Characterization of missing human genome sequences and copy-number polymorphic insertions

Jeffrey M Kidd[1], Nick Sampas[2], Francesca Antonacci[1], Tina Graves[3], Robert Fulton[3], Hillary S Hayden[1], Can Alkan[1], Maika Malig[1], Mario Ventura[4], Giuliana Giannuzzi[4], Joelle Kallicki[3], Paige Anderson[2], Anya Tsalenko[2], N Alice Yamada[2], Peter Tsang[2], Rajinder Kaul[1], Richard K Wilson[3], Laurakay Bruhn[2] & Evan E Eichler[1,5]

The extent of human genomic structural variation suggests that there must be portions of the genome yet to be discovered, annotated and characterized at the sequence level. We present a resource and analysis of 2,363 new insertion sequences corresponding to 720 genomic loci. We found that a substantial fraction of these sequences are either missing, fragmented or misassigned when compared to recent *de novo* sequence assemblies from short-read next-generation sequence data. We determined that 18–37% of these new insertions are copy-number polymorphic, including loci that show extensive population stratification among Europeans, Asians and Africans. Complete sequencing of 156 of these insertions identified new exons and conserved noncoding sequences not yet represented in the reference genome. We developed a method to accurately genotype these new insertions by mapping next-generation sequencing datasets to the breakpoint, thereby providing a means to characterize copy-number status for regions previously inaccessible to single-nucleotide polymorphism microarrays.

The human genome reference assembly is a mosaic of distinct haplotypes sampled from multiple individuals[1]. As a result of both gaps in the assembled sequence and the structural differences among different humans, individual genome projects are expected to uncover human sequences present in some (or all) individuals that are not represented in the assembly. Consistent with this prediction, the first sequences of individual genomes[2,3] revealed 23–29 Mb of sequence that does not map against the reference assembly. The short-read, high-throughput approaches currently being employed are also expected to uncover unrepresented insertions[4–7]. However, these sequences often assemble only as short (median length of 220–314 bp; ref. 7) contiguous sequences (contigs) that are difficult to anchor and incorporate into existing genome assemblies. Thus, although thousands of new sequences may be discovered over the next few years, their annotation and complete integration into the human genome will remain a significant bottleneck[8]. As genotyping and expression microarrays are fundamentally dependent upon the reference genome for array probe design, a small fraction of the human genome effectively cannot be assayed.

We recently reported efforts to systematically map and sequence human genome structural variation using a fosmid end-sequence pair mapping approach[9–11]. We fragmented genomic DNA from nine humans and subcloned 40-kb segments. Using standard capillary sequencing, we generated reads from both ends of each fragment (end-sequence pairs) and mapped clones to the human reference genome. Structural differences (inversions, deletions, insertions and translocations) between the reference genome assembly and the library source were identified on the basis of the mapped location of the end-sequence pairs. As the individual fosmid clones were retained, the procedure allowed simultaneous discovery and complete sequence characterization of a subset of structural variant loci including new insertion sequences common to most individuals but not represented in the human reference genome. Here we present a detailed sequence and copy-number analysis of these segments missing from the human reference genome.

RESULTS

Discovery

We systematically searched 9.7 million end-sequence pairs, corresponding to 92-fold physical coverage of the human genome, for sequences that did not map to the reference sequence (NCBI build 35). The end-sequence data set was derived from nine individual genomes (four Yoruba individuals from Ibadan, Nigeria (YRI), two individuals with European ancestry (CEU), two individuals with Han Chinese or Japanese ancestry (JPT+CHB), and one individual of unknown ethnicity). We distinguished clones that mapped onto the assembly with only one end (one-end anchored, or OEA, clones) and orphan clones for which neither end mapped. After eliminating low-quality sequence and obvious viral and bacterial contaminants, we identified 44,415 high-quality fosmid end sequences that do not map onto the genome reference sequence (NCBI build 35)[11]. This set includes individual sequences from 26,001 OEA clones and 9,207 orphan clones.

[1]Department of Genome Sciences, University of Washington School of Medicine, Seattle, Washington, USA. [2]Agilent Laboratories, Santa Clara, California, USA. [3]Washington University Genome Sequencing Center, School of Medicine, St. Louis, Missouri, USA. [4]Department of Genetics and Microbiology, University of Bari, Bari, Italy. [5]Howard Hughes Medical Institute, Seattle, Washington, USA. Correspondence should be addressed to E.E.E. (eee@gs.washington.edu).

RECEIVED 9 DECEMBER 2009; ACCEPTED 19 MARCH 2010; PUBLISHED ONLINE 18 APRIL 2010; DOI:10.1038/NMETH.1451

Table 1 | Assembling new sequence contigs

		Assembled sequences			Typed by array CGH			
Sample	Population	Contigs	Contig size (Mb)	Loci	Contigs	Contig size (Mb)	Loci	Sequenced clones
NA15510	–	768	0.904	345	387	0.512	177	9
NA18517	Yoruba	726	0.925	307	529	0.700	229	15
NA18507	Yoruba	1,386	1.752	534	904	1.208	363	22
NA18956	Japan	885	1.140	342	597	0.815	243	65
NA19240	Yoruba	1,034	1.295	400	682	0.910	295	44
NA18555	China	953	1.187	380	615	0.825	269	20
NA12878	CEPH	977	1.232	386	653	0.879	279	26
NA19129	Yoruba	990	1.277	359	678	0.932	266	13
NA12156	CEPH	996	1.278	377	667	0.914	266	8
Total (nonredundant)		3,963	4.465	1,182	2,363	2.834	720	192

Shown are the numbers of new sequence contigs, their sizes and the numbers of corresponding loci with contributions from each sample. Results are given for the initial set of 3,963 assembled contigs as well as for the 2,363 contigs that pass all filters. The sample origin of 222 sequenced clones (corresponding to 192 distinct loci) is also shown. –, unknown ethnicity.

Using Phrap (http://www.phrap.org), we initially assembled these individual sequences into 3,963 sequence contigs (total size = 4.47 Mb, N_{50} = 1,148 bp; **Table 1**), but after applying additional experimental and computational filters, we reduced this to 2,363 distinct sequence contigs (**Supplementary Note**).

Of the contigs, 40% (1,019 of 2,363) contain sequence contributed by at least one orphan clone, suggesting that these contigs represent segments longer than 40 kb (**Supplementary Table 1**). Using OEA anchoring information and the mate-pair relationships from the orphan clones, we identified 720 loci (400 of which have a mapped genomic position) corresponding to ~2.8 Mbp of sequence with a median contig size of 1 kb (**Supplementary Note**). Notably, 80 of the 400 anchored loci (20%) map within 5 Mb of the ends of a chromosome (a significant 2.9-fold subtelomeric enrichment, $P = 1.0^{-18}$, binomial test; **Supplementary Fig. 1** and **Supplementary Table 2**). In addition to these 720 loci, we identified 19,038 singleton OEA sequences (average length 790 bp) as well as 5,654 orphan clones that did not contribute to any contigs. By convention, we refer to these sequences as 'novel insertions' on the basis that they are not present within the public reference genome assembly.

Fluorescence *in situ* hybridization analysis

Our analysis distinguished two different types of new human sequences: 400 loci that were anchored within euchromatin, according to OEA assignments, and 320 unassigned loci for which a clear anchor position could not be identified. We explored the genomic distribution and assessed the accuracy of our assigned locations using individual fosmid clones as fluorescence *in situ* hybridization (FISH) probes. Although limited to larger regions, this analysis provided us valuable high-level mapping information with respect to the distribution of insertions in heterochromatin and euchromatin. We selected 33 contigs derived only from orphan clones (assigned to seven distinct unmapped loci) and mapped these loci to metaphase chromosomes by FISH. Three loci mapped separately to telomeric regions on chromosomes 10q, 7p and Xp; one locus mapped to 6q1; and three loci mapped to the p arms of the acrocentric chromosomes (**Supplementary Table 1**).

As a complement to these studies, we also tested an additional 68 large orphan contigs, which we constructed on the basis of a detailed fingerprint analysis of all orphan clones from a single individual human genome library (NA15510; **Supplementary Note**). After excluding 31 contigs assigned to genome assembly gaps[12], we found that 15 of the contigs mapped interstitially, with the remainder (22 of 37) mapping to telomeric, pericentromeric or acrocentric positions (**Supplementary Table 3** and **Supplementary Fig. 2**).

Finally, we considered sequence contigs that had been anchored by OEA clones, but also had contributions from at least one orphan clone, to positions in human euchromatin by using 37 fosmid clones (20 OEA and 17 orphan clones) as FISH probes. We found that 78% (29 of 37) of the clones support the predicted position, whereas 11% (4 of 37) map to a different interstitial location and 11% (4 of 37) map to the p arms of the acrocentric chromosomes. We additionally tested a limited number ($n = 3$) of smaller insertions (<30 kb), all three of which had been completely sequenced and confirmed by metaphase oligo-FISH. We found that two of three were copy-number polymorphic among the four individuals tested (**Supplementary Note**). Our FISH results indicate that megabases of uncharacterized sequence remain within the heterochromatin and euchromatin-heterochromatin transition regions of the human genome, but they also confirm the presence of missing euchromatic sequences that are copy-number polymorphic.

Assembly comparisons

We searched for evidence of the identified 2,363 sequence contigs in other human and nonhuman primate genome assemblies. Six hundred contigs (71 loci) have a match against the newest human reference genome assembly, GRCh37, and 1,467 contigs (54 loci) have a match against the HuRef assembly[2] (**Supplementary Note**). We find partial support for 1,700–2,000 of the contigs in sequence data from the JDW, YH and NA18507 genomes[3–5] (**Supplementary Note**). One of the genomes in our study, NA18507, had been sequenced previously to high coverage using the Illumina platform[4] and subjected to a SOAP *de novo* sequence assembly[7]. Notably, 94% of our smallest insertions identified from single unmapped reads (~790 bp) were not found to be part of the *de novo* assembly (**Supplementary Note**). Of our larger contigs, 32% had no representation and only 25% had complete sequence coverage (defined as more than 95% base-pair representation). When we restricted our analysis to insertions from sequenced NA18507 clones, we found that 52% (11 of 21 sequenced fragments) either were not present ($n = 4$) or mapped to different scaffolds ($n = 7$) in the *de novo* assembly. We found

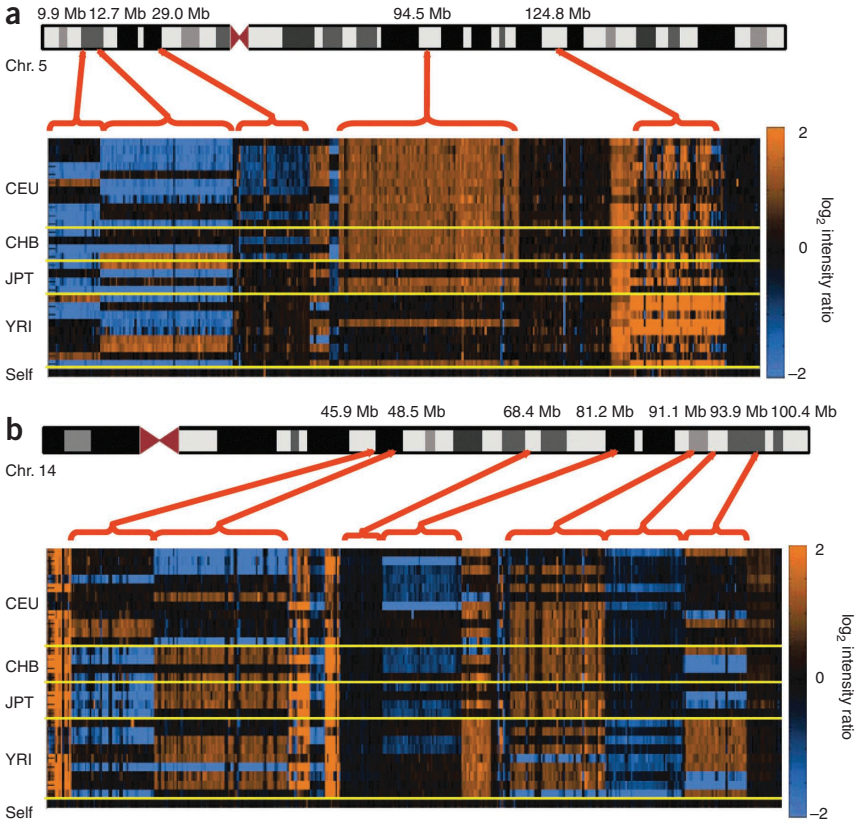

Figure 1 | Copy-number polymorphism of novel insertions. (**a**,**b**) Array CGH intensity data for new sequences ordered along chromosome 5 (**a**) and chromosome 14 (**b**) on the basis of anchored map locations (NCBI build 35 coordinates). Copy-number gains (orange) and losses (blue) are shown relative to the reference sample (NA15510). Each column in the heat map represents a probe on the array, and each row represents a sample ordered and separated (yellow lines) by corresponding HapMap population (CEU, CHB, JPT and YRI). The bottom row depicts a reference self-self hybridization as control. The red brackets group multiple contigs into loci that generally show a consistent hybridization pattern by array CGH.

that this fragmentation often corresponds to the presence of large common repeat sequences that disrupt the contiguity and complicate map assignment. Regions largely devoid of common repeats or segmental duplications showed the greatest correspondence in length and coverage.

To determine the ancestral state of each of these sequences, we also searched the 2,363 contigs against available whole-genome sequence data from chimpanzee and orangutan[13]. Seventy-four percent (1,745 of 2,363) of the contigs had a match against one of these data sets, with 68% (1,599 of 2,363) of the contigs identified within chimpanzee. We were concerned that these sequences might have characteristics leading to their underrepresentation in genome-sequencing data sets, so we performed a microarray-based comparative genomic hybridization (array CGH) experiment using DNA from a single chimpanzee and tested whether the DNA in fact hybridized. This experiment indicated that 84% (1,985 of 2,363) of the contigs were present in the single chimpanzee analyzed (**Supplementary Note**). This includes 624 contigs that do not have a match to the chimpanzee genome sequence data. In total, we found experimental or computational support for 94% (2,223 of 2,363) of the contigs in the chimpanzee and 96% (2,266 of 2,363) in either chimpanzee or orangutan. The absence of these new insertions in the current reference genome reflects either genome assembly errors or deletions that have emerged within the human lineage and are now copy-number polymorphic in our species.

Copy-number polymorphism

We designed two customized oligonucleotide microarrays to assess copy-number polymorphism among these novel insertion sequences. In the first, we designed a microarray targeting the 19,038 single OEA sequences that did not assemble into sequence contigs and tested them against the sample genomes used for discovery. After filtering additional contaminants, we found that 38% (7,240 of 19,038) of these unassembled sequences were represented by at least three probes with signal intensities detectable above the background. On the basis of a comparison of the intensity values for the eight analyzed samples, we estimate that 31% (2,228 of 7,240) of the assayable single OEA sequences are copy-number polymorphic (**Supplementary Table 4**).

In the second design, we investigated copy-number polymorphism for the 2,363 sequence contigs that had been assigned to 720 distinct loci and tested a larger collection of 28 unrelated HapMap individuals (9 CEU, 11 YRI and 8 JPT+CHB). These experiments clearly identified sets of sequences that are copy-number polymorphic or apparently fixed among the analyzed individuals (**Fig. 1**). We identified polymorphic contigs using two alternative calling schemes: a noise-multiplier approach that compares the median probe log-ratios for each contig with the results of a control self-self hybridization (using reference sample NA15510, **Supplementary Table 5**) and a clustering approach that assigns contigs to log-ratio clusters[14] that are then fitted to distinct, small-integer copy-number states (**Fig. 2** and **Supplementary Table 6**). The noise-multiplier approach identified 37% of the contigs as being copy-number polymorphic. We were able to fit 518 contigs to a copy-number state, of which 461 contigs are fitted to two or more distinct copy-number states; 443 contigs (18.7%) were identified as polymorphic by both approaches, an indication of the challenges in assigning discrete copy numbers to all copy number–variable loci.

We assessed the extent of population differentiation for these sequences using both the F_{ST} and V_{ST} statistics[15,16]. For 189 loci with a simple autosomal insertion-deletion variant, we found 20 loci having an F_{ST} greater than 0.35 (**Fig. 3**, **Supplementary Tables 7** and **8** and **Supplementary Fig. 3**). Among these, we identified a 3.9-kb insertion sequence within the first intron of the *lactase* gene (*LCT*; **Fig. 2**). Notably, this 3.9-kb insertion is prevalent among the YRI samples tested (allele frequency of 0.86) but is largely absent among European samples from the Centre d'Etude du Polymorphism Humain (CEPH) collection (allele frequency of 0.11), where it is in complete linkage disequilibrium

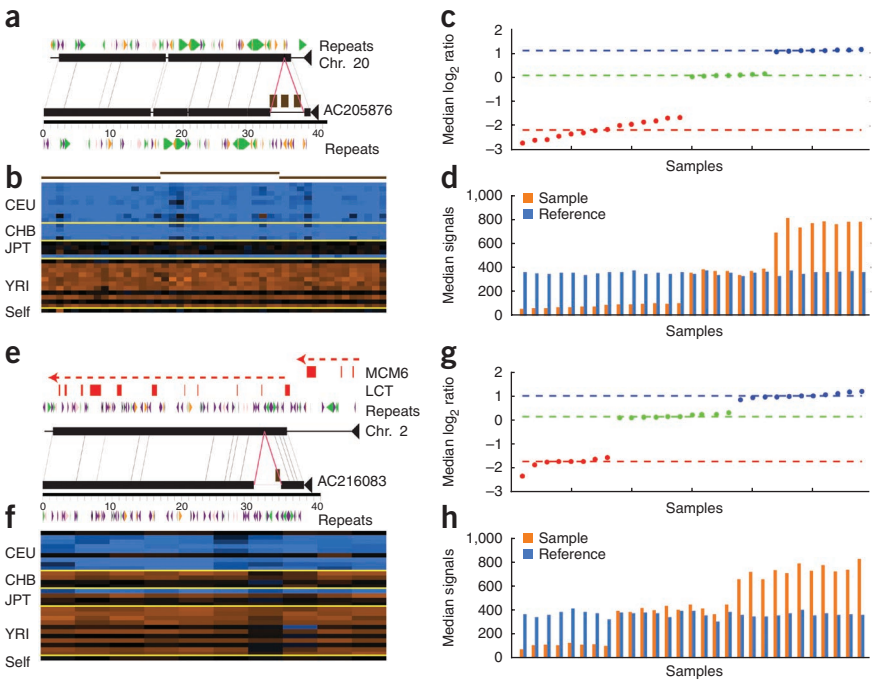

Figure 2 | Sequencing and genotyping insertions. (**a**) The complete sequence of a clone (AC205876) carrying a 4.8-kbp novel insertion sequence is compared to the corresponding segment from chromosome 20 using miropeats[27] (black lines connect segments of matching sequence; colored arrows correspond to common repeats; green, LINEs; purple, short interspersed repeat elements (SINEs); orange, long terminal repeat elements; pink, DNA elements). The magenta lines denote the insertion breakpoints. The brown boxes correspond to the mapped position of three assembled new sequence contigs. (**b**) Array CGH hybridization results represented as a heat map suggest that the deletion is fixed in the CEU and CHB populations. The brown-red lines correspond to the three sequence contigs depicted in **a** and are represented by 16, 15 and 18 array CGH probes, respectively. (**c**,**d**) The median $\log_2$ ratios (**c**) and single-channel intensities (**d**) for all probes matching AC205876. Note that the reference channel (blue bars) shows similar intensity across hybridizations. For this example, the reference sample is inferred to have a copy number of 1. The signals form three distinct clusters that are assigned integer copy-number states of 0, 1 and 2. The dotted red, green and blue lines in **c** correspond to the median intensities of each defined cluster. Using these genotypes, we calculated an F_{ST} of 0.70 for this insertion. (**e**–**h**) A second example, as described for **a**–**d**, showing a 3.9-kb insertion (AC216083) within the first intron of the *LCT* (*lactase*) gene (red boxes in **e** represent exons).

($D' = 1$) with the functional single-nucleotide polymorphism that has been associated with lactase persistence[17]. We repeated the analysis using the V_{ST} statistic for all 720 loci (**Supplementary Fig. 4**). We identified 27 loci that have a V_{ST} value greater than 0.35, with ten having a value greater than 0.5 (**Supplementary Table 9**). Fosmid clones corresponding to several of the most stratified loci have been completely sequenced, including a 4.8-kb insertion on chromosome 20 (AC205876, **Fig. 2**, V_{ST} = 0.73, F_{ST} = 0.70) and an 11.4-kb insertion on chromosome 1 near the *ATP6V1G3* gene (AC212752, V_{ST} = 0.48, F_{ST} = 0.37). These sites represent structures that show a high level of differentiation among human populations but are absent from the genome reference.

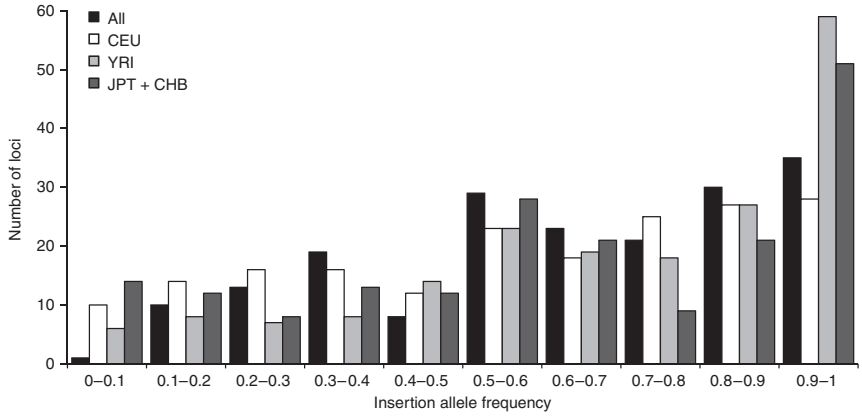

Figure 3 | Insertion allele frequency distribution. Frequencies of the insertion alleles are shown for 189 loci that are fitted to distinct copy numbers and are consistent with a simple autosomal insertion-deletion variant. Values are shown for all 28 individuals (black bars) and separately for each HapMap population, as indicated.

Sequencing and genotyping novel insertions

The complete sequence of insertions smaller than 40 kb can be directly obtained by sequencing an appropriate fosmid clone, whereas an iterative strategy is required to capture the sequence of larger insertions. We sequenced 222 fosmid clones (53 OEA clones and 169 spanned insertions) using a traditional capillary sequencing and assembly approach (**Supplementary Table 10** and **Supplementary Fig. 5**). The 222 clones correspond to 192 distinct genomic loci and contain a total of 1.67 Mb of inserted sequence (**Supplementary Fig. 6**) subsuming 475 of our original 2,363 contigs. Four of the completely sequenced insertions, ranging in size from 41 to 65 kb, were larger than a single clone insert (**Supplementary Note**). The sequenced insertions are similar in composition to segments sampled from the reference genome assembly, with a slight enrichment for common repeats, particularly long interspersed nuclear elements (LINEs; **Supplementary Table 11**). Only five of the 192 loci (**Supplementary Table 12**) have been updated in GRCh37; thus, the majority (97%) of these insertions await integration into the next version of the human genome.

We searched the sequenced insertions against the RefSeq gene database[18] to identify previously uncharacterized exons. We found that segments from 22 genes matched 21 of the insertions (**Supplementary Table 13**); the results included support for structures not represented in the build 36 assembly (for example, *MINK1*, *FSCN2*, *PECAM1* and

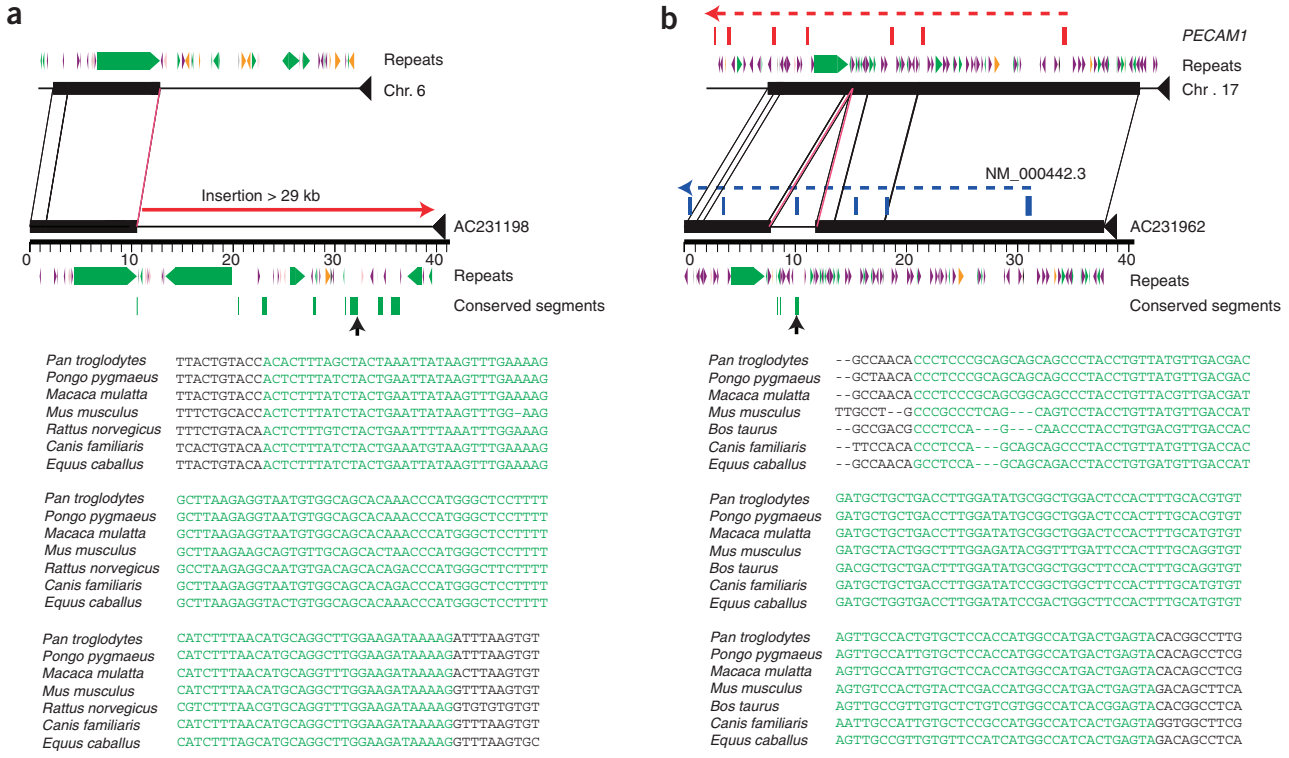

Figure 4 | Annotation of conserved and functional elements. (**a**) Using miropeats[27], we compared the complete sequence of an OEA clone carrying 29 kbp of new sequence to the reference genome. We identified a 95-bp conserved element within this sequence (green rectangles) as defined by a GERP analysis of eight species (see Online Methods). A multiple sequence alignment of one of these conserved elements (black arrow) is highlighted. (**b**) We predicted a new exon within the sequence of a 4.3-kbp insertion by comparison with the *PECAM1* transcript (NM_000442.3), as shown in blue. This alternative exon is supported by RNA-seq data and corresponds to a conserved element identified by alignment comparisons.

VPRBP genes; **Fig. 4**). We further searched for expressed elements using mRNA-seq data derived from multiple human tissues that do not map onto the build 36 genome assembly[19]. We mapped these previously unmapped reads onto the sequenced clones and found that 26 insertions contained segments supported by at least three mRNA-seq reads (**Supplementary Fig. 5** and **Supplementary Note**). We searched against an alignment of nine mammalian genomes to identify segments matching the

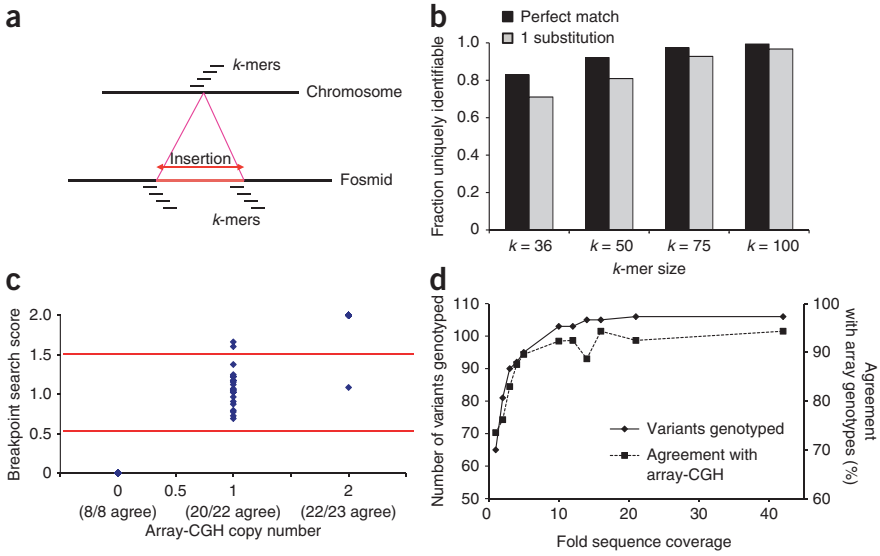

Figure 5 | Genotyping sequenced variants through unique *k*-mer matches. (**a**) We identified unique diagnostic *k*-mer sequences for each variant using sequence-resolved breakpoints. For the deletion breakpoint, we required *k*-mers to have a single match to the reference genome and no matches to the fosmid sequences. For the insertion breakpoints, we required *k*-mers to have no matches to the genome and a single match to the fosmid. To be uniquely identifiable, a variant must have at least one deletion *k*-mer and at least one insertion *k*-mer that meet these criteria. (**b**) Effect of *k*-mer length and search stringency on our ability to uniquely identify a variant. Of the sequenced sites, 71% (108 of 152) were uniquely identifiable with criteria of $k = 36$ and one substitution, and 97% (147 of 152) are assayable if *k*-mer length was increased to 100 bp. (**c**) A comparison of genotypes determined using array CGH and breakpoint *k*-mer matching is shown for sample NA18507. The search database consists of unique 36-mers (one substitution). We determined genotypes for 54 variants by both array CGH and breakpoint *k*-mer matching. Partitioning the breakpoint scores into distinct genotypes at 0.5 and 1.5 (red lines) resulted in 94.3% genotype agreement between the two methods. (**d**) Effect of sequence coverage on breakpoint *k*-mer genotyping. The number of variants genotyped (at least one matching read) and the percent agreement with array CGH results are shown at various sequence coverage levels (1–42×).

sequenced insertions (Ensembl Compara 51)[20,21]. Using these alignments, we identified 477 constrained elements from 104 different loci (**Fig. 4**), a signature that identifies segments of possible functional importance[22]. Six of the constrained elements intersect with mapped RefSeq exons, with the remainder having unknown functional importance. Using genomic evolutionary rate profiling (GERP) scores as a metric, we noted that the conserved elements found in the insertions show a level of constraint similar to that of the elements identified across the rest of the alignments (**Supplementary Fig. 7**).

High-quality sequence across the variant breakpoints permitted a detailed assessment of exact variant boundaries and associated sequences. We used the breakpoint sequence data obtained from 152 insertions spanned by individual fosmid clones to identify a set of unique, diagnostic *k*-mers specific to the insertion and deletion alleles of each variant (**Fig. 5**). We found that 108 of the sequenced loci could be uniquely identified using a *k*-mer length of 36 and a search stringency of one substitution. We were not able to uniquely identify 29% of the loci (44 of 152) using this approach. We note, however, that this method assumes that the genome reference assembly accurately represents the structure of the deletion allele and that all instances of the variant have identical breakpoints. If *k*-mer lengths increased to 100 bp, there would still be five loci that remained recalcitrant to analysis using this approach (**Fig. 5b**). We determined genotypes for 106 loci by searching Illumina sequence data from NA18507 against these diagnostic *k*-mers[4]. We observed agreement at 94.3% of the genotypes determined for this individual by array CGH (**Fig. 5c** and **Supplementary Table 14**). We simulated the effect of genome coverage by sampling subsets of the total sequence data from NA18507 (**Fig. 5d**). We found a rapid increase in the number and accuracy of the sites genotyped with increasing coverage, followed by a plateau of approximately 94% genotype agreement when sequencing coverage reached tenfold sequence coverage. This indicates that high-quality breakpoint sequence data can be used to genotype structural variants in samples that have been analyzed by next-generation sequencing (NGS).

DISCUSSION

Over the past five years the extent of structural variation among individual human genomes has become increasingly clear. Array-based approaches, for example, have systematically discovered and genotyped more than 50% of common copy-number polymorphic deletions[23,24]. Sequence-based approaches have begun to more fully explore the size spectrum, cataloging an increasing number of smaller deletions and moving toward personalized duplication maps for individual genomes[9,11,25,26]. The characterization of other classes of structural variation, including inversions and insertions, however, has lagged owing to technical biases in their discovery and difficulties associated with their validation. New insertions are limited, in particular, by the genetic community's reliance on a single mosaic reference genome, which at some positions represents rare structural configurations and entirely omits sequences that are found in the majority of individuals. The absence of these sequences from the reference genome hinders their functional characterization, leading to a less-than-complete understanding of the sequence content present in the majority of humans. We used a fosmid clone strategy to specifically focus on the characterization of human sequences that are not in the reference assembly and have therefore not been annotated for functional elements or systematically genotyped.

In this study we identified 720 distinct loci ranging in length from 1 to 20 kbp, as well as several thousand additional smaller segments <1 kbp in length. We determined that more than half map to the euchromatin, with a disproportionate fraction mapping within the last 5 Mbp of human chromosomes (**Supplementary Fig. 1**). A remarkable feature of these sequences is their degree of copy-number polymorphism. Array CGH analysis indicates that 18–37% of the assembled sequence contigs vary in copy number, with 80% of the genotyped variants having a minor allele frequency >10% among the 28 individuals surveyed (**Fig. 3**). Experimental and computational comparisons with chimpanzee DNA suggest that at least 94% arose as a result of deletions that occurred within the human lineage.

Many of the common insertions show striking differences in allele frequency among populations, a pattern suggestive of either selection or genetic drift since the migration of humans out of Africa (**Fig. 2** and **Supplementary Tables 8** and **9**). We observed that the average insertion allele frequency for the variable loci was significantly greater in African populations than among European or Asians (YRI versus CEU $P = 0.0003$, and YRI versus ASN $P = 0.005$, one-sided *t*-test). The 3.9-kb novel insertion within the first intron of the *LCT* gene is illustrative. Our initial survey suggests that this insertion sequence is prevalent among the Yoruba (86%) and Asian samples (63%) but is present at a much lower frequency among CEPH Europeans (11%). These findings raise the possibility that the additional sequence within this haplotype may have a role in regulating expression of this gene. The complete sequence of this haplotype (AC20193) now allows this hypothesis to be directly tested.

An important question going forward is how well *de novo* assembly methods using next-generation sequence data compare to the clone-based approach we have described here. We had the opportunity to compare an Illumina SOAP *de novo* assembly[7] against the clone-based discovery on the same individual genome (**Supplementary Note**). We found that many of the larger new contigs were only partially represented (50–60%) in a 30× *de novo* assembly, and in more than a third of studied cases new contigs were fragmented—mapping to two or more scaffolds instead of being placed in the same region. In many cases, the fragmentation corresponded to common repeats disrupting the contiguity of the new sequence. In regions largely devoid of retrotransposons, *de novo* sequence assemblies using NGS data sets perform quite well. These results highlight both the limitations of *de novo* sequence assembly using NGS and the value of high-quality clone-based data to resolve and integrate these sequences into the reference genome. Nevertheless, there are advantages to *de novo* assembly. The *de novo* sequence assembly identifies two to three times more new sequence per genome than our results from 0.3× sequence coverage per genome, suggesting that the methods are complementary. Notably, only 2.9% of our singletons from NA18507 (average size ~790 bp) were identified in the *de novo* assembly. As these smaller insertions require more characterization, the significance of this discrepancy is unclear.

The major benefit of our approach is the ability to directly obtain high-quality sequence for the insertion loci by complete sequencing of corresponding clone inserts at a quality commensurate with that of the human reference genome. Although no

complete missing genes were discovered, we did identify 477 elements that have been conserved over evolutionary time, including six that appear to correspond to exons from RefSeq genes, as well as 26 loci having support from multiple mRNA-seq reads. Moreover, we demonstrate that these high-quality sequences can be used to accurately genotype these regions using next-generation sequence sets produced from the 1000 Genomes Project and other projects. The complete sequence of these and other loci will facilitate their functional characterization, as they can now be incorporated into future genotyping platforms, expression microarrays and ultimately genome assemblies to provide a more accurate representation of the organization and genetic variation of the human genome.

METHODS

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

Accession codes. Gene Expression Omnibus: GSE20634. GenBank: assembled contigs, GU266782–GU269144; fully sequenced fosmid insert codes are given in **Supplementary Table 10**.

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

ACKNOWLEDGMENTS

We thank C. Campbell, G. Cooper, T. Marques-Bonet for thoughtful discussion, P. Sudmant for assistance with Illumina sequence data and members of the University of Washington and Washington University Genomes Centers for assistance with data generation. J.M.K. is supported by a US National Science Foundation Graduate Research Fellowship. This work was supported by the US National Institutes of Health grant HG004120 to E.E.E. E.E.E. receives funds as an Investigator of the Howard Hughes Medical Institute.

AUTHOR CONTRIBUTIONS

J.M.K., N.S., F.A., A.T., R.K. and E.E.E. analyzed data. N.S., P.A., A.T., N.A.Y., P.T. and L.B. performed array CGH and copy-number analysis. F.A., M.V. and G.G. performed FISH experiments. C.A. assembled contigs. T.G., R.F., H.S.H., M.M., J.K., R.K. and R.K.W. performed clone characterization and sequencing. J.M.K., R.K., L.B. and E.E.E. designed the study. J.M.K. and E.E.E. wrote the paper with contributions from the other authors.

COMPETING FINANCIAL INTERESTS

The authors declare competing financial interests: details accompany the full-text HTML version of the paper at http://www.nature.com/naturemethods/.

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

1. International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. *Nature* **431**, 931–945 (2004).
2. Levy, S. *et al.* The diploid genome sequence of an individual human. *PLoS Biol.* **5**, e254 (2007).
3. Wheeler, D.A. *et al.* The complete genome of an individual by massively parallel DNA sequencing. *Nature* **452**, 872–876 (2008).
4. Bentley, D.R. *et al.* Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456**, 53–59 (2008).
5. Wang, J. *et al.* The diploid genome sequence of an Asian individual. *Nature* **456**, 60–65 (2008).
6. McKernan, K.J. *et al.* Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two-base encoding. *Genome Res.* published online, doi:10.1101/gr.091868.109 (22 June 2009).
7. Li, R. *et al.* Building the sequence map of the human pan-genome. *Nat. Biotechnol.* **28**, 57–63 (2010).
8. Hormozdiari, F., Alkan, C., Eichler, E.E. & Sahinalp, S.C. Combinatorial algorithms for structural variation detection in high-throughput sequenced genomes. *Genome Res.* **19**, 1270–1278 (2009).
9. Tuzun, E. *et al.* Fine-scale structural variation of the human genome. *Nat. Genet.* **37**, 727–732 (2005).
10. Eichler, E.E. *et al.* Completing the map of human genetic variation. *Nature* **447**, 161–165 (2007).
11. Kidd, J.M. *et al.* Mapping and sequencing of structural variation from eight human genomes. *Nature* **453**, 56–64 (2008).
12. Bovee, D. *et al.* Closing gaps in the human genome with fosmid resources generated from multiple individuals. *Nat. Genet.* **40**, 96–101 (2008).
13. The Chimpanzee Sequencing and Analysis Consortium. Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* **437**, 69–87 (2005).
14. Perry, G.H. *et al.* The fine-scale and complex architecture of human copy-number variation. *Am. J. Hum. Genet.* **82**, 685–695 (2008).
15. Weir, B.S. *Genetic Data Analysis II* (Sinauer, Sunderland, Massachusetts, USA, 1996).
16. Redon, R. *et al.* Global variation in copy number in the human genome. *Nature* **444**, 444–454 (2006).
17. Enattah, N.S. *et al.* Identification of a variant associated with adult-type hypolactasia. *Nat. Genet.* **30**, 233–237 (2002).
18. Pruitt, K.D., Tatusova, T., Klimke, W. & Maglott, D.R. NCBI Reference Sequences: current status, policy and new initiatives. *Nucleic Acids Res.* **37**, D32–D36 (2009).
19. Wang, E.T. *et al.* Alternative isoform regulation in human tissue transcriptomes. *Nature* **456**, 470–476 (2008).
20. Paten, B., Herrero, J., Beal, K., Fitzgerald, S. & Birney, E. Enredo and Pecan: genome-wide mammalian consistency-based multiple alignment with paralogs. *Genome Res.* **18**, 1814–1828 (2008).
21. Paten, B. *et al.* Genome-wide nucleotide-level mammalian ancestor reconstruction. *Genome Res.* **18**, 1829–1843 (2008).
22. Cooper, G.M. *et al.* Distribution and intensity of constraint in mammalian genomic sequence. *Genome Res.* **15**, 901–913 (2005).
23. McCarroll, S.A. *et al.* Integrated detection and population-genetic analysis of SNPs and copy number variation. *Nat. Genet.* **40**, 1166–1174 (2008).
24. Conrad, D.F. *et al.* Origins and functional impact of copy number variation in the human genome. *Nature* **464**, 704–712 (2009).
25. Korbel, J.O. *et al.* Paired-end mapping reveals extensive structural variation in the human genome. *Science* **318**, 420–426 (2007).
26. Alkan, C. *et al.* Personalized copy number and segmental duplication maps using next-generation sequencing. *Nat. Genet.* **41**, 1061–1067 (2009).
27. Parsons, J.D. Miropeats: graphical DNA sequence comparisons. *Comput. Appl. Biosci.* **11**, 615–619 (1995).

ONLINE METHODS

Discovery and characterization of 'novel insertions'. We identified unmapped fosmid end sequences from nine individuals as previously described[11]. Sequence contigs were assembled using phrap (http://phrap.org/). We designed two custom Agilent oligonucleotide arrays to target these sequences and performed a series of hybridizations using genomic DNA from sample NA15510 as a reference. We removed artifactual and low-signal probes on the basis of an analysis of single-channel fluorescence intensity and the correlation of probe responses across experiments. Additional contigs having high-identity BLAST hits against sequences from nonprimate species were also removed. We assessed copy-number polymorphism through both direct comparisons of array intensity values and a modal clustering method that attempts to fit array intensity data to distinct integer copy-number states. Additional details of array design, probe quality analysis and polymorphism calling are described in the **Supplementary Note**.

Orphan clones were identified from the G248 (NA15510) clone library. We obtained restriction profiles using four enzymes and used them to link individual clones into contigs using the Contig Builder program[12].

We determined unmapped clone end sequences relative to the build 35 genome assembly, whereas completely sequenced clones were compared against the more recent build 36 assembly (**Supplementary Table 9**). Two sequenced clones (AC234849 and AC226835) were tested and determined to be artifactual and were omitted from all analysis. We identified approximate breakpoint regions using the program miropeats[27] and refined them on the basis of review of an alignment of sequences extracted from the corresponding breakpoint regions.

Gene analysis. We downloaded previously reported mRNA-seq data[19] from the short-read archive and mapped it against the sequenced novel insertion using mrsFAST (http://mrfast.sourceforge.net/). Only reads that did not map against the build 36 genome sequence were considered. Splicing was ignored, and all mapped positions with up to two mismatches were recorded. We identified all segments with a depth of three or more reads.

Comparative analysis. Conservation analysis was based on the Ensembl Compara 51 alignments of nine mammalian genomes (http://www.ensembl.org). We found segments corresponding to the human insertions by BLAST and identified conserved elements using GERP version 2.1 (http://mendel.stanford.edu/sidowlab/downloads/gerp/index.html). All conservation analysis was limited to the eight nonhuman genome sequences for which alignments were available.

Breakpoint genotyping. We identified diagnostic *k*-mers to genotype each variant by searching overlapping 36-mers derived from sequenced breakpoints against a database of sequenced insertions and the build 36 assembly. Only *k*-mers with a single hit (permitting one substitution) were retained. To be genotypeable, a variant had to have at least one deletion *k*-mer and one insertion *k*-mer that met these criteria. We then searched Illumina sequence reads against this set of diagnostic *k*-mers (requiring a perfect match). A breakpoint search score was computed for each variant on the basis of the number of reads that matched *k*-mers derived from the insertion or deletion alleles.

doi:10.1038/nmeth.1451