Mapping and Initial Analysis of Human Subtelomeric Sequence Assemblies

Harold Riethman,[1,4] Anthony Ambrosini,[1] Carlos Castaneda,[1,2] Jeffrey Finklestein,[1,3] Xue-Lan Hu,[1] Uma Mudunuri,[1] Sheila Paul,[1] and Jun Wei[1]

[1]*The Wistar Institute, Philadelphia, Pennsylvania 19104, USA*

Physical mapping data were combined with public draft and finished sequences to derive subtelomeric sequence assemblies for each of the 41 genetically distinct human telomere regions. Sequence gaps that remain on the reference telomeres are generally small, well-defined, and for the most part, restricted to regions directly adjacent to the terminal (TTAGGG)n tract. Of the 20.66 Mb of subtelomeric DNA analyzed, 3.01 Mb are subtelomeric repeat sequences (Srpt), and an additional 2.11 Mb are segmental duplications. The subtelomeric sequence assemblies are enriched >25-fold in short, internal (TTAGGG)n-like sequences relative to the rest of the genome; a total of 114 (TTAGGG)n-like islands were found, 55 within Srpt regions, 35 within one-copy regions, 11 at one-copy/Srpt or Srpt/segmental duplication boundaries, and 13 at the telomeric ends of assemblies. Transcripts were annotated in each assembly, noting their mapping coordinates relative to their respective telomere and whether they originate in duplicated DNA or single-copy DNA. A total of 697 transcripts were found in 15.53 Mb of one-copy DNA, 76 transcripts in 2.11 Mb of segmentally duplicated DNA, and 168 transcripts in 3.01 Mb of Srpt sequence. This overall transcript density is similar (within ~10%) to that found genome-wide. Zinc finger-containing genes and olfactory receptor genes are duplicated within and between multiple telomere regions.

[Supplemental material is available online at www.genome.org. Detailed maps, subtelomeric assemblies (FASTA format), and transcript annotations are also available at our laboratory Web site (http://www.wistar.upenn.edu/Riethman.]

Telomeres are extraordinarily dynamic chromosomal structures. They are essential for genome stability and faithful chromosome replication and mediate a host of key biological activities, including cell cycle regulation, cellular aging, movements and localization of chromosomes within the nucleus, and transcriptional regulation of subtelomeric genes (Blasco et al. 1999; Feuerbach et al. 2002). Specialized functions involving telomeric and subtelomeric DNA have evolved in a wide range of eukaryotes; for example, frequent subtelomeric gene conversion provides diversity for surface antigens in Trypanosomes (McCulloch et al. 1997), and rapidly evolving subtelomeric gene families confer selective advantages for closely related yeast strains (Carlson et al. 1985).

A conserved, (TTAGGG)n tract forms the DNA component of each chromosome terminus in humans (Moyzis et al. 1988). A specialized enzyme called telomerase can lengthen the telomere repeat motif by adding on motif-specific nucleotides in a DNA template-independent manner (Morin 1989). However, both telomerase-associated and telomerase-independent pathways for maintaining (TTAGGG)n repeats exist; the major telomerase-independent pathways are recombination based, sometimes involve coamplification of subtelomeric sequences along with the simple repeat tracts found at chromosome termini (Murnane et al. 1994; Bryan et al. 1995; Lundblad and Wright 1996; Henson et al. 2002) and can generate very long and heterogeneous stretches of (TTAGGG)n-containing repeats (Rizki and Lundblad 2001; Henson et al. 2002). Transcription of subtelomeric genes can be regulated by (TTAGGG)n tract length (Baur et al. 2001) and by subtelomeric repeat content and abundance, possibly by contributing specific sequence elements necessary for local silencing (Fourel et al. 1999; Pryde and Louis 1999) or by providing extended homology regions required for somatic pairing and heterochromatin formation (Donaldson and Karpen 1997).

Subtelomeric DNA, along with pericentromeric chromosome regions, are preferential sites of segmentally duplicated DNA. Estimated to comprise ~5% of the human genome, this class of low-copy repeat DNA is characterized by very high sequence similarity (90% to >99.5%) between homology tracts, and variable, but often very large tract lengths (1 kb to >200 kb). These large homology segments have complicated mapping and sequencing efforts, and caused a disproportionate number of assembly errors in the initial working draft sequence of the human genome. Segmental duplications can predispose associated chromosome segments to genetic instability and have been connected with several genetic diseases (Bailey et al. 2001). Evolutionarily recent duplicative transposition of these large DNA tracts has led to the generation of new gene families and to the formation of fusion transcripts with potentially new functions (Bailey et al. 2002). In this study, we define segmental duplications occurring in more than one subtelomeric region "subtelomeric repeats," and refer to all others simply as segmental duplications.

Large variant alleles of many human subtelomeric regions exist, and are believed to consist entirely of subtelomeric repeats (Wilkie et al. 1991; Macina et al. 1994, 1995; Trask et al. 1998; Mefford and Trask 2002). For example, Wilkie et al (1991) found that three alleles varying in length up to 260 kb exist at the 16p telomere. Trask et al. (1998) examined the structure and genomic distribution of a cosmid-sized block of segmentally duplicated subtelomeric DNA. They found that this block was consistently present at the 3q, 15q, and 19p telomeres in humans, was variably distributed at an additional subset of human telomeres, but

Present Addresses: [2]Program in Molecular Biophysics, Johns Hopkins University, Philadelphia, PA 19104, USA; [3]Cell and Molecular Biology Program, University of Pennsylvania, Baltimore, MD 21218, USA.
[4]Corresponding author.
E-MAIL Riethman@wistar.upenn.edu; FAX (215) 898-3868.
Article and publication are at http://www.genome.org/cgi/doi/10.1101/gr.1245004.

14:18–28 ©2004 by Cold Spring Harbor Laboratory Press ISSN 1088-9051/04; www.genome.org

was present in a single copy in nonhuman primate genomes. Similar studies have demonstrated more recently that the evolution of most primate subtelomeric regions has involved multiple, lineage-dependent duplications in recent evolutionary time (Martin et al. 2002; van Geel et al. 2002). The duplications have colonized many individual human subtelomeric regions in a variable fashion since the divergence of human and primate lineages.

A complete reference sequence for each human subtelomere region is an essential starting point for analysis of their function and evolution. Here, we report the mapping and initial analysis of a complete set of subtelomeric sequence assemblies. Comprised of both draft and finished public sequence accessions available as of August 1, 2003, the draft fragments are properly ordered and the assemblies are positioned relative to the respective telomere. These properties permit a comparison of subtelomeric sequence organization at each of the separate human telomeres, and the proper placement of transcripts relative to subtelomeric sequence elements and terminal (TTAGGG)n tracts.

RESULTS AND DISCUSSION

Preparation and Mapping of Subtelomeric Assemblies

Subtelomeric clones and sequence accessions that were identified and connected to telomeres previously (Riethman et al. 2001) were used in this study to nucleate the assembly of new and more complete subtelomeric draft/finished sequence contigs for these regions. Most of the sequence used in these assemblies (>98%) was acquired by the IHGSC as part of the finishing phase of the Human Genome Project, from clones contributed by our lab as well as from clones identified independently in the chromosome-specific projects and mapped relative to telomeres in our lab. Each sequence assembly is oriented from telomeric end (nucleotide position 1) to centromeric end. Maps and tables describing in detail the YAC, BAC, and cosmid clones supporting the sequence assemblies for each subtelomere region are provided as Supplemental material available online at www.genome.org (Suppl. Table 1; Suppl. Figs. 1ptel–Xqtel). Most of the subtelomeric sequences were ultimately derived from BAC sources (see Suppl. Table 1); ~1.6 Mb (7.7%) of the assembled sequence was derived solely from half-YACs.

Figure 1 summarizes the present status of sequence completion for each subtelomeric region. Finished or draft sequence extends to the terminal (TTAGGG)n tract of reference sequences for 19 telomeres (2p, 4p, 7q, 8p, 8q, 9p, 9q, 10q, 11p, 11q, 15q, 16p, 17p, 17q, 18p, 18q, 21q, Xp/Yp, Xq/Yq). For four of these (8p, 9q, 11p, 16p), the completed reference sequence is that of the smallest of several polymorphic allelic variants (each variant differs in size by hundreds of kilobases). It is important to note that the current reference sequence for a given telomere region represents only one of several possible subtelomeric variants in the population for many of the telomeres (see Table 1). The variant regions appear to be comprised largely or wholly of segmentally duplicated subtelomeric sequences (Wilkie et al. 1991; Trask et al. 1998; Bailey et al. 2002).

Assemblies mapping to <20 kb or between 20 and 70 kb from the respective telomere are available for most of the remaining telomeres (Fig. 1). One telomere (20q) is marked by a sequence assembly that extends from single copy into a subtelomeric repeat region, but the size of this subtelomeric repeat region has not been determined. The five telomeres from the p-arms of the acrocentric chromosomes, which contain mainly repetitive DNA, were not characterized as part of this study. Half-YACs recovered from these regions, although somewhat unstable mitotically, are currently being used to characterize sequences contained in these heterochromatic telomere regions.

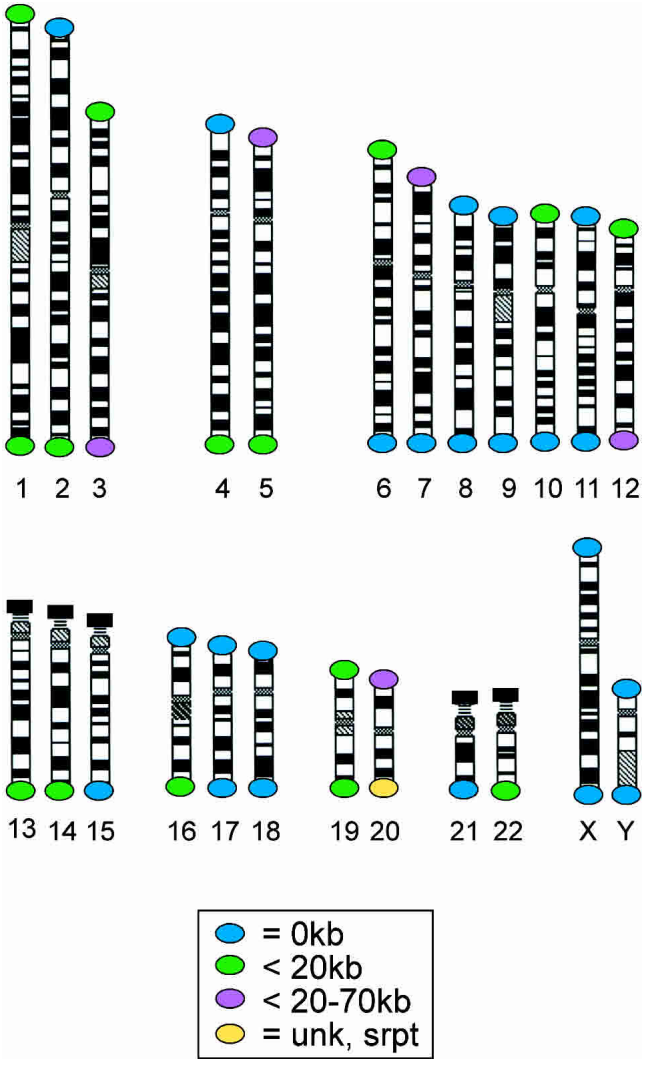

Figure 1 Telomere sequence gaps. Distance from terminal (TTAGGG)n tract to the subtelomeric sequence assemblies for each telomere is indicated. (Blue dot) Subtelomeric assemblies adjoin terminal (TTAGGG)n tract for a reference telomere; (green dot) subtelomeric assemblies end within 20 kb of the terminal (TTAGGG)n tract; (magenta dot) subtelomeric assemblies end between 20 and 70 kb from the terminal (TTAGGG)n tract; (yellow dot) a half-YAC clone has not been identified, the sequence assembly for this telomere extends from single-copy DNA into the subtelomeric repeat region, but the size of the subtelomeric repeat region has not been determined. (Black rectangles) The five telomeres from the p-arms of the acrocentric chromosomes, which contain mainly repetitive DNA unstable in yeast and bacteria, were not characterized as part of this study.

Sequence Organization of Subtelomeric DNA

The overall sequence organization of each subtelomeric assembly was evaluated initially in terms of subtelomeric repeats, segmental duplications, satellite sequences, and internal (TTAGGG)n—like sequence content. First, a BLAST-able database of the subtelomeric assemblies was created; Srpt sequences were defined as any nonself sequence match >90% identity within the subtelomeric sequence database. Second, sequence comparisons between the subtelomeric assemblies and public databases were used to define additional homology segments; segmental dupli-

cations were defined as nonself sequence matches in the NR or HTGS databases, but absent in the Srpt database, that were >90% identity and >1 kb in length. The Whole Genome Shotgun Sequence Detection (WSSD) Database (Bailey et al. 2002) detects both subtelomeric repeat regions and segmental duplications, and was used to add segmental duplication regions where those

Table 1. Summary of Subtelomeric Assemblies

Tel	Distance to telomere (kb)	Srpt region (kb)	Seg dups (kb)	One-copy region (kb)	Internal gaps (kb)	Evidence for tel linkage[a]	Large-scale variation[b]
1p	15	286	23	191	0	RARE, Srpt	Hi
1q	10	15	87	301	97	RARE, Srpt	Lo
2p	0	45	0	455	0	RARE, Srpt	Lo
2q	16	107	0	393	0	RARE, Srpt	Lo
3p	10	0	66	434	0	Srpt	unk
3q	50	127	63	310	0	Srpt	unk
4p	0	27	31	442	0	RARE, Srpt	unk
4q	10	127	266	107	0	Srpt	Hi
5p	70	53	72	375	0	RARE, Srpt	unk
5q	20	189	24	287	0	RARE, Srpt	Lo
6p	5	106	98	296	0	RARE, Srpt	Hi
6q	3	159	65	276	0	RARE, Srpt	Lo
7p	34	113	0	377	10	RARE, Srpt	Hi
7q	0	2	0	498	0	RARE, Srpt	Lo
8p	89 (A)	244	0	256	0	RARE, Srpt	Hi
	0 (B)					RARE, Srpt	
	3 (C)					Srpt	
8q	0	7	0	493	0	RARE, Srpt	Lo
9p	0	40	154	306	0	RARE, Srpt	unk
9q	unk (50), (A)	135	0	365	0	Srpt	Hi
	0 (B)					Srpt	Hi
10p	14	74	3	423	0	Srpt	unk
10q	0	89	25	386	0	RARE, Srpt	Hi
11p	150 (A)	137	2	361	0	RARE, Srpt	Hi
	0 (B)					RARE, Srpt	Hi
11q	0	0	55	445	0	Srpt	unk
12p	16	42	0	458	0	Srpt	Hi
12q	60	0	0	500	0	RARE, Srpt	Lo
13q	15	15	0	385	100	RARE, Srpt	Lo
14q	7	1	499	0	0	RARE, Srpt	Hi
15q	0	117	62	321	0	Srpt	Lo
16p	152 (A)	38	6	456	0	RARE, Srpt	Hi
	0 (B)					RARE, Srpt	Hi
16q	5	150	26	324	0	Srpt	Hi
17p	0	29	17	407	47	RARE, Srpt	unk
17q	0	44	0	456	0	RARE, Srpt	unk
18p	0	103	19	378	0	RARE, Srpt	Lo
18q	0	5	0	495	0	RARE, Srpt	Lo
19p	11	208	0	292	0	RARE, Srpt	Lo
19q	5	17	0	483	0	Srpt	Hi
20p	105 (A)	0	0	500		Srpt	unk
	55 (B)				0	Srpt	unk
20q	unk (50)	55	0	445	0	Srpt	Hi
21q	0	26	5	469	0	RARE, Srpt	Lo
22q	20	20	60	380	40	Srpt	unk
Xp/Yp	0	0	115	385	0	PPGE, Srpt	unk
Xq	0	30	76	394	0	PFGE, Srpt	unk
Yq	0	30	195	225	50	PFGE, Srpt	unk

[a]Where designated, mapping experiments using a site-specific cleavage method (RARE cleavage; Riethman et al. 1997) have been done to demonstrate colinearity of the half-YAC insert DNA with the cognate telomere. In the absence of RARE cleavage data, the presence of subtelomeric repeats adjacent to terminal (TTAGGG)n sequences in a half-YAC clone is taken as strong evidence for proximity to the telomere; this has been borne out by the RARE cleavage experiments carried out so far. The BAC clones used to mark two telomeres for which no half-YAC coverage exists (5p and 20q) were identified by their subtelomeric repeat-sequence content, the presence of an internal telomere repeat sequence, and their localization to the telomeric end of a distal contig in the Global BAC map. The cosmid used to mark the telomere of 19q contains subtelomeric repeats and an internal telomere repeat sequence, and forms the telomeric terminus of the 19q metric physical map (http://greengenes.llnl.gov/genome/). On the basis of the known sequence organization of other telomeres, only additional subtelomeric repeat sequence is likely to reside distal to the subtelomeric repeat segments contained in these clones, although the possibility of single-copy DNA distal to them cannot be formally excluded at present.
[b]Telomeres with a frequency of >10% large variant alleles in the small populations sampled are considered to have Hi polymorphism in the context of this paper, and those with <10% large variant alleles are considered to have Lo polymorphism. For the telomeres not listed, no molecular data are available with respect to large-scale variations and the available FISH data are inconclusive with respect to potential large-scale variation. The polymorphism frequencies detected by FISH are minimum numbers, as detection depends upon the variable presence/absence of only one specific FISH probe at the telomere. The size(s) of the polymorphisms cannot be determined by FISH, but are assumed to be at least the size of the probe used (on the basis of similar FISH signal intensities at all sites). Data on polymorphic telomeres are from Riethman, unpubl. results; Wilke et al. (1991); Ijdo et al. (1992); Cook et al. (1994); Macina et al. (1994, 1995); Martin-Gallardo et al. (1995); Reston et al. (1995); Monfouilloux et al. (1998); Trask et al. (1998); van Overveld et al. (2000).

sequences were not already identified by the BLAST analysis described above; for the most part, the regions identified by WSSD were consistent with our analyses, although each method missed a small percentage of the duplications. Finally, satellite-related and (TTAGGG)n-related sequences were identified using the high-sensitivity RepeatMasker parameters. TAR1 and SUBTEL are telomere-associated satellite sequences identified by RepeatMasker (other telomere-associated repeat sequences identified in the literature are components of the Srpt fraction of subtelomeric DNA). Each of these sequence elements are delineated on Figure 2; the total sizes of Srpt and segmental duplication within each assembly are indicated in Table 1.

The bulk of Srpt sequences are confined to the most distal regions of the subtelomere (Fig. 2), although there are several examples (2p, 2q, 5p, 7p, 8p, and 12p) where, in addition to a terminal block of Srpt, there are additional smaller segments interspersed within the adjacent one-copy DNA and segmentally duplicated DNA. Several of the incompletely sequenced telomeres lack Srpt in the assembled sequence (Table 1); because Srpts were identified in the half-YACs derived from these telomeres, a small Srpt region confined to close to the terminal (TTAGGG)n is anticipated for these telomeres. Segmental duplication blocks were often found adjacent to Srpts, but displayed a highly variable pattern of content and distribution at each chromosome end (Fig. 2; Table 1). Overall, 14.6% of the 20.66 Mb of subtelomeric DNA analyzed was comprised of Srpt and 10.2% of segmentally duplicated DNA, for a total of 24.8% segmental duplications of both types. Genome-wide, an estimated 5% of genomic DNA is believed to contain segmentally duplicated sequences (Bailey et al. 2002), indicating a fivefold enrichment of segmentally duplicated DNA in the subtelomeric regions analyzed. The nucleotide sequence similarity of duplicons in both the Srpt and segmental duplications varied from 90% to >99%, and occurred in sequence blocks that often, but not always, had sharply defined boundaries; more extensive comparative analysis of these regions and related sequences in nonhuman primate species (e.g., see Fan et al. 2002; Martin et al. 2002) are required to investigate their origin and evolution in detail.

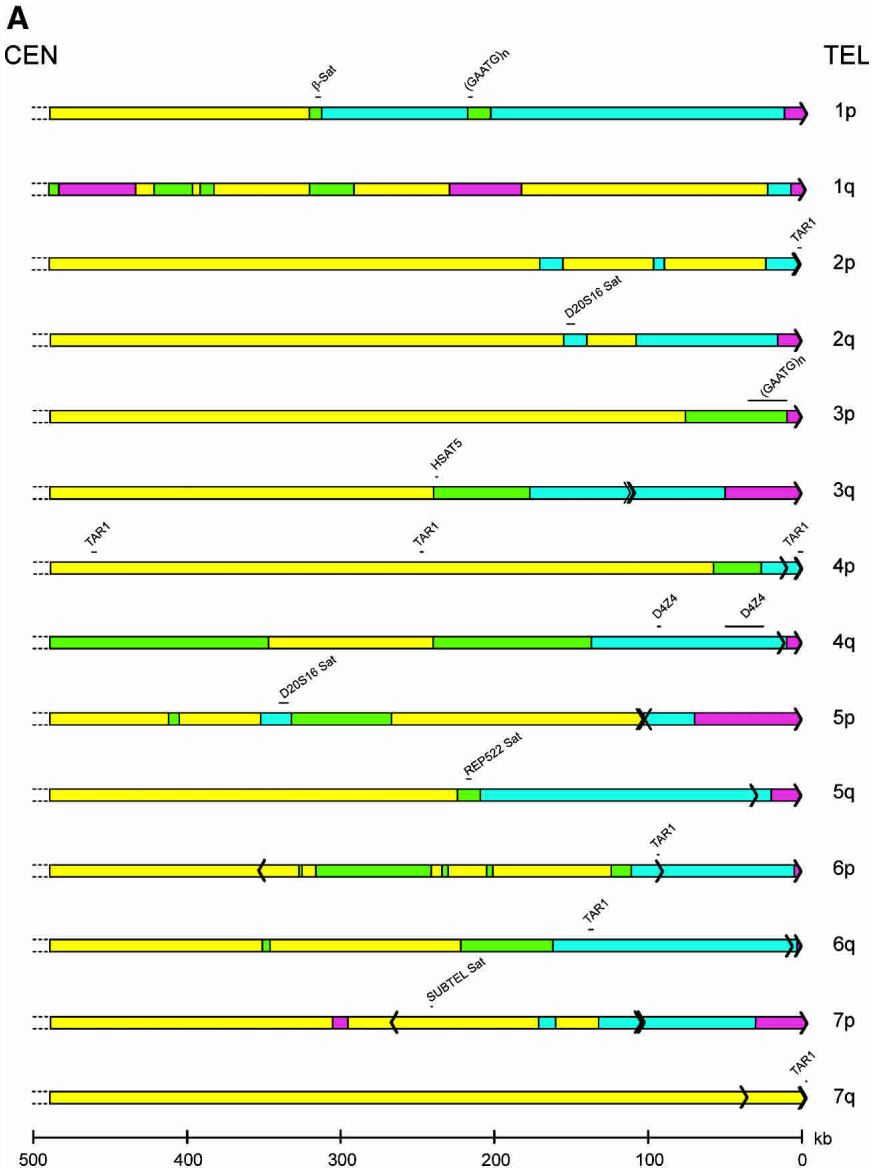

Figure 2. (Continued on next page)

Interstitial (TTAGGG)n-like sequence distribution was examined because of its potential role in subtelomeric recombination and telomere healing (Mondello et al. 2000; Azzalin et al. 2001; Ruiz-Herrera et al. 2002), and its hypothesized role as a boundary element for subtelomeric DNA compartments (Flint et al. 1997b). All significant RepeatMasker matches to the simple repeats (TTAGGG)n and (CCCTAA)n were counted as telomere-like sequence islands. A total of 114 matches were found within the subtelomeric sequence assemblies. The 5′–3′- orientation of the G-rich strand of the repeat is normally toward the telomere in the (TTAGGG)n tracts at the ends of chromosomes. Most of the telomere-like sequence islands followed this strand orientation (106/114 islands). Thirteen (TTAGGG)n islands corresponded to the beginning of terminal (TTAGGG)n tracts at the 2p, 4p, 7q, 9p, 10q, 11q, 16p, 17q, 18p, 18q, 21q, Xp/Yp, and XqYq telomeres; these were excluded from analysis of the interstitial, telomere-like sequence islands described below. The 5′–3′ (TTAGGG) orientation of individual islands is indicated by the direction of the arrows representing the sequence islands in the telomere diagrams (Fig. 2).

The 101 internal (TTAGGG)n-like sequence islands were analyzed in greater detail as shown in Figure 3. The sizes of (TTAGGG)n-like sequence islands (x-axis), number of occurrences for a given size of (TTAGGG)n tract (y-axis), similarity of (TTAGGG)n-like sequence islands to a perfect (TTAGGG)n tract (percent Divergence), and location of (TTAGGG)n-like sequence islands within the subtelomeric sequence organization as defined above (Srpt, one-copy, and boundary) are indicated in Figure 3. The internal subtelomeric (TTAGGG)n-like sequence islands ranged in size from 24 to 823 bp; most were in a rather tight size range of 151–200 bp. Those shorter than this size tended to be in one-copy sequence regions, those longer in Srpt sequence. The

boundary (TTAGGG)n islands ranged from 57 to 257 bp in size. There were 55 (TTAGGG)n-like sequence islands in Srpt, 0 in Segmental duplications, and 35 in one-copy regions. Eleven (TTAGGG)n-like sequence islands were at boundaries (two at SD/Srpt, nine at Srpt/one-copy). Four (TTAGGG)n-like islands that occurred at the allele boundaries were within the internal Srpt regions of long subtelomeric alleles (and were counted as such for this analysis), but mapped to the precise coordinates of the termini of shorter alleles for these same telomeres (8p, 9q, 11p, 16p; see Fig. 2). This suggests that the longer alleles of these telomeres might have been formed by simple addition of a terminal subtelomeric sequence segment to a pre-existing telomere.

A comparison of the number of interstitial (TTAGGG)n-like islands found in subtelomeric DNA with those found genome wide shows that, in a normalized comparison (occurrences per 20.66 Mb), (TTAGGG)n-like islands are highly enriched (>25-fold) in subtelomeric regions. In addition, they tend to be both longer and more similar to perfect (TTAGGG)n tracts in subtelomeric DNA compared with elsewhere in the genome (Fig. 3). From an evolutionary perspective, this suggests that most subtelomeric interstitial (TTAGGG)n tracts have arisen more recently than those found elsewhere in the genome, have originated via a separate mechanism than (TTAGGG)n islands found elsewhere (e.g., see Azzalin et al. 2001), or are under some selective pressure to maintain similarity to (TTAGGG)n (Flint et al. 1997b).

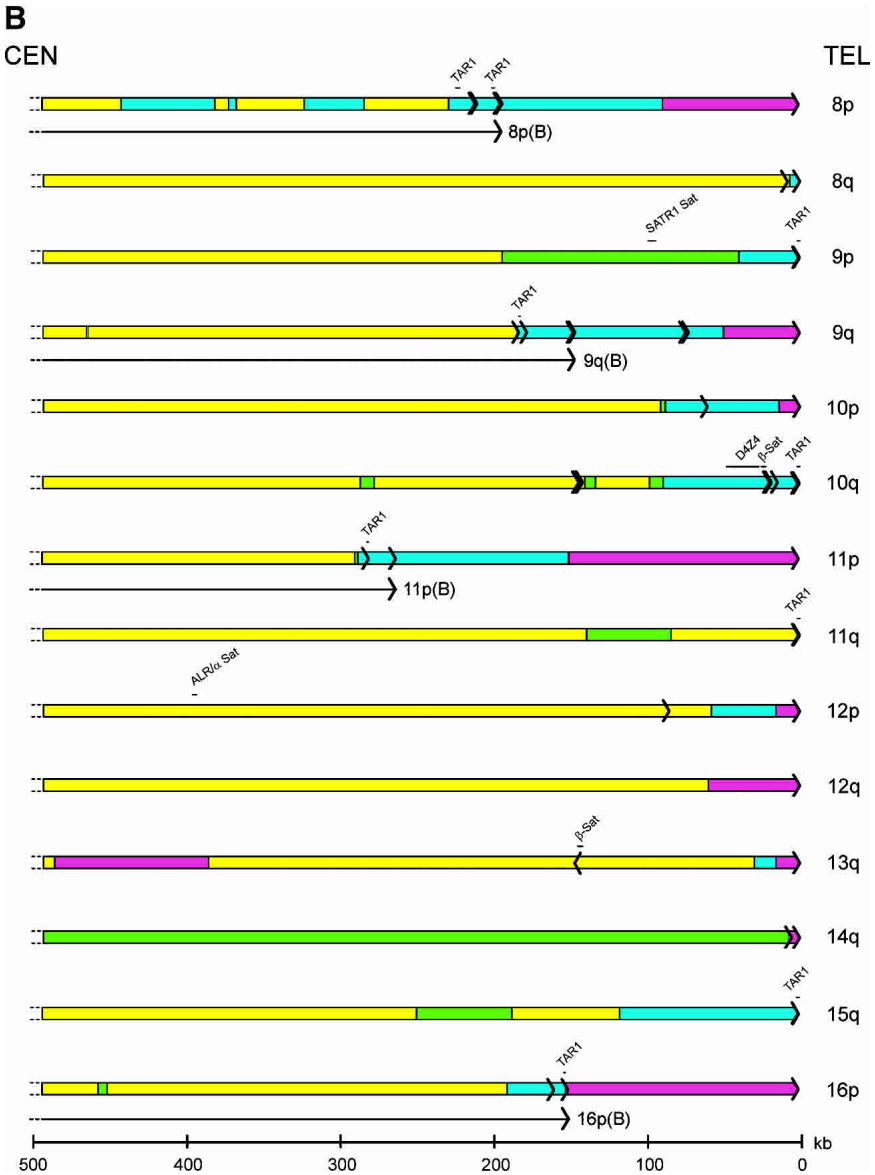

Figure 2. (Continued on next page)

GC Content and Interspersed Repeat Composition of Subtelomeric Sequence Assemblies

RepeatMasker (Smit and Green, RepeatMasker at http://ftp.genome.washington.edu/RM/RepeatMasker.html) was used to analyze the sequences for interspersed repeats and for GC content. The summary results of this analysis are shown in Figure 4, and the detailed breakdown is given in Supplemental Table 2. When taken as a whole, the subtelomeric one-copy regions had an elevated GC content (47.9%), whereas Srpt and segmentally duplicated regions had a slightly elevated GC content (44.0% and 43.0%, respectively), relative to the genome-wide average of 41.6%. However, there were wide fluctuations in GC content at individual telomeres, ranging from 62.5% GC content of the one-copy region of 1p to 37.5% GC content in the one-copy region of 3p (see Supplemental Table 2); several of the most GC-rich subtelomere regions contained one or more clusters of G-rich minisatellites. Similarly, interspersed repeat content, taken as a whole, did not display dramatic biases relative to the genome-wide averages (Fig. 4), but very large subtelomere-specific biases and sometimes strand biases were seen in LINE, SINE, LTR, and DNA repeat content (see Supplemental Table 2). However, no universal patterns of interspersed sequence composition emerged that clearly distinguish subtelomeric DNA from other regions of the human genome.

Transcript Content of Subtelomeric Assemblies

Transcripts were annotated in each subtelomeric assembly, noting their mapping coordinates relative to their respective telomere, and whether they originate in duplicated DNA or single-copy DNA. We used a database of unique transcripts representing each Unigene cluster (Schuler 1997; ftp://ftp.ncbi.nih.gov/repository/UniGene/; Hs.seq.uniq.Z file available from the Unigene build available July 1, 2003 containing transcript sequences representing ~124,000 Unigene clusters) for our initial annotation. Repeat-masked subtelomeric assemblies were analyzed by BLAST, and transcripts with matches >50 bp with 85% or greater identity were collected and parsed into a second database. Each transcript within this candidate database was compared with its cognate unmasked subtelomeric assembly using the program Spidey (Wheelan et al. 2001). Those with >95% sequence identity over at least 50% of the transcript length were displayed on the Genotator browser

(Harris 1997) and examined individually using Blixem (Sonnhammer and Durbin 1994). The same set of transcripts was displayed on the UCSC browser (Kent et al. 2002). The single transcript with the best nucleotide sequence match over the greatest proportion of the transcript in a given segment of the sequence was annotated. The complete set of transcripts with their corresponding coordinates within each subtelomere assembly, their percent identity within the matching sequence, and the proportion of the transcript covered by matching bases, is summarized in Table 2 and detailed in Supplemental Table 3.

A total of 941 subtelomeric transcripts were annotated in this manner, 697 from one-copy genomic regions and 244 from segmentally duplicated DNA and subtelomeric repeat DNA. Overall, the subtelomeric region is slightly enriched in Unigene transcripts (48 transcripts/Mb) relative to the genome-wide average (41 transcripts/Mb). The enrichment of transcripts in subtelomeric DNA is consistent with earlier studies, (Saccone et al. 1993; Flint et al. 1997a,b), although there is a great deal of variation in transcript concentration from telomere to telomere (Table 2).

Fifteen percent of the transcript matches localizing to one-copy regions either had apparent disruptions in their predicted ORFs or varied significantly (>1% in high-quality parts of the sequence) from the corresponding genomic sequence. These were designated "possible pseudogenes" (see Supplemental Table 3). However, given the frequency of sequence errors in the EST and mRNA database, as well as the draft nature of parts of the assemblies, these numbers are likely to change as experimental validation of the transcript annotations proceeds.

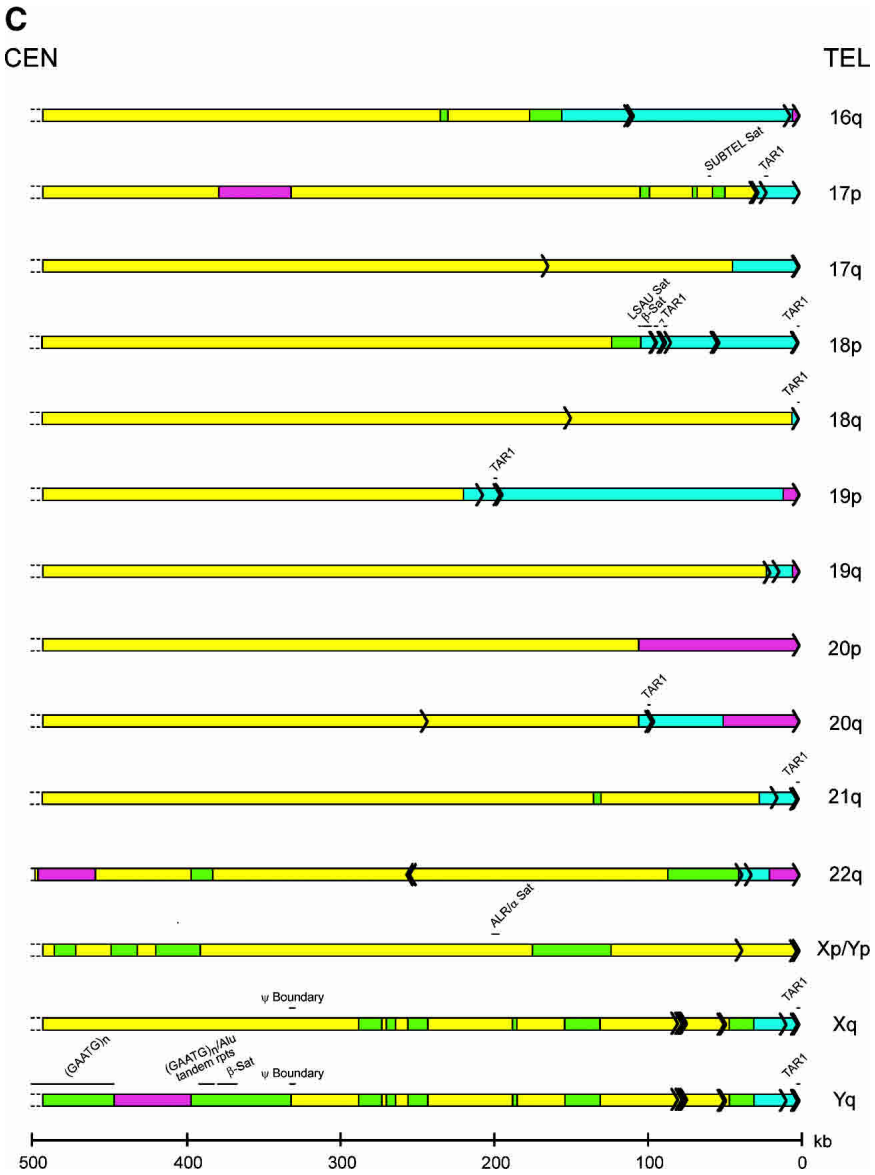

Figure 2 Sequence organization of subtelomeric assemblies. The terminal fragment of each chromosome arm that contains a subtelomeric sequence assembly is depicted. (Yellow) Single-copy DNA; (green) segmentally duplicated DNA unique to a single subtelomere region; (blue) segmentally duplicated subtelomeric repeat DNA; (magenta) unsequenced subtelomeric DNA. Each black arrow within the rectangles depicts a single (TTAGGG)n-like sequence, with the arrow pointing in the 5′ to 3′ direction for the G-rich strand. The position of satellite sequences detected by RepeatMasker software is indicated above each telomere diagram. The positions of the shortest reference alleles for the 8p, 9q, 11p, and 16p telomeres are indicated by a black line segment terminating with an arrow beneath the sequence assemblies for these telomeres.

Similarly, an unknown but significant fraction of the transcripts embedded within the segmental duplications and subtelomeric repeats are likely to be pseudogenes (e.g., see Kermouni et al. 1995; Amann et al. 1996; Flint et al. 1997a), whereas others are likely to be members of gene families with many closely related, but nonidentical functional transcripts (e.g., Flint et al. 1997b; Mah et al. 2001; Fan et al. 2002). In most cases, it is very difficult to clearly identify pseudogenes in Srpt regions; there are many large-scale structural polymorphisms involving hundreds of kilobases of subtelomeric DNA, and it is likely that many variant copies of subtelomeric repeat loci exist in the human population, but are currently absent from sequence databases. For example, genomic Srpt loci that encode partial transcripts in a particular reference sequence might have cognate, unsequenced variant loci in the human population that encode full transcript sequences. Similarly, ESTs obtained from subtelomeric regions of some individuals will necessarily have precise sequences slightly different from those in the reference sequences if the EST was transcribed from a variant subtelomere segment absent in the current assembly. Finally, transcribed pseudogenes, as well as noncoding transcripts can clearly have important biological roles, and it is important to catalog them where found. A great deal of additional, detailed work is required to sort through each of the potential gene/pseudogene families embedded in Srpt and segmentally duplicated DNA to identify the genomic origins of particular transcripts and to determine whether/what fraction of the transcripts might encode functional proteins. Therefore, at this early stage, we think it is prudent to annotate all of the transcript matches to properly lay the groundwork for more detailed analyses.

Supplemental Table 3 identifies each of these transcripts, and Tables 3A and 3B summarize the subset of transcripts in duplicated DNA that correspond to named

genes. Cross-boundary transcripts (Table 3B) contain part of a sequence from a duplicated genomic segment and part from a one-copy segment, or parts from a segmental duplication and from a subtelomeric repeat. These transcripts might represent transcribed pseudogenes generated by juxtaposition of progenitor transcript segments, or might generate new functionalities by virtue of exon shuffling upon duplication (Bailey et al. 2002; Fan et al. 2002); they include transcripts for an F-box protein, for a Zinc finger-containing protein, and for many unknown potential proteins (see Supplemental Table 3 for full list). It will ultimately be essential to acquire complete finished sequences for each distinct allele of each subtelomeric region in order to identify and analyze these genes and gene families, and to deconvolute the many instances of overclustered Unigenes and mRNAs derived from separate but highly similar duplicated genomic DNA fragments.

Subtelomeric gene families with members having nucleotide sequence similarity in the 70%–90% level include the immunoglobulin heavy-chain genes (found at 14q), olfactory receptor genes [one-copy regions of 1q, 5q, 10q, and 15q as well as previously characterized subtelomeric repeat DNA (1p, 6p, 8p, 11p, 15q, 19p, and 3q; Trask et al. 1998)], and zinc-finger genes (4p, 5q, 8p, 8q, 12q, 19q). Transcripts for multiple members of these gene families were found within many of the individual subtelomeric regions (see Supplemental Table 3). The abundance of gene families in subtelomeric regions is a common feature of most eukaryotes, and may reflect a generally increased recombination and tolerance of subtelomeric DNA for rapid evolutionary change.

Transcripts positioned closest to the telomere represent genes with the highest susceptibility to telomere deletions, rearrangements, and hypothesized position effects mediated by telomere (TTAGGG)n tract shortening and/or altered telomeric heterochromatin. Both the dosage (in the case of Srpt transcripts) and the true position of many of these genes relative to the telomere will be allele dependent, changing with different subtelomeric repeat composition and organization. Nonetheless, current data permit us to identify some representatives of most Srpt gene families, and nearly all of the most distal one-copy genes. The named one-copy transcripts closest (within 100 kb) of the telomeric end of each assembly are shown in Table 4. These distal one-copy transcripts, along with the Srpt and segmental duplication transcripts described above, should comprise the segment of the human transcriptome most susceptible to telomere truncations, rearrangements, and telomere-associated position effects.

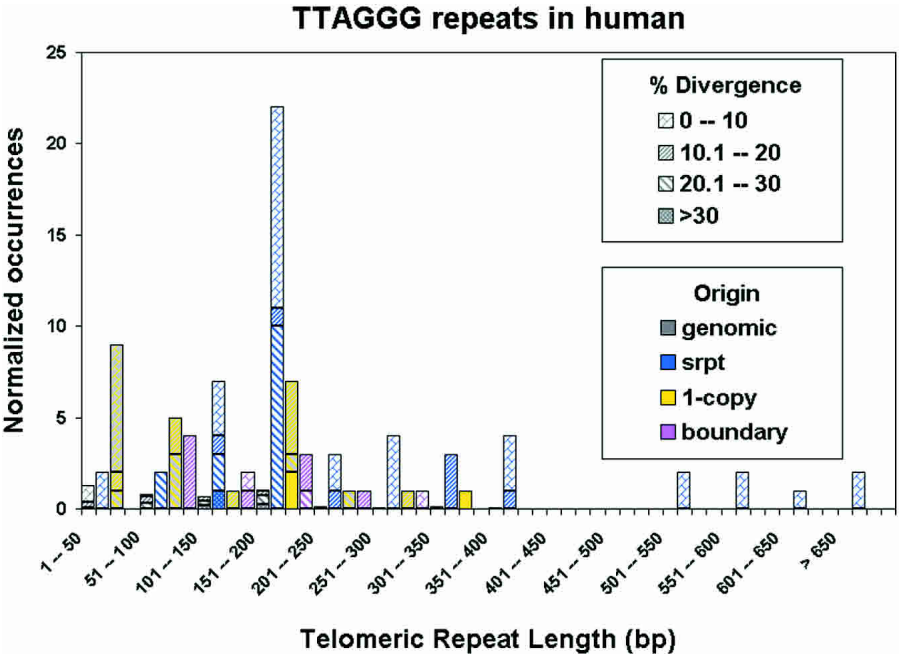

Figure 3 Characteristics of Interstitial (TTAGGG)n-like sequences in subtelomeric assemblies. (TTAGGG)n-like sequences detected using RepeatMasker were classified according to origin within the defined subtelomeric sequence classes (Subtelomeric repeat, blue; single-copy, yellow; boundary, magenta; none were found within Segmental Duplications), similarity to a perfect (TTAGGG)n sequence (percent divergence), the length of the (TTAGGG)n tract (x-axis), and the number of occurrences for each size class (y-axis). The short black histograms show the distribution and relative abundance of nonsubtelomeric (TTAGGG)n hits in the human genome (627 hits in 3098 Mb of DNA, including an unusual cluster of 166 hits in a 9-Mb region of Yq11.22–Yq11.23), normalized to the 20.66 Mb of subtelomeric DNA analyzed for comparison.

METHODS

Preparation of Subtelomeric Assemblies and Subtelomeric Maps

Each subtelomeric assembly was prepared by DNA sequence comparison of finished sequence accessions, draft sequence accession pieces, and half-YAC (Riethman et al. 1989; Kvaloy 1993) and cosmid end sequences (Riethman et al. 2001) from a given subtelomeric region. We extended the size of each subtelomeric region centromerically to include 500 kb of DNA, making use of public clone contigs and sequence overlaps of clones adjacent to the initial contig. Draft sequences were broken into their component pieces and imported along with all finished sequences into a telomere-specific Sequencher file containing all half-YAC-derived sequences and cosmid contig end sequences available.

Finished sequences for each telomere were used preferentially in the assemblies, with draft sequence fragments added as necessary to extend the assemblies. We used all or parts of NCBI assemblies from Build 34 first, then patched in draft sequences not included in the assembly. In regions in which NCBI Build 34 was inconsistent with our mapping data, we used individual accessions to complete the assemblies. The Sequencher assembler was used interactively to find and combine sequence overlaps among the imported pieces and between the half-YAC-derived sequences and the imported sequences. It was often necessary to break sequence fragments in VNTR-like regions and introduce a gap in one of the overlapping fragments (in effect, incorporating the larger sequence of a polymorphic VNTR) in order to obtain a contiguous assembly in this manner. Leftover draft sequence fragments were analyzed by BLAST to ensure that unique sequences were not missed in the assembly. A string of 100 Ns were placed between nonoverlapping, but adjacent draft sequence fragments. By use of the mapping data associated with the half-YAC-derived cosmid contigs, it was possible to uniquely orient and position most draft-sequence fragments. Subsequent comparison of each subtelomeric assembly against itself using PatternHunter software (Ma et al. 2002) revealed no instances of what appeared to be assembly generated duplications in the sequence.

We did not make any special effort to trim high-quality overlapping se-

Table 2. Summary of Subtelomeric Transcripts

	Total Srpt	Total SD	Total one-copy	Total	Known Srpt	Known SD	Known one-copy	Total
1ptel	18	2	23	43	5	1	6	12
1qtel	2		13	15			4	4
2ptel			15	15			2	2
2qtel	1		27	28			8	8
3ptel		2	8	10			2	2
3qtel	11	4	16	31	4		4	8
4ptel	1	1	14	16			2	2
4qtel	5	7	2	14	2	1		3
5ptel	4	5	12	21		1	5	6
5qtel	8		18	26	3		7	10
6ptel	4	3	14	21	3		2	5
6qtel	3	4	16	23	1	3	3	7
7ptel	9		20	29	2		2	4
7qtel			17	17			1	1
8ptel	14		15	29	3		0	3
8qtel			28	28			9	9
9ptel	6	18	8	32	2	3		5
9qtel	6		5	11	2		2	4
10ptel			16	16			4	4
10qtel	6		29	35	1		12	13
11ptel	6		34	40	3		17	20
11qtel		2	4	6		1	1	2
12ptel	2		17	19			5	5
12qtel			27	27			11	11
13qtel			12	12			4	4
14qtel		12		12		1		1
15qtel	8	1	4	13	3		1	4
16ptel	8		29	37	3		14	17
16qtel	9		19	28	5		9	14
17ptel	2		17	19	1		0	1
17qtel	1		21	22	1		1	2
18ptel	4	3	10	17	1		5	6
18qtel			20	20			5	5
19ptel	14		16	30	5		9	14
19qtel	2		40	42			15	15
20ptel			22	22			6	6
20qtel	2		30	32			13	13
21qtel	1	1	17	19			6	6
22qtel	1	5	35	41			14	14
Xp/Yptel		3	6	9		1	3	4
Xqtel	5	1	1	7	1			1
Yqtel	5	2		7	1	1		2
Total	168	76	697	941	52	13	214	279

PctID >95%. Coverage >50.

quence fragments (other than at the ends of overlapping draft fragments that were clearly error prone), but rather used the consensus from such overlap regions as our subtelomeric assembly. An N was placed in consensus positions, in which overlaps produced an ambiguous base (i.e., a SNP or a sequence error). Specific accessions as well as NCBI Build 34 contigs used in assembling each subtelomere sequence are indicated in Supplemental Table 1.

Analysis of Subtelomeric Sequence Composition and Organization

The sequence composition and organization of each subtelomeric assembly was analyzed in the following manner:

1. RepeatMasker (Smit and Green (http://ftp.genome.washington.edu/RM/RepeatMasker.html) was used to detect interspersed and satellite repeat sequences, as well as the simple repeat (TTAGGG)n and overall GC content. The high-sensitivity setting (which requires a minimum match of 8 and a minimum score of 250) was used.
2. Each repeat-masked subtelomeric assembly was used to query the NR, htgs, EST, and GSS divisions of GenBank (August 1, 2003).
3. Tandem Repeats were identified using Tandem Repeats Finder (Benson 1999).
4. GC content was determined and graphed using a sliding window of 500 bp.
5. A database comprised of all of the subtelomeric assemblies was prepared and queried with each individual subtelomeric assembly using BLAST (Altschul et al. 1997) to identify subtelomeric repeat sequences (Srpt).
6. To detect genes and potential genes, each masked assembly was used to query the NCBI database of sequences representative of Unigene clusters (Schuler 1997; Aug 1, 2003 database). Matches were mapped back to the unmasked assembly using Spidey (Wheelan et al. 2001) to generate gene models based upon these sequences.

The output of each of these analyses was consolidated on a single interactive Genotator (Harris 1997) browser to permit convenient visual displays of the different sorts of analysis for each region. BLAST hits displayed on Genotator were analyzed at the

Table 3A. Named Transcripts in Srpt DNA

Tel	Cov	Id	Transcript	Accession
1p	91.8	100	similar to GLP_26_54603_52153 (LOC345824), mRNA	XM_299071
1p	83.04	98	similar to bA476I15.3 (novel protein similar to septin) (LOC339493)	XM_290927
1p	55.26	96	similar to capicua (*Drosophila*) homolog (LOC343085)	XM_291399
1p	100	96	similar to β-tubulin 4Q (LOC348841), mRNA	XM_300862
3q	100	100	similar to β-tubulin 4Q (LOC348841), mRNA	XM_300862
3q	100	100	similar to FSHD Region Gene 2 protein (LOC351585), mRNA	XM_302081
3q	55.26	96	similar to capicua (*Drosophila*) homolog (LOC343085), mRNA	XM_291399
4q	100	99	double homeobox, 4 (DUX4), mRNA	NM_033178
4q	100	97	tubulin, β polypeptide 4, member Q (TUBB4Q), mRNA	NM_020040
5q	83.04	98	similar to bA476I15.3 (novel protein similar to septin) (LOC339493)	XM_290927
5q	55.26	96	similar to capicua (*Drosophila*) homolog (LOC343085), mRNA	XM_291399
6p	83.04	98	similar to bA476I15.3 (novel protein similar to septin) (LOC339493)	XM_290927
6p	55.26	96	similar to capicua (*Drosophila*) homolog (LOC343085), mRNA	XM_291399
6q	100	98	similar to 60S ribosomal protein L23a (LOC284315), mRNA	XM_209109
7p	100	98	similar to capicua (*Drosophila*) homolog (LOC350883), mRNA	XM_301212
8p	83.04	98	similar to bA476I15.3 (novel protein similar to septin) (LOC339493)	XM_290927
9p	100	99	CXYorf1 pseudoautosomal region gene (CXYorf1), mRNA	XM_088704
9q	97.14	97	similar to tubulin, β, 2, clone IMAGE: 4873024, mRNA	BC014971
10q	100	100	similar to double homeobox, 4; double homeobox protein 4 (LOC347758)	XM_300253
11p	100	99	ligand-binding protein RYD5 (RYD5), mRNA	NM_145651
11p	83.04	98	similar to bA476I15.3 (novel protein similar to septin) (LOC339493)	XM_290927
11p	100	95	similar to capicua (*Drosophila*) homolog (LOC350883), mRNA	XM_301212
15q	100	100	CXYorf1 pseudoautosomal region gene (CXYorf1) mRNA	XM_088704
15q	100	99	similar to seven transmembrane helix receptor (LOC163201), mRNA	XM_092071
16p	99.17	100	Hs interleukin 9 receptor (IL9R), transcript variant 2, mRNA	NM_176786
16p	100	98	CXYorf1 pseudoautosomal region gene (CXYorf1), mRNA	XM_088704
16q	97.14	99	Similar to tubulin, β, 2, clone IMAGE:4873024, mRNA	BC014971
16q	55.26	97	similar to capicua (*Drosophila*) homolog (LOC343085), mRNA	XM_291399
16q	83.04	97	similar to bA476I15.3 (novel protein similar to septin) (LOC339493)	XM_290927
17p	100	100	ligand-binding protein RYD5 (RYD5), mRNA	NM_145651
17q	100	96	similar to GLP_26_54603_52153 (LOC345824), mRNA	XM_299071
18p	97.14	97	similar to tubulin, β, 2, clone IMAGE:4873024, mRNA	BC014971
19p	100	100	similar to seven transmembrane helix receptor (LOC163201), mRNA	XM_092071
19p	83.04	97	similar to bA476I15.3 (novel protein similar to septin) (LOC339493)	XM_290927
19p	55.26	96	similar to capicua (*Drosophila*) homolog (LOC343085), mRNA	XM_291399
Xq	100	98	CXYorf1 pseudoautosomal region gene (CXYorf1), mRNA	XM_088704

Table 3B. Named Cross-Boundary and Segmental Duplication Transcripts

Tel	Region	Cov	Id	Transcript	Accession
2q	x-bndry	83.04	99	similar to bA476I15.3 (novel protein similar to septin) (LOC339493)	XM_290927
4p	x-bndry	99.18	98	similar to zinc finger protein 208, clone MGC:41929	BC043151
5p	x-bndry	100	99	arylhydrocarbon receptor repressor (AHRR), mRNA	NM_020731
5p	x-bndry	98.61	98	programmed cell death 6 (PDCD6), mRNA	NM_013232
6p	x-bndry	97.43	99	dual specificity phosphatase 22 (DUSP22), mRNA	NM_020185
7p	x-bndry	83.04	97	similar to bA476I15.3 (novel protein similar to septin) (LOC339493)	XM_290927
8p	x-bndry	71.74	99	F-box only protein 25 (FBXO25), mRNA	NM_012173
10q	x-bndry	100	100	similar to olfactory receptor MOR103-5 (LOC347752), mRNA	XM_301226
11p	x-bndry	89.29	97	outer dense fiber of sperm tails 3 (ODF3), mRNA	NM_053280
16p	x-bndry	100	100	polymerase (RNA) III (DNA directed) polypeptide K, 12.3 kD (POLR3K)	NM_016310
16p	x-bndry	99.17	100	axin 1 (AXIN1), transcript variant 1, mRNA	NM_003502
16q	x-bndry	98.37	99	PR domain containing 7 (PRDM7), mRNA	NM_052996
17p	x-bndry	99.7	99	double C2-like domains, β (DOC2B), mRNA	NM_003585
17p	x-bndry	99.61	99	rabphilin 3A-like (without C2 domains) (RPH3AL), mRNA	NM_006987
21q	x-bndry	100	100	disco-interacting protein 2 (*Drosophila*) homolog (DIP2), mRNA	NM_015151
22q	x-bndry	98.99	99	acrosin (ACR), mRNA	NM_001097
Xp/Yp	x-bndry	100	100	short stature homeobox (SHOX), transcript variant SHOXa, mRNA	NM_000451
Xq	x-bndry	99.17	100	interleukin 9 receptor (IL9R), transcript variant 2, mRNA	NM_176786
Xq	x-bndry	100	100	similar to transient receptor potential cation channel, subfam C, mem 6	XM_066809
Xq	x-bndry	99.5	100	synaptobrevin-like 1 (SYBL1), mRNA	NM_005638
Xq	x-bndry	100	99	sprouty homolog 3 (*Drosophila*) (SPRY3), mRNA	NM_005840
4q	SD	100	98	FSHD region gene 1 (FRG1), mRNA	NM_004477
5p	SD	99.96	99	succinate dehydrogenase complex, subunit A, flavoprotein (Fp) (SDHA)	NM_004168
6q	SD	99.08	100	programmed cell death 2 (PDCD2), transcript variant 2, mRNA	NM_144781
6q	SD	100	100	proteasome (prosome, macropain) subunit, β type, 1 (PSMB1)	NM_002793
6q	SD	98.88	98	TATA box binding protein (TBP), mRNA	NM_003194
9p	SD	100	100	forkhead box D4 (FOXD4), mRNA	XM_095746
9p	SD	98.85	98	3′ similar to gb:M83088 PHOSPHOGLUCOMUTASE (HUMAN)	H12280
9p	SD	99.33	98	similar to COBW-like protein, clone IMAGE:5287337, mRNA	BC043420
14q	SD	97.75	100	similar to immunogloblin heavy-chain variable region, clone MGC:45495	BC032733
Xp/Yp	SD	100	100	protein phosphatase 2A 48 kD regulatory subunit (PR48), mRNA	NM_013239
Yq	SD	74.59	98	B melanoma antigen variant d (BAGE1) mRNA	AF527552

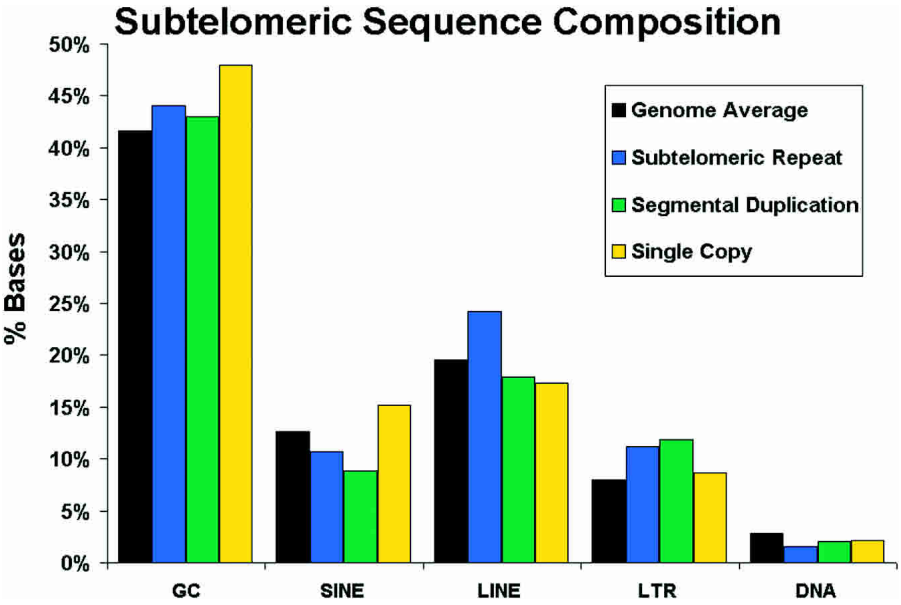

Figure 4 Sequence composition of subtelomeric assemblies. The GC percent and major interspersed repeat sequence composition of the subtelomeric assemblies are shown. The interspersed repeat classes were calculated independently for each strand, and the genome-wide averages were calculated from NCBI Build 34 of the human genome.

sequence level using Blixem (Sonnhammer and Durbin 1994). For regions in which transcript density was high, Spidey outputs were also downloaded onto the UCSC genome browser (Kent et al. 2002) to more easily compare multiple related transcripts across a given region.

ACKNOWLEDGMENTS

We thank the members of the International Human Genome Sequencing Consortium who participated in the sequencing of subtelomeric regions. Bob Moyzis, Jonathan Flint, and William Brown collaborated or provided reagents for the earlier stages of this work. John Rux and the Wistar Bioinformatics Facility provided programming and computational support. Financial support was provided by NIH HG00567 and CA 25874, and by the Commonwealth Universal Research Enhancement Program, PA Dept of Health.

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

REFERENCES

Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. 1997. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res.* **25:** 3389–3402.

Amann, J., Valentine, M., Kidd, V., and Lahti, J.M. 1996. Localization of Chl1-related helicase genes to human chromosome regions 12p11 and 12p13: Similarity between parts of these genes and conserved human telomere-associated DNA. *Genomics* **32:** 260–265.

Azzalin, C.M., Nergadze, S.G., and Giulotto, E. 2001. Human intrachromosomal telomeric-like repeats: Sequence organization and mechanisms of origin. *Chromosoma* **110:** 75–82.

Bailey, J.A., Yavor, A.M., Massa, H.F., Trask, B.J., and Eichler, E.E. 2001. Segmental duplications: Organization and impact within the current human genome project assembly. *Genome Res.* **11:** 1005–1017.

Bailey, J.A., Gu, Z., Clark, R.A., Reinert, K., Samonte, R.V., Schwartz, S., Adams, M.D., Myers, E.W., Li, P.W., and Eichler, E.E. 2002. Recent segmental duplications in the human genome. *Science* **297:** 1003–1007.

Baur, J.A., Zou, Y., Shay, J.W., and Wright, W.E. 2001. Telomere position effect in human cells. *Science* **292:** 2075–2077.

Benson, G. 1999. Tandem repeats finder: A program to analyze DNA

Table 4. Named Transcripts in Distal One-Copy Regions

Tel	Start	End	Strand	Cov	Id	Transcript	Unigene	Accession
1q	32654	19749	Minus	100	99	piggyBac transposable element derived 2 (PGBD2)	Hs.106773	NM_170725
12q	72288	43450	Minus	99.62	99	zinc finger protein 10 (KOX 1) (ZNF10), mRNA	Hs.104115	NM_015394
12q	80997	82631	Plus	100	100	similar to zinc finger protein 10 (KOX 1) (LOC341399), mRNA	Hs.257234	XM_292043
12q	122061	95510	Minus	99.5	98	zinc finger protein 140 (clone pHZ-39) (ZNF140), mRNA	Hs.154205	NM_003440
13q	62801	38596	Minus	99.08	100	UPF3 regulator of nonsense transcripts homolog A (yeast)	Hs.399740	NM_023011
13q	84949	71729	Minus	99.91	99	highly similar to Human CDC16Hs mRNA	Hs.430741	AK095082
13q	109505	71729	Minus	100	100	Homosapiens CDC16 cell division cycle 16 homolog	Hs.1592	NM_003903
16p	62591	48058	Minus	96.92	99	Homosapiens rhomboid family 1 (*Drosophila*) (RHBDF1), mRNA	Hs.57988	NM_002450
16p	69147	75841	Plus	100	99	Homosapiens N-methylpurine-DNA glycosylase (MPG), mRNA	Hs.79396	NM_002434
16p	128859	74273	Minus	100	99	Conserved gene telomeric to α globin cluster (CGTHBA)	Hs.19699	NM_012075
18q	13181	100758	Plus	100	99	par-6 partitioning defective 6 homolog γ (PARD6G)	Hs.164219	NM_032510
19q	38860	41434	Plus	96.08	99	zinc finger protein 42 (myeloid-specific retinoic acid-responsive)	Hs.169832	NM_003422
19q	45063	47761	Plus	99.46	96	similar to ubiquitin-conjugating enzyme E2M; UBC12 homolog	Hs.447116	XM_208901
19q	48381	51907	Plus	97.56	100	putative breast adenocarcinoma marker (32 kD) (BC-2)	Hs.12107	NM_014453
19q	59004	52758	Minus	99	100	tripartite motif-containing 28 TRIM28), mRNA	Hs.228059	NM_005762
19q	91495	109075	Minus	97.95	100	solute carrier family 27 (fatty acid transporter), member 5	Hs.111401	NM_012254
20p	8363	17174	Plus	100	100	Homosapiens defensin, β 125 (DEFB125)	Hs.380220	NM_153325
20p	63252	66392	Plus	100	100	Homosapiens defensin, β 126 (DEFB126)	Hs.124211	NM_030931
20p	78186	79804	Plus	100	98	Homosapiens defensin, β 127 (DEFB127)	Hs.99362	NM_139074
20q	82319	85971	Plus	55.26	96	similar to capicua (*Drosophila*) homolog (LOC343085), mRNA	Hs.453435	XM_291399
20q	169694	91917	Minus	100	99	myelin transcription factor 1 (MYT1), mRNA	Hs.279562	NM_004535
21q	64329	35038	Minus	99.67	99	HMT1 hnRNP methyltransferase-like 1 (*S. cerevisiae*) (HRMT1L1)	Hs.235887	NM_001535
21q	94900	101365	Plus	100	99	S100 calcium-binding protein, β (neural) (S100B), mRNA	Hs.83384	NM_006272
22q	85504	72929	Minus	100	99	SH3 and multiple ankyrin repeat domains 3 (SHANK3)	Hs.282076	XM_037493
Xp/Yp	112760	98476	Minus	99.41	100	Pseudoautosomal GTP-binding protein-like (PGPL); mRNA	Hs.372587	NM_012227

sequences. *Nucleic Acids Res*. **27:** 573–580.
Blasco, M.A., Gasser, S.M., and Lingner, J. 1999. Telomeres and telomerase. *Genes & Dev*. **13:** 2353–2359.
Bryan, T.M., Englezou, A., Gupta, J., Bacchetti, S., and Reddel, R.R. 1995. Telomere elongation in immortal human cells without detectable telomerase activity. *EMBO J*. **14:** 4240–4248.
Carlson, M., Celenza, J.L., and Eng, F.J. 1985. Evolution of the dispersed SUC gene family of *Saccharomyces* by rearrangements of chromosomal telomeres. *Mol. Cell Biol*. **5:** 2894–2902.
Cook, G.P., Tomlinson, I.M., Walter, G., Carter, N.G., Riethman, H.C., Winter, G., and Rabbitts, T.H. 1994. A map of the human immunoglobulin VH locus completed by analysis of the telomeric region of chromosome14q. *Nat. Genet*. **7:** 162–168.
Donaldson, K.M. and Karpen, G.H. 1997. Trans-suppression of terminal deficiency-associated position effect variegation in a *Drosophila* minichromosome. *Genetics* **145:** 325–337.
Fan, Y., Newman, T., Linardopoulou, E., and Trask, B.J. 2002. Gene content and function of the ancestral chromosome fusion site in human chromosome 2q13–2q14.1 and paralogous regions. *Genome Res*. **12:** 1663–1672.
Feuerbach, F., Galy, V., Trelles-Sticken, E., Fromont-Racine, M., Jacquier, A., Gilson, E., Olivo-Marin, J.C., Scherthan, H., and Nehrbass, U. 2002. Nuclear architecture and spatial positioning help establish transcriptional states of telomeres in yeast. *Nat. Cell Biol*. **4:** 214–221.
Flint, J., Thomas, K., Micklem, G., Raynham, H., Clark, K., Doggett, N.A., King, A., and Higgs, D.R. 1997a. The relationship between chromosome structure and function at a human telomeric region. *Nat. Genet*. **15:** 252–257.
Flint, J., Bates, G.P., Clark, K., Dorman, A., Willingham, D., Roe, B.A., Micklem, G., Higgs, D.R., and Louis, E.J. 1997b. Sequence comparison of human and yeast telomeres identifies structurally distinct subtelomeric domains. *Hum. Mol. Genet*. **6:** 1305–1313.
Fourel, G., Revardel, E., Koering, C.E., and Gilson, E. 1999. Cohabitation of insulators and silencing elements in yeast subtelomeric regions. *EMBO J*. **18:** 2522–2537.
Harris, N.L. 1997. Genotator: A workbench for sequence annotation. *Genome Res*. **7:** 754–762.
Henson, J.D., Neumann, A.A., Yeager, T.R., and Reddel, R.R. 2002. Alternative lengthening of telomeres in mammalian cells. *Oncogene* **21:** 598–610.
Ijdo, J.W., Lindsay, E.A., Wells, R.A., and Baldini, A. 1992. Multiple variants in subtelomeric regions of normal karyotypes. *Genomics* **14:** 1019–1025.
International Human Genome Sequencing Consortium (IHGSC). 2001. Initial sequencing and analysis of the human genome. *Nature* **409:** 860–921.
Kent, W.J., Sugnet, C.W., Furey, T.S., Roskin, K.M., Pringle, T.H., Zahler, A.M., and Haussler, D. 2002. The Human Genome Browser at UCSC. *Genome Res*. **12:** 996–1006.
Kermouni, A., Van Roost, E., Arden, K.C., Vermeesch, J.R., Weiss, S., Godelaine, D., Flint, J., Lurquin, C., Szikorza, J.P., Higgs, D.R., et al. 1995. The IL-9 receptor gene (IL9R): Genomic structure and chromosomal localization in the pseudoautosomal region of the long arm of the sex chromosomes, and identification of IL9R pseudogenes at 9qter, 10pter, 16pter, and 18pter. *Genomics* **29:** 371–382.
Kvaloy, K. 1993. "The long arm telomeres of the human sex chromosomes." Ph.D thesis, Wadham College, Department of Biochemistry, University of Oxford, UK.
Lundblad, V. and Wright, W.E. 1996. Telomeres and telomerase: A simple picture becomes complex. *Cell* **87:** 369–375.
Ma, B., Tromp, J., and Li, M. 2002. PatternHunter: Faster and more sensitive homology search. *Bioinformatics* **18:** 440–445.
Macina, R.A., Negorev, D.G., Spais, C., Ruthig, L.A., Hu, X-L., and Riethman, H.C. 1994. Sequence organization of the human chromosome 2q telomere. *Hum. Mol. Genet*. **3:** 1847–1853.
Macina, R.A., Morii, K., Hu, X.-L., Negorev, D.G., Spais, C., Ruthig, L.A., and Riethman, H.C. 1995. Molecular cloning and RARE cleavage mapping of human 2p, 6q, 8q, 12q, and 18q telomeres. *Genome Res*. **5:** 225–232.
Mah, N., Stoehr, H., Schulz, H.L., White, K., and Weber, B.H. 2001. Identification of a novel retina-specific gene located in a subtelomeric region with polymorphic distribution among multiple human chromosomes. *Biochim. Biophys. Acta*. **1522:** 167–174.
Martin, C.L., Wong, A., Gross, A., Chung, J., Fantes, J.A., and Ledbetter, D.H. 2002. The evolutionary origin of human subtelomeric homologies—or where the ends begin. *Am. J. Hum. Genet*. **70:** 972–984.
Martin-Gallardo, A., Lamerdin, J., Sopapan, P., Friedman, C., Fertitta, A.L., Garcia, E., Carrano, A., Negorev, D., Macina, R.A., Trask, B.J., et al. 1995. Molecular analysis of a novel subtelomeric repeat with polymorphic chromosomal distribution. *Cytogenet. Cell Genet*. **71:** 289–295.
McCulloch, R., Rudenko, G., and Borst, P. 1997. Gene conversions mediating antigenic variation in *Trypanosoma brucei* can occur on variant surface glycoprotein expression sites lacking 70-bp repeat sequences. *Mol. Cell Biol*. **17:** 833–843.
Mefford, H.C. and Trask, B.J. 2002. The complex structure and dynamic evolution of human subtelomeres. *Nat. Rev. Genet*. **3:** 91–102.
Mondello, C., Pirzio, L., Azzalin, C.M., and Giulotto, E. 2000. Instability of interstitial telomeric sequences in the human genome. *Genomics* **68:** 111–117.
Monfouilloux, S., Avet-Loiseau, H., Amarger, V., Balazs, I., Pourcel, C., and Vergnaud, G. 1998. Recent human-specific spreading of a subtelomeric domain. *Genomics* **51:** 165–176.
Morin, G.B. 1989. The human telomere terminal transferase enzyme is a ribonucleoprotein that synthesizes TTAGGG repeats. *Cell* **59:** 521–529.
Moyzis, R.K., Buckingham, J.M., Cram, S., Dani, M., Deaven, L.L., Jones, M.D., Meyne, J., Ratliff, R.L., and Wu, J.R. 1988. A highly conserved repetitive DNA sequence, (TTAGGG)n, present at the telomeres of human chromosomes. *Proc. Natl. Acad. Sci*. **85:** 6622–6626.
Murnane, J.P., Sabatier, L., Marder, B.A., and Morgan, W.F. 1994. Telomere dynamics in an immortal human cell line. *EMBO J*. **13:** 4953–4962.
Pryde, F.E. and Louis, E.J. 1999. Limitations of silencing at native yeast telomeres. *EMBO J*. **18:** 2538–2550.
Reston, J.T., Hu, X.-L., Macina, R.A., Spais, C., and Riethman, H. 1995. Structure of the terminal 300 kb of DNA from human chromosome 21q. *Genomics* **26:** 31–38.
Riethman, H.C., Moyzis, R.K., Meyne, J., Burke, D.T., and Olson, M.V. 1989. Cloning human telomeric DNA fragments into *Saccharomyces cerevisiae* using a yeast-artificial-chromosome vector. *Proc. Natl. Acad. Sci*. **86:** 6240–6244.
Riethman, H., Birren, B., and Gnirke, A. 1997. Preparation, manipulation, and mapping of high molecular weight DNA. In *Genome analysis: A laboratory manual, Volume 1: "Analyzing DNA"* (eds. B. Birren et al.), pp. 83–248. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.
Riethman, H.C., Xiang, Z., Paul, S., Morse, E., Hu, X.L., Flint, J., Chi, H.C., Grady, D.L., and Moyzis, R.K. 2001. Integration of telomere sequences with the draft human genome sequence. *Nature* **409:** 948–951.
Rizki, A. and Lundblad, V. 2001. Defects in mismatch repair promote telomerase-independent proliferation. *Nature* **411:** 713–716.
Ruiz-Herrera, A., Garcia, F., Azzalin, C., Giulotto, E., Egozcue, J., Ponsa, M., and Garcia, M. 2002. Distribution of intrachromosomal telomeric sequences (ITS) on *Macaca fascicularis* (Primates) chromosomes and their implication for chromosome evolution. *Hum. Genet*. **110:** 578–586.
Saccone, S., De Sario, A., Weigant, J., Raap, A.K., Della Valle, G., and Bernardi, G. 1993. Correlations between isochores and chromosomal bands in the human genome. *Proc. Natl. Acad. Sci*. **90:** 11929–11933.
Schuler, 1997. Pieces of the puzzle: Expressed sequence tags and the catalog of human genes. *J. Mol. Med*. **75:** 694–698.
Smit, A.F.A. and Green, P. RepeatMasker home page. http://ftp.genome.washington.edu/RM/RepeatMasker.html
Sonnhammer, E.L.L. and Durbin, R. 1994. A workbench for Large Scale Sequence Homology Analysis. *Comput. Applic. Biosci*. **110:** 301–307.
Trask, B.J., Friedman, C., Martin-Gallardo, A., Rowen, L., Akinbami, C., Blankenship, J., Collins, C., Giorgi, D., Iadonato, S., Johnson, F., et al. 1998. Members of the olfactory receptor gene family are contained in large blocks of DNA duplicated polymorphically near the ends of human chromosomes. *Hum. Mol. Genet*. **7:** 13–26.
van Geel, M., Eichler, E.E., Beck, A.F., Shan, Z., Haaf, T., van der Maarel, S.M., Frants, R.R., and de Jong, P.J. 2002. A cascade of complex subtelomeric duplications during the evolution of the hominoid and Old World monkey genomes. *Am. J. Hum. Genet*. **70:** 269–278.
van Overveld, P.G., Lemmers, R.J., Deidda, G., Sandkuijl, L., Padberg, G.W., Frants, R.R., and van Der Maarel, S.M. 2000. Interchromosomal repeat array interactions between chromosomes 4 and 10: A model for subtelomeric plasticity. *Hum. Mol. Genet*. **9:** 2879–2884.
Wheelan, S.J., Church, D.M., and Ostell, J.M. 2001. Spidey: A tool for mRNA-to-genomic alignments. *Genome Res*. **11:** 1952–1957.
Wilkie, A.O.M., Higgs, D.R., Rack, K.A., Buckle, V.J., Spurr, N.K., Fischel-Ghodsian, N., Ceccherini, I., Brown, W.R.A., and Harris, P.C. 1991. Stable length polymorphism of up to 260 kb at the tip of the short arm of human chromosome 16. *Cell* **64:** 595–606.

WEB SITE REFERENCES

ftp://ftp.ncbi.nih.gov/repository/UniGene/; database of best mRNA or EST sequence representative of each Unigene cluster.

Received February 6, 2003; accepted in revised form November 4, 2003.