**, 949–951 (2004).
4. Gonzalez, E. *et al.* The influence of CCL3L1 gene-containing segmental duplications on HIV-1/AIDS susceptibility. *Science* **307**, 1434–1440 (2005).
5. Stefansson, H. *et al.* A common inversion under selection in Europeans. *Nat. Genet.* **37**, 129–137 (2005).
6. International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. *Nature* **431**, 931–945 (2004).
7. She, X. *et al.* Shotgun sequence assembly and recent segmental duplications within the human genome. *Nature* **431**, 927–930 (2004).
8. Buckland, P.R. Polymorphically duplicated genes: their relevance to phenotypic variation in humans. *Ann. Med.* **35**, 308–315 (2003).
9. Small, K., Iber, J. & Warren, S. Emerin deletion reveals a common X-chromosome inversion mediated by inverted repeats. *Nat. Genet.* **16**, 96–99 (1997).
10. Colin, Y. *et al.* Genetic basis of the RhD-positive and RhD-negative blood group polymorphism as determined by Southern analysis. *Blood* **78**, 2747–2752 (1991).
11. 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).
12. Kruglyak, L. & Nickerson, D.A. Variation is the spice of life. *Nat. Genet.* **27**, 234–236 (2001).
13. Bailey, J.A. *et al.* Recent segmental duplications in the human genome. *Science* **297**, 1003–1007 (2002).
14. Locke, D.P. *et al.* Refinement of a chimpanzee pericentric inversion breakpoint to a segmental duplication cluster. *Genome Biol.* **4**, R50 (2003).
15. Brewer, C., Holloway, S., Zawalnyski, P., Schinzel, A. & FitzPatrick, D. A chromosomal duplication map of malformations: regions of suspected haplo- and triplolethality—and tolerance of segmental aneuploidy—in humans. *Am. J. Hum. Genet.* **64**, 1702–1708 (1999).
16. Lindsley, D.L. *et al.* Segmental aneuploidy and the genetic gross structure of the *Drosophila* genome. *Genetics* **71**, 157–184 (1972).
17. Snijders, A.M. *et al.* Assembly of microarrays for genome-wide measurement of DNA copy number. *Nat. Genet.* **29**, 263–264 (2001).
18. Horvath, J., Schwartz, S. & Eichler, E. The mosaic structure of a 2p11 pericentromeric segment: A strategy for characterizing complex regions of the human genome. *Genome Res.* **10**, 839–852 (2000).
19. Eichler, E.E. *et al.* Length of uninterrupted CGG repeats determines stability in the FMR1 gene. *Nat. Genet.* **8**, 88–94 (1994).
20. Badge, R.M., Alisch, R.S. & Moran, J.V. ATLAS: a system to selectively identify human-specific L1 insertions. *Am. J. Hum. Genet.* **72**, 823–838 (2003).
21. Wong, G.K., Yu, J., Thayer, E.C. & Olson, M.V. Multiple-complete-digest restriction fragment mapping: generating sequence-ready maps for large-scale DNA sequencing. *Proc. Natl. Acad. Sci. USA* **94**, 5225–5230 (1997).
22. Thompson, J.D., Higgins, D.G. & Gibson, T.J. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22**, 4673–4680 (1994).
23. Sprenger, R. *et al.* Characterization of the glutathione S-transferase GSTT1 deletion: discrimination of all genotypes by polymerase chain reaction indicates a trimodular genotype-phenotype correlation. *Pharmacogenetics* **10**, 557–565 (2000).
24. McLellan, R.A., Oscarson, M., Seidegard, J., Evans, D.A. & Ingelman-Sundberg, M. Frequent occurrence of CYP2D6 gene duplication in Saudi Arabians. *Pharmacogenetics* **7**, 187–191 (1997).
25. Aklillu, E. *et al.* Frequent distribution of ultrarapid metabolizers of debrisoquine in an ethiopian population carrying duplicated and multiduplicated functional *CYP2D6* alleles. *J. Pharmacol. Exp. Ther.* **278**, 441–446 (1996).
26. Koppens, P.F., Hoogenboezem, T. & Degenhart, H.J. Duplication of the *CYP21A2* gene complicates mutation analysis of steroid 21-hydroxylase deficiency: characteristics of three unusual haplotypes. *Hum. Genet.* **111**, 405–410 (2002).
27. Lee, E.J., Wong, J.Y., Yeoh, P.N. & Gong, N.H. Glutathione S–transferase-θ (*GSTT1*) genetic polymorphism among Chinese, Malays and Indians in Singapore. *Pharmacogenetics* **5**, 332–334 (1995).