Subtelomeric CTCF and cohesin binding site organization using improved subtelomere assemblies and a novel annotation pipeline

Nicholas Stong  
*University of Pennsylvania School of Medicine*

Zhong Deng  
*The Wister Institute*

Ravi Gupta  
*The Wister Institute*

Sufen Hu  
*The Wister Institute*

Shiela Paul  
*The Wister Institute*

See next page for additional authors

Follow this and additional works at: https://digitalcommons.wustl.edu/open_access_pubs

Please let us know how this document benefits you.

**Recommended Citation**

Stong, Nicholas; Deng, Zhong; Gupta, Ravi; Hu, Sufen; Paul, Shiela; Weiner, Amber K.; Eichler, Evan E.; Graves, Tina; Fronick, Catrina C.; Courtney, Laura; Wilson, Richard K.; Lieberman, Paul M.; Davuluri, Ramana V.; and Riethman, Harold, "Subtelomeric CTCF and cohesin binding site organization using improved subtelomere assemblies and a novel annotation pipeline." *Genome Research*. 24, 1039-1050. (2014).

https://digitalcommons.wustl.edu/open_access_pubs/2983
Authors
Subtelomeric CTCF and cohesin binding site organization using improved subtelomere assemblies and a novel annotation pipeline

Nicholas Stong, Zhong Deng, Ravi Gupta, et al.

Genome Res. 2014 24: 1039-1050 originally published online March 27, 2014
Access the most recent version at doi:10.1101/gr.166983.113

Supplemental Material
http://genome.cshlp.org/content/suppl/2014/04/16/gr.166983.113.DC1.html

References
This article cites 77 articles, 34 of which can be accessed free at:
http://genome.cshlp.org/content/24/6/1039.full.html#ref-list-1

Creative Commons License
This article is distributed exclusively by Cold Spring Harbor Laboratory Press for the first six months after the full-issue publication date (see http://genome.cshlp.org/site/misc/terms.xhtml). After six months, it is available under a Creative Commons License (Attribution-NonCommercial 4.0 International), as described at http://creativecommons.org/licenses/by-nc/4.0/.

Email Alerting Service
Receive free email alerts when new articles cite this article - sign up in the box at the top right corner of the article or click here.
Subtelomeric CTCF and cohesin binding site organization using improved subtelomere assemblies and a novel annotation pipeline

Nicholas Stong,1,2 Zhong Deng,2 Ravi Gupta,2 Sufen Hu,2 Shiela Paul,2 Amber K. Weiner,2 Evan E. Eichler,3 Tina Graves,4 Catrina C. Fronick,4 Laura Courtney,4 Richard K. Wilson,4 Paul M. Lieberman,2 Ramana V. Davuluri,2 and Harold Riethman2,5

Subtelomeric DNA is crucial for telomere (TTAGGG)n tract length regulation and telomeric chromatin integrity. A telomeric repeat-containing family of RNAs (TERRA) is transcribed from subtelomeres into the (TTAGGG)n tracts (Azzalin et al. 2007; Schoeftner and Blasco 2008; Porro et al. 2010) and forms an integral component of a functional telomere; perturbation of its abundance and/or localization causes telomere dysfunction and genome instability (Azzalin et al. 2007; Deng et al. 2009). Telomere dysfunction caused by critically short telomere DNA sequence or by disruption of telomeric chromatin integrity induces DNA damage response pathways that cause cellular senescence or apoptosis (depending on the cellular context) in the presence of a functional p53 tumor suppressor pathway (Palm and de Lange 2008). Only one or a few critically short telomeres in a cell are sufficient to induce DDR-mediated senescence or apoptosis (Zou et al. 2007; Meier et al. 2007). Senescence or apoptosis of somatic cells can disrupt tissue microenvironments, and senescence or apoptosis of stem cell populations can prevent proper replenishment of rapidly dividing cellular lineages, both impacting aging phenotypes and age-related diseases, including cancer (Coppé et al. 2010; Davalos et al. 2010; Jaksic et al. 2010; Sahin and Depinho 2010).

Subtelomeric DNA elements regulate both TERRA levels and haplotype-specific (TTAGGG)n tract length and stability (Graakjaer et al. 2003, 2006; Britt-Compton et al. 2006; Deng et al. 2009; Nergadze et al. 2009), with accumulating evidence for specific epigenetic modulation of these effects (Yehezkel et al. 2008; Casolini et al. 2009; Deng et al. 2009; Nergadze et al. 2009; Arnoult et al. 2012). Heterogeneously sized TERRA transcripts with as yet undefined transcription start sites and potential splice patterns originate in many, perhaps all, human subtelomere regions (Azzalin et al. 2008; Porro et al. 2010; Deng et al. 2012), with the sizes of the larger transcripts (>15 kb) suggesting structural overlap with some transcribed subtelomeric gene families (Riethman 2008a,b). While many details of the dynamic interplay between shelterin, telomere...
chromatin structure, TERRA expression, and telomere biology remain unclear, recent work from our group indicates that CTCF and cohesin are integral components of most human subtelomeres, and important for the regulation of TERRA transcription and telomere end protection (Deng et al. 2012).

The chromatin organizing factor CTCF has been implicated in numerous aspects of chromosome biology, including chromatin insulator, enhancer blocker, transcriptional activator and repressor, DNA methylation–sensitive parental imprinting, and DNA-loop formation between transcriptional control elements (Bushuy et al. 2008; Phillips and Corces 2009; Ohlsson et al. 2010). In addition to its role in TERRA regulation, CTCF has been implicated in the transcriptional repression of a subtelomeric D4Z4 macrosatellite repeat transcript – 30 kb from the telomere repeats of chromosome 4q (Ottaviani et al. 2011). At D4Z4, CTCF interacts with lamin A and tethers the chromosome 4q telomere to the nuclear periphery (Ottaviani et al. 2009a,b). A more general role for CTCF has been found in its ability to colocalize with cohesin subunits at many chromosomal positions (Parelo et al. 2008; Rubio et al. 2008; Stedman et al. 2008; Wendt et al. 2008). Cohesin is a multiprotein complex consisting of core subunits SMC1, SMC3, RAD21, and STAG1 or STAG2, which can form a ring-like structure capable of encircling or embracing two DNA molecules (Nasmyth and Haering 2005; Hirano 2006). Cohesin was originally identified as a regulator of sister chromatid cohesion, but subsequent studies in higher eukaryotes indicate functions in mediating long-distance interactions between DNA elements required for transcription regulation (Kagey et al. 2010; Dorsett 2011). Cohesin subunit STAG1 is recruited to telomere repeats by the shelterin protein TIN2, and this interaction is required for telomeric sister chromatid cohesion and efficient telomere replication (Canudas and Smith 2009; Remeseiro et al. 2012). STAG1 binds directly to telomere repeat DNA through a unique AT hook, and overexpression of STAG1 alone is sufficient to induce cohesion at telomeres independently of cohesin ring components (Bisht et al. 2013). In contrast, colocalized cohesin ring components and CTCF both contribute to subtelomeric TERRA transcriptional regulation and telomere end protection (Deng et al. 2012).

In humans, telomere regulation occurs in the context of subtelomeric DNA segmental duplications known as subtelomeric repeat elements (SREs), which comprise ~80% of the most distal 100 kb and 25% of the most distal 500 kb in human DNA (The International Human Genome Sequencing Consortium 2004; Riethman et al. 2004). SRE regions of human chromosomes contain mosaic patchworks of duplicons (Der-Sarkissian et al. 2002; Mefford and Trask 2002; Ambrosini et al. 2007) apparently generated by translocations involving the tips of chromosomes, followed by transposition of unbalanced chromosomal complements to offspring (Linaropoulou et al. 2005). Along with highly elevated sister chromatid exchange (SCE) rates in subtelomeres (Rudd et al. 2007), these studies indicate that human subtelomeres are duplication-rich hotspots of DNA breakage and repair.

Here, we have generated improved human subtelomere assemblies by sequencing additional subtelomeric clones and revising the reference sequence of distal subtelomere regions. A bioinformatic pipeline for annotation of the updated subtelomere assemblies using short-read data sets is developed and implemented. A public browser for the integrated display of short-read-based annotations relative to key subtelomere features such as the start of each terminal repeat tract, SRE identity and organization, and subtelomeric gene models is established and used to investigate cohesin and CTCF binding in SRE regions.

Results

Gap-filling and detection of distal telomeric structural variants

In order to fill remaining telomere-adjacent gaps from our previous reference subtelomere assembly (Ambrosini et al. 2007), we sampled telomere-adjacent DNA from deep fosmid clone libraries prepared from sheared genomic DNA samples (The International Human Genome Sequencing Consortium 2004; Kidd et al. 2008, 2010). Since each fosmid from these libraries had been end-sequenced using Sanger methods, we computationally searched for (CCCTAA)n sequence (the DNA sequence and orientation expected from fosmids ends located within telomere terminal repeat tracts) and selected the (CCCTAA)n-positive group of clones for further analysis. Each mate-pair read associated with a (CCCTAA)n read was mapped to our laboratory’s previous assembly (Ambrosini et al. 2007) to create a deep-coverage resource of mapped fosmid clones containing telomere-adjacent DNA. Using this mapping information, representative single clones that spanned gaps in the assembly were selected and sequenced (Table 1). Included in this group of clones were two structural variants identified in the mapping studies that, while capturing telomere-adjacent DNA for these chromosome ends, removed some SRE sequence from our previous assembly (analogous to the sequenced 16p allele relative to the longer mapped variant 16p alleles) (Flint et al. 1997). A second allele for the distal 4q subtelomere, which shared high sequence similarity with distal 10q (Lemmers et al. 2007), was also sequenced, as was a yeast artificial clone (YAC)–derived sequence we identified which filled a 12q gap. Finally, the mapped telomere fosmid resource was used to complete 8q and 18q telomere-adjacent sequences that contained sequence ambiguities and misassemblies immediately adjacent to the (TTAGGG)n tract in the previous assembly (Ambrosini et al. 2007); these errors were retained in hg19. Further details relating to fosmid library screening and characterization, the mapped telomere fosmid resources available from this work, and direct sequencing from distal telomere fosmids are provided in the Supplemental Material (Mapped Telomere Fosmid Resource; Supplemental Figs. 1, 2; Supplemental Tables 1–4).

Updated subtelomere assemblies

Rather than simply extending our previous assembly, we combined our new sequences with all other available fully sequenced subtelomere clones in NCBI to create an updated clone-based assembly of human subtelomere regions (Supplemental Table 5). We used, to the extent possible, contiguous segments of the existing hg19 assembly for the preparation of our 500-kb-sized subtelomere assemblies, only altering regions where our data indicated substantial change was required. The subtelomere regions that changed relative to hg19 are shown in Figure 1; 18 telomere-adjacent regions were altered: 15 by addition to or replacement of hg19 sequence and three by truncation of hg19 sequence. For all telomeres not showing change relative to hg19 in Fig 1, the distal-most telomere gaps and clone gaps (where they existed immediately adjacent to telomere gaps), represented in hg19 by a long string of Ns, were removed. Distal telomere tract sequence was also removed, so that coordinate 1 of each assembly corresponds to the start of the terminal repeat tract on the strand oriented toward the centromere (to maintain a consistent starting coordinate for subtelomere annotation). For the seven telomeres whose reference sequences do not extend to the terminal repeat (6p, 8p, 1p, 11p, 3q, 9q, 20p), coordinate 1 corresponds to the most distal base of the subtelomere assembly. The five acrocentric short arm telomeres are not represented in our assem-
Table 1. Subtelomeric sequences from telomeric clones

<table>
<thead>
<tr>
<th>Tel</th>
<th>Clone name</th>
<th>Accession</th>
<th>bp</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>10p</td>
<td>ABC7-43086900J1</td>
<td>AC215217</td>
<td>34335</td>
<td>Extends 10p ref sequence to terminal (TTAGGG)n tract</td>
</tr>
<tr>
<td>12p</td>
<td>ABC7-43239800N19</td>
<td>AC215219</td>
<td>35739</td>
<td>Extends 12p ref sequence to terminal (TTAGGG)n tract</td>
</tr>
<tr>
<td>13q</td>
<td>WI2-1528010</td>
<td>AC213859</td>
<td>28566</td>
<td>Extends 13q ref sequence to terminal (TTAGGG)n tract</td>
</tr>
<tr>
<td>14q</td>
<td>WI2-1019C11</td>
<td>AC213860</td>
<td>33970</td>
<td>Extends 14q ref sequence to terminal (TTAGGG)n tract</td>
</tr>
<tr>
<td>20q</td>
<td>ABC7-42391600012</td>
<td>AC215218</td>
<td>37776</td>
<td>Truncated variant allele of 20q to terminal (TTAGGG)n tract</td>
</tr>
<tr>
<td>22q</td>
<td>WI2-1161P17</td>
<td>AC213861</td>
<td>33328</td>
<td>Extends 22q ref sequence to terminal (TTAGGG)n tract</td>
</tr>
<tr>
<td>2q</td>
<td>ABC7-4304130019</td>
<td>AC215220</td>
<td>36897</td>
<td>Extends 2q ref sequence to terminal (TTAGGG)n tract</td>
</tr>
<tr>
<td>3p</td>
<td>ABC7-4283560016</td>
<td>AC215221</td>
<td>30142</td>
<td>Extends 3p ref sequence to terminal (TTAGGG)n tract</td>
</tr>
<tr>
<td>4q-1</td>
<td>WI2-30350022</td>
<td>AC225782</td>
<td>42093</td>
<td>Extends 4q ref sequence to terminal (TTAGGG)n tract</td>
</tr>
<tr>
<td>4q-2</td>
<td>ABC7-42391500016</td>
<td>AC215522</td>
<td>31434</td>
<td>Extends 4q ref sequence to terminal (TTAGGG)n tract (second allele)</td>
</tr>
<tr>
<td>7p</td>
<td>ABC7-41722F1</td>
<td>AC215522</td>
<td>33901</td>
<td>Truncated variant allele of 7p to terminal (TTAGGG)n tract</td>
</tr>
<tr>
<td>12q_gap</td>
<td>CA-2196C1 (from half-YAC)</td>
<td>AC226150</td>
<td>39835</td>
<td>Subcloned cosmid from half-YAC yRM2196; spans gap from AC026786.5 to a previously sequenced telomeric cosmid (CMF-21K2, AP006310), which contains start of 12q telomere terminal (TTAGGG)n tract.</td>
</tr>
</tbody>
</table>

The hybrid genome was used to annotate subtelomeric sequence features as described in the study by Ambrosini et al. (2007) and to map several ChlP (chromatin immunoprecipitation)-seq data sets of particular interest to subtelomere function (Deng et al. 2012). Figure 2 illustrates these annotations for the first 250 kb of the 19p subtelomere. Both coding and noncoding transcripts are abundant in SRE regions; while some are clearly functional, most are not well characterized (Linardopoulou 2001; Riethman et al. 2004; Linardopoulou et al. 2005; Ambrosini et al. 2007).

A paralogy map for SRE regions was prepared based upon the paralogy blocks defined previously (Linardopoulou et al. 2005) to facilitate graphic visualization of similar sequence segments occurring in multiple telomeres (see Fig. 1; Methods). Previously defined paralogy blocks covered most SRE regions, but we identified five new blocks and divided block 19 into two sub-blocks because of subtelomeric sequence not available to Linardopoulou et al. (2005). Paralogy blocks as defined by Linardopoulou et al. (2005) were developed as graphic visualization tools and have inexact borders with lower boundary resolution than the duplicons defined by Ambrosini et al. (2007). In addition, the paralogy blocks share slightly higher percentages of nucleotide sequence similarity than the duplicons defined by Ambrosini et al. (2007), because the paralogy blocks include high copy repeat sequence for this analysis, whereas the duplicon analysis of Ambrosini et al. (2007) uses only non-repeat-masked sequence for sequence comparisons.

The mapping of short-read data sets to human subtelomere regions requires special consideration because of the recent segmental duplication content. To deal with this challenge, we used a strategy of assigning a mapping likelihood (ml tag) to reads equal to the inverse of its genome-wide mapping positions; in effect, splitting up a read and mapping an equal portion of it to all of its possible sites of true mapping (Wang et al. 2010; Chung et al. 2011). By using this alternative mapping strategy, we then build fragment densities to display on enrichment tracks and to call peaks (see Methods). Concurrently, a track for each sample was built using only uniquely mapping reads (with an ml tag of 1) for comparison with the multiread track. The multiread tracks are shown in the figures; tracks for uniquely mapping reads can be found in the subtelomere browser (vader.wistar.upenn.edu/humansubtel). By using this pipeline, enrichment profiles for four of the ChIP-seq data sets (Genome Research 1041) were available and mapped to the entire hybrid genome using the multiread mapping approach and then displaying the short 500 kb on the subtelomere browser (see Methods). The same
subterminal binding enrichments for CTCF, SMC1A, RAD21, and POLR2A, interestingly, the sequences adjacent to this ITS share similar but nonidentical features with sequences adjacent to terminal (TTAGGG)n repeat tracts. The POLR2A peak is positioned over a degenerate version of the subterminal 29-mer element (Deng et al. 2012); this ITS-adjacent binding site corresponds to a 23-mer element that, like the 29-mer repeat, is CpG rich. The CTCF/cohesin peaks span an extended 61-mer repeat array (7.3 copies in the ITS-adjacent sequence, vs. between two and four copies at most subterminal sites), but only 44 of 61 bases on the consensus 61-mer sequences are shared between subterminal and internal copies. The pattern of CTCF, cohesin, and POLR2A binding to these internal sequences is nearly identical to that found adjacent to terminal repeats (Deng et al. 2012), even though the sequences have diverged substantially. In fact, the sequences adjacent to this ITS are more similar to several other subtelomeric ITS-adjacent sequences (90%) than they are to any subterminal copies (85%).

**SRE boundary enrichments**

Publicly available CTCF and cohesin subunit ChIP-seq data sets from human ES cells and primary diploid fibroblasts (IMR90) were mapped in the same fashion and compared with the LCL data. All of the data sets used in this study and their mapping characteristics are summarized in Supplemental Table 7. Broadly speaking, similar patterns of CTCF and cohesin binding to the terminal boundary regions [defined as within 3 kb of the (TTAGGG)n repeat tract] were observed in LCL, ES, and primary fibroblast (IMR90) cell types, although the relative peak heights sometimes varied substantially (Fig. 3; Supplemental Fig. 4). For example, terminal boundary RAD21 enrichment peaks were almost always visible at some level in the bedGraphs of the expected subtelomeres, but for some data sets, many of the peaks did not reach the MACS significance threshold set for peak-calling subtelomere-wide (P < 1.0 × 10^{-4}) (Supplemental Table 8). Many, but not all CTCF and cohesin sites across the SRE regions in LCLs were also detectable in the ES cell lines and in IMR90 (Fig. 3; Supplemental Fig. 4). Differences in library quality and depth, as well as differences in the antibodies used for ChIP-seq (Supplemental Table 7), are expected to have an effect on binding enrichments. However, easily discernible proportional differences in peak heights as well as clear instances of differential peak presence/absence between cell types may be indicative of true differential binding. These candidate differentially binding sites are easily detectable visually...
compare relative CTCF peak heights and relative RAD21 peak heights on distal 6q in Fig. 3, and in all subtelomeres in Supplementary Fig. 4 and on the subtelomere browser; vader.wistar.upenn.edu/humansubtel). Most CTCF binding sites have been thought to be invariant between cell types, but a recent study suggested significant plasticity in CTCF occupancy at a majority of sites genome-wide, with 41% of the variable occupancy sites linked to differential CpG methylation (Wang et al. 2012). Similarly, a subset of cohesin binding sites are known to display cell-type specificity, colocalizing with tissue-specific transcription factors (Merkenschlager and Odom 2013). In this context, it will be intriguing to follow up our initial annotations here with detailed studies of the differential CTCF and cohesin occupancy of binding sites in the telomere-adjacent regions, and their potential implications for telomere length and stability.

Visually noted apparent association of CTCF and cohesin peaks with some SRE boundaries was analyzed systematically using only significant peaks called for each data set by MACS (Table 2; Supplemental Table 8; Zhang et al. 2008). The terminal SRE boundary was defined as the start of the terminal (TTAGGG)n tract; since the CTCF and cohesin binding sites associated with terminal repeat tracts are consistently <3 kb from this boundary (Deng et al. 2012), we initially used a 3-kb window to scan all SRE boundaries for CTCF and cohesin subunit peaks. Peak association enrichments are the observed ratio of peaks in the boundary window regions to the expected peak number within these windows if the total number of peaks in the SRE regions were distributed evenly. Some boundaries are within the allowable window of each other; in these instances, a peak can be associated with more than one boundary, although no additional weighting is added to the boundary association of these peaks. To calculate a P-value for the enrichment of peaks in boundary regions, a one-sided binomial test was performed.
sites with ITSs in all data sets except for IMR90 RAD21 (Table 2). There were weaker and often statistically insignificant associations of CTCF and cohesin sites with internal SRE/SRE boundaries in the individual data sets from these cell types (Table 2). However, boundary analysis of just the strictly colocalized peaks for CTCF and cohesin subunits showed significant associations with SRE/SRE boundaries for LCLs and ES cells, but not for the primary fibroblast cell line IMR90 (Table 2). The positions of all colocalized CTCF and cohesin peaks occurring in at least one of these three cell types are shown relative to SRE organization in Figure 1.

Experimental validation of ChIP-seq peaks by ChIP-qPCR

Several recent reports have suggested that some human ITSs bind the shelterin components TERF1 and TERF2 (Simonet et al. 2011; Yang et al. 2011), which seems plausible given the demonstrated ability of TERF1 and TERF2 to interact with the very short TTAGGGTT motif in some contexts (Deng et al. 2003; Zhou et al. 2005). This could have important functional implications and suggest potential long-range interaction of ITSs with telomeres. We therefore mapped TERF1 and TERF2 ChIP-seq data sets we prepared from LCLs, as well as publicly available TERF1 and TERF2 ChIP-seq data sets from a transformed BJ fibroblast cell line (Simonet et al. 2011). Enrichment peaks localizing to many subtelomeric ITSs were initially found for both TERF1 and TERF2, in both cell types. However, in each case the mapped reads contributing to the peak did not have a normal distribution (Supplemental Fig. 3A), the consequence of a pile-up of reads mapped on both strands underneath a central peak region being extended to the ChIP fragment length, resulting in peak shoulders that do not correspond to true fragment ends (see Methods). The reads mapping to ITSs were composed of telomere-like repeat arrays. While these reads map “uniquely” according to sequence aligners, this is only in relation to the rest of the reference genome. Neither hg19 nor our hybrid genome includes proximal regions of terminal repeat tracts, known to contain extended regions of telomere-like sequences interspersed with pure TTAGGG repeat arrays (Baird et al. 1995). When telomere and telomere-like sequences were specifically removed from the data sets, peaks at all subtelomeric ITSs disappeared (Supplemental Fig. 3A). Examination of read orientations underneath a typical ITS peak compared with a true CTCF enrichment peak shows that reads responsible for ITS peaks are piled up in random orientation, whereas a true enrichment peak has reads oriented nonrandomly toward the peak of the enrichment (Supplemental Fig. 3B,C). In addition, true binding sites should be marked by noticeable enrichments in sequences flanking the central binding sites, but these enrichments were not found.

To test experimentally the computationally predicted subtelomeric CTCF and RAD21 colocalization sites in SRE regions and whether the called TERF1/TERF2 ChIP-seq peaks described above correlate with TERF1 and TERF2 binding, we carried out a series of ChIP-qPCR experiments summarized in Figure 4 and in Supplemental Figure 5. In Figure 4A, the colocalized CTCF and RAD21 sites in segments of the 6q and 16q SRE regions were examined; each of these sites was not called as peaks when only the uniquely mapping read sets were considered, but peaks were called at these positions using our multiread mapping pipeline. Each of the CTCF and RAD21 binding sites predicted by ChIP-seq mappings (primer positions 2, 4–6, 8–10) show the expected enrichments upon ChIP-qPCR relative to the control primer sets (3 and 7). In addition, the telomere-adjacent sites at primer positions 1 and 2 show the expected TERF1 and TERF2 enrichment very close to the terminal repeat tracts (Deng et al. 2012), whereas more distant subtelomeric sites at positions 3–10 show only background TERF1 and TERF2 levels. In Figure 4B, the expected CTCF and RAD21 en-
richments are also seen in assays corresponding to colocalized CTCF and RAD21 sites predicted by ChIP-seq (positions 2, 4--6, 8). The telomere-adjacent Xq sites at positions 1 and 2 detect the expected TERF1 and TERF2 enrichments (Deng et al. 2012), but the position at 17p corresponding to an ITS with called TERF1 and TERF2 ChIP-seq peaks (position 5) shows only background levels of TERF1 and TERF2 binding. The Xq ITS adjacent to position 3 lacked a ChIP-seq enrichment peak in the TERF1 and TERF2 data sets, yet the ChIP-qPCR showed slight enrichment for TERF2, possibly because it is relatively close (9 kb) to the Xq telomere. Additional ChIP-qPCR assays from 19p and 11p show no correlation between ITS-associated ChIP-seq peaks called in the TERF1 and TERF2 data sets and binding enrichment by ChIP-qPCR, while showing anticipated ChIP-qPCR enrichments at CTCF and RAD21 colocalization sites predicted by ChIP-seq (Supplemental Fig. 5). Thus, we conclude that CTCF and RAD21 binding sites in SRE regions predicted by ChIP-seq multiread mappings are true binding sites, but that the ITS-associated ChIP-seq peaks called in the TERF1 and TERF2 data sets cannot be used to predict true TERF1 and TERF2 binding.

CTCF data sets from additional primary and cancer cell lines

To test whether the terminal boundary and the ITS CTCF peak associations seen in the cell types described above are also seen in additional cell types, we mapped publically available CTCF ChIP-seq data sets from four primary cell lines (HMECs [human mammary epithelial cells], SAECs [small airway epithelial cells], HREs [human renal cortical epitheliums], and HRPEpic [retinal pigment epithelial cells]) and four immortal cell lines (MCF-7 [mammary gland adenocarcinoma], A549 [lung carcinoma], HEK 293 [embryonic kidney cells transformed by Adenovirus 5 DNA], and WERI-Rb-1 [a retinoblastoma line]). Boundary analysis indicated a similar number of subtelomeric CTCF binding sites and a similar range of \(P\)-values for terminal boundaries and ITS associations with peaks (Supplemental Table 9) as were found in ChIP-seq data sets for LCLs, human ES cells, and IMR90 (Supplemental Table 8). As with the individual CTCF data sets for LCLs, ES cells, and IMR90, nonterminal SRE/SRE boundary associations with just CTCF peaks were usually not significant in the cell lines. Cohesin ChIP-seq data sets were not available for most of these cell lines, so we could not determine colocalized CTCF and cohesin binding sites and test their boundary associations. While most of the same CTCF peaks were called near the terminal boundary and the ITSs, visual comparison of peaks showed clear differences in relative levels of peak enrichments between the cell lines (vader.wistar.upenn.edu/humansubtel), as well as some differentially called peaks. These preliminary observations merit follow-up with much larger data sets as well as experimental validation.

Discussion

With this work, we have revised and updated human subtelomere assemblies such that 34 of the 41 genetically distinct chromosome ends extend to the start of terminal repeat tracts (Fig. 1). This represents a significant advance over the previous human subtelomere assemblies (Riethman et al. 2004; Ambrosini et al. 2007). We also provide a multiread mapping pipeline that enables the systematic analysis of distal chromosome regions using short-read sequencing–based methods, leveraging the wealth of public genome-wide data sets available to help understand subtelomere and telomere function. We have also established a public browser (vader.wistar.upenn.edu/humansubtel) that integrates novel aspects of subtelomere sequence organization with short-read sequence–based annotations and displays this information in a manner optimized for understanding potential functional properties associated with the annotations relative to the telomere terminal repeat tract as well as subtelomeric sequence features. As additional annotation is added, we believe it will become an increasingly valuable resource for the telomere and chromosome biology communities.

Table 2. SRE boundary enrichments

<table>
<thead>
<tr>
<th>Cell line, ChIP-seq data set</th>
<th>SRE/SRE enrichment</th>
<th>(P)-value</th>
<th>Terminal enrichment</th>
<th>(P)-value</th>
<th>ITS enrichment</th>
<th>(P)-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>CTCF</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LCL_CTCF_ly</td>
<td>1.200</td>
<td>0.07314714</td>
<td>10.789</td>
<td>5.05593 (\times) 10(^{-19})</td>
<td>2.012</td>
<td>0.00019951</td>
</tr>
<tr>
<td>LCL_CTCF_W_Li</td>
<td>1.192</td>
<td>0.0762587</td>
<td>10.260</td>
<td>1.8887 (\times) 10(^{-18})</td>
<td>2.033</td>
<td>0.00013915</td>
</tr>
<tr>
<td>H1-hESC_CTCF_Be</td>
<td>1.189</td>
<td>0.1219057</td>
<td>13.011</td>
<td>1.39914 (\times) 10(^{-19})</td>
<td>2.308</td>
<td>4.3071 (\times) 10(^{-5})</td>
</tr>
<tr>
<td>H1-hESC_CTCF_My</td>
<td>1.419</td>
<td>0.00707443</td>
<td>15.530</td>
<td>1.15952 (\times) 10(^{-21})</td>
<td>1.771</td>
<td>0.00366826</td>
</tr>
<tr>
<td>IMR90_CTCF_Sn</td>
<td>0.967</td>
<td>0.6399912</td>
<td>8.880</td>
<td>2.33908 (\times) 10(^{-14})</td>
<td>1.725</td>
<td>0.00419184</td>
</tr>
<tr>
<td>Cohesin</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LCL_RAD21_My</td>
<td>1.200</td>
<td>0.3232894</td>
<td>4.247</td>
<td>0.001311649</td>
<td>2.298</td>
<td>0.00136006</td>
</tr>
<tr>
<td>LCL_SM1C_Ii</td>
<td>1.192</td>
<td>0.006035461</td>
<td>15.391</td>
<td>2.66361 (\times) 10(^{-23})</td>
<td>1.794</td>
<td>0.0021072</td>
</tr>
<tr>
<td>H1-hESC_RAD21_My</td>
<td>1.189</td>
<td>0.009682766</td>
<td>8.535</td>
<td>5.39758 (\times) 10(^{-14})</td>
<td>2.369</td>
<td>7.83 (\times) 10(^{-6})</td>
</tr>
<tr>
<td>H1-hESC_RAD21_Sn</td>
<td>1.617</td>
<td>0.000260144</td>
<td>5.980</td>
<td>5.41325 (\times) 10(^{-8})</td>
<td>2.614</td>
<td>0.00038859</td>
</tr>
<tr>
<td>IMR90_RAD21_Sn</td>
<td>0.967</td>
<td>0.2708427</td>
<td>4.407</td>
<td>0.012587796</td>
<td>0.963</td>
<td>0.53538354</td>
</tr>
<tr>
<td>Colocalized CTCF &amp; cohesin</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LCL_CTCF_ly &amp; LCL_RAD21_My</td>
<td>1.478</td>
<td>0.00664919</td>
<td>5.980</td>
<td>0.000158034</td>
<td>3.236</td>
<td>1.9703 (\times) 10(^{-5})</td>
</tr>
<tr>
<td>LCL_CTCF_W_Li &amp; LCL_SM1C_Ii</td>
<td>1.450</td>
<td>0.0309262</td>
<td>15.857</td>
<td>1.10817 (\times) 10(^{-21})</td>
<td>2.587</td>
<td>2.5738 (\times) 10(^{-6})</td>
</tr>
<tr>
<td>H1-hESC_CTCF_Be &amp; H1-hESC_RAD21_My</td>
<td>1.564</td>
<td>0.003089455</td>
<td>7.788</td>
<td>7.27985 (\times) 10(^{-6})</td>
<td>2.837</td>
<td>0.00036953</td>
</tr>
<tr>
<td>H1-hESC_CTCF_My &amp; H1-hESC_RAD21_Sn</td>
<td>1.485</td>
<td>0.040011074</td>
<td>15.305</td>
<td>5.16979 (\times) 10(^{-19})</td>
<td>1.581</td>
<td>0.01862117</td>
</tr>
<tr>
<td>IMR90_CTCF_Sn &amp; IMR90_RAD21_Sn</td>
<td>1.157</td>
<td>0.2708427</td>
<td>4.407</td>
<td>0.012587796</td>
<td>1.926</td>
<td>0.05884247</td>
</tr>
<tr>
<td>Tel repeat</td>
<td>1.545</td>
<td>0.000145563</td>
<td>All</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
</tbody>
</table>

Subtelomere annotation
The updated subtelomere reference assemblies are subject to caveats as are all regions of the reference human genome sequence; they are composed of DNA segments derived from multiple individuals, and for any sequenced clone, only one allele is represented. This means that the depicted reference allele sequences may not completely match that of corresponding subtelomere alleles from other source genomes. Much of the natural variation in human subtelomeres is due to differential placement of SRE regions at specific subsets of subtelomeres (Linardopoulou et al. 2005; Riethman 2008a,b), and this may complicate interpretation.

Figure 4. ChIP-qPCR analysis of subtelomeric DNA protein binding sites predicted by ChIP-seq data set mappings. Candidate sites of CTCF, cohesin, TERF1, and TERF2 binding were analyzed by ChIP-qPCR. Segments of the 6q and 16q (A) and the Xq and 17p (B) subtelomeres are shown, with the coordinates (in bp) shown at the top and the subtelomere paralogy regions indicated on the respective segments. The positions of ITSs are indicated by red rectangles extending from the segments; an ITS with called TERF1 and TERF2 ChIP-seq enrichment peaks is marked with a red asterisk. The positions of colocalized CTCF and cohesin (RAD21) peaks called in LCLs are shown as green dots (if not called in other cell types) and as blue dots (if also called in ES and/IMR90 cells). A diamond beneath a dot indicates a site where no ChIP-seq peak was called when only uniquely mapping reads were considered. Numbered ticks show the positions of primer sets used in the ChIP-qPCR experiments, and correspond to the numbered ChIP-qPCR results shown for CTCF, RAD21, and TERF1 and TERF2 graphed as the percentage of input DNA. The bar graphs represent the average of percentage input (mean ± SD) for each ChIP from three independent ChIP experiments. Ticks numbered 1 and 2 are qPCR assays for DNA immediately adjacent to the telomere, used here as positive controls for TERF1 and TERF2 binding (primer positions 1 and 2) and a positive control for a previously validated subtelomeric CTCF/RAD21 colocalization site (primer position 2).
of ChIP-seq signal strengths at specific high-similarity SRE sites when comparing data sets from nonsogenic source genomes. For example, a CTCF peak predicted by multiread mapping in a high-similarity SRE segment of the reference assembly is expected to have a higher enrichment level in a data set from a genome with more copies of the SRE segment than a data set with fewer copies of the SRE segment. Copy numbers of all known highly similar SRE blocks vary by a factor of two or less in the human population, although most vary by considerably less than twofold (Linardopoulou et al. 2005, 2007); depending on the SRE segment in question, a doubling or halving of an enrichment value at a peak may not be meaningful for a given data set. Prior knowledge of SRE copy number in the respective source genomes would help to mitigate this issue. Even with these limitations, the more complete sequence representation of our assemblies, especially in the distal subtelomere regions, has already permitted novel annotation leading to experimental validation and functional insights into telomere biology (Deng et al. 2012), which we have extended here.

As new technologies capable of adding complete alternative long-range subtelomere haplotypes to the reference assemblies are developed, these sequences will be annotated and incorporated into our browser.

The use of our multiread mapping approach for ChIP-seq short-read data sets had a very large impact on the annotation of candidate binding sites in SRE regions; most candidate CTCF and RAD21 binding sites in SRE regions were missed (between 70% and 90% of called peaks, depending on the data set) when only uniquely mapping reads were considered. This is illustrated dramatically in Figure 4, where all of the sites predicted by the multiread mapping in the SRE regions were missed in the analysis considering only uniquely mapping reads. Comparison of the multiread mapping tracks and the unique read mapping tracks on the bedGraphs in the subtelomere browser for the same experiment often revealed a small unique read peak corresponding to a much larger and robust multiread peak for SRE sites, indicating that some fraction of the reads were mapping uniquely to the site but that the unique enrichment peak was too weak to be called statistically significant. However, as we showed previously (Deng et al. 2012), in SRE regions with very high sequence similarity to paralogs, there was no detectable enrichment in the uniquely mapping data sets.

Because peaks detected using the multiread mapping method represent an average of enrichments over all genomic sites to which the reads map, there is the potential for prediction of false-positive peaks called due to extremely high true binding at one or a few sites, causing called peaks at all of them. This is a limitation of the approach and an important caveat to consider in the interpretation of the results. Short-read-based annotations in SRE regions or, for that matter, any region of the genome, are models. While perhaps revealing valuable insights into subtelomere biology, they ultimately require independent validation. The ChIP-qPCR results of the predicted CTCF and RAD21 peaks shown in Figure 4 provide strong validation of the ChIP-seq binding predictions in SRE regions; however, even here ChIP-qPCR primer sets in very high similarity duplicated regions sometimes cannot distinguish all individual copies (see Supplemental Table 10).

Somewhat to our surprise, we did not find evidence for specific TERF1 or TERF2 binding to ITS sites. Interestingly, however, we found evidence for enrichment of CTCF and cohesin subunit binding adjacent to ITS boundaries, in addition to the binding sites near terminal (TTAGGG)n sites noted previously (Table 2; Deng et al. 2012). When we considered only the CTCF and cohesin subunit peaks that colocalized exactly (see Fig. 1), the significance of association with telomere-adjacent DNA and ITSs typically increased, while the colocalized peak association with SRE/SRE boundaries reached significance for the ES and LCL lines but not for IMR90 (Table 2). Strong cohesin sites colocalizing with CTCF have been implicated in long-range chromosomal interactions (Merkenschlager and Odom 2013), suggesting colocalized cohesin/CTCF sites may mediate DNA looping and long-range DNA interactions as well as regulate transcription (Chien et al. 2011; Lee and Iyer 2012; Merkenschlager and Odom 2013). Even in the potential absence of direct shelterin interactions between ITSs and telomeres, it is possible that CTCF/cohesin interactions between binding sites associated with the terminal boundaries and internal binding sites such as the ITS-associated ones could mediate events impacting telomeres as well as the regulation of subtelomeric gene families. For example, long-range cohesin/CTCF-mediated interactions involving the telomere-adjacent cohesin/CTCF colocalization sites implicated in TERRA regulation (Deng et al. 2012) may provide a means to coordinate the regulated transcription of TERRA from subtelomeric loci, similar in principle to the coordinated regulation of other complex loci and multigene families by cohesin and CTCF (Merkenschlager and Odom 2013). Using our subtelomere browser and bioinformatics pipeline to leverage the rich public resource of additional short-read data sets for further annotation of these regions may point to focused experiments to test this hypothesis and help to tease out candidate functional sequences involved in subtelomere biology.

Methods

Fosmid library screening; fosmid end sequence mapping, gap-filling, and detection of telomeric structural variants; and directed sequencing of distal ends of terminal fosmids

Methods and materials for these experiments are described in text associated with Supplemental Tables 1 through 4 and Supplemental Figures 1 and 2.

Updated subtelomere assemblies

Supplemental Table 5 describes the complete clone-based subtelomere assemblies as well as their relationship to current clone-based tiling path files (TPFs) being used to update the human reference sequence. The hybrid genome was built by tying the updated subtelomere assemblies into hg19 at their connection point. These points were found by using BLAST (Altschul 1997) to align the most centromeric 10 kb of sequence from each subtelomere assembly with hg19 sequence. The BLAST results produced one perfect 10-kb hit in the expected orientation, forward for p arm subtelomeres and reverse for q arm subtelomeres. The positions of these hits were then used to extract the nonsubtelomeric portion of the hybrid genome using BEDTools (Quinlan and Hall 2010). The sequence of each 500-kb subtelomere assembly is provided as a concatenated FASTA file in the Supplemental Material. The joining coordinates for connecting hg19 to the subtelomere assemblies are listed in Supplemental Table 6.

Sequence feature annotation

SRE and SD annotation were carried out as described previously (Ambrosini et al. 2007). Duplkon boundaries were defined as the end positions of duplkon blocks. Boundaries within 40 bp of each other were combined at a position corresponding to the weighted
average of the number of boundaries they incorporate and were
declared a single boundary for analysis purposes. Paralogy
tracks were generated by first comparing the representative blocks iden-
tified by Linardopoulou et al. (2005) with the updated assemblies
and then adding blocks corresponding to new SRE segments
shared in the manner described by Linardopoulou et al. (2005).
Existing Block 19 was broken into two separate blocks based upon
the SRE/1-copy boundary generated by 17q sequence, which was
not available to Linardopoulou et al. (2005). Representative se-
quences for paralogy blocks 19a, 19b, 45, 46, 47, 48, and 49 are
provided as a concatenated FASTA file in Supplemental File S2.

Subtelomere sequence assemblies were analyzed with Repeat-
Masker (Smit et al. 1996 at http://repeatmasker.org) and Tandem
Repeats Finder (Benson 1999). Ensembl transcripts (Flicek et al.
2012) and RefSeq genes (Pruitt et al. 2012) were aligned to sub-
telomerases using Spidey (Wheelan et al. 2001).

Short-read–based annotation pipeline

Data sets analyzed in this study are listed with their specific sample
and control GEO accessions, as well as the specific antibodies used
and their sources, in Supplemental Table 7. The LCL-associated
data sets for CTCF, RAD21, and SMC1 were the same as described
previously (Deng et al. 2012). Additional data sets were down-
loaded as raw data FASTQ files from the ENCODE Project (The
ENCODE Project Consortium 2011) through the UCSC portal. H1-
hESC_CTCF_Be and HMEC_CTCF_Be are from the Bernstein lab-
oratory (GSE29611 series) at the Broad Institute. IMR90_CTCF_Sn,
IMR90_RAD21_Sn, IMR90_POLR2A_Sn, and H1-hESC_RAD21_Sn
correspond to the Snyder laboratory data from Stanford (GSE31477
series). H1-hESC_RAD21_My, H1-hESC_CTCF_My, and H1-hESC-
POLR2A_My correspond to the Myers data from HudsonAlpha
(GSE32465 series).

Reads were aligned to the hybrid genome using BWA 0.6.2
(Li and Durbin 2009), allowing multimapping up to 101 locations
(-n 101). BWA does not prioritize multimapping reads, and alter-
nate mapping locations are not included as reads but instead are
listed in an XA tag. Alternate positions were then expanded from
the XA tag to one mapping position per line. A mapping likelihood
(ml) tag was added as the inverse of the number of mapping lo-
cations. It is still possible to only consider uniquely mapping reads
by analyzing only those reads with an ml tag equal to one. Fragment
length was estimated by cross-correlation implemented in the
SPP ChIP-seq mapping program (Kharchenko et al. 2008).
bedGraph coverage files were created from the mapping positions
by extending read mappings to the estimated fragment size. Frag-
ment coverage for each position was calculated as the sum of ml
values of fragments overlapping that position and then averaged
over a 20-bp sliding window. Adjacent positions were given the
same value if the coverage was within 0.1. To simplify fold change
calculations, values less than one were given a pseudo count to be
equal to one. Fold enrichment tracks were built between control
(Input or IgG) and sample to be used as a signal track, normalizing
the control data set to the size of the sample. Negative values were
used to show stronger signal in the control. A pseudo count of one
was used in locations where there was no mapping for the sample or
control. A smoothing window of 500 bases was used on all control
data sets. Peak calls were made using MACS 2.0.10 using the sample
and control bedGraphs. First, bdcmp –m ppois was called, setting
ppois as the method and calculating P-value tracks. Peaks were
called using bdgpeakcall –1L0 –c 4, setting minimum peak length to
50 and a P-value significance cut of 4 (10^-4) (Zhang et al. 2008).
Overall quality and mapping metrics for the data sets were deter-
mined as previously described (Landt et al. 2012) and are included in
Supplemental Table 7.

Subtelomere browser

The subtelomere browser can be found on a mirror site of the UCSC
Genome Browser maintained by the Wistar Bioinformatics Facility
(vader.wistar.upenn.edu/humansubtel). The entire subtelomere
region of interest is displayed by typing it in the format chrNp:1-
500000 or chrNq:1-500000. The subtelomere browser has similar
navigation and mapped data set selection functionalities as the
UCSC Genome Browser (Kent et al. 2002). The updated subtelomere
assemblies in FASTA format are found in Supplemental File S2 and
can be found on the Riethman laboratory website (http://www.
wistar.org/sites/default/files/protected/htel_1-500K_1_10_12_v4_3_
12fasta.TXT).

Peak/boundary association enrichment calculation

Peak/boundary association enrichments were defined as the ratio
of the number of peaks observed in defined boundary window
regions (across all SRE sequence space) to the expected number of
peaks within these window regions if the total number of peaks in
the SRE sequence space were distributed evenly. Some boundaries
were within the allowable window of each other; in these instances,
a peak was associated with more than one boundary, although no
additional weighting was added to the boundary association of
these peaks. To calculate a P-value, a one-sided binomial test was
performed, using the expected percentage as the probability of
success, the associated number of peaks as the number of successes,
and the total number of peaks in the SRE as the number of trials.
Terminal boundaries and their associated peaks were excluded when
calculating P-values for peak association with ITSs.

ChIP assay

ChIP assays were performed with the protocol provided by Milli-
pore with minor modifications as described previously (Deng et al.
2009). Briefly, LCLs were crosslinked in 1% formaldehyde with
shaking for 15 min, and DNA was sheared to between 200- and
400-bp fragments by sonication with a Diagenode Bioruptor.
Quantification of ChIP DNA at subtelomeric regions was deter-
mimed using quantitative PCR (qPCR) with the ABI 7900 sequence
detection system (Applied Biosystems). qPCR was performed in
triplicates from three independent ChIP experiments, and PCR data
were normalized to input values. Primer sequences used for qPCR
were designed using Primer Express (Applied Biosystems), and listed
in Supplemental Table 10. Each primer set was validated by using
melting curve analysis, in which one major dissociation peak was

TERFI and TERF2 data sets

Publicly available data sets from Simonet et al. 2011 (GSE26005)
were downloaded and analyzed. These are color space reads mapped
on the AB SOLiD System 3.0. The color space reads were mapped
using SHRMP 2.2.3 (Rumble et al. 2009), allowing for reads mapping
up to 104 mapping positions (-o 102). Once mapping positions
were examined, the pipeline followed was the same as other ChIP-
seq data sets. However, cross-correlation analysis failed at finding
a fragment size, so the selected fragment size of 200 bases was used
(Supplemental Table 7). Additional TERFI and TERF2 ChIP-seq data
sets were generated for LCL as described previously (Lu et al. 2012),
using rabbit antibodies to TERFI and TERF2, which were generated
against recombinant protein and affinity purified. The 100-bp Illumina
reads in these data sets were trimmed from both the 3’ and
5’ ends up to the first high-quality base (>=Phred 30). Telomere and
telomere-like simple repeats were identified by RepeatMasker (Smit
et al. 1996).

Peak/boundary association enrichment calculation

Peak/boundary association enrichments were defined as the ratio
of the number of peaks observed in defined boundary window
regions (across all SRE sequence space) to the expected number of
peaks within these window regions if the total number of peaks in
the SRE sequence space were distributed evenly. Some boundaries
were within the allowable window of each other; in these instances,
a peak was associated with more than one boundary, although no
additional weighting was added to the boundary association of
these peaks. To calculate a P-value, a one-sided binomial test was
performed, using the expected percentage as the probability of
success, the associated number of peaks as the number of successes,
and the total number of peaks in the SRE as the number of trials.
Terminal boundaries and their associated peaks were excluded when
calculating P-values for peak association with ITSs.

ChIP assay

ChIP assays were performed with the protocol provided by Milli-
pore with minor modifications as described previously (Deng et al.
2009). Briefly, LCLs were crosslinked in 1% formaldehyde with
shaking for 15 min, and DNA was sheared to between 200- and
400-bp fragments by sonication with a Diagenode Bioruptor.
Quantification of ChIP DNA at subtelomeric regions was deter-
mimed using quantitative PCR (qPCR) with the ABI 7900 sequence
detection system (Applied Biosystems). qPCR was performed in
triplicates from three independent ChIP experiments, and PCR data
were normalized to input values. Primer sequences used for qPCR
were designed using Primer Express (Applied Biosystems), and listed
in Supplemental Table 10. Each primer set was validated by using
melting curve analysis, in which one major dissociation peak was
observed. ChIP DNA at telomeres was assayed by dot blotting with γ\(^{[32P]}\)ATP-labeled probes specific for telomere (4 × TAGGG) or Alu repeats (ggaggtcgtcttgccgacgctgctgga). After hybridization, the blot was developed with a Typhoon 9410 imager (GE Healthcare) and quantified with ImageQuant 5.2 software (Molecular Dynamics). Antibodies used in ChIP assay include rabbit polyclonal antibodies to CTCF (Millipore 07-729) and RAD21 (abcam ab992). Rabbit antibodies to TERF1 and TERF2 were generated against recombinant protein and affinity purified.

**Data access**

DNA sequence for gap-filling clones and fragment components were submitted to the NCBI GenBank (https://www.ncbi.nlm.nih.gov/genbank/) under accession numbers AC215217, AC215219, AC213859, AC213860, AC215218, AC213861, AC215220, AC215221, AC225782, AC215524, AC215522, AC226150, KF477190, KF477189, KF477188, KF477185, KF477187, KF477186, and KF477184 (see Table 1). The TERF1 and TERF2 ChIP-seq data sets generated as part of this study were submitted to the NCBI Genome Expression Omnibus (GEO; https://www.ncbi.nlm.nih.gov/gds) under accession numbers GSM1328844 and GSM1328845. Each of the 500-kb subtelomere reference assemblies are available as a concatenated FASTA file in Supplemental File S1. New SRE paralogy blocks 45–49, 19a, and 19b are available as a concatenated FASTA file in Supplemental File S2. The subtelomere browser link is vader.wistar.upenn.edu/humansubtel.

**Acknowledgments**

We acknowledge contributions from the Wistar Cancer Center Core facilities in Bioinformatics and Genomics, especially Priyankara Wickramasinghe for maintenance and updating of the subtelomere browser in the Wistar Bioinformatics Core facility, and Brett Taylor Wickramasinghe for maintenance and updating of the subtelomere facilities in Bioinformatics and Genomics, especially Priyankara Wickramasinghe. Work done by T.G., C.C.F., L.C., and R.K.W. was supported by R01LM011297. E.E.E. was supported by HG004120 and R21HG007205) and P.L. (RO1CA140652). Z.D. was supported by the Commonwealth Universal Research Enhancement Program, PA University award (N.S.) and by NIH grants to H.R. (R21CA143349 and 3P50CA105133-14). J.C. was supported by an American Heart Association grant (11SDG5330017) and P.L. (RO1CA140652). Z.D. was supported by the Wistar Cancer Center core grant (P30 CA10815) and by NIH grants to H.R. (R21CA143349 and R01LM011297).

**References**


