nature genetics

Closing gaps in the human genome with fosmid resources generated from multiple individuals

Donald Bovee[1], Yang Zhou[1], Eric Haugen[1], Zaining Wu[1], Hillary S Hayden[1], Will Gillett[1], Eray Tuzun[2], Gregory M Cooper[2], Nick Sampas[3], Karen Phelps[1], Ruth Levy[1], V Anne Morrison[2], James Sprague[2], Donald Jewett[1], Danielle Buckley[1], Sandhya Subramaniam[1], Jean Chang[1], Douglas R Smith[4], Maynard V Olson[1,2], Evan E Eichler[2,5] & Rajinder Kaul[1]

The human genome sequence has been finished to very high standards; however, more than 340 gaps remained when the finished genome was published by the International Human Genome Sequencing Consortium in 2004. Using fosmid resources generated from multiple individuals, we targeted gaps in the euchromatic part of the human genome. Here we report 2,488,842 bp of previously unknown euchromatic sequence, 363,114 bp of which close 26 of 250 euchromatic gaps, or 10%, including two remaining euchromatic gaps on chromosome 19. Eight (30.7%) of the closed gaps were found to be polymorphic. These sequences allow complete annotation of several human genes as well as the assignment of mRNAs. The gap sequences are 2.3-fold enriched in segmentally duplicated sequences compared to the whole genome. Our analysis confirms that not all gaps within 'finished' genomes are recalcitrant to subcloning and suggests that the paired-end-sequenced fosmid libraries could prove to be a rich resource for completion of the human euchromatic genome.

The finished human genome sequence contains 2.85 billion nucleotides in euchromatic regions interrupted by 340 gaps, including 250 euchromatic gaps estimated to span 25 Mb, 33 heterochromatic gaps spanning over 200 Mb and 58 unfinished clones spanning 2.9 Mb[1]. Our detailed analysis of sequences surrounding the human genome gaps in the July 2003 build suggested that the euchromatic gap sequences may not be recalcitrant to cloning[2]. We noted that 56% of the unfinished segments and 52% of the clone gaps were flanked by segmentally duplicated sequences[3]. We thus found segmental duplication content to be the sequence feature that best predicted the location of a gap within the human genome. Comparison of the Celera and

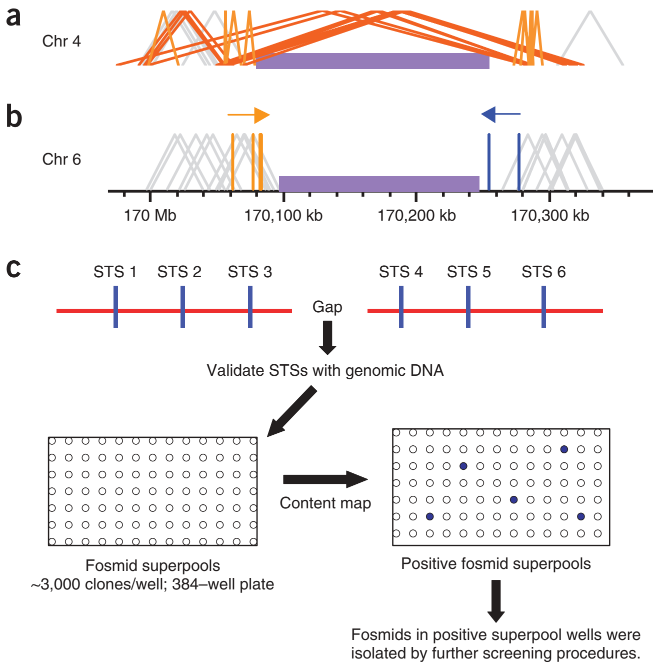
a
Chr 4
b
Chr 6
170 Mb
170,100 kb
170,200 kb
170,300 kb
c
STS 1
STS 2
STS 3
STS 4
STS 5
STS 6
Gap
Validate STSs with genomic DNA
Content map
Fosmid superpools
~3,000 clones/well; 384–well plate
Positive fosmid superpools
Fosmids in positive superpool wells were isolated by further screening procedures.

Figure 1 Strategies for closing or extending into gaps in the human genome. (**a**) Schematic for identification of fosmids spanning a gap by end-sequence-pair placement (red and orange) flanking a gap (purple). The presence of other discordant pairs (highlighted by the green bars) suggests structural variation. (**b**) Identification of clusters of singleton fosmids (yellow and blue) flanking a gap on chromosome 6. One end of each of these singletons maps in proximity to the gap, but the other end does not map to the human genome, despite the presence of high-quality sequence. (**c**) STS mapping strategy whereby fosmid libraries stored as superpools of 3,000 clones per well were interrogated with STS around gaps. Selected superpools were negative for STS 1 or STS 6, but not necessarily for STS 2 or STS 5. The fosmid clones spanning or extending into gaps were isolated by methods described earlier[27].

[1]Division of Medical Genetics, Department of Medicine, and [2]Department of Genome Sciences, University of Washington, Seattle, Washington 98195, USA. [3]Molecular Technology Laboratory, Agilent Technologies, Santa Clara, California 95051,USA. [4]Agencourt Sequencing Center, Agencourt Bioscience Corporation, 500 Cummings Center, Beverly, Massachusetts 01519, USA. [5]Howard Hughes Medical Institute, University of Washington, Seattle, Washington 98195, USA. Correspondence should be addressed to R.K. (rkkaul@u.washington.edu).

Received 12 March; accepted 19 September; published online 23 December 2007; doi:10.1038/ng.2007.34

public sequence genome assemblies showed that at least some of the gaps could be filled or extended into by the Celera data; however, in most cases, there was ambiguity about the final sequence because of the complex nature of the regions and the heterogeneity of donor DNA samples[4]. Structural variation is increasingly recognized as a significant contributor to human genetic disease and disease susceptibilities[5–20], and recent studies have also shown that structural variation is enriched in the gap regions of the genome[15]. Such structurally polymorphic regions, especially in duplicated regions, might also have contributed to creating gaps in the human genome assembly, particularly if the adjacent clones around a gap originated from two structural-variant haplotypes. In this study, we focused on closing the euchromatic gaps, and we identified previously unknown sequences that closed 26 of 250 gaps and extended into 67 other euchromatic gaps throughout the genome.

We used two different approaches to recover clones mapping near gaps in the human genome sequence (**Fig. 1** and Methods) based on the construction of genomic fosmid libraries from unrelated human DNA sources (**Supplementary Table 1** online). Complete insert sequencing of 36 clones from various fosmid libraries led to closure of 26 gaps on 14 human chromosomes (**Table 1**). For 21 of those, we obtained previously uncharacterized human sequences ranging in size from 404 bp to 64,987 bp, resulting in a total of 363,114 bp of previously unknown sequences (average gap size = 13,966 bp). Five other closed gaps apparently were not missing any sequence *per se*; rather, they represented situations in which the sequences in gaps were abutting with 0 bp of overlap or in which an apparent gap existed as a result of the haplotypic differences between clones on either side of the gap, even though no sequence was actually missing. Closure of gap g06_01 is an example of a potential haplotypic difference in genome structure (**Fig. 2**). In this case, simple sequencing of a gap-spanning fosmid clone could not unambiguously close the gap. Instead, it was necessary to remap the entire region in overlapping clones of the same haplotype; therefore, we sequenced 149,528 bp to resolve the gap, even though no new sequence was added in the process. We found a similar situation resulting from haplotypic differences around gap g19_02 on chromosome 19. The accessioned clones AC092315.2 and AC145203.1 flanking the left end of gap g19_02 are apparently from a haplotype different from that of the accessioned clone AC130468.2 that flanks the right end of the gap. Replacing clones AC092315.2 and AC145203.1 on the left end of gap g19_02 with the accessioned sequence AC196354 from the current study would close this putative

Table 1 Human euchromatic gaps closed using clones from fosmid libraries and identification of polymorphism

Chr	Gap	Left accession	Right accession	Gap size (bp)	Polymorphic (current study)	Detection method (current study)	Polymorphic (other studies, reference)	Accession
1	g01_28	AL604028.15	BX664740.3	8,683				AC160854
1	g01_48	AL356957.27	AL109948.9	16,503	Yes	PES_MCD[a]	10, 14, 15	AC157321
2	g02_01	AC144527.3	AC140476.2	4,547				AC161035
2	g02_02	AC141930.2	AC144450.2	34,307				AC160855, AC160856[b]
2	g02_11	AC099646.6	AC092683.3	13,729	Yes	PES_MCD	15	AC156159
2	g02_20	AC144525.3	AC149644.1	31,487				AC174049
3	g03_04	AC091633.22	AC069513.28	26,497	Yes	PES_MCD		AC160857
4	g04_03	AC147876	AC143342.3	0[c]	Yes	PES_MCD	15	AC157210
4	g04_08	AC142281.2	AC118275.4	404				AC195454
4	g04_13	AC134919.3	AC009570.13	0[c]				AC160858
6	g06_01	BX255934.7	AL138884.10	0[c]	Yes	Sequencing[d]/ PES_MCD/CGH[e]		AC160851, AC174439, AC174441, AC174438
9	g09_44	AL683798.22	AL669970.10	16,449				AC156789
10	g10_01	BX276080.7	AL365356.13	0[c]			15	AC156158
10	g10_21	AL691429.17	AL445199.37	11,055	Yes	CGH	15	AC158211
11	g11_06	AADB01066164	AADD01116830	64,987				AC154114, AC160860, AC154091
12	g12_07	AC073578.17	AC140063.13	2,638				AC155072
17	g17_08	AC060780.18	AC109326.11	0[c]				AC160862
17	g17_09	AC134407.6	AC145343.3	8,866				AC174157
19	g19_01	AC125387.2	AC126754.2	27,138				AC195455, AC196636
19	g19_02	AC130469.2	AC130468.2	13,660[c]	Yes	Sequencing		AC196364
20	g20_05	AL499627.23	AL449263.6	27,849				AC161429
22	g22_05	CR936488.1	CR954960.1	3,758				AC174156
22	g22_10	CR790389.1	CR932344.3	1,131				AC174074
X	gxx_02b	BX901949.9	BX908382.8	32,331	Yes	CGH		AC188045, AC188046
X	gxx_19	CR376833.3	CR394566.2	10,121				AC186562
X	gxx_21	BX936347	BX936365	6,974				AC160849
14	**26**			**363,114**	**8**			

Accessioned left and right clones surrounding each gap define the gap locations. Gaps with 0 bp of new sequence are cases where the newly sequenced clone(s) provided the necessary overlap to validate an end-to-end join of adjacent accessioned clones, or, alternatively, closed the gap by using clones from the same haplotypes, as was the case for gap g06_01. [a]Representative concordant and discordant fosmid clones from specific library were identified by PES span across closed gaps, and were analyzed further by MCD fingerprinting with *Eco*RI, *Bgl*II, *Hin*dIII and *Nsi*I restriction enzymes to estimate their average insert sizes[7]. The data are shown in **Supplementary Table 2**. [b]Multiple overlapping clones were sequenced to span the gap. [c]CGH was not carried out for gaps g04_03, g04_13, g06_01, g10_01 and g17_08. g19_02 sequence was not included in CGH analysis. [d]Fosmid clones sequenced to close the gap were found to be polymorphic when compared to the reference genome sequences flanking the gaps in build hg17. [e]CGH: Previously unknown sequences identified in gap closing experiments were tiled as oligonucleotide microarray. Comparative genome hybridization (CGH) was carried out using WIBR2 source DNA as control against genomic DNA for each of the ABC7–ABC12 libraries. Data are shown in **Supplementary Figure 3**.

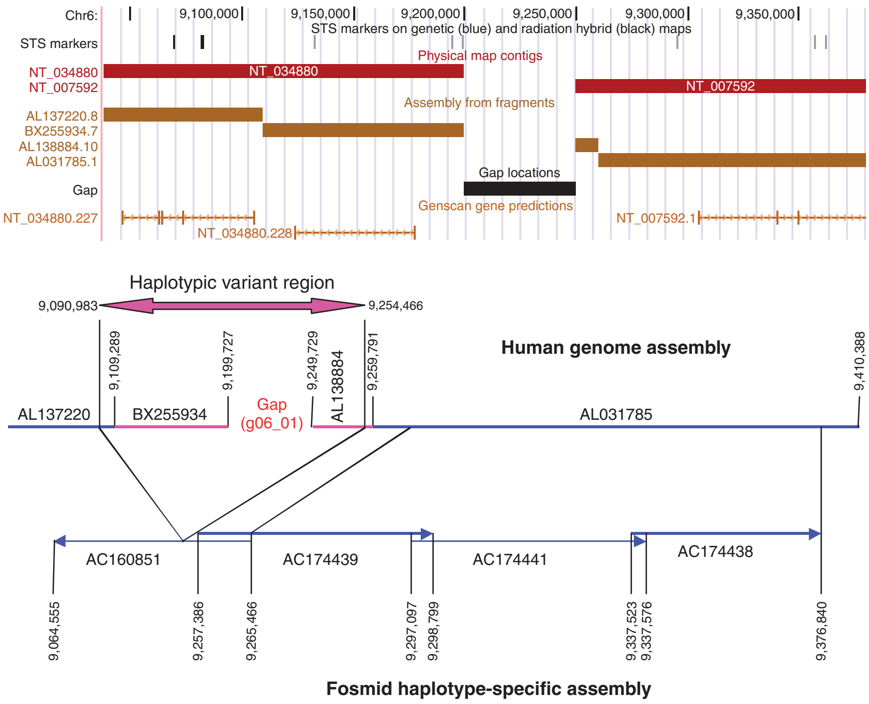

Figure 2 Closing euchromatic gaps using haplotype-specific sequence assemblies. Human chromosome 6 gap between 9,199,727 and 9,249,729 (build hg17). Top, a snapshot of human genome assembly build hg17 around gap g06_01. Bottom, a schematic representation of the four fosmid clones derived from the WIBR2 fosmid library that were sequenced to create a 149,528 bp sequence contig that spanned across gap g06_01. The assembled contig is haplotypically different from the build hg17 or hg18 human genome assemblies. The gap was found to be created as a result of structural differences in BACs around this gap that were derived from different source libraries (top panel). Segmental duplication sequences in BAC clones further confounded the sequence assembly. The sequenced fosmid clones closed the gap without adding any new sequence to the current human-genome assembly. The fosmid sequences were submitted to GenBank and also communicated to chromosome coordinators at sequencing centers.

13,660 bp gap in the current human genome assembly build ‘hg18’ (see URLs section in Methods). Of course, our method strongly biased success toward gaps that could be spanned by a single fosmid; 13 of the 26 closed gaps required sequencing of a single fosmid clone. Gap-closing sequences for gaps g01_28, g01_48, g02_02, g02_11, g09_44, g10_21, g12_07 and gxx_21 have already been incorporated in current build hg18 and can be accessed through the University of California Santa Cruz browser.

We found gaps g01_48, g02_11, g03_04 and g04_03 to be polymorphic by paired-end-sequence (PES) and multiple-complete digestion (MCD) analysis[7,21,22] of the ABC7–ABC12 fosmid libraries[21], whereas we found gaps g10_01 and gxx_2b to have copy number variation (CNV) by oligonucleotide microarray comparative genome hybridization (CGH) analysis of the source DNA samples used for ABC7–ABC14 fosmid libraries[21] (**Table 1**, **Fig. 3** and **Supplementary Table 2** online). Non-overlaps in the detection of structural variants by different technologies are well recognized[9]. The g01_48, g02_11 and g03_04 sequences were composed entirely of complex segmental duplications, and their polymorphism was thus not detectable by CGH; g04_01 was not included in the CGH analysis. The gap g06_01 sequence version presented here seemed to be the haplotype represented in the analyzed DNA samples and is different from that represented in build hg17 or 18. Including the g19_02 gap described earlier, 30.7% (8 of 26) of the closed gaps thus seemed to be structurally polymorphic; some of these regions were reported to be polymorphic by other investigators[10,14,15] (**Table 1**). In addition, gap g10_01 has also been reported to have CNV[15], making a total of 39% of the closed gaps polymorphic.

We extended sequences into 67 euchromatic gaps by identifying fosmids where one end sequence was oriented and anchored adjacent to a gap but where the other end did not map against the human genome despite the presence of high-quality sequence (**Fig. 1b** and **Supplementary Tables 1** and **3** online). Most (92 of 110) of the fosmids identified were from the paired-end sequenced WIBR2 fosmid library. In all, 2,125,728 bp of previously uncharacterized sequence across 20 chromosomes were added, with an average extension of 31,727 bp per gap. For 16 gaps, we obtained sequence from both proximal and distal ends, and through recurrent mapping of end-sequences, it became possible to build sequence contigs in some of the gaps. Gaps g06_05 and g07_01 had the longest extensions, of 320,525 and 223,365 bp, respectively; however, neither gap is as yet closed.

We validated the origin and structure of these previously uncharacterized human sequences by BLAST similarity searches against the chimpanzee whole-genome shotgun trace archive (see URLs section in Methods). All gap sequences were well represented, with an average of 5.96-fold sequence coverage (excluding the regions highly enriched in segmental duplications). The gap sequences also mapped to the expected syntenic regions of the chimpanzee genome build 2.1 (see URLs section in Methods), with sequence identity ranging from 96% to 98% (data not shown). To validate the long-range sequence integrity, we analyzed all 26 gap-closure regions by mapping paired-end sequences from the ABC7–ABC12 human fosmid libraries generated as part of the Human Genome Structural Variation Project[21]. Overall, 208–489 fosmid clones per library (**Table 2**), and a total of 2,158 fosmids (**Supplementary Table 4** online), were identified that spanned the 26 closed gaps. The structure and organization of each gap region has been confirmed in at least one additional human genome source. Notably, the depth of clone coverage across the sequenced gaps varies considerably, suggesting biases in subcloning efficiency and/or structural variation among different humans. In addition to their utility in finding structural variants in the human genome[7,23], these data highlight the utility of paired-end sequence as clone resources to close gaps in the human genome. Analyses so far of the ABC libraries have identified fosmids that would potentially span 20 additional gaps not previously covered by the WIBR2 fosmid library (unpublished results); 16 of these gaps have been reported to reside in copy number variant regions[5–8,14,15]. It thus seems that the extent of polymorphism in euchromatic gaps in the human genome may have been underappreciated.

For each fosmid clone that extended into a gap, we extracted flanking sequence from both sides of the remaining gap, whereas for regions that spanned a gap, we analyzed sequence within the gap for the presence of segmentally duplicated sequences. We estimate that 11.8% (294 of 2,488 kb) of the gap sequences correspond to segmental duplications when a sequence-identity threshold of >95% is used (**Supplementary Table 5** online). This level of segmental duplication represents a ~2.3-fold increase with respect to the genome-wide average of 5.2%. It has been previously reported that compared to a genome-wide average of 5% of the human genome consisting of segmental duplications, 54% of the gaps are flanked by segmental duplications[2]. Ten of the 87 gap sequences consisted almost entirely of duplications, whereas the duplicated portions in each of the four other gaps exceeded 60%. These corresponded to duplications with high average sequence identity (98.7%) and were most frequently embedded within complex blocks of segmental duplication[24]. Our estimate of duplication content is conservative, as many of the gaps in the human genome are relatively short (median 24.5 kb) and, therefore, below the limit of duplication detection using the whole-genome shotgun sequence-detection (WSSD) method.

By BLAST analysis of the gap sequences against the EST database, we identified several previously uncharacterized spliced ESTs (**Supplementary Table 6** online). Seven gap sequences had sequence similarity to spliced ESTs with at least two potential exonic sequences; one gap sequence, g06_08, contained three exonic sequences. BLASTX comparison of repeat-masked sequence against the protein database shows sequence matches of gap g02_08 to the gene encoding the immunoglobulin kappa chain V-I region protein, and the g02-20 sequence matched the gene encoding twist-related protein 2 (Dermo-1). None of the other five gap sequences with matches against the spliced EST database corresponded to entries in the protein database. We next

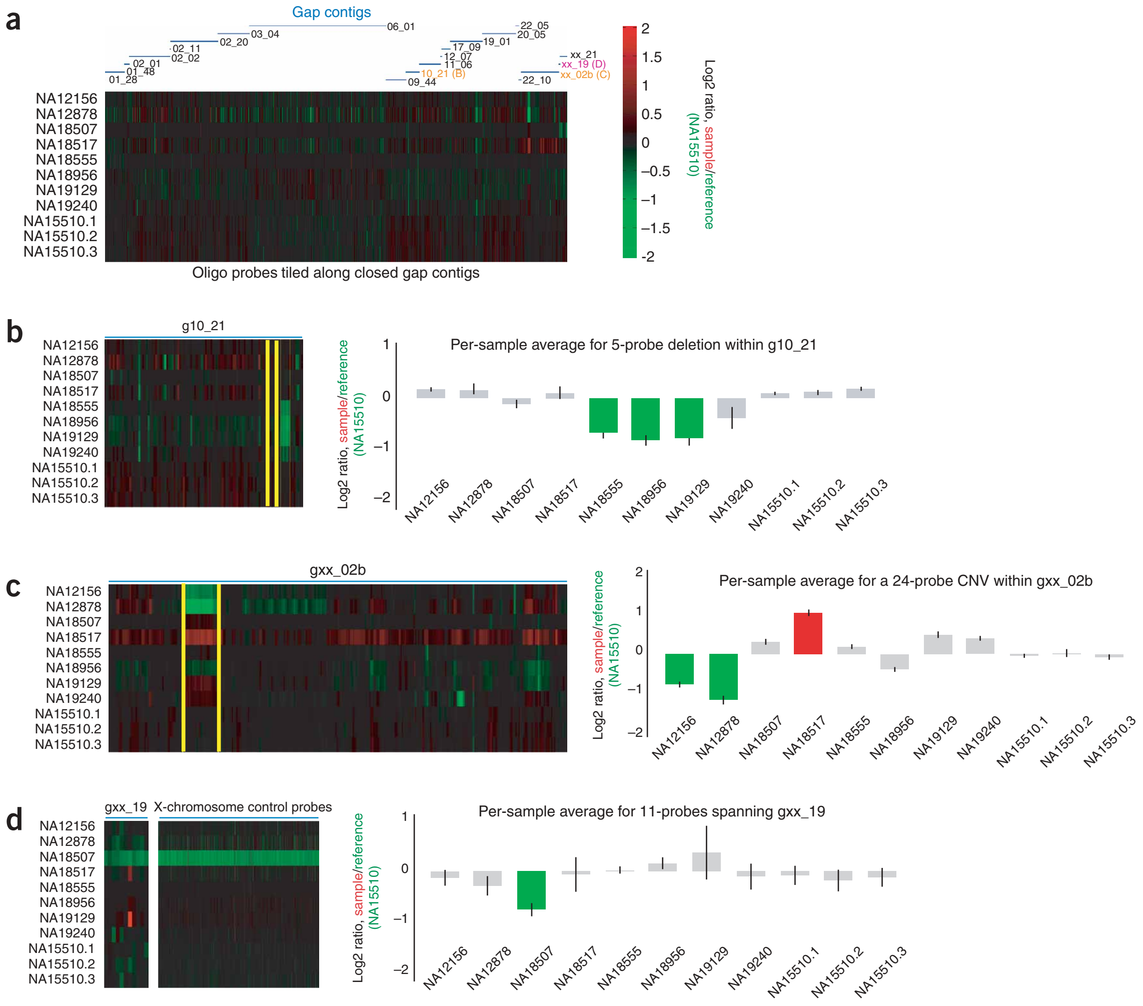

Figure 3 Overview of CGH experiments for the closed gap-contigs. DNA from NA15510 (a female) was used as reference against the eight HapMap samples. (**a**) Normalized log2 ratios of sample to reference for all probes within closed gap contigs are presented as a heatmap. The color scale, shown on the right, is scaled between –2 and +2 with excess sample indicated as red. The bottom three rows show data for the three self-self experiments carried out using NA15510 as both 'sample' and 'reference'. Each closed gap is indicated along the top as a horizontal bar. Those gaps showing strong evidence for harboring copy number polymorphisms are colored orange; a gap contig (gxx_19) showing male hemizygosity on the X chromosome is labeled purple. (**b**) Detailed view for g10_21. A likely deletion is highlighted with vertical yellow bars, and quantification of the average log2 ratios for each sample is shown below the heatmap. Error bars correspond to 3 s.e.m. units above and below the mean. Samples that are significantly depleted with respect to the reference NA15510 are colored green. (**c**) Similar to **b**, except for gxx_02b located on the X chromosome. There are both losses (green) and gains (red) with respect to NA15510 within the highlighted region. NA18507 (a male) carries the same number of copies as the reference NA15510 (a female), suggesting that the entire interval shows CNV in the human population. The persistent excess (red color) indicated along most of this interval in NA18517 (mean log2 ratio of the whole contig is 0.43, s.e.m. = 0.017) corroborates this, although it falls just short of our established threshold for calling a 'gain'. (**d**) The closed gap gxx_19 on the X chromosome and ~1,300 control probes scattered along the X chromosome show the expected 'loss' signal for NA18507 (a male) compared to NA15510 (a female).

examined whether any of the gap sequences mapped within the introns of assigned human genes, mRNA or spliced ESTs (**Supplementary Table 7** online). Of the 76 gaps thus analyzed, 29 gaps had annotations for known genes, mRNA or spliced ESTs that spanned across the gap. Ten of these 29 gaps have been closed, thereby completing the genomic sequence of the genes they lie within. The remaining 19 gaps that are spanned by annotated genes or spliced ESTs have not yet been closed.

It has previously been estimated that some 6% of the human sequences transform poorly into *Escherichia coli* and argued that transformation-associated recombination (TAR) in yeast cells be used for cloning such sequences from the human genome[25]. It was further suggested that the loss of some BAC sequences cloned in *E. coli* might result in misassemblies in the human genome[25]. TAR technology has been used to close four gaps in human chromosome 19 in build hg15 (ref. 26). Arguably, these gaps were recalcitrant to cloning in large-insert *E. coli*–based vectors. We examined whether or not the four chromosome 19 gaps closed[26] in build hg15 are spanned by fosmids derived from the currently available paired-end sequenced fosmid libraries[7,21]. We found that fosmids from the WIBR2, ABC7 and ABC8 libraries spanned three of these four gaps in build hg15, whereas for the fourth gap closed by TAR cloning (build hg18 coordinates 14,565,069 to 14,582,756), only fosmid clones from the ABC7 library spanned the region. Hence the inability to detect large-insert clones following extensive searches in multiple libraries is often a poor criterion from which to assume the nonclonability of such sequences in *E. coli*. We believe that the analysis of paired-end sequence from additional fosmid libraries would be a useful resource for resolving gap regions of the human genome that are difficult, structurally variant or both.

Table 2 Distribution of fosmid clones from six PES fosmid libraries that span the sequences listed in Table 1 and reported in this study

Gap	ABC7 NA18517	ABC8 NA18057	ABC9 NA18956	ABC10 NA19240	ABC11 NA18555	ABC12 NA18555	Total fosmids
g01_28	2	8	9	5	4	8	36
g01_48	32	54	56	51	45	40	278
g02_01	8	13	7	3	5	4	40
g02_02	5	26	13	15	19	18	96
g02_11	10	30	13	12	15	21	101
g02_20	11	22	16	13	13	19	94
g03_04	18	65	37	46	44	37	247
g04_03	10	11	8	6	10	11	56
g04_08	0	2	2	1	1	2	8
g04_13	2	2	2	2	0	2	10
g06_01	21	50	25	42	29	28	195
g09_44	3	13	4	26	7	12	65
g10_01	6	28	4	15	22	28	103
g10_21	17	27	29	31	31	37	172
g11_06	0	1	1	0	1	1	4
g12_07	9	29	18	22	18	2	98
g17_08	3	11	2	7	19	3	45
g17_09	4	14	8	10	6	6	48
g19_01	1	0	1	5	1	0	8
g19_02	10	24	8	23	16	20	101
g20_05	10	24	8	23	16	20	101
g22_05	0	1	5	4	2	2	14
g22_10	4	6	9	4	10	16	49
gxx_02b	15	19	10	7	8	1	60
gxx_19	3	3	3	8	6	7	30
gxx_21	5	14	17	20	18	34	108
Total	**208**	**489**	**489**	**390**	**369**	**380**	**2,158**

Fosmid libraries and their source DNAs are both listed in the column headers. The tabulated fosmid clones spanning closed gaps, along with their corresponding paired-end-sequences trace archive accession numbers, are listed in **Supplementary Table 4**.

The current study reports closure of 10% of euchromatic gaps and identifies 30.7% of the closed gaps as polymorphic. By our estimates, at least 40 other euchromatic gaps have been closed in build hg18 by others. At present about 184 euchromatic gaps remain in build hg18; of these, about 68 are flanked by segmentally duplicated sequences. The balance of 116 euchromatic gaps includes 67 gaps for which partial sequences have been provided in the current study. The slow progress in closing human genome gaps reflects the complexity of the task, and the efforts to close additional gaps will undoubtedly become increasingly difficult as more tractable targets are exhausted. We suggest that their resolution, in part, will require identification of haplotype-specific tiling paths, including the resequencing of regions bracketing the gap. The paired-end-sequenced fosmid libraries[7,21] will potentially be helpful for identification of fosmid clones that would close or further extend into these gaps. Closure of these gaps would allow for creating the appropriate gene models, a basic step in the analysis of gene functions. The missing sequences in gaps could also contain regulatory or other conserved elements with functionally important biological roles. Closure of the remaining euchromatic gaps remains an unmet, yet tractable, challenge in defining the 'book of life' for humans.

METHODS

Gap closure and sequence extension. Fosmid clones derived from different human individuals were used in this study (**Supplementary Table 1**). We used two independent fosmid-based approaches to identify clones that would either span gaps, or, alternatively, allow extension of DNA sequences into gaps (**Fig. 1**). Three unique STSs from DNA sequences flanking either end of the gap were designed for 28 gaps. We designed the STSs near the gap ends to be within 500 bases or as close to the gap ends as possible. The other two STSs at either end of the gap were designed to be within 5–15 kb of the first STS. The STSs were validated and used for isolating unique gap-spanning fosmid clones or clones that extend into the gap as described earlier[27]. Alternatively, we used paired-end sequence data from human fosmid libraries WIBR2, ABC7 and ABC8 to identify clones that spanned a gap[7,21] or to identify clones where one end sequence was anchored within the human genome assembly but the other end sequence did not map to the human reference sequence, despite the presence of high-quality sequence data. Clusters of such singletons at the edges of gaps were considered candidates for extensions into gaps (**Fig. 1b**). As sequences from clones extending into gaps became available, we used the sequences to identify additional fosmids from paired-end sequence data for further sequence extensions.

Oligonucleotide microarray comparative genome hybridization. Eight HapMap samples, NA12156 (ABC14), NA12878 (ABC12), NA18057 (ABC8), NA18517 (ABC7), NA18555 (ABC11), NA18956 (ABC9), NA19129 (ABC13) and NA19240 (ABC10), that are currently being analyzed by fosmid pair end-sequence analysis[21] were analyzed via microarray comparative genomic hybridization using Agilent Technologies oligonucleotide array-CGH platform. We tiled a total of 3,850 isothermal oligonucleotide probes (T_m = 80 °C, size range 45–60 bp) across 20 closed gap-completion contigs, including the gap g06_01 where no new sequence was added but a haplospecific tiling path yielding 149,528 bp was sequenced to close the gap. The average and median spacings between these probes were 100 bp and 47 bp, respectively. We used a variation of Agilent's probe design algorithm (see URLs section below) to generate candidate probes; however, as these target sequences do not exist within the reference assembly, the methodology was altered accordingly. We generated a minimum of five probes (g02_11) and a maximum number of 1,134 probes (g06_01), with the median being 112 probes per gap sequence. The sample NA15510 (WIBR2 fosmid library[7]) was used as the reference sample in all experiments. We carried out pairs of dye-reversed hybridizations for each sample relative to the reference, and we did three self-self (NA15510 used as both 'sample' and 'reference') analyses for each set of hybridization reactions (see URLs section below). The normalization procedure considered all probes on the array (~240,000, 88% of which were tiled to regions of the reference assembly), not just the 3,850 considered here. Nevertheless, the gap interval probes are somewhat more noisy and seem to be more susceptible to subtle dye-bias effects than the bulk of the reference assembly probes (**Fig. 3**; compare **a** to the X-chromosome control probes in **d**), which may be a consequence of the duplication-rich nature of the gap-fill sequences. Although future efforts may yield improved normalization, we considered a sample to show significant evidence for a loss (or gain) in a given interval only if the absolute value of the mean sample-to-reference log2 ratio exceeded 0.4 by at least 3 standard error units (s.d. of probe measurements within the interval divided by the square root of the number of probes). With this threshold, there are no calls within the three self-self comparisons for any of the gap contigs, establishing this as a stringent empirically derived threshold.

Sequence analyses. We estimated the duplication content of gap sequence using the whole-genome shotgun sequence-detection strategy[4]. A database of 16,222,458 human whole-genome shotgun-sequence reads (approximately threefold sequence coverage of the human genome) was obtained from the National Center for Biotechnology Information trace repository (see URLs section below) and used to establish autosomal-duplication thresholds of 53.1 reads per 5-kb window (mean 30.3 ± 7.6 kb). Statistically significant departures in the corresponding depth of coverage by these reads (two of three 5-kb windows exceeding a 3 s.d. threshold) were used to estimate duplication content (>94% sequence identity). We assessed the completeness of annotated genes or spliced ESTs in proximity to each gap that was closed or extended. Gap sequences were searched against the human EST database by BLASTN sequence similarity. Repeat-masked spliced ESTs were searched by BLASTX against the protein database to discover potential protein-encoding domains (blastall -d /mnt/gpfs/blastdb/swissprot -e 1e-15 -p blastx).

URLs. Chimpanzee whole-genome shotgun trace archive, http://www.ncbi.nlm.nih.gov/blast/Blast.cgi; Chimpanzee Genome Resources, http://www.ncbi.nlm.nih.gov/projects/genome/guide/chimp/; human genome assembly Build hg18, http://genome.ucsc.edu/cgi-bin/hgTracks; Agilent probe design algorithm, https://www.opengenomics.com/webcastinfo.aspx?wid = 21; details on the hybridization and normalization procedures, http://www.chem.agilent.com/scripts/literaturePDF.asp?iWHID = 39980 and http://opengenomics.com/pdf/pn_gs_feature_extraction.pdf; National Center for Biotechnology Information trace repository, http://www.ncbi.nlm.nih.gov/Traces.

Accession codes. National Center for Biotechnology Information Gene Expression Omnibus: Microarray data from array-CGH analysis have been deposited with accession code GSE9075. All other DNA sequence data in the study have been submitted to GenBank, and the accession codes are listed in **Table 1** and **Supplementary Table 3**.

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

ACKNOWLEDGMENTS

The authors would like to acknowledge the University of Washington Genome Center staff and A.Yamada at Agilent Technologies for technical assistance. This work was supported by US National Institutes of Health grants 3 U54 HG002043 to M.V.O. and HG004120 to E.E.E. and M.V.O. E.E.E. is an investigator of the Howard Hughes Medical Institute.

AUTHOR CONTRIBUTIONS

R.K. and M.V.O. designed the overall study. R.K. oversaw the overall data production and analysis. R.K. and E.E.E. carried out data analysis and wrote the manuscript with comments from M.V.O. D. Bovee was responsible for data curation, identification of clones from paired-end-sequenced fosmid libraries and incorporation of publicly available sequence data. E.H. and D.J. provided the informatics support and submission of the sequence data to Genbank. Y.Z. and Z.W. were responsible for finishing the fosmid clones. J.C. was responsible for custom fosmid library constructions, and R.L., D. Buckley and S.S. generated shotgun-sequencing data. H.S.H., W.G. and K.P. generated and analyzed the MCD fingerprint data. G.M.C. and N.S. designed and analyzed the aCGH data; E.E.E., E.T., V.A.M. and J.S. initially identified and cherry-picked the fosmid clones from G248 and ABC libraries. D.R.S. was responsible for generating ABC fosmid libraries and their paired-end-sequence data.

Published online at http://www.nature.com/naturegenetics
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. Eichler, E.E., Clark, R.A. & She, X. An assessment of the sequence gaps: unfinished business in a finished human genome. *Nat. Rev. Genet.* **5**, 345–354 (2004).
3. Bailey, J.A. *et al.* Recent segmental duplications in the human genome. *Science* **297**, 1003–1007 (2002).
4. Istrail, S. *et al.* Whole genome shotgun assembly and comparison of human genome assemblies. *Proc. Natl. Acad. Sci. USA* **101**, 1916–1921 (2004).
5. Sebat, J. *et al.* Large-scale copy number polymorphism in human genome. *Science* **305**, 525–528 (2004).
6. Iafrate, A.J. *et al.* Detection of large-scale variation in the human genome. *Nat. Genet.* **36**, 949–951 (2004).
7. Tuzun, E. *et al.* Fine-scale structural variation of the human genome. *Nat. Genet.* **37**, 727–732 (2005).
8. Sharp, A.J. *et al.* Segmental duplications and copy number variation in human genome. *Am. J. Hum. Genet.* **77**, 78–88 (2005).
9. Eichler, E.E. Widening the spectrum of human genetic variation. *Nat. Genet.* **38**, 9–11 (2006).
10. Conrad, D.F., Andrews, T.D., Carter, N.P., Hurles, M.E. & Pritchard, J.K. A high-resolution survey of deletion polymorphism in the human genome. *Nat. Genet.* **38**, 75–81 (2006).
11. Freeman, J.L. *et al.* Copy number variation: new insights in genome diversity. *Genome Res.* **16**, 949–961 (2006).
12. Hinds, D.A., Kloek, A.P., Jen, M., Chen, X. & Frazer, K.A. Common deletions and SNPs are in linkage disequilibrium in the human genome. *Nat. Genet.* **38**, 82–85 (2006).
13. Khaja, R. *et al.* Genome assembly comparison identifies structural variants in the human genome. *Nat. Genet.* **38**, 1413–1417 (2006).
14. McCarroll, S.A. *et al.* International HapMap Consortium. Common deletion polymorphisms in the human genome. *Nat. Genet.* **38**, 86–92 (2006).
15. Redon, R. *et al.* Global variation in copy number in the human genome. *Nature* **444**, 444–454 (2006).
16. Feuk, L., Carson, A.R. & Scherer, S.W. Structural variation in the human genome. *Nat. Rev. Genet.* **7**, 85–97 (2006).
17. Aitman, T.J. *et al.* Copy number polymorphism in *Fcgr3* predisposes to glomerulonephritis in rats and humans. *Nature* **439**, 851–855 (2006).
18. Gonzalez, E. *et al.* The influence of CCL3L1 gene-containing segmental duplications on HIV-1/AIDS susceptibility. *Science* **307**, 1434–1440 (2005).
19. Stefansson, H. *et al.* A common inversion under selection in Europeans. *Nat. Genet.* **37**, 129–137 (2005).
20. Fellermann, K. *et al.* A chromosome 8 gene-cluster polymorphism with low human beta defensin 2 gene copy number variations predisposes to Crohn's disease of the colon. *Am. J. Hum. Genet.* **79**, 439–448 (2006).
21. Eichler, E.E. *et al.* Completing the map of human genetic variation. A plan to identify and integrate normal structural variation into the human genome sequence. *Nature* **447**, 161–165 (2007).
22. 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).
23. Newman, T.L. *et al.* High throughput genotyping of intermediate-size structural variation. *Hum. Mol. Genet.* **15**, 1159–1167 (2006).
24. She, X. *et al.* A preliminary comparative analysis of primate segmental duplications shows elevated substitution rates and a great-apes expansion of intrachromosomal duplications. *Genome Res.* **16**, 576–583 (2006).
25. Kouprina, N. *et al.* Segments missing from the draft human genome sequence can be isolated by transformation-associated recombination closing in yeast. *EMBO Rep.* **4**, 257–262 (2003).
26. Leem, S.-H. *et al.* Closing the gaps on human chromosome 19 revealed genes with a high density of repetitive tandemly arrayed elements. *Genome Res.* **14**, 239–246 (2004).
27. Raymond, C.K. *et al.* Targeted haplotype resolved resequencing of long segments of the human genome. *Genomics* **86**, 759–766 (2005).