AN ASSESSMENT OF THE SEQUENCE GAPS: UNFINISHED BUSINESS IN A FINISHED HUMAN GENOME

Evan E. Eichler, Royden A. Clark and Xinwei She

Biological research increasingly depends on 'finished' genome sequences. Deducing what is absent from these sequences is not trivial. More than 99% of the euchromatic portion of the human genome is now represented as a high-quality finished sequence with each base ordered and oriented. However, two principal types of gap remain: heterochromatic (estimated to be ~200 Mb) and euchromatic (23.0 Mb) gaps. Here, we use various global sources of data to help understand the nature of the gaps in the finished human genome. Not all gaps are recalcitrant to subcloning, nor are most heterochromatic. The presence of recent segmental duplications is the most important predictor of gap location in euchromatic sequences. The resolution of these regions remains an important challenge for the completion of the human genome, gene annotation and SNP assignment.

HETEROCHROMATIN
Parts of chromosomes with an unusual degree of contraction and that consequently have different staining properties from the euchromatin at nuclear divisions. Largely composed of repetitive DNA, heterochromatin forms dark bands after Giemsa staining.

Department of Genetics, Center for Computational Genomics, Case Western Reserve University School of Medicine and University Hospitals of Cleveland, BRB720, 10900 Euclid Avenue, Cleveland, Ohio 44106, USA.
Correspondence to E.E.E. e-mail: eee@cwru.edu

doi:10.1038/nrg1322

The completion of the human genome in April 2003 marked one of the most significant accomplishments in biology[1,2]. More than 99% of the euchromatic portion and ~94% of the total genome is now represented as high-quality finished sequence with each base ordered and oriented. What can be said about the remainder of the sequence? Simplistically, there are two types of gap in the genome: HETEROCHROMATIC (estimated to be ~200 Mb) and EUCHROMATIC (24.4 Mb) gaps (TABLE 1; International Human Genome Sequencing Consortium, manuscript in preparation). Heterochromatic regions were never intended to be targets of the Human Genome Project. Gaps in this *sequence non grata,* therefore, were fully anticipated in the early phases of the project[3]. Centromeric SATELLITE DNA and ACROCENTRIC portions of human chromosomes were included in this category. Despite the functional implication of the term 'heterochromatin', heterochromatic regions assumed a *de facto* sequence definition as those large regions of the genome that are populated almost exclusively by tandem repeats. By contrast, euchromatic portions of the genome contained the genes. Of course, such operational definitions became more and more wooly as the sequencing of chromosomes neared completion, and the transition between euchromatin and heterochromatin became impossible to delineate in an assortment of genes, large tracts of duplicated sequence and islands of intermittent satellite sequence. In the end, workers in the genome centres operationally recognized only two types of gap: those that could be closed within existing cloning vectors but that might be difficult to assemble ('finishing gaps'), and those that could not be traversed in bacterial artificial chromosome (BAC)- and cosmid-cloning vectors ('clone gaps'). Here, we present a detailed analysis of the gaps in the July 2003 assembly of the human genome and review our current understanding of the nature of these gaps. Although the euchromatic regions remain active targets of directed finishing, it is still a matter of debate as to whether there is sufficient need to justify the cost of sequencing all heterochromatic regions. Despite their intractable nature, the available data are revealing important glimpses into the complex biology and pathology of these unfinished regions.

Sequence gaps and segmental duplications

One of the surprising findings from the analysis of the initial working draft sequences was that 5% of our genome consists of segmental duplication[4,5]. Segmental

Table 1 | **Gap representation (Mb)**

Gaps	Total	Clone gaps*	Finishing gaps‡
Heterochromatin§	~200	~200	0
Telomere‖	~1.0	~1.0	0
Euchromatin¶	24.4	19.7	4.6
Unique	9.7	7.9	1.8
Duplicated	14.6	11.8	2.8

*Clone gaps are considered to be recalcitrant to subcloning and/or sequencing. ‡Finishing gaps are gaps for which a spanning clone that is being sequenced has been identified. §Heterochromatin was defined as higher-order arrays of centromeric α-satellite DNA — that is, tandem arrays of satellite DNA that are found in secondary constrictions and acrocentric portions of the genome. Gaps that are contiguous with heterochromatin were classified as heterochromatic. In this analysis, variable C-staining pericentromeric regions (19q12, 3q11.2, and so on) were classified as heterochromatic. Their median size is unknown and so their contribution to the overall genome size can only be estimated (~6 Mb). ‖Telomere denotes gaps at the ends of chromosomes that did not traverse into the TTAGGG sequence. ¶Duplication content of gaps was estimated on the basis of the gap size and whether the gap was flanked on either side by a duplicated sequence. Gap regions that were not flanked on either side by a duplication were deemed unique.

duplications have been defined as fragments of genomic sequence with high sequence identity (>90% and 1 kb) that map to multiple regions. It was clear that the largest (>100 kb) and most identical (>99%) segmental duplications would complicate the sequencing and assembly of the human genome in predictable ways[6]. Both experimental and computational analyses of the working draft sequence[4,7–9] confirmed that such regions could be inadvertently collapsed, leading to misassembly and concomitant gaps in other parts of the genome assembly. More fundamentally, such regions were initially underrepresented among the underlying clone-ordered reference sequences[7,10]. Later, it was discovered that some of these areas could be subject to large structural rearrangements (frequent inversions, amplifications and deletions) that range from a few kb to hundreds of kb in size[11–14]. The sequencing of these regions was further complicated by the fact that not only did duplicated copies need to be distinguished, but in some cases, continuity within the same haplotype would be required to achieve closure. Extra investment and resources were levied to resolve these regions in the last two years of the project[15–20]. An assessment of segmental duplication content by an assembly-independent approach[5] reveals that the finished genome assembly has correctly identified the location of most of these segmental duplications (International Human Genome Sequencing Consortium, manuscript in preparation). On the basis of our analysis of segmental duplications in the human genome assembly, we examined the nature of sequence gaps in the human genome with respect to duplications.

***In silico* analysis.** We considered all gaps (5 kb–3 Mb) in the human genome (according to the July 2003 assembly) and their distribution by chromosome and location (euchromatin/heterochromatin). We counted a total of 379 gaps in the assembled sequence. This includes 9 heterochromatic/acrocentric regions (136 Mb), 25 telomeric regions (estimated to be 798 kb), 24 centromeres (65 Mb) and the remaining 321 gaps in putative euchromatic regions of the genome (24.4 Mb) (TABLE 1). These estimates are slightly discrepant with the published genome analysis owing to differences in the accounting and classification of gaps in heterochromatin (TABLE 1). Whether or not duplications flanked the euchromatic gaps was assessed using two complementary methods: whole-genome assembly comparison and whole-genome shotgun sequence detection. A sequence assembly gap was scored positive for duplication if a PARALOGOUS sequence was identified within 5 kb from a gap. By count, ~54% (173/321) of the sequence gaps are flanked by segmental duplications (FIG. 1). By sequence length, 41% of the 33.9 Mb of the 5 kb of DNA that flank the gaps consists of duplicated sequence (>90% identity). No other sequence property that we assessed showed such a strong association (TABLE 2). As duplications only account for ~5% of the genome, duplicated regions are enriched eightfold for gaps, indicating that such areas remain problematic for finishing.

It has been shown that subtelomeric and pericentromeric regions are significantly enriched for duplication content (three–fivefold). We considered these regions of the genome independently, defining subtelomeric DNA as DNA within 1 Mb of the most distally placed sequence CONTIG, and pericentromeric DNA as 5-Mb regions on either side of the centromeric gap. As expected, pericentromeric and subtelomeric DNA show an increased frequency of gaps (78 and 25, respectively). Most of these gaps (100/103) are associated with segmental duplication. Not all pericentromeric regions show evidence for large blocks of duplication. Such regions (Xp11, 8p11, 3p11, 4q11, 5p11 and 19q11) are conspicuously devoid of gaps, despite the fact that they make transitions into centromeric satellite DNA. We conclude that the increased density of gaps in pericentromeric DNA is a property of recent segmental duplications and not simply a property of proximity to satellite DNA.

In a final analysis, we distinguished gaps that were classified as recalcitrant to subcloning (clone gaps) from those that were traversed by clones but had proved to be difficult to finish (finishing gaps). Clone gaps constitute the bulk of these by number and by mass. Once again, we examined the duplication content that flanks the interstitial euchromatic content. Most finishing gaps (56% (44/79)) and clone gaps (52% (126/241)) are enriched for segmental duplications that flank the gaps. If we assume that all gaps that are flanked by segmental duplications are completely composed of duplicated sequence, we can estimate that 14.6 Mb (~10%) of duplicated sequence remains to be sequenced and/or assembled as part of the human euchromatin. In general, regions with larger and more homologous segmental duplications harbour the largest proportion of gaps. For example, an analysis of chromosome 1 shows that 46/66 duplications are flanked by segmental duplications, in which the individual alignments share more than 98% sequence identity. In the case of chromosome 9, 37/44 of the duplications occur within an ~8-Mb pericentromeric region that is composed almost entirely of segmental duplications (FIG. 2).

EUCHROMATIN
Parts of chromosomes that show the normal cycle of condensation and normal staining properties at nuclear divisions. Euchromatin generally contains active or potentially active genes and falls within light bands after Giemsa staining.

SATELLITE DNA
Various classes of highly repetitive DNA that are tandemly repeated and most often associated with centromeric or pericentromeric regions of the genome. α-Satellite DNA is a class of centromeric satellite in which the monomeric unit is 171 bp. Higher-order structures of this satellite define the DNA component of human centromeres. β-Satellite DNA is a class of pericentromeric satellite in which the basic repeat unit is a 68-bp monomer.

ACROCENTRIC
A chromosome in which the centromere is located subterminally, and concomitantly the chromosome arms are unequal in length.

PARALOGOUS
The quality of having sequence similarity as a result of duplication.

CONTIG
A set of contiguous overlapping clones that span a genomic region.

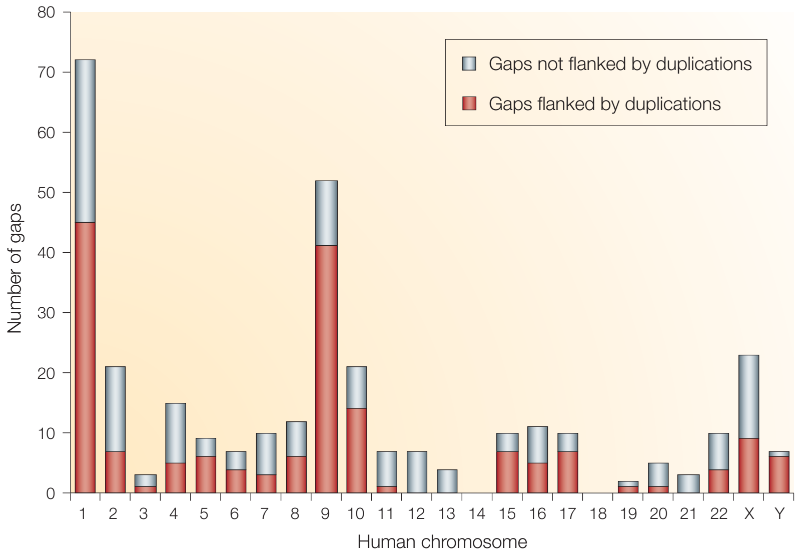

Figure 1 | **Chromosomal distribution of sequence gaps.** The number of euchromatic sequence gaps that are flanked by segmental duplications and by unique sequence are shown per chromosome. A total of 321 sequence gaps of the 379 gaps were classified as interstitial euchromatin. Although euchromatic and heterochromatic DNA cannot be simply defined at the sequence level, for the purpose of this analysis we considered pericentromeric DNA (distal to the most proximal block of satellite DNA) as euchromatic, because there is evidence of transcription in these regions. We did not consider the 58 gaps that are largely repetitive and for which there is little evidence of transcription (including telomeric ends, acrocentric portions or centromeric (primary or secondary constriction) satellites of human chromosomes).

FISH analysis. Highly duplicated regions of the human genome are readily characterized by multiple signals if fluorescence *in situ* hybridization (FISH) is used to analyse human metaphase chromosomes[4,7,9]. MULTISITE SIGNALS have proved to be instructive in directing efforts to close gaps in some highly paralogous regions of the genome in which no sequence homology could be detected. Indeed, FISH has proved to be an invaluable tool in detecting sequences that are not present in the assembly (FIG. 3). We examined the multisite distribution of 161 BAC clones, the underlying sequence of which was confirmed as duplicated[7,9]. Using low-stringency criteria (≥90% sequence identity ≥5,000 bp), we used similarity searches to simulate the potential location of multisite signals in the finished genome assembly. Although many false positives (alignments >90% identity with no experimental match by FISH) would be expected as a result of this low level of sequence divergence, this threshold would minimize the number of false-negative regions in the assembly that contain sequence — undetected by BLAST — that was positive by FISH. We therefore considered these criteria as conservative. FISH analysis of these 161 multisite BAC clones identified a total of 708 chromosome signals. Of these, 74.4% (521/708) could be confirmed by *in silico* analysis of the finished human genome (TABLE 3). This represents a marked improvement from previous analyses of the working draft sequence assembly, which predicted less than ~50% concordancy[7,9]. At the level of cytogenetic GIEMSA bands (G-bands), the concordancy level drops to 58.6% (572/976). This is not unexpected, as the correspondence between *in silico* and experimental FISH-cytogenetic G-band location is often ambiguous. We examined the distribution of experimental FISH matches that could not be verified by *in silico* analysis. Interestingly, the distribution of these signals closely matches the chromosomal gap distribution that is observed by sequence analysis, including a preponderance of signals near the pericentromeric and subtelomeric regions of the genome. In particular, chromosomes 1q12 and 9p11-q12 show a strong correlation between potential duplication-induced gaps and computationally unverified multisite FISH signals.

The nature of duplicated gaps. Do the gaps, and particularly those that are embedded within duplicated sequence, represent regions of the genome that cannot be subcloned into BACs, or are the difficulties in sequencing and assembly primarily the result of biological complexity that is associated with highly homologous duplications? Several recent analyses indicate that the latter might not be uncommon. Using a pericen-

Table 2 | **Sequence properties flanking gaps**

Property	**5-kb-flanking gap**		**Genome**	
	bp	**% flanking the gap**	**bp**	**% of the genome**
Duplications	1386592	40.85%	150940541	5.3%
Gene content	144493	4.26%	876738325	30.6%
Repeat	1514187	44.61%	1392119068	48.6%
LINE	613767	18.08%	606569326	21.2%
SINE	426172	12.56%	390431934	13.6%
LTR	321235	9.46%	248348409	8.7%
DNA	65485	1.93%	86316578	3.0%
Satellite	730	0.02%	12332522	0.4%
Simple	54876	1.62%	26378348	0.9%
Other	10671	0.31%	4028684	0.1%
Total sequence	3394350		2865069170	

Sequence properties that flank (5 kb) interstitial euchromatin gaps (*n* = 321) are compared to the genome average. Duplication content was determined according to the Human Paralogy Server Segmental Duplication Database (http://humanparalogy.gene.cwru.edu). Gene content corresponds to the transcribed portion (heterogenous nuclear RNA) but might be an underestimate as it requires the identification of at least two exons within 5 kb. Repeat content is based on RepeatMasker classification[73](June, 2003). LINE, long interspersed elements; LTR, long terminal repeats; SINE, short interspersed repeats.

MULTISITE SIGNALS
Multiple fluorescence *in situ* hybridization (FISH) signals on metaphase chromosomal preparations.

GIEMSA
A cytogenetic stain that is applied to metaphase chromosomes after limited digestion with trypsin. Individual chromosomes are distinguished on the basis of a characteristic banding pattern of dark and light bands.

tromeric interspersed repeat as a marker to identify duplicated sequence, a recent study characterized large-insert BAC clones that mapped to specific pericentromeric regions but that had not been incorporated into the final assembly of the Human Genome Project[15,19]. Considerably more sequence diversity in duplicated regions remains in GenBank (see online links box) when compared with the final assembly of the human genome. The most proximal portion of the published version of human chromosome 14 (REF. 20) currently lacks ~270 kb of duplicated sequence that shares 99% identity with human chromosome 22 (REF. 10; FIG. 2). Interestingly, BAC clones that correspond to this region have been sequenced but were apparently discarded in later assemblies probably owing to the high degree of sequence identity to chromosome 22. During the final phases of the sequencing and assembly of human chromosome 7, large, highly identical segmental duplications that are associated with the breakpoints of the Williams-Beuren syndrome region proved to be particularly problematic. Excessive redundancy of sequencing from RPCI-11 BACs was required to resolve this structure[18]. The sequence in these regions required assembly within a single chromosomal haplotype for the genome to be properly assembled to avoid errors that result from structural variants between haplotypes. Similar difficulties have been encountered in other duplicated regions of the human genome. Ample experimental data indicate that large-scale structural variants are common near duplications and in pericentromeric regions of the genome[13,14,21–27]. The conclusion for these and other regions has been that structural variation between chromosomal haplotypes complicates the assembly of duplicated regions. This leads to the formation of *de facto* gaps if two structurally different haplotypes have been incorporated into the assembly (FIG. 4). Similar issues of large-scale structural polymorphism and large blocks of highly identical sequence complicated the sequencing and assembly of the Y chromosome. The final sequencing and assembly of the Y chromosome (which is unusually enriched for segmental duplications) was achieved largely owing to the fact that all the “BAC

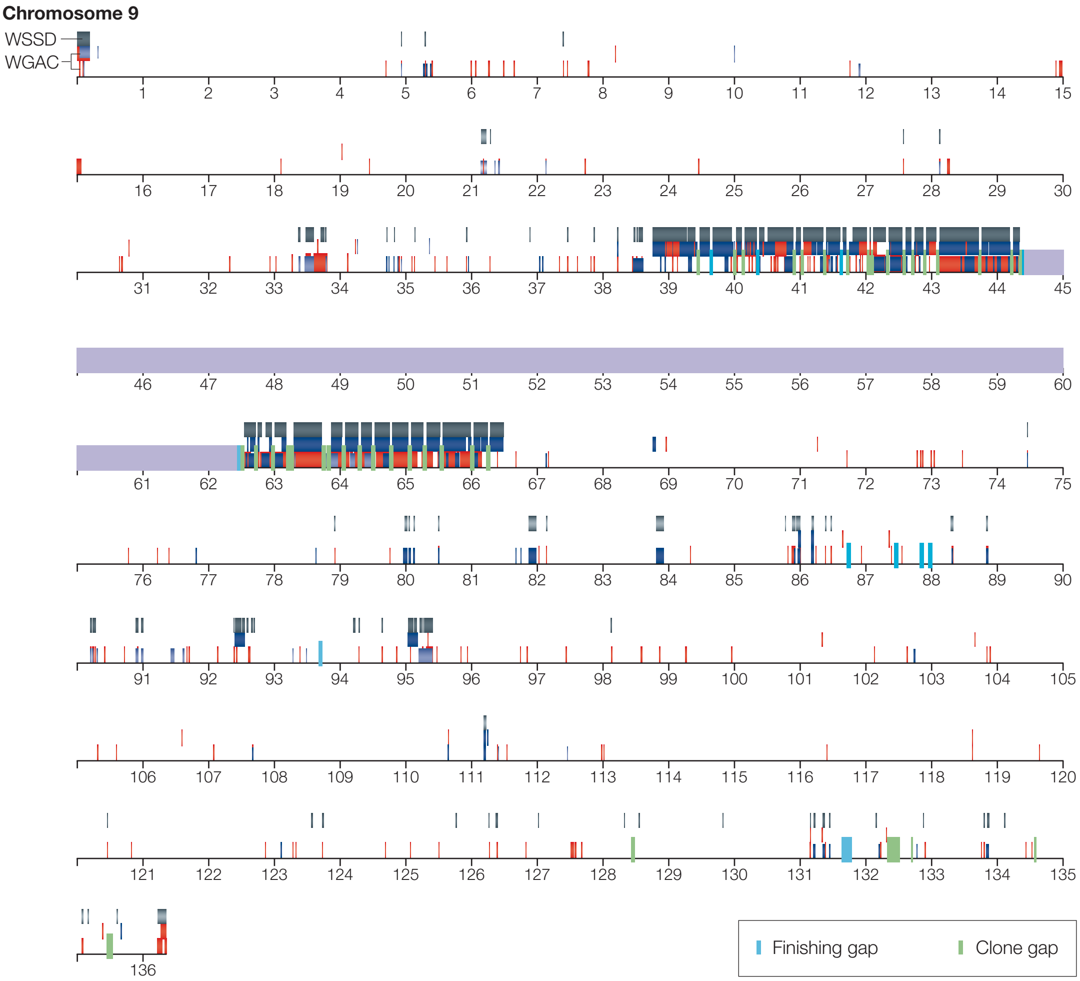

Figure 2 | **Duplications and sequence gaps.** Duplication content for chromosome 9 is shown as coloured (red and blue) or dark grey bars above the horizontal line (determined by WGAC and WSSD methods, respectively). The position of euchromatin sequence gaps is shown as light blue (finishing gaps) or light green vertical bars (clone gaps) that traverse the horizontal line. Centromeric satellite sequences are depicted in purple. Most (90%) of the sequence gaps associate with segmental duplications. WGAC, whole-genome analysis comparison; WSSD, whole-genome shotgun sequence detection.

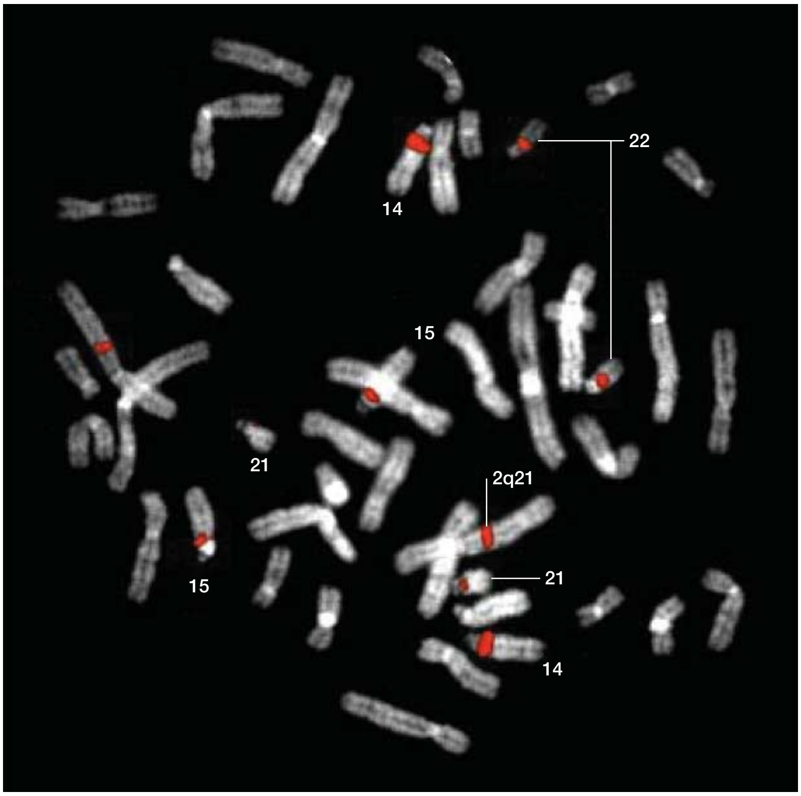

Figure 3 | ***In silico* versus cytogenetic analysis of the human genome.** *In situ* hybridization of a bacterial artificial chromosome (BAC) probe (RP11-140M6), which has been mapped to a highly duplicated region of 22q11 (REF. 10), is shown. Multiple signals correspond to sites of duplication. Particularly strong signals are detected on 14q11, corresponding to multiple copies of this locus as confirmed by extensive sequence analysis of chromosome 14 MONOCHROMOSOMAL HYBRIDS[10]. No evidence of this duplication can be detected on the basis of the sequence analysis of the genome. Fluorescence *in situ* hybridization (FISH) results pinpoint potential gaps in the finished genome. Reproduced with permission from REF. 10 © University of Chicago Press (2002).

MONOCHROMOSOMAL HYBRID
A cell line that carries a single, intact human chromosome in a rodent somatic cell background.

clones came from one man's Y chromosome"[17]. Unlike the X and Y chromosomes[28], the assembly of autosomal duplications requires the extra effort of resolving haplotype differences that result from the diploid nature of the underlying BAC library.

α-Satellite and centromeric transition regions

A total of 43 pericentromeric transitions would have been expected to have been traversed during the course of the Human Genome Project (acrocentric transition regions were not targeted)[3]. All human centromeres are characterized by large blocks (1–3 Mb) of tandem higher-order α-satellite repeat arrays[29,30]. Several reports have shown that these blocks of higher-order repeats are flanked by monomeric tracts (92 kb) of α-satellite DNA as well as other types of pericentromeric satellite DNA[29–32]. We examined the 2 Mb that flank either side of the putative centromere in each human chromosome for the presence of large tracts (>10 kb) of flanking α-satellite DNA. A total of 29/43 (67%) of the targeted chromosomal regions show blocks of α-satellite DNA that is positioned in the most proximal portion of the sequence contig. Interestingly, only 9 pericentromeric regions showed a near-perfect match with higher-order α-satellite DNA (chromosomes 1, 2, 5, 7, 8, 9, 19, X and Y) (>96%), which indicates that classically defined centromeric DNA or degenerative higher-order repeat sequences might have been attained for only a subset of the chromosomes[31]. A further four chromosomes (1q12, 4p11, 10p11 and 16q11) show evidence for other large tracts of pericentromeric satellite DNA (for example, HSATI, HSATII and GAATG). For at least three of these chromosomes[30], these large tracts of sequence represent putative transition regions to secondary constrictions and therefore delineate reasonable points of termination for the Human Genome Project. Therefore, 33/43 (76.7%) pericentromeric regions are represented by genome sequence that is typical of the model euchromatic/heterochromatic transition region. Of the remaining ten chromosomes that show no evidence of a satellite/euchromatin transition in the most proximal portion of the chromosomal arm sequence, all correspond to highly duplicated regions of the human genome.

Acrocentric DNA

The short arms of human chromosomes 13, 14, 15, 21 and 22 were not targeted as part of the Human Genome Project. Together with the centromeric satellite DNA, these regions contain highly repetitive DNA, they are generally regarded as heterochromatic and, consequently, they constitute another portion of the *sequence non grata*. Although the length of these regions might vary considerably between individuals, acrocentric DNA has been estimated to account for ~67 Mb of the genome (International Human Genome Sequencing Consortium, manuscript in preparation). Our understanding of the organization of these regions stems from early *in situ* and DNA-satellite repeat studies. On the basis of this work, the model structure of an acrocentric arm consists of relatively large tracts of satellite sequences near the centromeres and tandem arrays of 28S- and 18S-ribosomal DNA in the centre of the arms that are flanked on either side by variable blocks of β-satellite DNA[33–34]. Despite the fact these regions are not an official target of the Human Genome Project, yeast-artificial chromosome (YAC) maps of these regions have been successfully constructed[36,37]. Several BACs and cosmids have been sequenced as part of an initial assessment of their sequence structure[38–40]. The limited sequence analysis of these regions indicates that our model of the acrocentric arms might be too simplistic. A considerable number of additional acrocentric segmental duplications has been documented in recent years[41–44]. Their mosaic structures, size and high degree of sequence identity are reminiscent of other pericentromeric duplications in the long arms of human chromosomes. In addition, a few genes and gene families are embedded in these duplications between acrocentric and non-acrocentric chromosomes. Most notably, the testis-expressed transmembrane tyrosine phosphatase with tensin homology (*TPTE*) gene family has been described as the first non-ribosomal gene with a possible endocrine or spermatogenic function[45]. Is it possible that the acrocentric portions of the genome harbour many more genes that remain to be discovered?

FOSMID
A low-copy vector for the construction of stable genomic libraries that uses the *Escherichia coli* F-factor origin for replication.

NON-ROBERTSONIAN TRANSLOCATION
A chromosomal rearrangement that is characterized by the fusion of different chromosomes, but does not involve the fusion of whole long arms of two acrocentric chromosomes (chromosomes 13, 14, 15, 21 and 22 in humans).

PALINDROME
DNA sequence in which the 5′–3′ composition is identical on each strand with respect to its midpoint.

Table 3 | **Cytogenetic versus *in silico* analysis of human segmental duplications**

Threshold	Concordant chromsomes	Concordant bands	Discordant chromsomes	Discordant bands
20k95	416	429	278	527
20k90	445	475	245	475
10k95	462	483	246	493
10k90	509	550	199	426
5k95	478	501	230	475
5k90	527	572	181	404
1k95	493	522	215	454
1k90	551	599	157	377

Duplication content of the genome was analysed at various thresholds (that is, 20k95 = segmental duplications >20 kb and >95% sequence identity). The *in silico* multisite distribution pattern was then compared with 161 bacterial artificial chromosome (BAC) clones, which generated multiple chromosomal signals by fluorescence *in situ* hybridization (FISH) as described previously by Cheung *et al.*[10]. Only those multisite BACs in which the underlying sequence was determined to be duplicated, and in which the best placement of the BAC and the cytogenetic location were concordant, were considered. A total of 956 cytogenetic signals that mapped to 708 chromosomes were analysed. On the basis of the experimental results, the numbers of concordant and discordant chromosomes and Giemsa-bands (G-bands) were considered. The number for G-bands is substantially higher owing to ambiguities in G-band assignment both experimentally and by *in silico* approximation.

It is unlikely. An analysis of available gene sequences indicates that few, if any, of the unplaced genes map to the short arms of acrocentric chromosomes (International Human Genome Sequencing Consortium, manuscript in preparation).

Muted gaps and structural polymorphisms

The term 'muted gap' has been applied to regions of the genome in which apparent sequence continuity has been achieved in the assembly but that also contain a hidden gap (that is, additional sequence might be present in the human population but not in the assembly). Although relatively few of these have been documented[46–48], two apparent root causes for such muted gaps emerge: clonal instability in *Escherichia coli* vectors and/or structural polymorphism in the human population. Such muted gaps have also been inferred from the complete genome sequence of human chromosomes. During the sequencing of chromosome 14, for example, known genes that mapped to the chromosome were not identified despite the fact that no mapped gaps remain[46].

It is currently unclear exactly how many muted gaps there are in the finished human genome, although the overlay of paired-end sequences from whole-genome shotgun sequence and FOSMID subclones is beginning to flag potential sites for further investigation (International Human Genome Sequencing Consortium, manuscript in preparation). Representational difference analysis, which is based on subtractive hybridization of reference and target DNA, has been successfully used to enrich for such sequence differences in complex genomes[47,49]. In one study, this approach recovered a 9.1-kb deletion that was not initially present in the genome reference for human chromosome 22 but was found to correspond to a polymorphic insertion/deletion with a worldwide distribution pattern[47]. Subsequent library hybridizations recovered large-insert BAC libraries that contained both the insertion and the deletion. It is probable that the human genome contains a large number of small insertion/deletion polymorphisms, some of which might have clinical consequences[48,50–52].

One of the best-characterized examples of both a muted and an unclonable gap came from a human genetics study to clone the breakpoints of the most frequent NON-ROBERTSONIAN TRANSLOCATION between human chromosomes 11 and 22 (REFS 46,52). The precise region of the translocation breakpoint corresponded to a PALINDROMIC AT-rich repeat sequence (PATRRs) that was consistently deleted from BACs that mapped to chromosome 11 and that represented an unclonable sequence gap in chromosome 22. The nature of the sequence was ultimately deduced by direct sequencing of PCR products that were derived from hybrid material obtained from t(11:22) constitutional translocations[46,52]. Although the PATRR sequence in chromosome 11 was eventually subcloned (albeit frequently in a truncated form), repeated attempts failed to recover the corresponding segment in chromosome 22. The unclonability of this portion of chromosome 22 has been proposed to be the result of the size and/or length of the PATTR sequences in these regions[51]. Computational analysis indicates that such palindromic sequences have the potential to form unusual hairpin structures that are unstable in *E. coli* because of the preference to promote double-strand

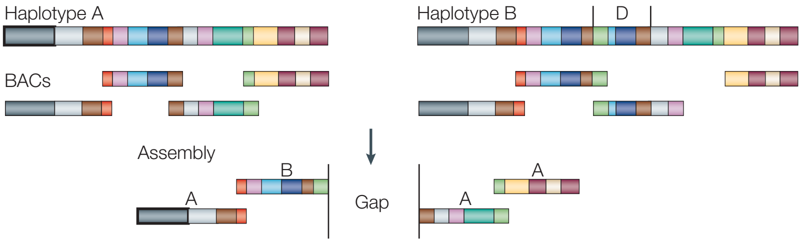

Figure 4 | **Gaps, duplications and structural variation.** A hypothetical large-scale structural variant (D) is shown in a highly duplicated region (duplicated segments are shown as coloured bars). Sequencing of targeted bacterial artificial chromosomes (BACs) from a diploid source leads to inconsistencies when assembly is attempted from both haplotypes (A and B). Sequence-assembly overlap from both haplotypes creates a *de facto* gap (green). It is impossible to assemble this region unless single haplotype continuity is obtained.

breaks or targeted deletions. Similarly, difficulties were encountered during the cloning and sequencing of CGG-triplet repeats[53]. In both cases, it is noteworthy that clonal instability corresponded to sites of genomic instability.

Gene and SNP annotation

The first step towards functional genomics is genome annotation. Owing to its obvious biomedical relevance, complete gene annotation is one of the top priorities of the Human Genome Project. The presence of gaps, *de facto* or muted, complicates this process. Three different approaches are being implemented to improve the human annotation set: *ab initio* predictions[54], characterization of ORTHOLOGOUS genes between distantly related species and analysis of cDNA source material (ESTs and full-length mRNA)[55,56]. In a recent, second-generation analysis of chromosome 22 (REF. 57), some important lessons were learned. Hand curation is key and the process is time-consuming — an estimated four years' of analysis was needed after the finishing of chromosome 22 (REF. 58). One of the surprising findings of the chromosome-22 analysis was that nearly one-third of all annotations (234/780) involved unprocessed pseudogenes. A further 32/546 of the putative protein-encoding annotations corresponded to partial-gene duplications known as 'pre-pseudogenes'. In these cases, no evidence for transcription could be found, but there was also no evidence that the partial open reading frame was in fact disrupted. One of the most difficult tasks of gene annotation is to discriminate between true genes and these genomic decoys that are often embedded within duplicated sequences. Similar difficulties have been reported for other organisms, albeit to a lesser extent[59]. The more finished sequence that is available, the less ambiguous this task becomes. Even among regions that are 99% identical, it is often possible to assign cDNA source material to its best location by match and by sequence identity when complete ascertainment of genomic sequence is available. If a particular genomic reference sequence is not represented and transcripts map to multiple locations with ~98% sequence identity, gene annotation is compromised. Inclusion of the additional locus, which is 99% identical at the sequence level, has been shown to help resolve annotation as pseudogene or gene[60]. It is perhaps not surprising in the light of the association of segmental duplications and gaps that 74% (209/282) of the distinct mRNA that are missing, that are partially represented or that are ambiguous to assign correspond to duplicated regions of the human genome.

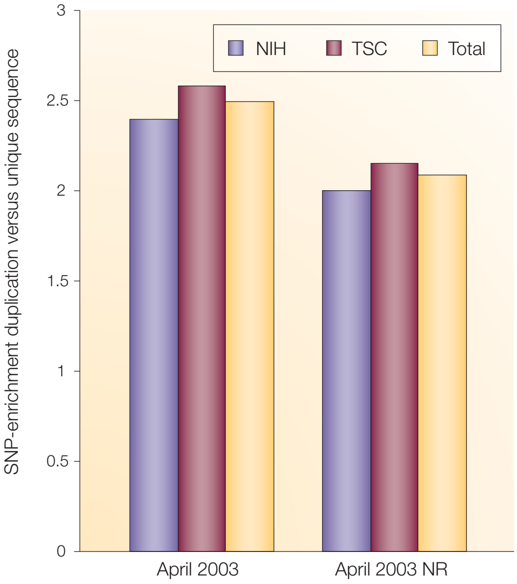

Figure 5 | **SNP enrichment in duplicated sequences.** SNP density between duplicated regions and unique regions was assessed in the finished genome (April 2003). 429,999 SNPs mapped to 154.0 Mb of duplicated sequence compared with 3.02 million SNPs that were assigned to 2.87 Mb of unique sequence. Based on the density of SNPs, this represents a 2.5-fold enrichment in duplicated regions of the genome. If SNPs are only allowed to be placed once in the genome (that is, multisite SNPs are counted once), this enrichment drops to 2.16-fold. SNP data are categorized on the basis of source (NIH or TSC). NIH, National Institutes of Health; NR, non-redundant; TSC, The SNP Consortium).

The same issue also applies to other forms of genome annotation, such as the assignment of SNPs. The mapping of a SNP to its best location is relatively straightforward if the reference genome is complete. Gaps in the genome sequence, and particularly those that surround duplicated regions, however, can create havoc. Paralogous sequence variants can be incorrectly assigned as putative SNPs — being discovered only during genotyping when all individuals are found to be heterozygous for the variant in question. Several studies have reported that the density of SNPs increases in duplicated regions of the genome, indicating that such errors in assignment might be common[4,61]. One solution to this problem has been to disregard all reads that surpass a specified SNP-density threshold[62,63]. This has the danger, of course, of normalizing the human genome to an *a priori* expectation of allelic variation. Regions with high levels of allelic diversity, such as those under BALANCING SELECTION, might in theory be eliminated. The other, more favourable, solution is to improve the quality of finished sequence in paralogous regions of the genome. We compared SNP annotation in duplicated and unique regions in the finished genome and found a significant enrichment (2.5-fold) of SNPs in duplicated regions (FIG. 5). If SNPs that are annotated as redundant are excluded, the enrichment drops to 2.16-fold. These data indicate that there is either more variation in duplicated regions or that a fraction of the most highly homologous duplications remain to be captured. Closure of the sequencing gaps and experimental analysis will be necessary to distinguish between these two possibilities.

Strategies to finish the 'finished genome'

There have been several success stories related to the sequencing and assembly of difficult regions of the human genome. In the case of pericentromeric regions and the human Y chromosome, single haplotype genome

ORTHOLOGOUS
The quality of having sequence similarity as a result of speciation.

BALANCING SELECTION
Natural selection in which heterozygotes have increased evolutionary fitness with respect to either homozygous condition.

Box 1 | Filling the gaps

Although segmental duplications associate significantly with sequence gaps, nearly 45% of the gaps show no apparent association with these low-copy repeat sequences. There are some indications that bacterial artificial chromosome (BAC) cloning vectors show subcloning biases against AT-rich and GC-rich DNA, particularly against those that can form secondary structures. Several other technologies have been implemented to begin to resolve and validate the assembly of these problematic regions of the genome.

Transformation-associated recombination cloning
The cloning of large DNA sequences in circular yeast-artificial chromosomes on the basis of recombination between a target anchor sequence (that is, human reference sequence) in the vector and transformed human DNA fragments in a yeast background. It allows for selective isolation of DNA fragments, particularly near gaps, without the need for the construction of an entire genomic library.

Representational difference analysis
Subtractive hybridization of reference and target DNA to enrich for differences between complex genomes. It might be used to identify, isolate and clone genomic variation both within and between species.

Optical mapping
A technology for developing high-resolution restriction maps by capturing single molecules of DNA and visualizing restriction fragments by fluorescence microscopy. The analysis of multiple overlapping single molecules generates physical maps of genomic DNA without the subcloning bias that is associated with large-insert clones.

Paired-end sequence analysis
A technology that uses pairs of sequence reads as a mapping reagent to connect sequence contigs. Mate-pairs or sequence-tag connectors might be used as an integral component of whole-genome assembly of genomes to close gap regions and/or to identify potential sites of inconsistency or structural variation in a genome. The resolution depends on the average insert size of the genomic library. The development of capillary-fluorescent technology, which effectively eliminated sample lane-tracking errors, was an essential component of this technology.

resources have been exploited to develop sequence continuity within highly duplicated and structurally variant portions of the genome. In these cases, paralogous sequence variants or sequence family signatures are used as mapping tools to exclude or confirm sequence overlaps. So, misassembly in duplicated sequences might be averted. The anticipated construction of a human hydatidiform mole BAC library (a white paper proposal that was approved in 2003; see online links box) will eliminate potential problems that are associated with large-scale structural polymorphisms and duplications by allowing sequencing and assembly in a single human haplotype. Specialized cloning vectors such as the half-YAC[64] have successfully been used to recover the terminal portions of human chromosomes. Nearly three-quarters of all subtelomeric sequences are now represented in the finished genome largely as a result of this strategy[65] and the use of chromosome-specific cosmid libraries. In the future, it is probable that production of sheared genomic libraries will more effectively eliminate restriction enzyme biases in BAC libraries, therefore allowing terminal regions as well as euchromatin/heterochromatin transitions to be more readily recovered. Alternative subcloning methodologies, such as transformation-associated recombination[66], that target specific genomic regions following homologous recombination in yeast SPHEROPLAST FUSION, show considerable promise to capture centromeric satellite sequences as well as sequences that are generally unstable in standard *E. coli* vectors[48,67]. Of course, simply capturing a sequence in a subclone does not guarantee its proper assembly. Advances in the assembly of repetitive and/or unusually GC-rich DNA[68,69] are required to facilitate this process. Finally, methods that independently assess the quality of finished assembled sequences (that is, optical mapping, paired-end sequence analysis and representation difference analysis; BOX 1) will be vital in validating the quality of the assembly. Although the nature of the gaps might vary, it is noteworthy that there can be considerable overlap in the implementation of strategies to resolve these regions. For example, pericentromeric and subtelomeric regions are enriched for segmental duplications, are subject to extensive polymorphism, harbour repetitive sequences and create subcloning biases owing to their proximity to α-satellite DNA and chromosome termini. Sheared or TRANSFORMATION-ASSOCIATED RECOMBINATION-BASED GENOMIC LIBRARIES from single haplotype source material, such as monochromosomal hybrids or hydatidiform mole source material, could resolve several problems simultaneously. Such resources might be considered as supplementary material for genomes that are slated for finished quality sequence.

Perspective and conclusions

Ultimately, it might take 5, 10 or even 20 years before we can confidently say that the last gap of the genome has been traversed. Should we invest the effort? From an immediate practical perspective, the sequences that remain in the gaps confound SNP assignment and unambiguous gene annotation, and therefore deserve special attention. From the biological/clinical perspective, the available data indicate that these regions represent sites of large-scale structural polymorphism, chromosomal rearrangement that is associated with disease, genomic instability, rapid chromosomal evolution and that these regions are the source for rapidly evolving genes and gene families that are specific to the hominoid lineage[4,24,52,70–72]. Sequencing these regions would provide fundamental insights into the relationship between chromosome structure and function. Our complete understanding of human disease, evolution and gene annotation requires that these difficult regions be resolved. Progress on chromosome 22, the first chromosome to be finished, is sobering. Only two of the thirteen gaps have been partially filled since its finished status was declared in 1999. It is clear that this might seem to have diminishing returns for considerable effort — many more appealing targets have justifiably emerged to attract the attention of funding agencies. So, can such efforts to fill in the gaps be consolidated? On the basis of the few success stories, finishing the 'finished genomes' requires a different *modus operandi* from what can effectively be applied to the bulk of the genome. We propose a three-component plan. The first step involves the construction of specialized resources for genomes in which finished sequence is a priority (for example, sheared genomic libraries,

SPHEROPLAST FUSION
A method for transferring DNA into cells or between cells whereby the host cell wall is removed (spheroplast) before polyethylene glycol fusion.

TRANSFORMATION-ASSOCIATED RECOMBINATION-BASED GENOMIC LIBRARIES
Genomic libraries that are constructed by a transformation-associated recombination (TAR) cloning system (see box 1).

half-YAC libraries, and so on). Second, specialized researchers with expertise and interest in exceptional regions should be encouraged to develop new technology, strategies or software to target these areas. Third, these individuals should work closely with large-scale sequencing centres that can provide the requisite infrastructure for the high-quality sequencing and assembly of these regions. In the case of the Y chromosome, pericentromeric and subtelomeric regions, this model has been partially successful. This model runs counter to the prevailing wisdom of large-scale genome projects that have tended to centralize resources and eliminate the specialist biologists in an effort to cut costs.

1. International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature* **409**, 860–920 (2001).
The first description and analysis of a publicly released assembly of the human genome.
2. Collins, F. S., Green, E. D., Guttmacher, A. E. & Guyer, M. S. A vision for the future of genomics research. *Nature* **422**, 835–847 (2003).
3. Collins, F. S. *et al.* New goals for the U.S. Human Genome Project: 1998–2003. *Science* **282**, 682–689 (1998).
4. Bailey, J. A. *et al.* Recent segmental duplications in the human genome. *Science* **297**, 1003–1007 (2002).
A global analysis of the organization and properties of recent segmental duplications in the human genome using whole-genome shotgun sequence data.
5. Green, P. Against a whole-genome shotgun. *Genome Res.* **7**, 410–417 (1997).
6. Eichler, E. E. Masquerading repeats: paralogous pitfalls of the human genome. *Genome Res.* **8**, 758–762 (1998).
7. Bailey, J. A., Yavor, A. M., Massa, H. F., Trask, B. J. & Eichler, E. E. Segmental duplications: organization and impact within the current Human Genome Project assembly. *Genome Res.* **11**, 1005–1017 (2001).
8. Cheung, J. *et al.* Genome-wide detection of segmental duplications and potential assembly errors in the human genome sequence. *Genome Biol.* **4**, R25 (2003).
9. Cheung, V. G. *et al.* Integration of cytogenetic landmarks into the draft sequence of the human genome. The BAC Resource Consortium. *Nature* **409**, 953–958 (2001).
10. Bailey, J. A. *et al.* Human-specific duplication and mosaic transcripts: the recent paralogous structure of chromosome 22. *Am. J. Hum. Genet.* **70**, 83–100 (2002).
11. Kehrer-Sawatzki, H., Schwickardt, T., Assum, G., Rocchi, G. & Krone, W. A third neurofibromatosis type 1 (NF1) pseudogene at chromosome 15q11. 2. *Hum. Genet.* **100**, 595–600 (1997).
12. Kehrer-Sawatzki, H. *et al.* Molecular characterization of the pericentric inversion that causes differences between chimpanzee chromosome 19 and human chromosome 17. *Am. J. Hum. Genet.* **71**, 375–388 (2002).
13. Barber, J. C., Reed, C. J., Dahoun, S. P. & Joyce, C. A. Amplification of a pseudogene cassette underlies euchromatic variation of 16p at the cytogenetic level. *Hum. Genet.* **104**, 211–218 (1999).
14. 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).
15. Horvath, J. E. *et al.* Using a pericentromeric interspersed repeat to recapitulate the phylogeny and expansion of human centromeric segmental duplications. *Mol. Biol. Evol.* **20**, 1463–1479 (2003).
16. 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).
17. Kuroda-Kawaguchi, T. *et al.* The AZFc region of the Y chromosome features massive palindromes and uniform recurrent deletions in infertile men. *Nature Genet.* **29**, 279–286 (2001).
18. Hillier, L. W. *et al.* The DNA sequence of human chromosome 7. *Nature* **424**, 157–164 (2003).
19. Horvath, J. E., Bailey, J. A., Locke, D. P. & Eichler, E. E. Lessons from the human genome: transitions between euchromatin and heterochromatin. *Hum. Mol. Genet.* **10**, 2215–2223 (2001).
20. Heilig, R. *et al.* The DNA sequence and analysis of human chromosome 14. *Nature* **421**, 601–607 (2003).
21. Giglio, S. *et al.* Olfactory receptor-gene clusters, genomic-inversion polymorphisms, and common chromosome rearrangements. *Am. J. Hum. Genet.* **68**, 874–883 (2001).
22. Osborne, L. R. *et al.* A 1. 5 million-base pair inversion polymorphism in families with Williams-Beuren syndrome. *Nature Genet.* **29**, 321–325 (2001).
Provides evidence that large-scale structural polymorphisms might increase the risk of recurrent chromosomal structural rearrangements among offspring.
23. Gimelli, G. *et al.* Genomic inversions of human chromosome 15q11-q13 in mothers of Angelman syndrome patients with class II (BP2/3) deletions. *Hum. Mol. Genet.* **12**, 849–858 (2003).
24. Giglio, S. *et al.* Heterozygous submicroscopic inversions involving olfactory receptor-gene clusters mediate the recurrent t(4;8)(p16;p23) translocation. *Am. J. Hum. Genet.* **71**, 276–285 (2002).
25. Ritchie, R. J., Mattei, M. G. & Lalande, M. A large polymorphic repeat in the pericentromeric region of human chromosome 15q contains three partial gene duplications. *Hum. Mol. Genet.* **7**, 1253–1260 (1998).
26. Barber, J. C. *et al.* Neurofibromatosis pseudogene amplification underlies euchromatic cytogenetic duplications and triplications of proximal 15q. *Hum. Genet.* **103**, 600–607 (1998).
27. Fantes, J. A. *et al.* Organisation of the pericentromeric region of chromosome 15: at least four partial gene copies are amplified in patients with a proximal duplication of 15q. *J. Med. Genet.* **39**, 170–177 (2002).
28. Skaletsky, H. *et al.* The male-specific region of the human Y chromosome: a mosaic of discrete sequence classes. *Nature* **423**, 825–837 (2003).
29. Alexandrov, I., Kazakov, A., Tumeneva, I., Shepelev, V. & Yurov, Y. α-Satellite DNA of primates: old and new families. *Chromosoma* **110**, 253–266 (2001).
A thorough overview of the various classes of α-satellite DNA and their evolutionary properties.
30. Lee, C., Wevrick, R., Fisher, R. B., Ferguson-Smith, M. A. & Lin, C. C. Human centromeric DNAs. *Hum. Genet.* **100**, 291–304 (1997).
31. Schueler, M. G., Higgins, A. W., Rudd, M. K., Gustashaw, K. & Willard, H. F. Genomic and genetic definition of a functional human centromere. *Science* **294**, 109–115 (2001).
Functional and structural characterization of a euchromatin–heterochromatin transition region on the X chromosome.
32. Horvath, J. *et al.* Molecular structure and evolution of an α/non-α satellite junction at 16p11. *Hum. Mol. Genet.* **9**, 113–123 (2000).
33. Worton, R. *et al.* Human ribosomal RNA genes: orientation of the tandem array and conservation of the 5′ end. *Science* **239**, 64–68 (1988).
34. Greig, G. & Willard, H. β-Satellite DNA: characterization and localization of two subfamilies from the distal and proximal short arms of human acrocentric chromosomes. *Genomics* **12**, 573–580 (1992).
35. Choo, K. H., Vissel, B. & Earle, E. Evolution of α-satellite DNA on human acrocentric chromosomes. *Genomics* **5**, 332–344 (1989).
36. Korenberg, J. R. *et al.* A high-fidelity physical map of human chromosome 21q in yeast artificial chromosomes. *Genome Res.* **5**, 427–443 (1995).
37. Wang, S. Y. *et al.* A high-resolution physical map of human chromosome 21p using yeast artificial chromosomes. *Genome Res.* **9**, 1059–1073 (1999).
38. Gonzalez, I. L. & Sylvester, J. E. Complete sequence of the 43-kb human ribosomal DNA repeat: analysis of the intergenic spacer. *Genomics* **27**, 320–328 (1995).
39. Gonzalez, I. L. & Sylvester, J. E. Incognito rRNA and rDNA in databases and libraries. *Genome Res.* **7**, 65–70 (1997).
40. Gonzalez, I. L. & Sylvester, J. E. Human rDNA: evolutionary patterns within the genes and tandem arrays derived from multiple chromosomes. *Genomics* **73**, 255–263 (2001).
41. Wohr, G., Fink, T. & Assum, G. A palindromic structure in the pericentromeric region of various human chromosomes. *Genome Res.* **6**, 267–279 (1996).
42. Eisenbarth, I., Konig-Greger, D., Wohr, G., Kehrer-Sawatzki, H. & Assum, G. Characterization of an alphoid subfamily located near p-arm sequences on human chromosome 22. *Chromosome Res.* **7**, 65–69 (1999).
43. Hattori, M. *et al.* The DNA sequence of human chromosome 21. *Nature* **405**, 311–319 (2000).
44. Cserpan, I. *et al.* The chAB4 and NF1-related long-range multisequence DNA families are contiguous in the centromeric heterochromatin of several human chromosomes. *Nucleic Acids Res.* **30**, 2899–2905 (2002).
45. Guipponi, M. *et al.* Genomic structure of a copy of the human *TPTE* gene which encompasses 87 kb on the short arm of chromosome 21. *Hum. Genet.* **107**, 127–131 (2000).
46. Kurahashi, H., Shaikh, T. H. & Emanuel, B. S. Alu-mediated PCR artefacts and the constitutional t(11;22) breakpoint. *Hum. Mol. Genet.* **9**, 2727–2732 (2000).
47. Robledo, R. *et al.* A 9.1-kb gap in the genome reference map is shown to be a stable deletion/insertion polymorphism of ancestral origin. *Genomics* **80**, 585–592 (2002).
48. Kouprina, N. *et al.* Segments missing from the draft human genome sequence can be isolated by transformation-associated recombination cloning in yeast. *EMBO Rep.* **4**, 257–262 (2003).
49. Frohme, M. *et al.* Directed gap closure in large-scale sequencing projects. *Genome Res.* **11**, 901–903 (2001).
50. Siniscalco, M. *et al.* A plea to search for deletion polymorphism through genome scans in populations. *Trends Genet.* **16**, 435–437 (2000).
51. Kurahashi, H., Shaikh, T., Takata, M., Toda, T. & Emanuel, B. S. The constitutional t(17;22): another translocation mediated by palindromic AT-rich repeats. *Am. J. Hum. Genet.* **72**, 733–738 (2003).
52. Kurahashi, H. & Emanuel, B. S. Long AT-rich palindromes and the constitutional t(11;22) breakpoint. *Hum. Mol. Genet.* **10**, 2605–2617 (2001).
Sequence characterization of a gap in the human genome and its association with recurrent chromosomal instability.
53. Verkerk, A. J. *et al.* Identification of a gene (*FMR-1*) containing a CGG repeat coincident with a breakpoint cluster region exhibiting length variation in fragile X syndrome. *Cell* **65**, 905–914 (1991).
54. Kasukawa, T. *et al.* Development and evaluation of an automated annotation pipeline and cDNA annotation system. *Genome Res.* **13**, 1542–1551 (2003).
55. Furuno, M. *et al.* CDS annotation in full-length cDNA sequence. *Genome Res.* **13**, 1478–1487 (2003).
56. Mouse Genome Sequencing Consortium. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**, 520–562 (2002).
57. Collins, J. E. *et al.* Reevaluating human gene annotation: a second-generation analysis of chromosome 22. *Genome Res.* **13**, 27–36 (2003).
A careful re-examination of gene annotation on chromosome 22 that identifies common sources of error on the basis of genome structure and limitations of EST/gene databases.
58. Dunham, I. *et al.* The DNA sequence of human chromosome 22. *Nature* **402**, 489–495 (1999).
59. Mounsey, A., Bauer, P. & Hope, I. A. Evidence suggesting that a fifth of annotated *Caenorhabditis elegans* genes may be pseudogenes. *Genome Res.* **12**, 770–775 (2002).
60. Collins, J. E., Mungall, A. J., Badcock, K. L., Fay, J. M. & Dunham, I. The organization of the γ-glutamyl transferase genes and other low copy repeats in human chromosome 22q11. *Genome Res.* **7**, 522–531 (1997).
61. Estivill, X. *et al.* Chromosomal regions containing high-density and ambiguously mapped putative single nucleotide polymorphisms (SNPs) correlate with segmental duplications in the human genome. *Hum. Mol. Genet.* **11**, 1987–1995 (2002).
62. Reich, D. E. *et al.* Linkage disequilibrium in the human genome. *Nature* **411**, 199–204 (2001).
63. Reich, D. E. *et al.* Human genome sequence variation and the influence of gene history, mutation and recombination. *Nature Genet.* **32**, 135–142 (2002).

64. Riethman, H. C. *et al.* Integration of telomere sequences with the draft human genome sequence. *Nature* **409**, 948–951 (2001).
65. Riethman, H. C. *et al.* Mapping and initial analysis of human subtelomeric sequence assemblies. *Genome. Res.* (in the press).
Describes the sequence organization of human subtelomeric regions by implementing a half-YAC strategy to resolve these complex regions of the genome.
66. Larionov, V. *et al.* Specific cloning of human DNA as yeast artificial chromosomes by transformation-associated recombination. *Proc. Natl Acad. Sci. USA* **93**, 491–496 (1996).
67. Kouprina, N. *et al.* Cloning of human centromeres by transformation-associated recombination in yeast and generation of functional human artificial chromosomes. *Nucleic Acids Res.* **31**, 922–934 (2003).
68. Tammi, M. T., Arner, E. & Andersson, B. TRAP: Tandem Repeat Assembly Program produces improved shotgun assemblies of repetitive sequences. *Comput. Methods Programs Biomed.* **70**, 47–59 (2003).
69. Pevzner, P. A., Tang, H. & Waterman, M. S. An Eulerian path approach to DNA fragment assembly. *Proc. Natl Acad. Sci. USA* **98**, 9748–9753 (2001).
70. Paulding, C. A., Ruvolo, M. & Haber, D. A. The *Tre2* (*USP6*) oncogene is a hominoid-specific gene. *Proc. Natl Acad. Sci. USA* **100**, 2507–2511 (2003).
71. Johnson, M. E. *et al.* Positive selection of a gene family during the emergence of humans and African apes. *Nature* **413**, 514–519 (2001).
72. Lupski, J. R. Genomic disorders: structural features of the genome can lead to DNA rearrangements and human disease traits. *Trends Genet.* **14**, 417–422 (1998).
73. RepeatMasker documentation. *Index of RM* [online], <http://repeatmasker.genome.washington.edu/RM/> (1997).

Acknowledgements

We would like to thank F. Collins, A. Pelsenfeld, B. Waterston and E. Lander for helpful comments in the preparation of this manuscript. This work was supported, in part, by grants from the National Institutes of Health to E.E.E.

Competing interests statement

The authors declare that they have no competing financial interests.

Online links

DATABASES

The following terms in this article are linked online to:

LocusLink: http://www.ncbi.nlm.nih.gov/LocusLink

TPTE

FURTHER INFORMATION

GenBank Database:
http://www.psc.edu/general/software/packages/genbank/genbank.html

Human Paralogy Server Segmental Duplication Database:
http://humanparalogy.gene.cwru.edu

Proposal for Construction of a Human Haploid BAC Library from Hydatidiform Mole Source Material:
http://www.genome.gov/Pages/Research/Sequencing/BACLibrary/HydatidiformMoleBAC021203.pdf

University of California Santa Cruz Genome Bioinformatics:
http://genome.ucsc.edu

Access to this interactive links box is free online.