Human subtelomeric copy number variations

H. Riethman

The Wistar Institute, Philadelphia, PA (USA)

Accepted in revised form for publication by H. Kehrer-Sawatzki and D.N. Cooper, 23 June 2008.

Abstract. Copy number variation is a defining characteristic of human subtelomeres. Human subtelomeric segmental duplication regions ('Subtelomeric Repeats') comprise about 25% of the most distal 500 kb and 80% of the most distal 100 kb in human DNA. Huge allelic disparities seen in subtelomeric DNA sequence content and organization are postulated to have an impact on the dosage of transcripts embedded within the duplicated sequences, on the transcription of genes in adjacent single copy DNA regions, and on the chromatin structures mediating telomere functions including chromosome stability. In addition to the complex duplicon substructure and huge allelic variations in extended subtelomere regions, both copy number variation and alternative sequence organizations for DNA char-

acterize the sequences immediately adjacent to terminal (TTAGGG)_n tracts ('subterminal DNA'). The structural variation in subterminal DNA is likely to have important consequences for expression of subterminal transcripts such as a newly-discovered gene family encoding actin-interacting proteins and a non-coding telomeric repeat containing RNA (TERRA) transcript family critical for telomere integrity. Major immediate challenges include discovering the full extent and nature of subtelomeric structural and copy number variation in humans, and developing methods for tracking individual allelic variants in the context of total genomic DNA.

Copyright © 2009 S. Karger AG, Basel

Terminal (TTAGGG)_n tracts and immediate subterminal sequences

The DNA at each human chromosome terminus is a simple repeat sequence tract (TTAGGG)_n, typically 5 kb to 15 kb in length in somatic cells (Moyzis et al., 1988). The average lengths of terminal repeat tracts in humans are dynamically modulated by a balance of exonuclease- and replication-associated attrition and telomerase- and/or recombination-mediated lengthening (Harley et al., 1990; Reddel, 2003; Blackburn, 2005). The average bulk telomere length is believed to reflect a partly inherited telomere length trait

(Slagboom et al., 1994), as well as both the replicative history of the cell lineage and environmental factors such as exposure to stress and oxidative damage (Wright and Shay, 2002; Aviv et al., 2003; Epel et al., 2004). A single-stranded, G-rich overhang produced as a result of the RNA priming requirement for DNA replication or by post-replicative processing steps (Sfeir et al., 2005) is involved in the formation of a t-loop secondary structure (Griffith et al., 1999) believed to be required for proper telomere function. Disruption of t-loop or associated telomeric chromatin structures by interference with Shelterin protein components required for their formation and maintenance (Karlseder et al., 2002), by loss of the G-overhang itself, or by a critically short tract of (TTAGGG)_n can each lead to telomere dysfunction (de Lange, 2005). Depending on the cellular context (especially the presence or absence of a functional p53 pathway), such telomere loss or uncapping can trigger cellular senescence, apoptosis, and/or genome instability.

The replicative capacity of cells is limited by telomere length, implicating short telomere length in age-related diseases (Harley et al., 1990; Lee et al., 1998). Supporting this

This work was supported by NIH grants HG00567 and HG2933, and by the Commonwealth Universal Research Enhancement Program, PA Dept. of Health.

Request reprints from Harold Riethman
The Wistar Institute, 3601 Spruce St.
Philadelphia, PA 19104 (USA)
telephone: +1 215 898 3872; fax: +1 215 898 3868
e-mail: Riethman@wistar.org

notion, heterozygous germline mutations in the telomerase protein component (hTert; encoded by *TERT*) or RNA component (hTR; encoded by *TERC*) are associated with abnormally short telomeres in humans and can lead to autosomal dominant forms of the diseases dyskeratosis congenita (characterized by aplastic anemia and bone marrow failure, as well as predisposition to cancers of the skin, hematopoietic system, and oral mucosa) and idiopathic pulmonary fibrosis (leading to respiratory failure) (Vulliamy et al., 2001; Armanios et al., 2007). Because anticipation occurs in autosomal dominant dyskeratosis congenita (more severe and earlier occurring phenotypes are seen in successive generations), the short telomeres must cause the disease rather than the telomerase mutation itself (Armanios et al., 2005). Thus, short telomeres limit tissue renewal capacity in somatic tissues, even in the presence of telomerase (Hiyama and Hiyama, 2007), and can cause both congenital and age-related disease in humans.

The lengths of (TTAGGG)*n* tracts have been shown to vary from telomere to telomere within individual cells (Lansdorp et al., 1996; Zijlmans et al., 1997; Graakjaer et al., 2004) and between alleles at the same telomere (Baird et al., 2003; der-Sarkissian et al., 2004). These individual-specific patterns of relative telomere-specific (TTAGGG)*n* tract lengths are regulated in part by *cis*-acting factors (Graakjaer et al., 2003, 2006; Britt-Compton et al., 2006), and these patterns appear to be defined in the zygote and maintained throughout life (Graakjaer et al., 2004). Since the immediate effects of (TTAGGG)*n* tract loss on cell viability and chromosome stability are attributable to the shortest telomere(s) in a cell, rather than to average telomere length (Hemann et al., 2001; der-Sarkissian et al., 2004; Meier et al., 2007), individual-specific patterns of allele-specific (TTAGGG)*n* tract lengths may be crucial for the biological functions of telomeres and the effects of telomere attrition and dysfunction associated with aging, cancer, stress and coronary artery disease (Wright and Shay, 2002; Aviv et al., 2003; Epel et al., 2004).

Cis-acting mediators of allele-specific (TTAGGG)*n* tract length are likely to depend upon subtelomeric sequences immediately adjacent to terminal (TTAGGG)*n* tracts. These sequences have been determined for the reference sequences of some chromosome arms (Riethman et al., 2004), and a model of human terminal and subterminal DNA based upon these completed reference sequences is shown in Fig. 1. All subterminal human DNA fragments cloned in bacteria (fosmids, cosmids) and yeast (half-YACs) have lost the terminal portion of the (TTAGGG)*n* tract and retain 300 to 800 bp of the centromere-adjacent region of the simple repeat tract; the cloning-associated deletions appear to be specific for the distal (TTAGGG)*n* regions and do not involve rearrangement of subterminal sequences (Riethman, unpublished). Most of the known human (TTAGGG)*n*-adjacent sequences are related in some fashion to the telomere-associated repeat (TAR) sequences originally described by Brown and co-workers (1990). Parts of the canonical TAR1 sequence are present within 2 kb of the beginning of nearly all sequenced (TTAGGG)*n*-adjacent DNA, although the

Fig. 1. Model of terminal and subterminal human DNA. The single-stranded, G-rich 3' overhang is shown in red; black represents double-stranded (TTAGGG)*n* repeat DNA (5'-(TTAGGG)*n* -3' towards the telomere end, 5-(CCCTAA)*n* -3' strand toward the centromere). The lighter gray blocks interspersed with black represent the most centromeric part of the terminal repeat tract, where blocks of variant repeats related to (TTAGGG)*n* are usually present and interspersed with canonical (TTAGGG)*n* repeats in allele-specific patterns that are in linkage disequilibrium with the 2–4 kb of adjacent subterminal DNA. The pink region represents variably-sized (0–2 kb) sequence segments bearing similarity to the TAR1 repeat family (Brown et al., 1990), and the blue region represents subterminal DNA which can be either one-copy (six telomeres) or fall into one of the six families of subterminal repeat homology we have recently characterized (Ambrosini et al., 2007). The minimum length of the human terminal repeat tract has been estimated at 2 kb using bulk telomere measurements in senescing cells (Levy et al., 1992), but the higher-resolution single telomere length analysis (STELA) found a wide variability of telomere lengths at individual chromosome ends in senescing fibroblasts, with a minimum modal terminal tract length of 300 bp found on an allele of 17p in senescent cell populations (Britt-Compton et al., 2006). Human germline terminal tract lengths display a wide individual-specific variation range (from <9 kb to greater than 17 kb), with a substantial frequency of outliers with very short telomeres found in the sperm DNA samples measured (Baird et al., 2006). Non-canonical terminal repeats found in the centromeric part of the tract include (TGAGGG)*n*, (TCAGGG)*n* and (TTGGGG)*n* (Baird et al., 1995, 2000; Coleman et al., 1999) as well as many others (e.g., (TTCGGG), (TTAGGGG)*n*, (TCGGGG)*n*, and others; Xu and Blackburn, 2007; Riethman, unpublished), and can extend telomerically for at least 1 kb into the tract (Baird et al., 1995). The green arrows represent the position and orientation of subterminal transcripts (see Table 1) including telomeric repeat-containing RNA (TERRA) molecules, for which both the 5' and 3' end positions remain undefined (Azzalin et al., 2007).

part and degree of similarity can vary substantially. TAR1 similarity can also be found adjacent to many of the internal (TTAGGG)*n*-like sequences in Srpt regions, and more distantly related copies of TAR1-like sequence are often found in pericentromere regions (Ambrosini et al., 2007). Interestingly, one known case of a naturally-occurring segment of (TTAGGG)*n*-adjacent DNA lacking TAR1 similarity is an allele of 17p shown by single telomere ligation-associated PCR (STELA) to be shorter than most other telomeres (Britt-Compton et al., 2006). CNVs in these duplicated subterminal sequences are known to exist (Riethman, 2008) and are likely to affect the allele-specific *cis*-regulation of (TTAGGG)*n* tracts.

Table 1. Subterminal transcripts

Name	Accession	kb to tel	Comments	Orientation
TERRA	DB453573	1	contains (UUAGGG) _n repeats	3' end distal
	DB073084	1	spliced EST, testis	5' end distal
	CB989481	1	testis	5' end distal
	AK125998	1	spliced EST, placenta	5' end distal
	DB079246	1	unspliced transcript, testis cDNA	5' end distal
	CN412081	1		5' end distal
	DB248987	2	spliced EST, testis	5' end distal
	DA948568	2	spliced EST, spleen	5' end distal
	BM920887	2	pooled brain, lung, testis	5' end distal
	BM920886	2	pooled brain, lung, testis	5' end distal
	BI909979	2	spliced EST, leuk	5' end distal
DEAD/H	BC032353	2	spliced EST	5' end distal
	AI807076	2	subtr lib	5' end distal
	DB519034	3	spliced EST, testis	5' end distal
	BX089548	3	testis	5' end distal
	BC070227	3	spliced EST	5' end distal
	AA620647	3	testis	5' end distal
	N45203	4	MS libr	3' end distal
	BG216713	4	RAGE library	3' end distal
	BC048328	4–6	<i>cXYorf</i> , med form, alt splice	3' end distal
	NM_182905	4–6	<i>FAM39E</i> (short isoform)	3' end distal
WASH gene	CR612421	4–6	<i>FAM39B</i> (long isoform)	3' end distal
WASH gene		4–6	many additional subterminal WASH transcripts ending 4–6 kb from tel	3' end distal
Additional WASH transcripts		4–6		
<i>RPL23AP7</i> -like	NM_203302	7	similar to <i>RPL23AP7</i> , variant 1	3' end distal
<i>RPL23AP7</i> -like	BC065480	8		3' end distal
<i>RPL23AP7</i> -like	BC065556	9		3' end distal
Additional <i>RPL23AP7</i> -like transcripts		9	many additional subterminal <i>RPL23a</i> -like transcripts ending 7–9 kb from tel	3' end distal

Subterminal transcripts

Telomeric DNA is transcribed. Indeed, a recently-discovered noncoding subterminal transcript that includes sequences from the terminal repeat itself may prove to be central to understanding telomere biology. The telomeric repeat-containing RNA (TERRA) molecules found in humans range in size from about 100 bases to at least 9 kb, are transcribed almost exclusively from the C-strand of subterminal DNA from multiple (perhaps all ?) telomeres, and associate with telomeric chromatin (Azzalin et al., 2007). The association of TERRA molecules with telomeric chromatin is regulated by protein effectors of the nonsense-mediated RNA decay pathway, which include the human homologs of the yeast Est1p protein; depletion of effectors increased association of TERRA molecules with telomeric chromatin and caused loss of telomere DNA. A similar telomeric RNA molecule found in mouse was shown to be transcribed by RNA polymerase II, polyadenylated, and under developmental regulation (Schoeftner and Blasco, 2008). Evidence has been presented suggesting that high levels of the mouse telomeric repeat-containing RNA could inhibit telomerase activity (Schoeftner and Blasco, 2008). Noncoding RNAs have been shown to play roles in heterochromatin formation (Sugiyama et al., 2007) and X-inactivation (Wutz and

Gribnau, 2007). Similarly, TERRA molecules may be involved in the modulation of telomeric chromatin structure and telomeric heterochromatin formation.

Several additional transcript families are embedded within or extend into subterminal DNA (Fig. 1, Table 1). Although many members of these transcript families appear to be non-coding pseudogenes, some contain open reading frames and encode protein. One particularly interesting subterminal transcript family (originally named *cXYorf1*-related; Gianfrancesco et al., 2001) is associated with a particular subterminal duplicon block (Ambrosini et al., 2007); the 3' ends of this transcript family end within 4–6 kb of the start of the (TTAGGG)_n tract (Fig. 1, Table 1). Very recently, this *cXYorf1*-related transcript family has been shown to correspond to the 3' ends of genes from a new subclass of the Wiscott-Aldrich Syndrome Protein (WASP) family, termed *WASH* genes (Linardopoulou et al., 2007).

The *WASH* protein family is old and conserved. It is single-copy in most lower organisms and in early primates, and was shown to be essential for survival in *Drosophila* (Linardopoulou et al., 2007). However, it is multicopy in gorilla, chimps, and humans, with the highest copy numbers in human. *WASH* proteins appear to be involved in cytoskeletal organization and signal transduction, and can function in the regulation of actin filament polymerization. The human

WASH gene family structures predict short forms and long forms of the protein; the 3' exons encoding the short form are embedded in subterminal duplicon block C (Ambrosini et al., 2007), whereas the full-length forms start near a CpG island roughly 20 kb from telomeres within a polymorphically distributed segmental duplication (Martin-Gallardo et al., 1995). The *WASH* family varies widely in both dosage and telomere distribution in individual genomes, and usually terminates less than 5 kb from the start of the terminal (TTAGGG)_n tract (Table 1). Individual telomeric transcription sites for this family might be differentially susceptible to position effects depending upon both local telomeric chromatin/heterochromatin status and on chromosome-specific telomere lengths. Similarly, CNV either disrupting the gene structure or affecting *WASH* dosage could readily affect the number of alleles encoding functional protein within a genome. The subtelomeric/subterminal genomic location, as a hotspot of DNA break and repair (Linardopoulou et al., 2005; Rudd et al., 2007), has likely contributed to the production of many copies and combinations of the short and long forms of *WASH* gene family members in very recent primate and human evolution. It is intriguing to speculate on potential roles this gene family may have played in the evolution of primate and human-specific phenotypes, as well as potential roles in the somatic evolution of cancer genomes, where telomere loss or dysfunction is believed to play an initiating role (Meeker et al., 2004; Baird et al., 2006; Capper et al., 2007).

A second major subterminal transcript family consists of members that bear highest similarity to a ribosomal protein pseudogene, *RPL23AP7*. As with the *WASH* transcript family, transcripts with both interrupted and open reading frames co-exist within the genome, and the 3' ends of most transcripts extend into subterminal duplicon blocks (in this case, related blocks A, B, and D; Ambrosini et al., 2007). No information is yet available as to whether proteins corresponding to the open reading frames from this transcript family are made. Curiously, transcription of both the *RPL23AP7* and *WASH* transcript families is in the same orientation and their 3' end positions coincide with where TERRA RNAs are expected to initiate, begging the question of whether these abundant transcript families are regulated in a coordinate fashion.

A collection of apparently less abundant transcripts are transcribed in opposite orientation to the TERRA, *WASH*, and *RPL23AP7*-like families; these are also listed in Table 1. There seems to be an overrepresentation of testis transcripts in this group of RNAs, the 5' ends of which originate within 1–4 kb of the start of the (TTAGGG)_n tract. Little else is known regarding these RNAs.

Telomere epigenetics

Mouse studies have shown that mammalian telomere length and stability can be regulated by epigenetic factors (Blasco, 2007); heterochromatic epigenetic marks are present near mouse telomeres, a loss of some of those marks oc-

curs upon reduction of telomere (TTAGGG)_n tract lengths to a critically short length in the absence of telomerase, and de-regulation of telomere tract length and stability is evident in mouse knockouts affecting enzymes responsible for normally maintaining the heterochromatic marks. Loss of heterochromatic marks at mouse telomeres was found to be associated with recombination-associated instability.

Human subterminal DNA is hypomethylated in the germline, methylated de novo in somatic tissues, and may contain non-canonical DNA modifications (de Lange et al., 1990; Steinert et al., 2004). Hypomethylation of subterminal DNA in the human germline (and suppression of meiotic recombination in the most distal 2–4 kb of telomere-adjacent DNA) are characteristics of satellite heterochromatin, but the active transcription of genes within a few kb of some human (TTAGGG)_n tracts and the small telomere-adjacent LD region suggest a very small telomeric heterochromatic region and perhaps a precisely regulated heterochromatic-euchromatic boundary region in human subterminal DNA. Human subterminal DNA sequences typically contain clusters of very CpG-rich sequences which are arranged in variably sized and organized stretches among separate telomeres (Riethman, 2008); these sequences may be among those differentially methylated in germline versus differentiated cells, and their methylation patterns may reflect the telomeric heterochromatin-euchromatin transition for specific telomere alleles. Epigenetic modification of subterminal sequences could play an important role in regulating telomere replication and in recombination-associated phenomena such as elevated levels of telomere-associated sister-chromatid exchange (Rudd et al., 2007). Indeed, one could speculate that allele-specific (TTAGGG)_n tract length regulation might be mediated by differential positioning of the euchromatin/heterochromatin boundary within subterminal sequences.

Extended human subtelomeric DNA regions

Human subtelomeric segmental duplications ('subtelomeric repeats') comprise about 25% of the most distal 500 kb and 80% of the most distal 100 kb in human DNA (International Human Genome Sequencing Consortium, 2004; Riethman et al., 2004). The overall size, sequence content, and organization of subtelomeric segmental duplications relative to the terminal (TTAGGG)_n repeat tracts and to subtelomeric single-copy DNA is different for each subtelomere. Subtelomeric regions of human chromosomes have long been known to contain mosaic patchworks of duplicons (e.g., see der-Sarkissian et al., 2002; Mefford and Trask, 2002; Riethman et al., 2004). Recent genome-wide analyses of these regions have revealed that this sequence organization appears to arise from translocations involving the tips of chromosomes, followed by transmission of unbalanced chromosomal complements to offspring (Linardopoulou et al., 2005). A recent analysis of SCEs revealed highly elevated rates near telomeres (1,600× greater in the most distal 10 kb, including subterminal sequences and the telomere re-

Fig. 2. Duplicon analysis for example subtelomeres. The most distal regions of the indicated reference sequences are shown oriented with the telomeric end on the left. Duplicated genomic segments are identified by chromosome (color) and whether they are subtelomeric (bounded rectangles), non-telomeric (unbounded rectangles), or intra-chromosomal (located above the subtelomere coordinates). The arrows on the coordinate line represent the positions of (TTAGGG) n and (TTAGGG) n -like internal sequences in each subtelomere. Each rect-

angle represents a separate duplicon, and each of the pairwise alignments generated by the duplicon definitions was tabulated and analyzed. The line segments ending in arrows above each subtelomere indicate the positions of named transcript families found in these subtelomere regions. Our web site (<http://www.wistar.upenn.edu/Riethman>) contains analogous figures and underlying sequence alignments for the duplicon organization of each subtelomere region.

peat itself), and 160 $\times$ greater in the adjacent 100-kb subtelomere regions (Rudd et al., 2007). Together these studies indicate that subtelomeres are hotspots of DNA breakage and repair; these features are likely to be responsible for the generation of the complex interchromosomal duplication patterns and the rapid evolution of these genomic regions, as well as the prevalence of large CNVs near telomeres.

A detailed analysis of subtelomeric duplicons in the human reference sequence was completed recently (Ambrosini et al., 2007). Duplicon blocks were defined by sequence similarity between segments of subtelomeric DNA from single telomeres and the assembled human genome. Each duplicon module was defined by a set of pairwise alignments with the query subtelomere sequence, and a % nucle-

otide sequence identity for the non-masked parts of each chained pairwise alignment was derived from the BLAST alignments. This analysis is summarized for a sampling of telomeres in Fig. 2; the full set of modules for all telomeres, including the coordinates of their genomic alignments, is presented on our web site (<http://www.wistar.org/Riethman>).

Several key features of duplicon organization were revealed by this analysis (Ambrosini et al., 2007). First, the internal (TTAGGG) n sequences are usually oriented towards the telomere and almost always co-localize to duplicon boundaries. In yeast, interstitial telomere-like sequence islands in subtelomeric DNA are capable of playing important roles in subtelomeric recombination and transcription,

telomere maintenance via the recombination-based ALT (Alternative Lengthening of Telomeres) pathway, and telomere healing (Stavenhagen and Zakian, 1994; Lundblad, 2002). Several studies have suggested similar roles for internal human (TTAGGG)_n-like sequence islands (e.g., Mondello et al., 2000; Azzalin et al., 2001; Ruiz-Herrera et al., 2002), including a hypothesized role as a boundary element for subtelomeric DNA compartments (Flint et al., 1997). (TTAGGG)_n-like islands are enriched about 25-fold in human subtelomeric regions (Riethman et al., 2004). In addition, they tend to be both longer and more similar to perfect (TTAGGG)_n tracts in subtelomeric DNA compared to elsewhere in the genome. From an evolutionary perspective, this might suggest 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 and a minimum tract length (Flint et al., 1997). Given the known functions of internal telomere sequences in yeast and the demonstrated ability of TRF1, TRF2, and hRAP1 (encoded by *TERF1*, *TERF2*, and *TERF2IP*, respectively) to interact with very short TTAGGGTT motifs (Deng et al., 2003; Zhou et al., 2005), it is important to investigate further the abundance, allelic variation and potential functions of internal (TTAGGG)_n sequence islands in human subtelomere regions. It is not known whether TERRA-like molecules are transcribed near internal (TTAGGG)_n-like sequences.

Second, the depth of coverage in duplicons for a given subtelomere region (corresponding to duplicon copy number for specific subtelomere coordinates) varies widely across subtelomere regions (Ambrosini et al., 2007). For example, large segments of the 18p subtelomere are duplicated two to six times, whereas other segments are present in much higher copy numbers (20 to 25). Some blocks of duplicated DNA are located exclusively to subtelomeres (e.g., coordinates 0–30 kb of the 17p subtelomere sequence; Fig. 2), whereas the majority of the duplicated subtelomere DNA contains non-subtelomere copies as well (e.g., coordinates 0–45 kb of the 18p subtelomere; Fig. 2).

Third, duplicons defined in this manner that occupy subtelomeric sequences are generally both larger and more abundant than those occurring elsewhere in the genome, consistent with the notion that subtelomeric location in humans is permissive for and/or somehow promotes large duplication events (Ambrosini et al., 2007). Although smaller and fewer, non-subtelomeric copies of duplicons tended to cluster at relatively few pericentric and interstitial loci; these include 2q13→q14 (at the site where ancestral primate telomeres fused to form modern human chromosome 2), 1q42.11→q42.12, 1q42.13, 1q43→q44, 3p12.3, 3q29, 4q26, 7p13, 9q12→q13, and Yq11.23. These sites had also been documented previously in genome-wide analyses of segmental duplications (Bailey et al., 2002) and represent sites that were apparently susceptible to either donation or acceptance of these duplicated chromosome segments in recent evolutionary time.

Subtelomeric transcripts

In addition to the *WASH* (Linardopoulou et al., 2007) and the *RPL23AP7*-like gene families described for subterminal DNA sequences (Table 1) and the TERRA RNAs that are transcribed from the regions spanning the subterminal DNA/(TTAGGG)_n tract region (Azzalin et al., 2007), many other subtelomeric gene families have been identified that are embedded in the duplicated DNA encompassing more extended subtelomere regions. These include the immunoglobulin heavy chain genes (found at 14q) (Cook et al., 1994) and olfactory receptor genes (1p, 6p, 8p, 11p, 15q, 19p, and 3q) (Trask et al., 1998). Figure 2 shows the positions of named transcript families on each of the selected subtelomeres. In most cases, pseudogene copies of the respective gene family co-exist amongst the duplicons. Thus, all duplicon copies contributing to subtelomere CNVs are not equivalent, and most current methods for detecting and quantitating CNVs are incapable of distinguishing duplicons carrying functional versus nonfunctional genes. This remains a key analytical challenge for CNVs near telomeres. In addition, it is important to note that only a single reference sequence has been sampled so far, and given the abundant large-scale variation in these regions there are certain to be many additional members of most of these gene families yet to be discovered in the human population. A great deal of work clearly lies ahead in terms of deciphering the functions, evolution and population dynamics of individual members of subtelomeric gene families.

Subtelomeric structural and copy number variation and its functional consequences

Large variant regions of subtelomeric DNA detected at many human telomeres were among the first examples of structural variation and Copy Number Variation (CNV) (Wilkie et al., 1991; Ijdo et al., 1992; Cook et al., 1994; Macina et al., 1994, 1995; Martin-Gallardo et al., 1995; Monfouilloux et al., 1998; Trask et al., 1998; van Overveld et al., 2000), which were subsequently found genome-wide (Iafrate et al., 2004; Sebat et al., 2004). The ‘small’ subtelomere alleles sometimes differ from the ‘large’ alleles by hundreds of kilobases of additional subtelomeric DNA, leading to huge allelic disparities in subtelomeric DNA sequence content and consequent separation of single-copy genes from terminal (TTAGGG)_n tracts. These large allelic disparities are postulated to have an impact both on the dosage (and hence expression levels) of transcripts embedded within the duplicated sequences, but also upon transcription of genes in adjacent 1-copy regions. Although these polymorphisms are often due mainly to differential Srpt DNA content and organization on particular variant alleles, there are also cases of deletion/insertion polymorphisms of single copy DNA adjacent to the Srpt regions. Figure 3 illustrates several types of polymorphic subtelomere structures characterized to date. Any of these large allelic disparities might be expected to significantly affect subtelomeric transcript identity or abundance, either via copy number changes, by alteration of transcriptional regulation, or by rendering particular

Fig. 3. Models of large-scale variation leading to length and copy number variations (CNVs) seen at individual human telomeres. The models are based on published mapping studies as well as our unpublished results. In **A** the alleles vary by addition/deletion of a large SrpT segment at an existing telomere, with the breakage/joining site correlating with an internal (TTAGGG) n -like island on the long allele. Variants at 16p, 11p, 8p, 1p, and 20q (Wilkie et al., 1991; Riethman et al., 2004; Riethman, unpublished) appear to have this structure. In **B** the alleles vary in both SrpT content and SrpT organization, without a simple relationship to each other in the variant region. Some of the 6p and 8p variants appear to follow this model, as does a 19p variant (Riethman et al., 2004; Riethman, unpublished). In **C** the two alleles diverge before the SrpT region begins, leading to a large insertion/deletion polymorphism in distal one-copy DNA. Evidence for this sort of variation exists at the 2q telomere (Macina et al., 1994; also see Ravn et al., 2006, for similar examples of subtelomeric CNVs). In **D** the two alleles are similar except for an insertion/deletion polymorphism contained entirely within a single class of duplicated DNA. Evidence for this exists near the 14q telomere where duplicated IgG VH genes are located (Cook et al., 1994; Redon et al., 2006). A specialized version of the model in **D** is supported by studies of the 4q and the 10q subtelomere, where alleles vary by insertion/deletion polymorphism of tandemly repeated D4Z4 DNA tracts located within the SrpT region (van Overveld et al., 2000). Combinations of these five classes of alternative subtelomeric sequence organization are likely, as are additional types not yet characterized.

transcripts nonfunctional by insertion, deletion, or truncation.

The existence of large-scale subtelomere polymorphisms detected in these earlier studies was recently confirmed and extended in the much larger and comprehensive studies of Redon et al. (2006). This group characterized the genomic DNA of 269 individuals (already genotyped for SNPs as part of the HapMap project) for copy number variation using two genome-scale array CGH platforms; the 500K Affymetrix SNP chip, and a Whole Genome Tiling Path (WGTP) Array of BACs. Copy number variation regions were detected at almost every subtelomere, and in most cases these CNV regions were adjacent to and/or were co-ex-

tensive with the previously identified segmental duplication regions. Given our current understanding of the large and complex segmental duplication regions near telomeres, including duplicons occupying multiple telomeres as well as internal loci, this result is not surprising. The resolution of these and most current CNV detection platforms is generally >50 kb, and most subtelomere regions are under-represented on these platforms, so the level of detailed subtelomere CNV information one can obtain from currently available CNV databases is extremely limited. Fosmid end sequence mapping strategies (Tuzun et al., 2005) applied to subtelomere regions will provide a much more powerful method for exploring and characterizing this variation, and will yield cloned fragments corresponding to the sites of variation to analyze at high resolution. For subtelomere alleles with large and complex duplications, existing and planned BAC libraries (Eichler et al., 2007) will be needed to fully characterize these variants; however, the fosmid libraries can be used to define and characterize the boundaries of the large variant regions.

Structural variation is abundant in subterminal DNA, and is likely to have important consequences for expression of subterminal genes and TERRA RNAs as well as for regulation of (TTAGGG) n tract lengths, especially individual-specific patterns of allele-specific (TTAGGG) n tract lengths that may be crucial for the biological functions of telomeres. Both copy number variation and alternative sequence organizations for DNA immediately adjacent to terminal (TTAGGG) n tracts are major components of this structural variation.

One important near-term goal is to better understand the extent and nature of subterminal structural variation. Recently described structural variation libraries will go a long way towards addressing this gap. The construction and end-sequencing of fosmid libraries from the DNA of unrelated HapMap individuals as part of the Human Structural Variation initiative (Eichler et al., 2007) has resulted in a remarkable resource for investigation of the highly variable subtelomeric regions of the human genome. Each Fosmid End Sequence (FES) library has roughly 12-fold clone coverage of the source genome. Importantly, because the libraries were prepared using sheared source DNA, terminal fosmids with one mate pair corresponding to part of the terminal (TTAGGG) n sequence tract can be identified. Preliminary studies in our lab have demonstrated the feasibility of isolating novel subterminal structural variants using this approach and have shown that, while distal terminal (TTAGGG) n tracts are truncated in the clones, subterminal DNA remains intact (Riethman, unpublished). Many (TTAGGG) n -adjacent sequences are still missing from the reference sequence itself, and a large number of structural variants carrying new common subterminal alleles remain to be discovered and characterized. Given the importance of subterminal sequences for telomere chromatin assembly and telomere length regulation, an effort to systematically identify and characterize subterminal sequences from common telomere alleles is warranted. Such an effort will fill a major gap in our knowledge of subtelomere structures in the

human genome, and will permit development of new tools for studying telomere biology.

Developing methods for tracking individual subtelomeric allele variants in the context of total genomic DNA is still a huge challenge. PCR-based approaches to distinguish similar but non-identical copies of duplicons and groups of adjacent duplicons have had limited success so far, mainly because the sequences and duplicon organization of most common subtelomere alleles in human populations are still unknown. Once more data are available, multiplex ligation-dependent probe amplification (MLPA)-like (Schouten et al., 2002) approaches for duplicon identification and quantitation in case-control cohorts can be envisioned. Similarly, STELA-based approaches (Britt-Compton et al., 2006) for analyzing single-telomere length and stability will likely be expanded as the sequences for a more complete set of subterminal alleles become available. Perhaps the most intriguing approach, however, may be to attempt specific fractionation of subtelomeric DNA from individual genomes using a microarray capture method (Albert et al., 2007; Hodges et al., 2007; Okou et al., 2007; Porreca et al., 2007), then use deep sequencing of the enriched subtelomere frac-

tion and analysis of paired-end reads (Korbel et al., 2007; Campbell et al., 2008) to re-construct subtelomeric scaffolds of the target genome. While high similarity subtelomere duplicons would certainly pose challenges for this method, the potential depth of coverage in paired-end reads and the judicious use of sequence landmarks such as adjacent single copy DNA regions, known associations of adjacent duplicons within subtelomeric segmental duplication regions (Linardopoulou et al., 2005; Ambrosini et al., 2007), and the unique positions of pure (TTAGGG)_n tracts at chromosome termini (Fig. 1) might enable efficient re-sequencing of key parts of subtelomeric and subterminal regions. Such a method would provide unprecedented levels of comprehensive detail with respect to sequence organization and structural variation for this complex but critical region of the human genome.

Acknowledgement

I thank Sheila Paul for helping with the preparation of the manuscript.

References

- Albert TJ, Molla MN, Muzny DM, Nazareth L, Wheeler D, et al: Direct selection of human genomic loci by microarray hybridization. *Nat Methods* 4:903–905 (2007).
- Ambrosini A, Paul S, Hu S, Riethman H: Human subtelomeric duplicon structure and organization. *Genome Biol* 8:R151 (2007).
- Armanios M, Chen JL, Chang YP, Brodsky RA, Hawkins A, et al: Haploinsufficiency of telomerase reverse transcriptase leads to anticipation in autosomal dominant dyskeratosis congenita. *Proc Natl Acad Sci USA* 102:15960–15964 (2005).
- Armanios MY, Chen JJ, Cogan JD, Alder JK, Ingersoll RG, et al: Telomerase mutations in families with idiopathic pulmonary fibrosis. *N Engl J Med* 356:1317–1326 (2007).
- Aviv A, Levy D, Mangel M: Growth, telomere dynamics and successful and unsuccessful human aging. *Mech Ageing Dev* 124:829–837 (2003).
- Azzalin CM, Nergadze SG, Giulotto E: Human intrachromosomal telomeric-like repeats: sequence organization and mechanisms of origin. *Chromosoma* 110:75–82 (2001).
- Azzalin CM, Reichenbach P, Khoraiuli L, Giulotto E, Lingner J: Telomeric repeat containing RNA and RNA surveillance factors at mammalian chromosome ends. *Science* 318:798–801 (2007).
- Bailey JA, Gu Z, Clark RA, Reinert K, Samonte RV, et al: Recent segmental duplications in the human genome. *Science* 297:1003–1007 (2002).
- Baird DM, Jeffreys AJ, Royle NJ: Mechanisms underlying telomere repeat turnover, revealed by hypervariable variant repeat distribution patterns in the human Xp/Yp telomere. *EMBO J* 14:5433–5443 (1995).
- Baird DM, Coleman J, Rosser ZH, Royle NJ: High levels of sequence polymorphism and linkage disequilibrium at the telomere of 12q: implications for telomere biology and human evolution. *Am J Hum Genet* 66:235–250 (2000).
- Baird DM, Rowson J, Wynford-Thomas D, Kipling D: Extensive allelic variation and ultrashort telomeres in senescent human cells. *Nat Genet* 33:203–207 (2003).
- Baird DM, Britt-Compton B, Rowson J, Amso NN, Gregory L, Kipling D: Telomere instability in the male germline. *Hum Mol Genet* 15:45–51 (2006).
- Blackburn EH: Telomeres and telomerase: their mechanisms of action and the effects of altering their functions. *FEBS Lett* 579:859–862 (2005).
- Blasco MA: The epigenetic regulation of mammalian telomeres. *Nat Rev Genet* 8:299–309 (2007).
- Britt-Compton B, Rowson J, Locke M, Mackenzie I, Kipling D, Baird DM: Structural stability and chromosome-specific telomere length is governed by *cis*-acting determinants in humans. *Hum Mol Genet* 15:725–733 (2006).
- Brown WR, MacKinnon PJ, Villasante A, Spurr N, Buckle VJ, Dobson MJ: Structure and polymorphism of human telomere-associated DNA. *Cell* 63:119–132 (1990).
- Campbell PJ, Stephens PJ, Pleasance ED, O'Meara S, Li H, et al: Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. *Nat Genet* 40:722–729 (2008).
- Capper R, Britt-Compton B, Tankimanova M, Rowson J, Letsolo B, et al: The nature of telomere fusion and a definition of the critical telomere length in human cells. *Genes Dev* 21:2495–2508 (2007).
- Coleman J, Baird DM, Royle NJ: The plasticity of human telomeres demonstrated by a hypervariable telomere repeat array that is located on some copies of 16p and 16q. *Hum Mol Genet* 8:1637–1646 (1999).
- Cook GP, Tomlinson IM, Walter G, Riethman H, Carter NP, et al: A map of the human immunoglobulin VH locus completed by analysis of the telomeric region of chromosome 14q. *Nat Genet* 7:162–168 (1994).
- de Lange T: Shelterin: the protein complex that shapes and safeguards human telomeres. *Genes Dev* 19:2100–2110 (2005).
- de Lange T, Shiue L, Myers RM, Cox DR, Naylor SL, et al: Structure and variability of human chromosome ends. *Mol Cell Biol* 10:518–527 (1990).
- Deng Z, Atanasiu C, Burg JS, Broccoli D, Lieberman PM: Telomere repeat binding factors TRF1, TRF2, and hRAP1 modulate replication of Epstein-Barr virus OriP. *J Virol* 77:11992–12001 (2003).
- Der-Sarkissian H, Vergnaud G, Borde YM, Thomas G, Londono-Vallejo JA: Segmental polymorphisms in the proterminal regions of a subset of human chromosomes. *Genome Res* 12:1673–1678 (2002).
- Der-Sarkissian H, Bacchetti S, Cazes L, Londono-Vallejo JA: The shortest telomeres drive karyotype evolution in transformed cells. *Oncogene* 23:1221–1228 (2004).
- Eichler EE, Nickerson DA, Altshuler D, Bowcock AM, Brooks LD, et al: Completing the map of human genetic variation. *Nature* 447:161–165 (2007).
- Epel ES, Blackburn EH, Lin J, Dhabhar FS, Adler NE, et al: Accelerated telomere shortening in response to life stress. *Proc Natl Acad Sci USA* 101:17312–17315 (2004).
- Flint J, Bates GP, Clark K, Dorman A, Willingham D, et al: Sequence comparison of human and yeast telomeres identifies structurally distinct subtelomeric domains. *Hum Mol Genet* 6:1305–1313 (1997).
- Gianfrancesco F, Falco G, Esposito T, Rocchi M, D'Urso M: Characterization of the murine orthologue of a novel human subtelomeric multi-gene family. *Cytogenet Cell Genet* 94:98–100 (2001).

- Graakjaer J, Bischoff C, Korsholm L, Holstebro S, Vach W, et al: The pattern of chromosome-specific variations in telomere length in humans is determined by inherited, telomere-near factors and is maintained throughout life. *Mech Ageing Dev* 124:629–640 (2003).
- Graakjaer J, Pascoe L, Der-Sarkissian H, Thomas G, Kolvraa S, et al: The relative lengths of individual telomeres are defined in the zygote and strictly maintained during life. *Aging Cell* 3: 97–102 (2004).
- Graakjaer J, Der-Sarkissian H, Schmitz A, Bayer J, Thomas G, et al: Allele-specific relative telomere lengths are inherited. *Hum Genet* 119: 344–350 (2006).
- Griffith JD, Comeau L, Rosenfield S, Stansel RM, Bianchi A, et al: Mammalian telomeres end in a large duplex loop. *Cell* 97:503–514 (1999).
- Harley CB, Futcher AB, Greider CW: Telomeres shorten during ageing of human fibroblasts. *Nature* 345:458–460 (1990).
- Hemann MT, Strong MA, Hao LY, Greider CW: The shortest telomere, not average telomere length, is critical for cell viability and chromosome stability. *Cell* 107:67–77 (2001).
- Hiyama E, Hiyama K: Telomere and telomerase in stem cells. *Br J Cancer* 96:1020–1024 (2007).
- Hodges E, Xuan Z, Balija V, Kramer M, Molla MN, et al: Genome-wide in situ exon capture for selective resequencing. *Nat Genet* 39:1522–1527 (2007).
- Iafraite AJ, Feuk L, Rivera MN, Listewnik ML, Donahoe PK, et al: Detection of large-scale variation in the human genome. *Nat Genet* 36:949–951 (2004).
- IHGSC: Finishing the euchromatic sequence of the human genome. *Nature* 431:931–945 (2004).
- Ijdo JW, Lindsay EA, Wells RA, Baldini A: Multiple variants in subtelomeric regions of normal karyotypes. *Genomics* 14:1019–1025 (1992).
- Karlseder J, Smogorzewska A, de Lange T: Senescence induced by altered telomere state, not telomere loss. *Science* 295:2446–2449 (2002).
- Korbel JO, Urban AE, Affourtit JP, Godwin B, Grubert F, et al: Paired-end mapping reveals extensive structural variation in the human genome. *Science* 318:420–426 (2007).
- Lansdorp PM, Verwoerd NP, van de Rijke FM, Dragowska V, Little MT, et al: Heterogeneity in telomere length of human chromosomes. *Hum Mol Genet* 5:685–691 (1996).
- Lee HW, Blasco MA, Gottlieb GJ, Horner JW 2nd, Greider CW, DePinho RA: Essential role of mouse telomerase in highly proliferative organs. *Nature* 392:569–574 (1998).
- Levy MZ, Allsopp RC, Futcher AB, Greider CW, Harley CB: Telomere end-replication problem and cell aging. *J Mol Biol* 225:951–960 (1992).
- Linaropoulou EV, Williams EM, Fan Y, Friedman C, Young JM, Trask BJ: Human subtelomeres are hot spots of interchromosomal recombination and segmental duplication. *Nature* 437: 94–100 (2005).
- Linaropoulou EV, Parghi SS, Friedman C, Osborn GE, Parkhurst SM, Trask BJ: Human subtelomeric WASH genes encode a new subclass of the WASP family. *PLoS Genet* 3:e237 (2007).
- Lundblad V: Telomere maintenance without telomerase. *Oncogene* 21:522–531 (2002).
- Macina RA, Negorev DG, Spais C, Ruthig LA, Hu XL, Riethman HC: Sequence organization of the human chromosome 2q telomere. *Hum Mol Genet* 3:1847–1853 (1994).
- Macina RA, Morii K, Hu XL, Negorev DG, Spais C, et al: Molecular cloning and RARE cleavage mapping of human 2p, 6q, 8q, 12q, and 18q telomeres. *Genome Res* 5:225–532 (1995).
- Martin-Gallardo A, Lamerdin J, Sopapan P, Friedman C, Fertitta AL, et al: Molecular analysis of a novel subtelomeric repeat with polymorphic chromosomal distribution. *Cytogenet Cell Genet* 71:289–295 (1995).
- Meeker AK, Hicks JL, Iacobuzio-Donahue CA, Montgomery EA, Westra WH, et al: Telomere length abnormalities occur early in the initiation of epithelial carcinogenesis. *Clin Cancer Res* 10:3317–3326 (2004).
- Mefford HC, Trask BJ: The complex structure and dynamic evolution of human subtelomeres. *Nat Rev Genet* 3:91–102 (2002).
- Meier A, Fiegler H, Munoz P, Ellis P, Rigler D, et al: Spreading of mammalian DNA-damage response factors studied by ChIP-chip at damaged telomeres. *EMBO J* 26:2707–2718 (2007).
- Mondello C, Pirzio L, Azzalin CM, Giulotto E: Instability of interstitial telomeric sequences in the human genome. *Genomics* 68:111–117 (2000).
- Monfouilloux S, Avet-Loiseau H, Amarger V, Balazs I, Pourcel C, Vergnaud G: Recent human-specific spreading of a subtelomeric domain. *Genomics* 51:165–176 (1998).
- Moyzis RK, Buckingham JM, Cram LS, Dani M, Deaven LL, et al: A highly conserved repetitive DNA sequence, (TTAGGG)_n, present at the telomeres of human chromosomes. *Proc Natl Acad Sci USA* 85:6622–6626 (1988).
- Okou DT, Steinberg KM, Middle C, Cutler DJ, Albert TJ, Zwick ME: Microarray-based genomic selection for high-throughput resequencing. *Nat Methods* 4:907–909 (2007).
- Porreca GJ, Zhang K, Li JB, Xie B, Austin D, et al: Multiplex amplification of large sets of human exons. *Nat Methods* 4:931–936 (2007).
- Ravnan JB, Tepperberg JH, Papenhausen P, Lamb AN, Hedrick J, et al: Subtelomere FISH analysis of 11,688 cases: an evaluation of the frequency and pattern of subtelomere rearrangements in individuals with developmental disabilities. *J Med Genet* 43:478–489 (2006).
- Reddel RR: Alternative lengthening of telomeres, telomerase, and cancer. *Cancer Lett* 194:155–162 (2003).
- Redon R, Ishikawa S, Fitch KR, Feuk L, Perry GH, et al: Global variation in copy number in the human genome. *Nature* 444:444–454 (2006).
- Riethman H: Human telomere structure and biology. *Annu Rev Genomics Hum Genet* 9:1–19 (2008).
- Riethman H, Ambrosini A, Castaneda C, Finklestein J, Hu XL, et al: Mapping and initial analysis of human subtelomeric sequence assemblies. *Genome Res* 14:18–28 (2004).
- Rudd MK, Friedman C, Parghi SS, Linaropoulou EV, Hsu L, Trask BJ: Elevated rates of sister chromatid exchange at chromosome ends. *PLoS Genet* 3:e32 (2007).
- Ruiz-Herrera A, Garcia F, Azzalin C, Giulotto E, Egozcue J, et al: Distribution of intrachromosomal telomeric sequences (ITS) on *Macaca fascicularis* (Primates) chromosomes and their implication for chromosome evolution. *Hum Genet* 110:578–586 (2002).
- Schoeftner S, Blasco MA: Developmentally regulated transcription of mammalian telomeres by DNA-dependent RNA polymerase II. *Nat Cell Biol* 10:228–236 (2008).
- Schouten JP, McElgunn CJ, Waaijer R, Zwijnenburg D, Diepvens F, Pals G: Relative quantification of 40 nucleic acid sequences by multiplex ligation-dependent probe amplification. *Nucleic Acids Res* 30:e57 (2002).
- Sebat J, Lakshmi B, Troge J, Alexander J, Young J, et al: Large-scale copy number polymorphism in the human genome. *Science* 305:525–528 (2004).
- Sfeir AJ, Chai W, Shay JW, Wright WE: Telomere-end processing the terminal nucleotides of human chromosomes. *Mol Cell* 18:131–138 (2005).
- Slagboom PE, Droog S, Boomsma DI: Genetic determination of telomere size in humans: a twin study of three age groups. *Am J Hum Genet* 55: 876–882 (1994).
- Stavenhagen JB, Zakian VA: Internal tracts of telomeric DNA act as silencers in *Saccharomyces cerevisiae*. *Genes Dev* 8:1411–1422 (1994).
- Steinert S, Shay JW, Wright WE: Modification of subtelomeric DNA. *Mol Cell Biol* 24:4571–4580 (2004).
- Sugiyama T, Cam HP, Sugiyama R, Noma K, Zofall M, et al: SHREC, an effector complex for heterochromatic transcriptional silencing. *Cell* 128:491–504 (2007).
- Trask BJ, Friedman C, Martin-Gallardo A, Rowen L, Akinbami C, et al: 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 (1998).
- Tuzun E, Sharp AJ, Bailey JA, Kaul R, Morrison VA, et al: Fine-scale structural variation of the human genome. *Nat Genet* 37:727–732 (2005).
- van Overveld PG, Lemmers RJ, Deidda G, Sandkuijl L, Padberg GW, et al: Interchromosomal repeat array interactions between chromosomes 4 and 10: a model for subtelomeric plasticity. *Hum Mol Genet* 9:2879–2884 (2000).
- Vulliamy T, Marrone A, Goldman F, Dearlove A, Bessler M, et al: The RNA component of telomerase is mutated in autosomal dominant dyskeratosis congenita. *Nature* 413:432–435 (2001).
- Wilkie AO, Higgs DR, Rack KA, Buckle VJ, Spurr NK, et al: Stable length polymorphism of up to 260 kb at the tip of the short arm of human chromosome 16. *Cell* 64:595–606 (1991).
- Wright WE, Shay JW: Historical claims and current interpretations of replicative aging. *Nat Biotechnol* 20:682–688 (2002).
- Wutz A, Gribnau J: X inactivation Xplained. *Curr Opin Genet Dev* 17:387–393 (2007).
- Xu L, Blackburn EH: Human cancer cells harbor T-stumps, a distinct class of extremely short telomeres. *Mol Cell* 28:315–327 (2007).
- Zhou J, Chau CM, Deng Z, Shiekhhattar R, Spindler MP, et al: Cell cycle regulation of chromatin at an origin of DNA replication. *EMBO J* 24:1406–1417 (2005).
- Zijlmans JM, Martens UM, Poon SS, Raap AK, Tanke HJ, et al: Telomeres in the mouse have large inter-chromosomal variations in the number of T2AG3 repeats. *Proc Natl Acad Sci USA* 94: 7423–7428 (1997).