TELOMERE AND PROXIMAL SEQUENCE ANALYSIS USING HIGH-THROUGHPUT SEQUENCING READS

Nicholas Stong

A DISSERTATION

in

Genomics and Computational Biology

Presented to the Faculties of the University of Pennsylvania

in

Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy

2014

Supervisor of Dissertation

_____

Harold Riethman, Ph.D.
Associate Professor,
Molecular and Cellular Oncogenesis Program,
The Wistar Institute

Graduate Group Chairperson

_____

Li-San Wang, Ph.D
Associate Professor of Pathology and Laboratory Medicine, University of Pennsylvania

Dissertation Committee:
Junhyong Kim, Ph.D.
Endowed Professor, Department of Biology, University of Pennsylvania
Lyle Ungar, Ph.D.
Associate Professor, CIS, University of Pennsylvania
Bradley Johnson, M.D. Ph.D.
Associate Professor of Pathology and Laboratory Medicine, University of Pennsylvania
Hongzhe Li, Ph.D.
Professor of Biostatistics, University of Pennsylvania

TELOMERE AND PROXIMAL SEQUENCE ANALYSIS USING HIGH-THROUGHPUT SEQUENCING READS

Dedicated to my loving wife Laura

ACKNOWLEDGMENT

The Riethman lab was a wonderful environment for me to be exposed to the day to day wet lab work used to study telomere biology and have the computational resrouces to gain experience in bioinformatics. Harold has always been a kind mentor and thought of my personal achievments and development as a scientist. The Lieberman lab provided much of the support I needed to complete the biological story surrounding my computational findings. Particularly Zhong Deng who always had the right questions that would help me frame what I was working on along with helping me to validate many of my computational predictions. Amber Weiner was a fantastic undergrad that was always willing to work hard and learn everything I gave her.

I also need to thank my friends and family who have supported me over the past six years. The friends I've made through my years at Penn have been great companions through challenges and triumphs. My family has always been incredibely supportive of me. Without my parents ongoing encouragement I couldn't have made it to this point. My wife Laura is constantly helping with all aspects of my life and has made my work possible.

ABSTRACT

TELOMERE AND PROXIMAL SEQUENCE ANALYSIS USING HIGH-THROUGHPUT SEQUENCING READS

Nicholas Stong

Harold Riethman, Ph.D.

The telomere is a specialized simple sequence repeat found at the end of all linear chromosomes. It acts as a substrate for telomere binding factors that in coordination with other interacting elements form what is known as the shelterin complex to protect the end of the chromosome from the DNA damage repair machinery. The telomere shortens with each cell division, and once critically short is no longer able to perform this role. Short dysfunctional telomeres result in cellular senescence, apoptosis, or genome instability. Telomere length is regulated by many factors including cis-acting elements in the proximal sequence which is known as the subtelomere. The Riethman lab played a pivotal role in generating the reference sequence of the subtelomere in both the human and mouse genomes, providing an essential resource for this work. Short high throughput sequencing (HTS) reads generated from the simple repeat containing telomere or the segmental duplication rich subtelomere cannot be aligned to a reference genome uniquely. They are filtered and excluded from many HTS analysis methods. A ChIP-Seq analysis pipeline was developed to incorporate these multimapping reads to study DNA-protein interactions in the subtelomere. This pipeline was employed to search for factors regulating the expression TERRA, an essential long non-coding RNA, and to better characterize their transcription start sites. ChIP-seq analysis in the human subtelomere found colocalization of CTCF and Cohesin directly adjacent to the telomere and throughout the subtelomere specific repeats. Follow up functional studies showed this binding regulated TERRA transcription at these sites. Extending these analyses in the mouse genome showed very different patterns of CTCF and cohesin binding, with no evidence of binding at apparent sites of TERRA transcription. Mouse subtelomere sequence analysis showed the co-occurence of two repeats at sites of putative TERRA expression, MurSatRep1 and MMSAT4, one of which was previously shown to

be expressed in lincRNAs. The Telomere Analysis from SEquencing Reads(TASER) pipeline was developed to capture telomere information from HTS data sets and used to investigate telomere changes that occur in prostate cancer. TASER analysis of 53 paired prostate tumor and normal samples revealed an overall decrease in telomere length in tumor samples relative to matched paired normal tissue, especially sequence containing the exact canonical telomere repeat. Multimapping reads contain important information, that when used properly, help elucidate understanding of telomere biology, cancer biology, and genome regulation and stability.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF ILLUSTRATIONS

CHAPTER 1: BACKGROUND

1.1 Telomere Biology

1.1.1 Telomere Structure

In 1938 Herman Muller first hypothesized the presence of a terminal gene with the special function of sealing the end of a linear chromosome, which he termed the telomere. Drawing on years of work trying to understand the mechanisms of chromosome rearrangement after X-ray irradiation in *Drosophila*, he reasoned that if chromosomal breakage were to occur before rearrangement initiated, then it would be possible for the broken fragments to not find each other before the acentric fragment was lost in mitosis. However having never observed any such fragments he concluded that the telomere was required for chromosome stability[1]. It was also observed in *Zea Mays* that chromosome breaks created by the divergent pulling of centromeres in a dicentric chromosome results in unstable ends, which go through repeated breakage-fusion-bridge cycles[2]. The chromosome ends resulting from breaks are highly unstable and have a propensity to fuse leading to chromosomal rearrangements. Telomeres found at natural chromosome ends are required for genome stability.

It was several decades before the structure of the telomere began to be unraveled with the advent of DNA sequencing. Elizabeth Blackburn, a former graduate student of Fred Sanger, sequenced the telomeres of rDNA in *Tetrahymena thermophilia* by use of restriction endonuclease digestion and found a tandemly repeated nucleotide sequence 5'$(CCCCAA)_n$ 3' repeated 20 to 70 times[3]. The telomeres of other ciliates with a large polyploid macronucleus were also found to contain G rich tandem repeats[4,5]. Eventually it was found that the human genome[6,7] and all vertebrates[8] have a telomere repeat of TTAGGG. The number of tandem repeats and therefore the total telomere length varies from chromosome to chromosome and between individuals. The length of telomeres in the human genome ranges from 3-20kb[6,9,10],

and is considerably longer in many other verterbrates, including in *Mus musculus*[9]. A universal feature of telomere sequences is a short single stranded overhang of the 3' G rich strand[5,11]. In the human genome this overhang is on average 200bp[12] This genomic sequence is responsible for differentiating the ends of linear chromosomes from ends resulting from internal double strand DNA breaks.

Even before the sequence structure of the telomere was known, it was abstracted that the ends of the linear chromosomes could not be replicated completely. The mechanics of the semiconservative DNA replication are unable to fully duplicate template stands of DNA. DNA polymerase requires an RNA primer to begin DNA synthesis, which is later removed by RNAse. This combined with the directionality of DNA polymerase from 5' to 3' means the terminal 3' end of both template strands cannot be duplicated [13–15]. Alexey Olovnikov and Jim Watson both independently described this as the end replication problem. This results in the shortening of the resulting chromosome by at least 12-15bp, if the RNA primer is placed at the last base of the 3' end of the template strand. Olovnikov went further to argue that marginotomy of DNA (the shortening of replicated DNA in respect to the template), and the total loss of telogenes, leads to the elimination of aged cells, and speculated that it was responsible for the depletion of cell populations in the body and therefore a primary cause in disorders of aging[14]. The end replication problem leads to a 3' overhang characteristic of telomere ends however only at the shortened end of the chromosome. In addition the shortening due to the end replication problem is lesser than the observed average 200bp overhang in the human genome[12]. This indicates that telomeric single stand 3' overhangs are maintained through exonucleic degradation[16,17]. Telomere shortening occurs with DNA replication due to both the end replication and post replicative processing. This has been observed after cell divisions in both *in vitro* and *in vivo*[18,19].

The telomere is able to form a structure that protects it from being recognized as a double strand break by the DNA damage repair machinery through both the physical conformation of the telomere sequence and an association with a specialized protein complex.

The G rich tandem repeat of the telomere matches the sequence motif necessary to form G quadruplexes, a stable structure in which guanine residues form hydrogen bonds in a helical structure of one strand of DNA[20]. Given the single stand overhang at the telomere this conformation is energetically favorable[21]. There is extensive evidence for these structures occurring in vivo[22,23]. The single strand overhang is also able to invade the double stranded region of the telomere, forming a small single stranded displacement loop (D-loop) where the overhang disrupts the DNA duplex, and a large lasso-like loop (T-loop) at the chromosome end[24]. The formation of this loop depends on an interaction with telomere associated proteins[25].

1.1.2 Telomere DNA Interactions

A number of proteins have been shown to interact directly and in complex with the telomere sequence. Two related proteins, telomeric repeat binding factor (TRF) 1 and 2 were found to bind the double stranded DNA fragments containing the telomere repeat[26–29]. TRF 1 and 2 share a C' terminal Myb related motif called the teleobox which is capable of recognizing the telomere repeat and differ in their N' terminal dimerization domains[28]. A third protein able to directly interact with the telomere is protection of telomeres 1 (POT1), which binds directly to single stranded telomeric DNA[30,31]. hPOT1 is able to bind to both the single stranded 3' telomeric overhang, and the single stranded DNA displaced in the D-loop[32]. TRF2 and POT1 are required to maintain the structure of the T-loop[25,33]. Other telomere associated factors were discovered through their association with these telomere binding proteins, TIN2 interacting with TRF1[34] and Rap1 interacting with TRF2[35]. TPP1 was found to interact with both TIN2[36] and POT1[37,38]. These 6 proteins are found together in nuclear fractions[39] and together form what is known as the shelterin complex[40] which is required for telomere end capping. A separate complex that associates with the telomere is known as the CST complex, made up of conserved telomere protection component 1 (CTC1), suppressor of CDC thirteen 1 (STN1), and telomeric pathway with STN1 (TEN1)[41,42]. The CST complex interacts with the single stranded telomere overhang, however unlike in yeast where these genes were named, the

CST complex in human cells is required for DNA replication, restarting stalled replication forks[43,44].

While shelterin protects the telomere from being recognized as a double strand break normal telomere shortening occurs with each cell division. In addition accelerated telomere shortening can occur due to oxidative stress[45] or polymerase pausing at sites of single stranded DNA damage leading to premature termination of DNA replication in the telomere[46]. To overcome these mechanisms of shortening the cell is able to extend telomeres. Telomerase is a ribonucleoprotein that like the telomere sequence itself was first identified in *Tetrahymena*. Telomerase has terminal transferase activity that is able to directly add the telomeric repeat on to existing telomeres[47]. Telomerase is able to do this because it has an RNA component which contains the priming sequence for the addition of the whole telomere repeat on to the chromosome end[48]. Telomerase is also responsible for lengthening telomeres in the human genome[49]. The human RNA component of telomerase, hTR, contains an 11 nucleotide template region, (5'-CUAACCCUAAC), which is complementary to the human telomere sequence[50]. The protein component of telomerase is catalytically active and drives the reverse transcriptase activity of telomerase which adds the telomeric sequence[51,52]. This human telomerase reverse transcriptase (hTERT) subunit is the limiting factor in telomere lengthening. While hTR is expressed in all human tissues[53], hTERT is not expressed significantly in somatic cells[54,55].

In cells with insufficient or negligible telomerase expression telomere length decreases with each cell division. Before the discovery of telomeres it was already known that human fibroblasts have a limited replicative capacity[56] that is dependent on the number of population doublings[57]. It was shown early on that telomeres play a key role in senescence, as replicative capacity is highly predictable from telomere length[58]. Once a telomere is critically short it can no longer associate with enough shelterin components to form the structures necessary to protect the end of the chromosome[59]. In a normal cellular context this induces replicative senescence, preventing the cell from dividing further[60]. Senescent human cells have a DNA damage

response including increased p53 activity[61]. In mammalian cells this response to DNA damage is mediated through two protein kinases ATM and ATR. ATM recognizes double strand breaks and ATR recognizes a single strand from a resected double strand break[62]. As part of the damage response ATM and ATR phosphorylate histone H2AX (known as gamma-H2AX)[63,64], promoting the accumulation of damage response factors 53BP1, MDC1, and the MRN complex. This accumulation of damage response elements can be found at individual critically short telomeres[65–67]. ATM and ATR also phosphorylate Chk1 and Chk2 respectively which can cause G1 and G2 phase arrest and are involved in the activation of p53 which inhibits cell cycle progression through downstream activation of p21[68]. In certain cell types and conditions this same p53 mediated signaling cascade can also lead to apotosis[65,69].

In cells with shortening telomeres where cell cycle checkpoints are defective the repair machinery will act on telomeres to resolve the perceived damage through nonhomologous end joining (NHEJ) or homologous recombination. The NHEJ factors DNA ligase IV and Ku70/80 repair dysfunctional telomeres by creating end-to-end fusions[70–72]. It is also possible in the absence of the repressive presence of Ku70/80 for alternative micro homology mediated NHEJ to occur through PARP1 and ligase III[73,74]. This is readily seen in mouse cells when TRF2 is deleted in a p53 null background causing 30-50% of the telomeres to fuse, creating long trains of chromosomes[75]. The fusion of two chromosome creates a dicentric chromosome which leads to genome instability in a cycling cell. During mitosis the dicentric chromosome can be pulled apart creating a bridge between daughter cells. This bridge can itself be lethal if it compromises the formation of a cellular or nuclear membrane. As the fused chromosome is pulled in different directions it is likely to break, leaving broken ends which are then repaired. This leads to either random stable translocations where daughter cells can gain or lose genetic information, or produce another dicentric chromosome which will go through another breakage-fusion-bridge cycle[76–78].

Since the telomere contains long stretches of identical sequence it is possible for aberrant homologous recombination (HR) to occur. HR can occur where the end of the telomere

is interacting to form a t-loop, resulting in a truncated telomere and leaving the cleaved t-loop as telomere circle (t-circle). The end of the telomere and internal double stranded telomere can form a Holliday junction which is cleaved. The formation of t-circles is dependent on XRCC3 and NBS1, consistent with known mechanisms of HR[79,80]. HR can also occur between nascent chromosomes termed telomere sister chromatid exchange (T-SCE). This unequal exchange of telomere sequence results in one telomere being elongated at the expense of a shortened telomere. Some cells in crisis recover by activating a telomerase independent lengthening mechanism known as alternate lengthening of telomeres (ALT)[81,82]. ALT cells have increased rates of T-SCE[83]. ALT cells have high levels of t-circles and heterogeneous telomere lengths, including some very long telomeres, consistent with a HR mechanism[80,84].

1.1.3 Telomeres and Disease

Dysfunctional shortened telomeres have a drastic effect on genome integrity and as such play an important role in a number of diseases. The most drastic effects are those in which a monogenic lesion has a direct effect on telomere length. The first of these disorders to be discovered was dyskeratosis congenita (DC). Family studies originally found a mutation in a gene which was subsequently named dyskerin (DKC1)[85], a small nucleolar protein that binds and stabilizes RNA such as TR. All patients with DC were found to have short telomeres[86]. As affected families exhibited differing modes of inheritance alternative mutations were found in TR[87,88]. Most cases of DC are due to mutations in DKC1, TR, TERT[89], and the shelterin component TIN2[90] however cases have been reported with mutations in four additional telomerase and telomere associated genes[91]. Patients have defects in highly regenerative tissues, such as skin and bone marrow. In most cases organ failure occurs first in the bone marrow leading to aplastic anemia[92]. The related and more severe diseases Hoyeraal-Hreidarsson and Revesz syndrome are due to the same deficiencies in telomere maintenance with a more severe telomere shortening and phenotype[92]. Patients with coats plus syndrome were found to have biallelic mutations in the CST complex gene CTC1[93]. These patients also have short telomere length[94]. Adult onset disease can also be attributed to telomerase

deficiencies caused by mutations in TR and TERT. Mutations are found in 8-15% of familial and in 1-3% of sporadic idiopathic pulmonary fibrosis (IPF)[95,96], 3-5% of aplastic anemia[97–99], and some familial cases of liver cirrhosis[100]. These adult onset conditions are also complications that can occur in patients with DC [91].

These conditions represent a spectrum of disease in which the most severe cases have the shortest telomeres[101]. As expected this results in deterioration or in the manifestation of symptoms over time as telomere length continues to erode. Mutations that occur in TR or TERT are inherited in an autosomal dominant manner which is due to telomerase haploinsufficieny[102]. As telomerase is unable to sufficiently extend telomeres in germ cells affected families display genetic anticipation in which subsequent generations have more severe disease[89]. This results in the spectrum of disease occurring in one family, with earlier generations suffering from IPF later in life, and successive generations having complications of bone marrow failure occur at a younger age, and dykeratosis congentina in further generations[103]. Genetic anticipation is also seen in telomerase knockout mouse models. In founder mice which have long (~50kb) telomeres severe defects in highly proliferative tissues only occur in the 5^{th} and 6^{th} generations[104–107].

As telomeres shorten with each cell division it has long been hypothesized that they act as a mitotic clock, and are responsible for the effects of aging[108]. Age related telomere loss has been shown in diverse tissue types[109,110]. In stem cell pools it is apparent that this shortening occurs despite expression of telomerase, and leads to stem cell exhaustion[111,112]. This loss of proliferative capacity has been linked to disease[112,113]. Telomerase knockout mice have a premature aging syndrome causing hair greying and loss, delayed wound healing, and short lifespan[106,114]. Short telomere lengths in endothelial progenitor cells[115] and leukocytes[116] are risk factors for coronary heart disease. Short telomere lengths in leukocytes also correlate with coratid artery thickening[117] and risk of myocardial infarction[118,119]. Cells with shortening telomeres have been shown to secrete factors that represent biomarkers of aging and disease[120], independent of factors secreted by senescent cells[121] which also accumulate

due to telomere shortening and other factors leading to senescence. In a model system of telomere shortening, the telomerase knockout mouse, continued replicative capacity granted through p53 mutation abates the telomere dysfunction phenotype, restoring reproductive capacity and allowing later generations of telomerase deficient mice[122]. However the loss of proliferative capacity is a suppressor of carcinogenesis.

Overcoming the effects of telomere shortening allows for limitless replicative potential of a cell, which occurs in cell immortalization and in developing neoplasms. In a cell population that is not dividing as a result of telomere shortening mutation of p53 and RB by viral oncoproteins or other mechanisms allows further cell division resulting in further telomere erosion, genomic instability and likely cell death, a state known as crisis[123,124]. Cells that escape crisis have an activated mechanism to extend telomeres, usually through activation of telomerase. Over 90% of cancers and immortalized cell lines express active telomerase[125,126]. In those immortalized cell lines and cancer cells that do not have activated telomerase telomere lengthening is achieved by the ALT mechanism[81,82,127]. Enabling a mechanism of telomere elongation is necessary for a developing cancer, as it is the only way to acquire limitless replicative potential, one of the six hallmarks of cancer[128]. Telomere lengthening alleviates genomic instability, resulting in near elimination of chromosome end-to-end fusions[129] and FISH signal free ends[130] observed in crisis cells. Immortalized human fibroblasts that escape crisis have a shorter average telomere length than cells in crisis[129]. Telomerase most efficiently extends critically short telomeres[131] which leads to telomere length stabilization, not necessarily an average increase, resolving the telomere dysfunction caused by the an individual critically short telomere[132]. In cancer cells telomerase elongation is distributed across telomeres adding 50-100 bp to maintain telomere length[133].

Carcinogenesis is a multistep process in which cells sequentially acquire mutations which contribute to the ability of resulting cell population to continue its uncontrolled proliferation[128,134,135]. Genomic instability can accelerate this process by increasing the number of mutational events leading to an increase in proliferatively beneficial mutations[136]. In

the context of mutational deficiencies in DNA damage response elements, cells experiencing telomere dysfunction are able to avoid replicative senescence or apoptosis[122]. A model of this situation exists in telomerase knockout mice compounded with a homozygous or heterozygous p53 knockout. In late generation mice with telomere dysfunction this continued cellular division creates bridge-fusion-breakage cycles leading to abnormal karyotypes. These mice have accelerated rates of tumor formation, in different types of tissues. In the heterozygous p53 mutant mice carcinomas, mainly skin, breast, and gastrointestinal tract, are the main tumor type[137]. Analysis of these tumors shows non-reciprocal translocations between non-homologous chromosomes, which can lead to copy number changes[138], and are a major cause of oncogene duplication and tumor suppressor deletion[139,140]. There is also evidence of non-reciprocal translocation in human fibroblasts in response to telomere damage due to TRF2 depletion leading to NHEJ[141].

There is extensive evidence of telomere shortening in human cancers. Early measurements such as telomere restriction fragment length, which uses southern blot to find an average bulk telomere length of a sample on a gel, showed breast and colon tumor samples had shorter telomere lengths than adjacent normal tissue[142–144]. More accurate measurement techniques have been developed through the use fluorescent in situ hybridization. Using this technique PNA probe fluorescence is measured from individual telomere ends in an individual cell. Short telomeres can be measured in invasive breast, prostate, and pancreatic cancers[145–148]. Short telomeres can also be found in a majority of pre-invasive cancerous growths in bladder, cervix, colon, esophagus, and oral cavity, indicating that telomere shortening occurs in the early steps of carcinoma development[149]. Telomere length can also be measured using less laborious PCR based methods. Average telomere length can be assayed by a quantitative PCR method which compares the amount of telomere sequence signal as measured by qPCR to the amount of signal generated from a single copy locus in the genome[150]. Individual telomeres can be directly PCR amplified by the use of the single telomere length assay (STELA),

however this is limited to telomeres on chromosomes with a unique subtelomeric priming site[151].

1.1.4 Subtelomere Structure

Despite the crucial role telomeres play in genome stability and their contribution to cancer and disease, telomere length varies widely in the human population. qFISH measurements of individual telomeres have shown that telomere length is heterogeneous[152,153], even between homologous chromosomes[154]. However the length profile is largely consistent in discrete tissues of an individual, and the profile is predominantly heritable[155]. Telomere length profiles are set in the zygote and maintained throughout life[156]. This telomere length regulation is controlled in part by cis-acting factors in the telomere adjacent sequence[157–159]. This directly adjacent sequence, the subtelomere, contains the telomere associated repeat (TAR) repeats which were described early on as being telomere adjacent in the human genome[160,161]. Parts of the TAR1 repeat are found within 2kb of nearly all telomeres, and similar sequences are also found in the pericentromere[162]. While the TAR repeats are found directly adjacent to the telomere, a larger proximal region, the first 500kb adjacent to each telomere makes up the subtelomere sequence. In the human genome the subtelomere is a hotspot of recombination, with sister chromatid exchange enriched 160 fold in the terminal 100kb [163,164]. As such the subtelomere contains a high degree of segmental duplication. These duplicated sequences are termed subtelomere duplicon blocks. Patterns of duplicated blocks in the first 25kb of the subtelomere are shared between related subtelomeres, defining six subterminal duplicon families (A-F). There are seven distinguishable single copy subtelomeres (7q, 8q, 11q, 12q, 14q, 18q, XpYp), and the rest are identifiable to their pertaining family[162].

The subtelomere repeat elements (SRE) are segmental duplications of sequence from other subtelomeres found on separate chromosome ends. SRE sequence makes up 25% of the subtelomere and 80% of the most distal 100kb[165]. There is also extensive segmental duplication (SD) of sequences copied from loci in the internal genome. The source of these duplicated sequences tend to be clustered around specific sites in the genome such as

pericentromic regions, and the ancestral chromosome fusion on chromosome two[166]. The segmental duplications in the subtelomere are larger and more abundant than elsewhere in the genome[167]. The subtelomere is also enriched in internal telomere sequence (ITS) sites. These are (TTAGGG)n-like sequences that are found through the genome, however they are enriched 25 fold in the subtelomere and tend to be longer and more similar to the perfect canonical telomere repeat[165]. In the yeast genome these ITS sites play a role in subtelomere transcriptional regulation and recombination, especially in the context of ALT telomere maintenance[168,169]. A similar role for these sequences has been hypothesized in the human genome[170,171]. The subtelomere also varies in the human population, it was one of the first genomic regions to be identified to contain copy number variants[172,173]. Variant combinations of subtelomere duplicons define diverse subtelomere alleles which can have an effect on subtelomere regulation and transcription[174].

1.1.5 Subtelomere Transcripts

The telomere and subtelomere were once thought to be maintained in a repressive heterochromatic state, as is the case in drosophila[175] and budding yeast[176] due to what is known as the telomere position effect; however studies exploring this effect in mammalian genomes have been inconclusive. The subtelomere contains many transcripts including throughout the SRE. While some appear to be noncoding or pseudogene copies, some contain open reading frames that can encode proteins[162]. An important subtelomere transcript in the human genome is a subclass of the Wiscott-Aldrich Syndrome Protein family, Wiscott-Aldrich Syndrome Protein and Scar homolog (WASH). The WASH gene family is made up of different duplicated isoforms throughout the subtelomere, most of which end within 5kb of the telomere tract. The WASH gene is an ancient conserved protein whose function is as an actin polymerization regulation factor and is an essential gene in Drosophila[177]. The subtelomere also contains large families of genes such as the immunoglobulin heavy chain genes, olfactory receptor genes, and zinc finger genes. A large set of gene families within the subtelomere is a feature found in many eukaryotic genomes, in combination with the fact that the subtelomere is a

hot spot of recombination; it may be a mechanism by which new genes are generated[167,178]. The subtelomere has also been shown to contain the transcription start sites of TERRA.

Telomeric repeat containing RNA (TERRA) is a long non-coding RNA that is transcribed from the subtelomere through the telomeric sequence. It is highly conserved and has been detected in *Homo Sapiens*, *Mus musculus*, *Danio rerio*, plants, and yeasts[179,180]. TERRA transcripts contain both subtelomeric sequences and the canonical telomere repeat, but are heterogeneous in length, ranging from 100bp to 9kb[181,182]. Inhibition of RNA Polymerase II (RNAP2) abolishes TERRA transcription in human U2OS cells[182]. TERRA transcripts are 5' capped by 7-methylguanosine[183], however only 7% of transcripts are polyadenylated[181,182]. Subtelomere fragments that were previously isolated from 10q and XqYq[161] were used to show a CpG island (61-29-37 repeats) within the SRE sequence acted as a promoter sequence and was capable of transcribing a reporter gene placed between the 61-29-37 repeat and the telomere[184]. TERRA localizes to the nucleus where it forms discrete foci at telomeres[181,182]. TERRA associates with shelterin components TRF1 and TRF2 and heterochromatin found at telomeres[185,186]. In yeast TERRA overexpression has been shown to lead to shortening of the transcribed telomere through exonucleic activity[187], or DNA replication dependent loss[188]. TERRA has been shown to act in coordination with hnRNP A1 to help loading of POT1 on single stranded telomere overhangs[189]. Knockdown of both TERRA and hnRNP A1 lead to significant increases in telomere dysfunction induced foci (TIF)[185,190,191]. TERRA expression is deregulated in human cancers, with some studies showing cancer cell lines and certain cancer types have decreased TERRA expression[192,193], and others showing increased TERRA expression overall, with loss of expression from certain subtelomeres[194]. TERRA transcription is tightly regulated in healthy cells.

1.2 Cohesin and CTCF

Cohesin is a protein complex that is required for cohesion between sister chromatids. Cohesin has three main subunits, SMC1, SMC3, and RAD21, which form a ring structure that

holds distant or disjointed regions of the genome within the ring[195,196]. Cohesin also associated with accessory proteins SA 1, 2, and 3[197]. There is evidence that cohesin has roles beyond sister chromatid cohesion, and is involved in transcriptional regulation. In yeast SMC mutants have disrupted insulator function[198] and in *Xenopus* and human cells cohesin is associated with chromatin at points in the cell cycle with no sister chromatid cohesion[199–201], and cohesin subunits are expressed in post-mitotic murine neurons where there will be no more sister chromatid cohesion[202,203]. Cohesin has been found to frequently colocalize with the zinc finger domain protein CCCTC binding factor (CTCF) in the mouse and human genomes, and these sites of colocalization are enriched in regions within 2kb of genes[202,204–206].

CTCF was originally found as a protein bound to the promoter region of the *MYC* oncogene[207]. CTCF has been identified as a transcriptional activator and repressor, as it has been found at sites between transcriptionally active and inactive chromatin[208–210]. CTCF binding is inhibited by CpG island methylation[211], and conversely CTCF binding inhibits denovo CpG methylation[212]. It has been postulated that CTCF is the main factor controlling spatial positioning of genomic segments, grouping disparate segments in condensing or decondensing nuclear territories[213]. Chromsome confirmation capture (3C) experiments have shown long range interactions play an important role in the expression of other repetitive sequences such as olfactory genes[214,215] and CTCF was shown to be involved in regulating gene expression through paternal imprinting at the IGF2/H19 locus by controlling long range interactions[216–218].

1.3 High Throughput Sequencing

1.3.1 Technology

The advent of high throughput sequencing (HTS) has transformed how we are able to study the genome, allowing in depth observation of variation, expression, regulatory function and epigenetics at a global scale. Second generation HTS has relied on optical detection of fluorescently labeled nucleotides or luminescence from millions of individual DNA fragments simultaneously. These small DNA fragments are generated from random fragmentation of the

genome either physically through sonication or enzymatically through restriction digests. In order to generate a large enough signal suitable for detection identical copies of individual DNA fragments are generated and fixed at a single physical site that can be imaged. To accomplish the amplification of these individual fragments universal priming sites are first ligated to the fragments before they are PCR amplified simultaneously. The details of the PCR reaction conditions and how the fragments are fixed differentiate the existing technologies.

The first HTS method that was developed was based on pyrosequencing, which was commercialized by 454. This technology employs emulsion PCR (emPCR) to amplify individual fragments. The fragments are captured by beads containing the complement of the annealed universal primer under conditions such that you would expect at most one fragment per bead. The PCR reaction can then amplify each fragment in an individual microreactor, coating the bead with copies[219]. Pyrosequencing takes advantage of the release of inorganic phosphate during DNA synthesis. When nucleotides are added to a nascent strand of DNA they release an inorganic phosphate which can be measured by its conversion to visible light through enzymatic reactions [220,221]. Beads are sequestered in wells on a chip and washed with consecutive washes of dNTPs. Fluorescent signals are captured, with homopolymers generating an increased signal[222]. As pyrsoquencing does not rely on recurrently terminating DNA synthesis it is able to generate the longest reads currently widely available.

Life technologies platform, SOLiD, also employs the use emPCR, however the sequencing reaction does not rely on the measurement of bases incorporated by DNA polymerase, instead it measures the ligation of dye labeled complementary probes[223]. Once probes have been ligated the fluorescent signals are measured, the probes are cleaved from the fragment and new probes that interrogate the next base position are ligated. The SOLiD platform has the advantage of using dinucleotide probes, therefore interrogating each base position twice[224].

The most popular HTS platform is technology was developed by Solexa which was later acquired by Illumina. Instead of binding fragments to a bead as in ePCR, fragments are bound to primers that are covalently attached to a glass slide. These fragments are then are then amplified directly on the slide by bridge amplification, generating clusters of amplified DNA distributed on the chip. Illumina sequencing relies on reversible terminator chemistry, permitting the addition of one fluorescently labeled nucleotide in each cycle. After unincorporated nucleotides are washed away, fluorescence is measured through imaging, and the terminating group is chemically cleaved, allowing for the next cycle of nucleotide addition[225].

Further advances in HTS are allowing for direct sequencing of single molecules to eliminate the noise introduced by amplification (PacBio, Nanopore), and allowing for nonoptical measurements of base interrogation to allow for higher throughput (IonTorrent, Nanopore). The second generation HTS technologies generate reads between 36 and 250 bases in length. These reads are too short and biased to find stretches of overlapping fragments for de-novo assembly of complex genomes, instead these reads are aligned to the existing reference sequence.

1.3.2 Alignment

Alignment of sequences is a long standing problem in bioinformatics to enable to comparison of similar sequences throughout and between genomes. The algorithms developed for fast alignment depend on a breakdown of a reference sequence into a specialized data structure, creating an index of the reference. The most widely used long sequence aligners starting with BLAST use a hash table index[226]. This index maps the position of each k-mer sequence in the reference, which can then be used to find the positions of seed sequences which are then extended and joined and refined by a Smith-Waterman alignment[227,228]. Short read aligners have implemented different strategies to improve the speed and accuracy of aligning short query sequences to a large reference genome of the same species. A spaced seed, with mismatches within the seed, is able to improve the sensitivity of hits[229]. The first Illumina alignment tool ELAND, SOAP, SeqMap, and MAQ use variations of this strategy[230–232].

Another strategy to allow both mismatches and gaps within the seed is to use q-gram filtering, which is implemented in SHRiMP[233]. Building a hash table of an entire reference genome takes a significant amount of space and holding it in memory for searching can be taxing on computational resources. A different strategy for aligning sequences relies on suffix trees. These have the advantage of being able to be represented by more efficient data structures, a prefix array[234], and further compressed in structures that are based on the Burrows-Wheeler Transform[235], a FM-index[236]. Using an FM-index the human genome reference can be built to take up 2-8Gb of memory. The most popular short read alignment algorithms, bowtie, SOAP2, and BWA use an FM-index strategy[237–239]. All these mapping strategies allow for mismatches and gaps in the final alignment to account for sequencing errors, and variation and mutation found in sample sequences compared to the reference. However a major issue that arises in aligning these sequences is the presence of repetitive sequence in the reference genomes, for example nearly half of the human genome is repetitive sequence[240,241]. Reads generated from these parts of the genome that are completely identical at the length of the read generated cannot be mapped unambiguously. The standard analysis pipelines do not consider these multimapping reads and instead focus only those reads which map uniquely. This limits the application of the data sets in analyzing the telomere and subtelomere.

1.3.3 Applications

Whole genome sequencing (WGS) has allowed for in depth observations of the variants and mutations that exist in the human population. Many software tools have been developed to call single nucleotide polymorphisms (SNP)[242], small indels[243], and copy number variants (CNV)[244]. This has allowed us to find rare variants and denovo mutations, and study rare Mendelian disorders with small numbers of affected individuals. It has also allowed for the study of cancer genomics and acquired mutations that are common to certain cancers, and the changes that occur in tumors over time as subclonal populations of cells expand and outgrow others[245]. This allows us to track the mutations driving the growth of the cancer, identifying potential targets for therapy.

RNA-seq allows analysis of whole transcriptomes by sequencing of cDNA made from reverse transcription of poly(A) selected, or ribosome repeat depleted RNA[246]. The analysis of these datasets is similar to WGS, except that reads must either be mapped to the known transcriptome, or allow for large gaps within the genome mapping to account for transcript splicing. Gene expression levels are then estimated by calculating the number of reads per kilobase of transcript per million reads. The problem of multi-mapping reads is exacerbated in RNA-seq data as there are both gene families made up of similar psuedogenes, and a large degree of transcript heterogeneity in the form of alternatively transcribed and spliced isoforms of genes. This problem has been addressed in most implementations of RNA-seq analysis, and a common strategy is to assign partial reads to all of its possible mapping positions. ERANGE assigns location weighting based on the expression level of adjacent unique regions[246]. Methods that quantify individual isoform levels, such as RSEM, go beyond this to assign partial reads to different isoforms representing the same mapping position[247,248].

ChIP-seq is a method to measure protein-DNA interaction and epigenetic marks genome wide. In Chromatin immunoprecipitation (ChIP) experiments DNA binding proteins are cross linked to DNA *in vivo* by treating cells with formaldehyde before the genome is fragmented. Antibodies targeted to the selected protein of interest, or directly targeting chromatin marks, immunoprecipitate the DNA-protein complex. Once the crosslinks are reversed the resulting DNA fragments are used to build a HTS library that is enriched for the targeted protein or histone mark. The resulting sequenced reads are aligned to the reference genome in the same way as WGS datasets. An important caveat of ChIP-seq dataset is that they are only enriched for the chosen protein or mark. As such they are compared to an input or ChIP (IGG) control. Regions statistically enriched for the sample are shown to interact with the selected protein. All the standard ChIP-seq analysis tools only consider uniquely mapping reads, ignoring the repetitive parts of the genome.

1.4 Outline of Dissertation

In chapter 2 I present the initial mapping and analysis of ChIP-seq datasets to the 15kb of sequence proximal to the telomere in the human genome, which was my key contribution to the Deng et al 2012 paper I co-authored. CTCF and cohesin colocalization within 2kb of the telomere was found on the majority of human subtelomeres. The results of follow-up functional studies carried out by Zhong Deng in the Lieberman lab for this paper are summarized briefly, and demonstrate the importance of CTCF and cohesin binding for TERRA transcription and telomere stability. In chapter 3 I show the results of extending this analysis to the entire 500kb human subtelomere. A complete multi-mapping ChIP-seq analysis pipeline was developed. ChIP-seq results for a number of datasets were generated and are all available on a Wistar mirror of the UCSC genome browser focused on the subtelomere.

In chapter 4 I extended the ChIP-seq pipeline for use on murine samples. I describe the sequence characteristics of the recently completed mouse subtelomere sequence, and the differences between the human and mouse subtelomere repeat structure. A number of ChIP-seq datasets were analyzed to provide an online resource similar to the human subtelomere browser. Additionally RNA-seq datasets were analyzed to look for evidence of TERRA transcription. Differences in human and mouse TERRA regulation and transcription are explored.

In chapter 5 I present Telomere Analysis from SEquencing Reads (TASER), a pipeline to capture telomere sequence information from HTS data sets. This pipeline was used to analyze 53 paired tumor normal samples from the prostate cancer genome sequencing project. The results of the analysis were used to show the feasibility of the strategy and show a trend in the telomere changes that occur in prostate cancer. The telomere status was then used to classify

samples as tumor or normal. In chapter 6 I conclude with an overall discussion of the significance of this thesis work.

CHAPTER 2: PROXIMAL 15KB ANALYSIS

2.1 Introduction

The ends of eukaryotic chromosomes form specialized chromatin structures that are essential for chromosome stability and genome maintenance[1]. The terminal TTAGGG repeats of mammalian telomeres bind to a set of proteins that are nucleated by the DNA-binding proteins TRF1, TRF2, and Pot1, and are collectively referred to as shelterin[2,3] or telosome[4,5]. These terminal repeat binding factors regulate telomere length homoeostasis and DNA damage repair processing at the chromosome termini[6]. Loss or damage of the terminal repeats can initiate a DNA damage response and trigger cellular replicative senescence[7,8]. DNA damage and senescence can also be elicited by mutation or depletion of telomere repeat binding proteins[9]. Dynamic remodelling of telomere repeat factors and telomere DNA conformation is also required for normal telomere length regulation and telomerase accessibility[10–13].

In addition to shelterin and telomerase, telomere maintenance depends on the proper assembly and regulation of telomeric chromatin[5,14,15]. Traditionally, telomeres have been thought of as highly heterochromatic structures associated with condensed chromatin and transcriptional silencing[14,16–18]. More recent studies have revealed that many eukaryotic telomeres, including human and yeast, can be transcribed, indicating that telomeric silencing is incomplete and telomere chromatin is dynamic[19–23]. The chromatin structure of telomeres is further complicated by the variations in the subtelomeric DNA structures, suggesting that telomeric heterochromatin structure and regulation may vary among different chromosomes[24–26]. In budding yeast, telomeric silencing is mediated by Sir proteins that interact with telomere repeat binding factor Rap1[27]. In mammalian telomeres, nucleosomal arrays commonly associated with heterochromatin appear to be irregularly spaced or disrupted by telomere repeat binding factors[28,29]. Numerous interactions between shelterin and chromatin regulatory factors suggest that telomere repeat factors contribute to telomeric chromatin structure[30–35]. We have

previously shown that TRF2 can bind directly to telomeric repeat-containing RNA (TERRA) to recruit heterochromatin proteins including ORC and HP1 and maintain histone H3K9me3 enrichment at telomeres[33]. TERRA expression is itself dependent on histone H3K4 methyltransferase MLL[36], as well as DNA methylation status and CpG-island promoter found in many subtelomeric regions [37–39]. In fission yeast, the expression of TERRA and other subtelomeric transcripts are subject to diverse regulation by chromatin regulatory factors[40,41]. The dynamic interplay between shelterin, telomere chromatin structure, TERRA expression, and telomere biology appears to be an essential and universal component of chromosome stability.

The chromatin organizing factor CTCF has been implicated in numerous aspects of chromosome biology, including chromatin insulation, enhancer blocking, transcriptional activation and repression, DNA methylation-sensitive parental imprinting, and DNA-loop formation between transcriptional control elements[42–44]. CTCF has been implicated in the transcriptional repression of the D4Z4 macrosatellite repeat transcript found ~30 kb from the telomere repeats of chromosome 4q[45]. At D4Z4, CTCF interacts with lamin A and tethers the chromosome 4q telomere to the nuclear periphery [46,47]. A more general role for CTCF has been found in its ability to colocalize with cohesin subunits at many chromosomal positions[48–51]. Cohesin is a multiprotein complex consisting of core subunits SMC1, SMC3, Rad21, and SCC3 (referred to as SA1 or SA2 in humans), which can form a ring-like structure capable of encircling or embracing two DNA molecules[52,53]. Cohesin was originally identified as a regulator of sister-chromatid cohesion, but subsequent studies in higher eukaryotes indicate that they have functions in mediating long-distance interactions between DNA elements required for transcription regulation[54,55]. Cohesin subunit SA1 is recruited to telomere repeats by the shelterin protein Tin2, and this interaction is required for telomeric sister-chromatid cohesion and efficient telomere replication[56,57]. Tin2 can also promote heterochromatin formation through an interaction with heterochromatin protein HP1γ, but how this relates to sister-chromatid cohesion and cohesin function is not completely clear[34]. It is also not known whether CTCF can associate with

telomeres or subtelomeres in addition to binding the D4Z4 gene repeat, nor if it can interact with cohesin at these locations.

The chromosome region immediately adjacent to the terminal repeats has been referred to as the subtelomere. In humans, the distal subtelomeres consist of a variety of degenerate repeat elements with a few discrete gene transcripts interspersed at various distances from the terminal TTAGGG repeat tracts[24–26,58,59]. TERRA transcription initiates from within the subtelomeres, and a promoter containing a CpG-island and subtelomeric 29- and 61-bp repeat element has been identified in plasmid reconstitution assays[38]. DNA methylation and DNA methyltransferases have been shown to inhibit TERRA expression since TERRA levels are highly elevated in cells where DNMTs have been genetically disrupted or depleted [38,60], as well as in Immunodeficiency-Centromeric instability-Facial abnormalities (ICF) Syndrome cells that are genetically defective in DNA methyltransferase 3B (DNMT3B) [37,39,61]. CTCF binding is known to be DNA methylation sensitive but it is not yet known whether CTCF associates with transcriptional regulatory elements important for TERRA regulation or telomere maintenance. Herein, we investigate the role of CTCF and cohesin at human subtelomeres and their role in regulating TERRA expression, telomere chromatin organization, and telomere DNA end protection.

2.2 Methods

2.2.1 ChIP-Seq data

ChIP-Seq was performed using 1 × 107 BCBL1 cells per assay with either rabbit anti-cellular SMC1 (Bethyl A300-055A-3) or CTCF antibody (Millipore 07–729), or control rabbit IgG (Santa Cruz Biotechnology), using Illumina-based sequencing as described[249].

The public CTCF data are from ENCODE data series GSE19622[250]. The RNAPII data are from data series GSE19484[251]. The Rad21 data are from ENCODE/HAIB data series GSE32465. All three datasets used were from LCL lines.

2.2.2 Mapping ChIP-Seq data to human subtelomeres

The human subtelomere reference assemblies used for the mapping studies represent the most distal 15 kb of DNA sequence adjacent to the (TTAGGG)n terminal repeat tract for the indicated telomeres. Each assembly is oriented with the telomere end on the left with nucleotide position 1 corresponding to the first (CCCTAA) of the tract, which continues to the left of this position but was truncated for mapping consistency purposes. Some of these sequences were available in HG19 [165] whereas others were assembled by merging new fosmid sequence data with HG19 to bridge remaining gaps. In several instances, structural variants corresponding to alternative subtelomere alleles were also included in the set of subtelomere assemblies used here because they differed substantially from the original reference telomere. The full set of subtelomere assemblies are described in detail in Chapter 3. All of the sequences in the described orientation are available in FASTA format in Supplementary File S1.

Reads were mapped to the 15kb subtelomere reference using bowtie[238]. Many subtelomeres are duplicon rich with duplicon-specific nucleotide sequence similarities ranging from 90 to 99% between individual members of duplicon families that occur on separate subtelomeres[162,164,252,253]. To deal with this issue, we required a perfect match to retain a read, and all perfect matches of a given read to positions within the reference assemblies were recorded. Multiply mapping reads were dealt with as described previously[254], by assigning weights to reads such that multiple mapping positions sum to one read. Mapping likelihood was added to the reads as the inverse of the number of mapping positions. Picard (picard.sourcefourge.net) was used to mark and remove pcr duplicates. Coverage maps were then constructed using the mapping likelihood as a weight and extending the reads to the appropriate fragment length in the data set. The coverage map was calculated at single base resolution. Enrichment profiles were made from comparing RPM values between sample and IgG control. RPM=(coverage at position)/(total reads in library/10^6). The complete read mapping statistics (including Unique versus Multimapping Reads) are available for each of the data sets

used in Supplementary File S2. All figures were generated on the subtelomere reference genome hosted at the Wistar mirror of the UCSC genome browser, http://vader.wistar.upenn.edu.

2.3 Results

2.3.1 CTCF, cohesin, and RNAPII binding to the CpG-island promoters in human subtelomeres

Genome-wide analyses of CTCF, cohesin, and RNA polymerase II (RNAPII) have been performed in several different cell lines from various laboratories, including those generating the human ENCODE database [1–5]. In these published studies, the complete human subtelomeric DNA was not available for ChIP-Seq data mapping, with gaps immediately adjacent to the start of terminal repeat tracts for many telomeres[6]. We have generated complete assemblies of human subtelomeres for most of these chromosome ends and here we use these reference assemblies to map the read sequences from data sets, including our own, for CTCF, Rad21, SMC1, and RNAPII (Figure 2.1;Supplementary Figures 2.3 and 2.4). We found that most but not all human chromosome ends have a major CTCF-binding site within 1–2 kb from the TTAGGG repeat tracts. These CTCF sites consistently mapped to a region just upstream (centromeric) to the CpG-islands and 29 bp repeats, often overlapping 61 bp repeat element (Supplementary Figures 2.3 and 2.4). In the few exceptions to this pattern, CTCF sites were observed at positions ~10 kb from the TTAGGG repeats (7p, XYp) or several CTCF-binding sites with relatively low peak scores (3p, 7q, 8q, and 12q) (Supplementary Figure 2.4). We refer to these two different subtelomeres as type I (with major CTCF peaks at ~1–2 kb) or type II (lacking obvious CTCF peaks proximal to the telomere repeat tracts). In almost all cases, including those of type II, we observed an overlap of CTCF-binding sites with cohesin subunit Rad21 (Figure 2.1; Supplementary Figure 2.3). We confirmed that CTCF and cohesin peaks overlap in different cell lines by performing an independent CTCF and SMC1 ChIP-Seq experiment in a B-lymphoma cell line (Supplementary Figures 2.3 and 2.4). Our ChIP-Seq showed a nearly perfect overlap of CTCF and SMC1 in these cells, and a strong correlation with CTCF and Rad21 binding in

multiple cell types. In contrast to CTCF and cohesin, RNAPII bound as a more diffuse peak most commonly at a position immediately telomeric to the CTCF-binding sites and more directly overlapping the CpG-islands. A schematic summary of the average binding pattern of CTCF, Rad21, and RNAPII relative to CpG-island, 29 and 61 bp repeats is shown inFigure 2.1B.

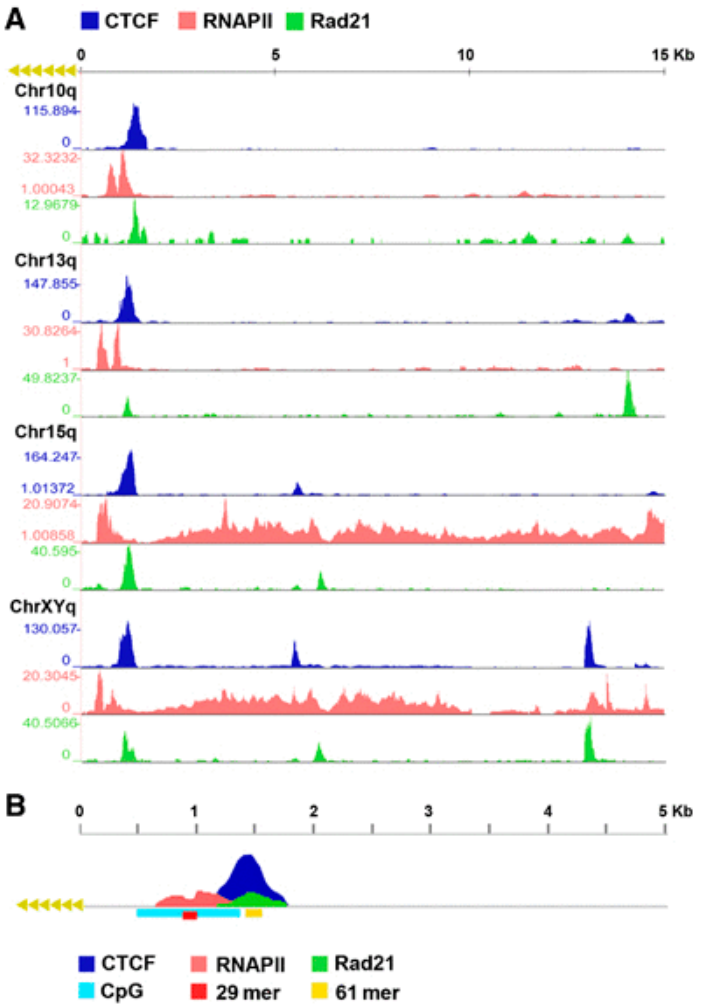

Figure 2.1 – Enrichment profiles for ChIP-Seq analysis of CTCF, cohesin, and RNAPII binding to human subtelomeres. Fragment density profiles were generated for samples and a matched IgG control as described in Materials and methods. The fold enrichment of sample over IgG is shown. (A) CTCF, RNAPII, and Rad21 binding in the first 15 kb subtelomeres of chromosome arms 10q, 13q, 15q, and XYq. The *y*-axis for each track is auto-scaled to highest peak in each chromosome region shown. (B) Model enrichment profile with peaks within the first 5 kb of the telomere tract. The CTCF peak is just centromeric to the CpG-island, typically centred over a 61-mer repeat. The RNA Pol II tract is centred over the 29-mer repeat. The exact position of these peaks varies with the positioning of these genomic features relative to the start of the terminal repeat tract on each chromosome arm.

Due to the complex duplications in subtelomeric sequence, we permitted multimapping signals weighted according to the number of perfect subtelomeric mapping sites to contribute, along with uniquely mapping reads, to subtelomeric ChIP-seq signals. We found that the remaining unique signals recapitulated the ChIP-Seq peak positions in most cases when multiple mappings were eliminated (Supplementary Figure 2.3E), suggesting that most of the binding sites can be uniquely assigned to specific subtelomeres. Some unique signals are lost, as expected for perfect duplications. This was sometimes the case with the 29-mer repeats over which RNAPII signal is centred and which a portion of the CTCF and cohesin read peaks was formed at many subtelomeres. Supplementary Figure 2.3D illustrates this effect for the example subtelomeres shown in the Supplementary Figure 2.3E. At the same time, Supplementary Figure 2.3D also shows the clear enrichment of RNAPII ChIP-seq reads mapping to the 29-mer variable number tandem repeat (VNTR) over the IgG controls, a true binding peak that would have been missed if multimapping signal contributions were disallowed.

2.3.2 CTCF binds directly upstream of the CpG-island and 29 repeat element found in subtelomeres

To verify the ChIP-Seq data for CTCF, cohesin, and RNAPII, we performed conventional ChIP-qPCR with primers spanning the first 3 kb of the XYq (Figure 2.2A and B) and 10q (Supplementary Figure 2.5) subtelomeres in a B-cell lymphoma-derived cell line used for ChIP-seq. As a control, we assayed TRF1 and TRF2 ChIP. As expected, TRF1 and TRF2 were enriched at positions closest to the TTAGGG repeats (primer set 1). ChIP assays with CTCF and cohesin subunits Rad21 and SMC1 revealed strong enrichment at the CpG-islands (primer set 2), while RNAPII was enriched at regions closest to the TTAGGG repeats (primer set 1), consistent with ChIP-Seq data indicating that RNAPII bound to broad peaks between CTCF–cohesin and the TTAGGG repeats. To determine if CTCF bound directly to subtelomere DNA, we assayed the ability of purified recombinant CTCF protein to bind candidate recognition sites in vitro by electrophoretic mobility shift assay (EMSA) (Figure2. 2C and D). Human CTCF protein was expressed and purified from baculovirus (Figure 2.2C). Candidate CTCF-binding sites from the

ChIP-seq peaks in subtelomere XYq, 10q, and 7p, as well as control oligonucleotides containing substitution mutations in the putative CTCF consensus sites, ΔXYq, Δ10q, and Δ7p, were synthesized as 46 mers for EMSA probes (Supplementary Table S3). Purified CTCF protein bound efficiently to the XqYq and 7p probes, less efficiently to 10q probe, but not to the mutated ΔXYq, Δ10q, and Δ7p probes (Figure 2.2D), indicating that these subtelomere ChIP-Seq peaks contain bonafide CTCF recognition sites. The relative binding affinities of these subtelomeric CTCF-binding sites was further quantified by a fluorescence polarization based competitor assay (Figure 2.2E). The wild-type CTCF-binding sites from XYq, 10q, and 7p showed robust competition against a FAM6-labelled probe containing a CTCF-binding site with high similarity to the consensus motif as defined previously[255]. Inhibitory constants (Ki) for each binding sites were equal to 11.82, 20.67, and 10.88 nM, respectively. On the other hand, the mutant ΔXYq, Δ10q, and Δ7p probes show linear relationship to increasing competitor with no plateau, suggesting a nonspecific inhibition of CTCF binding (Figure 2.2E). These findings indicate that the subtelomeric CTCF-binding sites have relatively high affinities for CTCF in vitro.

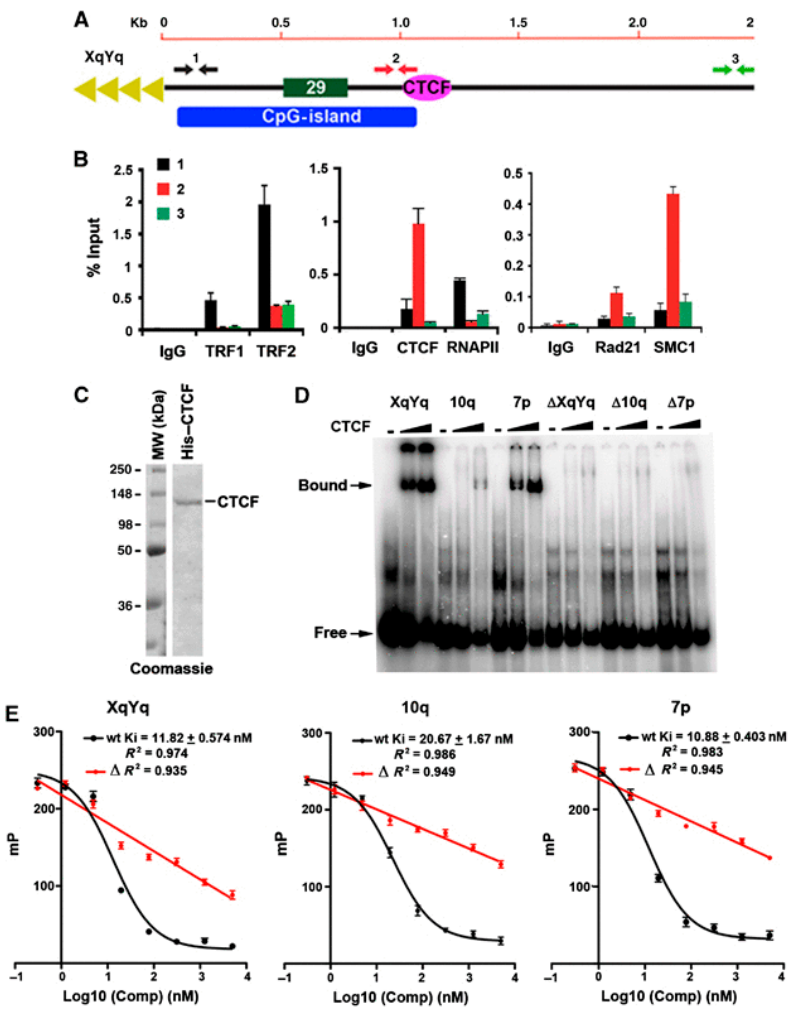

Figure 2.2 – Identification of CTCF-binding site elements in the 61-bp element of human subtelomeres. (A) Schematic of the type I subtelomere showing the relative positions of the 29- and 61-bp repeat element, CpG-island, and TTAGGG terminal repeats. (B) ChIP-qPCR for TRF1, TRF2, CTCF, RNAPII, Rad21, and SMC1 relative to IgG controls using primers for the XYq subtelomere at positions close (~150 bp) to TTAGGG repeat (black), at CpG-island (red), or ~3 kb from terminal repeats (green). Bar graph represents the average value of percentage of input for each ChIP from three independent PCR reactions (mean±s.d.). (C) Purified recombinant CTCF protein analysed by Coomassie staining of SDS–PAGE gel. (D) EMSA with CTCF protein binding to DNA oligonucleotide probes containing putative binding sites from subtelomere XYq, 10q, or 7p, as well as with oligonucleotides containing point mutations in CTCF recognition sites designated ΔXYq, Δ10q, and Δ7p. Free probe and bound probe were indicated with arrow. (E) Inhibitory constants (Ki) were calculated by titrating the same DNA probes used in EMSA against a FAM6-labelled probe with a known dissociation constant and measuring changes in CTCF binding via fluorescence polarization. Mutant (Δ) sites show a linear binding isotherm over the same concentration range of competitor, suggesting only nonspecific competition.

2.3.3 Summary of Results of Dr. Zhong Deng's functional experiments in Dr. Paul Lieberman's lab that were included in the Deng et al., 2012 publication

Using the CTCF and cohesin binding to subtelomeres suggested by initial mapping and analysis experiments as a starting point, Dr. Deng led a series of functional studies establishing the importance of subtelomeric CTCF and cohesin binding for TERRA transcription and genome integrity. These experiments indicated that (1) CTCF recuits RNAPII to subtelomeres; (2) CTCF and cohesin stabilize TRF binding to subtelomere; and (3) CTCF and cohesin are required for protection of telomeres and prevention of telomere DNA damage signaling. A subtelomere construct was used to show that both mutation of the CTCF binding site, and siRNA knock down of CTCF results in a decrease in TERRA transcription. CTCF and RAD21 knock downs were also used to show a subsequent loss of RNAPII binding in the subtelomere. Telomere damage foci were observed as a result of the depleted TERRA transcription. Dr. Deng's experiments are detailed in Deng et al. 2012 and are not included in this dissertation.

2.4 Discussion

2.4.1 A foundation for a chromatin atlas of the human subtelomeres

Genome-wide studies on chromatin structure and histone modification patterns have been incomplete near human telomeres, due both to remaining gaps in the reference sequence adjacent to the start of the (TTAGGG)n tracts and to subtelomeric segmental duplication families near many telomeres. In this work, we provide a foundation for a more complete analysis of the human genome by examining regions of the human subtelomeres that had previously not been included in human genome-wide studies. Using new sequence data to complete most of the gaps adjacent to (TTAGGG)n tracts and stringent read mapping criteria (both described in detail in Chapter 3), we have established a human subtelomere map and genome browser for next-generation DNA sequence analyses, including ChIP-Seq and RNA-Seq. Here, we mapped

several ChIP-Seq data sets to the most distal parts of human subtelomeres (Figure 2.1; Supplementary Figures 2.3 and 2.4). We focused on CTCF and cohesin subunits because of their general importance in chromosome organization throughout vertebrate evolution. We found that CTCF and cohesin colocalized at a position immediately adjacent to the CpG-islands implicated in TERRA promoter regulation[184] (Figures 2.1 and 2.2). We confirmed this binding by generating a new experimental data set for CTCF and SMC1 ChIP-Seq in a B-lymphoma cell lines. In addition, we mapped RNAPII binding and found that it localized more broadly across the subtelomeres, but had an average enrichment at the telomeric side of the CpG-island promoter for TERRA expression. CTCF and cohesin bound just centromeric to the CpG-island, and were further investigated for their role in TERRA expression and telomere end protection. The genome browser and methods established for mapping next-generation sequence data to the subtelomere provides a foundation for building a more complete atlas of epigenetic marks and chromatin organization at human subtelomeres.

2.4.2 Supplemental Figures

Large multipage figure available as figure S1 at
http://onlinelibrary.wiley.com/doi/10.1038/emboj.2012.266/suppinfo

Figure 2.3 – Summary of ChIP-Seq analysis of CTCF, cohesin, and RNAPII binding to type I human subtelomeres. Fragment density profiles were generated for samples and matched IgG controls as described in Methods. The fold enrichment of sample over IgG is shown. The Y axis for each track is auto-scaled to highest peak in each chromosome region shown. Subtelomere identity is indicated in the top left of each panel. (A-D) CTCF_W and SMC1_W were newly generated ChIP-Seq data using human pleural effusion lymphoma cell line BCBL1. CTCF, RNAPII, and Rad21 were derived from human encode data sets using B-lymphoblastoid cell lines. (E) Example enrichment profiles for ChIP-Seq analysis, comparing the standard multi-mapping method used vs allowing only unique mappings. The top track of each dataset pair permitted multimapping in the indicated ChIP and control IgG dataset, and the bottom (designated by _U) permitted only unique mappings in both ChIP and IgG control dataset. Binding in the first 15 kb subtelomeres of chromosome arms 10q, 13q, 15q, and XYq are shown.

Large multipage figure available as figure S2 at
http://onlinelibrary.wiley.com/doi/10.1038/emboj.2012.266/suppinfo

Figure 2.4 – Summary of ChIP-Seq analysis on type II human subtelomeres. Same as in Figure S1, except for subtelomeres that lack an obvious CTCF binding peak proximal to the terminal repeat tracts.

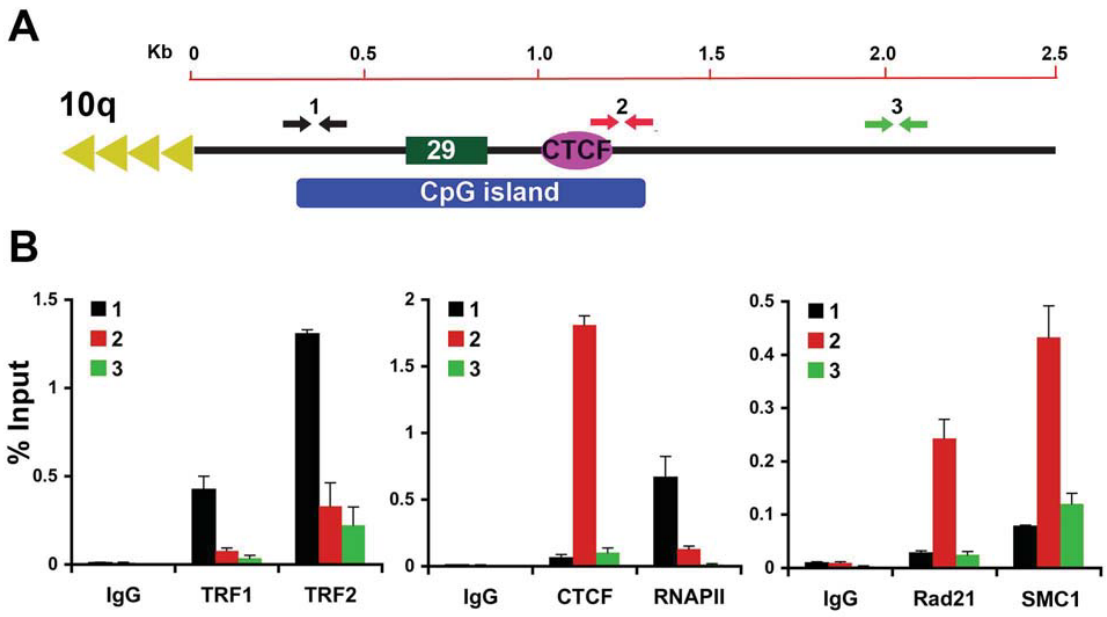

Figure 2.5 – Validation of CTCF binding site at 10q human subtelomeres in BCBL1 cells. (A) Schematic of the 10q subtelomere showing the relative positions of the 29 bp repeat element, CpG island, and TTAGGG terminal repeats. (B) ChIP-qPCR for TRF1, TRF2, CTCF, RNAPII, Rad21, and SMC1 relative to IgG controls using primers for the 10q subtelomere at positions close (~400 bp) to TTAGGG repeat (black), at CpG island (red), or ~2 kb from terminal repeats (green). Bar graph represents the average value of percentage of input for each ChIP from three independent PCR (Mean + SD).

CHAPTER 3: HUMAN SUBTELOMERE ANALYSIS

3.1 ABSTRACT

Mapping genome-wide data to human subtelomeres has been problematic due to the incomplete assembly and challenges of low-copy repetitive DNA elements. Here, we provide updated human subtelomere sequence assemblies that were extended by filling telomere-adjacent gaps using clone-based resources. A bioinformatic pipeline incorporating multi-read mapping for annotation of the updated assemblies using short-read datasets was developed and implemented. Annotation of subtelomeric sequence features as well as mapping of CTCF and cohesin binding sites using ChIP-seq datasets from multiple human cell types confirmed that CTCF and cohesin bind within 3 kb of the start of terminal repeat tracts at many, but not all subtelomeres. CTCF and cohesin co-occupancy was also enriched near Internal Telomere-like Sequence (ITS) islands and the non-terminal boundaries of subtelomere repeat elements (SREs) in transformed lymphoblastoid cell lines (LCLs) and human embryonic stem cell (ES) lines, but not significant in the primary fibroblast IMR90 cell line. Subtelomeric ITS islands were found to be frequent sites of artifactual mappings using short-read datasets due to the similarity of their sequences to those in terminal repeat tracts; TERF1 and TERF2 ChIP-seq peaks called at ITS sites could not be confirmed by ChIP-qPCR analysis of those sites. By contrast, subtelomeric CTCF and cohesin sites predicted by ChIP-seq using our bioinformatics pipeline (but not predicted when only uniquely mapping reads were considered) were consistently validated by ChIP-qPCR. The co-localized CTCF and cohesin sites in SRE regions are candidates for mediating long-range chromatin interactions in the transcript-rich SRE region. A public browser for the integrated display of short-read sequence-based annotations relative to key subtelomere features such as the start of each terminal repeat tract, SRE identity and organization, and subtelomeric gene models was established (vader.wistar.upenn.edu/humansubtel).

3.2 INTRODUCTION

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 [181,183,182] and forms an integral component of a functional telomere; perturbation of its abundance and/or localization causes telomere dysfunction and genome instability [181,185]. 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 [256]. Only one or a few critically short telomeres in a cell are sufficient to induce DDR-mediated senescence or apoptosis [66,257]. 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 [258–261].

Subtelomeric DNA elements regulate both TERRA levels and haplotype-specific (TTAGGG)n tract length and stability [157–159,184,185], with accumulating evidence for specific epigenetic modulation of these effects [184,185,262–264]. Heterogeneously-sized TERRA transcripts with as yet ill-defined transcription start sites and potential splice patterns originate in many, perhaps all human subtelomere regions [181,183,265], with the sizes of the larger transcripts (greater than 15 kb) suggesting structural overlap with some transcribed subtelomeric gene families [167,266]. 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 [265].

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[213,267,268]. 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 [269]. At D4Z4, CTCF interacts with lamin A and tethers the chromosome 4q telomere to the nuclear periphery [270,271]. A more general role for CTCF has been found in its ability to colocalize with cohesin subunits at many chromosomal positions [202,204–206]. 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[272,273]. 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 [274,275]. Cohesin subunit STAG1 is recruited to telomere repeats by the shelterin protein TINF2, and this interaction is required for telomeric sister chromatid cohesion and efficient telomere replication [276,277]. 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 [278]. By contrast, co-localized cohesin ring components and CTCF both contribute to subtelomeric TERRA transcriptional regulation and telomere end protection [265].

In humans, telomere regulation occurs in the context of subtelomeric DNA segmental duplications known as Subtelomeric Repeat Elements (SRE), which comprise about 80% of the most distal 100 kb and 25% of the most distal 500 kb in human DNA [165,279]. SRE regions of human chromosomes contain mosaic patchworks of duplicons [162,280,281] apparently generated by translocations involving the tips of chromosomes, followed by transmission of unbalanced chromosomal complements to offspring [164]. Along with highly elevated sister

chromatid exchange (SCE) rates in subtelomeres [163], 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 datasets 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.

3.3 METHODS

3.3.1 Fosmid library screening,

Methods and materials for these experiments are described in text associated with Supplementary Tables 3.3-3.6 and Supplementary Figures 3.5 and 3.6.

3.3.2 Updated subtelomere assemblies.

Supplementary Table 3.5 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 [282] to align the most centromeric 10kb of sequence from each subtelomere assembly with hg19 sequence. The blast results produced one perfect 10kb 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 non-subtelomeric portion of the hybrid genome using BEDTools [283]. The sequence of each 500 kb subtelomere assembly is provided as a concatenated FASTA file in Supplementary Materials. The joining coordinates for connecting hg19 to the subtelomere assemblies are listed in Supplementary Table 3.8.

3.3.3 Sequence Feature Annotation.

SRE and SD annotation were carried out as described previously [162]. Duplicon boundaries were defined as the end positions of duplicon 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 declared a single boundary for analysis purposes. Paralogy tracks were generated by first comparing the representative blocks identified by Linardopoulou et al. (2005) with the updated assemblies, and then adding blocks corresponding to new SRE segments shared in the manner described by [164]. 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 sequences for paralogy blocks 19a, 19b, 45, 46, 47, 48, and 49 are provided as a concatenated FASTA file in supplementary file 2. Subtelomere sequence assemblies were analyzed with RepeatMasker [284] and Tandem Repeats Finder [285]. Ensembl transcripts [286], and RefSeq genes [287] were aligned to subtelomeres using Spidey [288].

3.3.4 Short-read-based Annotation Pipeline.

Datasets 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 Supplementary Table 3.9. The LCL-associated datasets for CTCF, RAD21 and SMC1 were the same as described previously [265]. Additional data sets were downloaded as raw data FASTQ files from the ENCODE project [289] through the UCSC portal. H1-hESC_CTCF_Be and HMEC_CTCF_Be are from the Bernstein lab (GSE29611 series) at the Broad Institute. IMR90_CTCF_Sn, IMR90_RAD21_Sn, IMR90_POLR2A_Sn, and H1-hESC_RAD21_Sn correspond to the Snyder lab data from Stanford (GSE31477 series). H1-hESC_RAD21_My, H1-hESC_CTCF_My, and H1-hESC_POLR2A_My correspond to the Myers data from Hudson Alpha (GSE32465 series).

Reads were aligned to the hybrid genome using BWA 0.6.2 [239], allowing multimapping up to 101 locations (-n 101). BWA does not prioritize multimapping reads and alternate 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 locations. 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 [290]. bedGraph coverage files were created from the mapping positions by extending read mappings to the estimated fragment size. Fragment coverage for each position was calculated as the sum of ml values of fragments overlapping that position, and then averaged over a 20bp 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 dataset 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 were there was no mapping for the sample or control. A smoothing window of 500 bases was used on all control datasets. Peak calls were made using MACS 2.0.10 using the sample and control bedgraphs. First bdgcmp –m ppois, was called setting ppois as the method, calculating p value tracks. Peaks were called using bdgpeakcall –l 50 –c 4, setting minimum peak length to 50, and a p value significance cut of 4 (10^{-4}) [291]. Overall quality and mapping metrics for the datasets were determined as described [292] and are included in Supplementary Table 3.9.

TERF1 and TERF2 datasets. Publicly available datasets 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 SHRiMP 2.2.3 [233], allowing for reads mapping up to 101 mapping positions (-o 102). Once mapping positions were determined 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 (Supplementary Table 3.9). Additional TERF1 and TERF2 ChIP-seq datasets were generated for LCL as described in [249], using rabbit antibodies to TERF1 and TERF2 which were generated against

recombinant protein and affinity purified. The 100bp Illumina reads in these datsasets 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 RepeatMasker [284].

3.3.5 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 dataset selection functionalities as the UCSC Genome Browser [293]. The updated subtelomere assemblies in FASTA format are found in Supplementary File 2 and can be found on the Riethman lab web site (http://www.wistar.org/sites/default/files/protected/htel_1-500K_1_10_12_v4_3_12fasta.TXT).

3.3.6 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.

3.3.7 Chromatin Immunoprecipitation (ChIP) assay.

ChIP assays were performed with the protocol provided by Millipore with minor modifications as described previously[185]. Briefly, LCLs were crosslinked in 1% formaldehyde with shaking for 15 minutes, and DNA was sheared to between 200- to 400-bp fragments by

sonication with a Diagenode Bioruptor. Quantification of ChIP DNA at subtelomeric regions was determined using quantitative PCR (qPCR) with 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 supplementary Table 3.12. Each primer sets 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 g-[^{32}P]ATP labeled probes specific for telomere (4 x TTAGGG) or *Alu* repeats (cggagtctcgctctgtcgcccaggctggagtgcagtggcgcga). 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 and RAD21 (Millipore). Rabbit antibodies to TERF1 and TERF2 were generated against recombinant protein and affinity purified.

3.4 RESULTS

3.4.1 Gap-filling and detection of distal telomeric structural variants

In order to fill remaining telomere-adjacent gaps from our previous reference subtelomere assembly [43], we sampled telomere-adjacent DNA from deep fosmid clone libraries prepared from sheared genomic DNA samples [39,46,47]. 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 lab's previous assembly [43] 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 3.1). 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 [48]). A second allele for the distal 4q subtelomere, which shared high sequence similarity with distal 10q [49] 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 mis-assemblies immediately adjacent to the (TTAGGG)n tract in the previous assembly [43]; 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 is provided in Supplementary Materials (Mapped Telomere Fosmid Resource, Supplementary Figures 3.5 and 3.6, Supplementary Tables 3.3-3.6).

Table 1. Subtelomeric sequences from telomeric clones

Tel	Clone name	Accession	bp	Comment
10p	ABC7-43086900J11	AC215217	34335	Extends 10p ref sequence to terminal (TTAGGG)n tract
12p	ABC7-42389800N19	AC215219	35739	Extends 12p ref sequence to terminal (TTAGGG)n tract
13q	WI2-1528O10	AC213859	28566	Extends 13q ref sequence to terminal (TTAGGG)n tract
14q	WI2-1019G11	AC213860	33970	Extends 14q ref sequence to terminal (TTAGGG)n tract
20q	ABC7-42391600O12	AC215218	37776	Truncated variant allele of 20q to terminal (TTAGGG)n tract
22q	WI2-1161P17	AC213861	33328	Extends 22q ref sequence to terminal (TTAGGG)n tract
2q	ABC7-4304130019	AC215220	36897	Extends 2q ref sequence to terminal (TTAGGG)n tract
3p	ABC7-4028360016	AC215221	30142	Extends 3p ref sequence to terminal (TTAGGG)n tract
4q-1	WI2-3035O22	AC225782	42093	Extends 4q ref sequence to terminal (TTAGGG)n tract
4q-2	ABC7-42391500H16	AC215524	31434	Extends 4q ref sequence to terminal (TTAGGG)n tract (second allele)
7p	ABC7-481722F1	AC215522	33901	Truncated variant allele of 7p to terminal (TTAGGG)n tract
12q_gap	CA-2196C1 (from half-YAC)	AC226150	39835	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.
8q	ABC8-41019700A20	KF477190	5885	Distal end of telomeric fosmid
	ABC14-50184800C17	KF477189	8401	Distal end of telomeric fosmid
	ABC8-43258800E7	KF477188	8383	Distal end of telomeric fosmid
18q	ABC8-41174800P2	KF477185	7812	Distal end of telomeric fosmid
	ABC14-50923700D9	KF477187	7819	Distal end of telomeric fosmid
	ABC14-952514J11	KF477186	7789	Distal end of telomeric fosmid
	ABC8-2608140D9	KF477184	7991	Distal end of telomeric fosmid

Table 3.1 – Subtelomeric sequences from telomeric clones

3.4.2 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 (Supplementary Table 3.7). 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 3.1; eighteen telomere-adjacent regions were altered, 15 by addition to or replacement of hg19 sequence and 3 by truncation of hg19 sequence. For all telomeres not showing change relative to hg19 in Fig 3.1, the distal-most telomere gaps and clone gaps (where they existed immediately adjacent to telomere gaps), represented in hg19 by a long string of N's, 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 towards 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 assemblies; while they are known to contain a characteristic SRE organization closely related to distal 4p [50], they cannot be distinguished from each other and assemblies adjacent to them are unavailable. Thirty-five of the telomere assemblies extend to the start of the terminal telomere repeat tract, and those that do not can be defined relative to the start of the terminal repeat tract by comparison with known SRE organizations and independent mapping data [43,44].

Figure 3.1 shows the distal parts of the assemblies, encompassing all SRE regions. The one-copy DNA at the centromeric end of each assembly corresponded to and was connected to hg19 at the coordinates shown in Supplementary Table 3.8. In a few cases large segments of hg19 subtelomeric sequence were removed in our assemblies (e.g., removal of about 520 kb of distal hg19 sequence at the 1p subtelomere), but in most cases the updated assemblies were similar to those in hg19 with the exception of the most distal DNA segments. The resulting "hybrid genome", comprised mostly of hg19 sequence but modified by incorporation of our new subtelomere assemblies, allowed consistent genome-wide annotation that takes into account the entire reference sequence. The subtelomere browser described below displays only the first 500 kb of each chromosome arm from the annotated hybrid genome. It is important to note that the

subtelomere assemblies are not from single haplotypes. The hg19 genome assembly is comprised of clones from the DNA of many individuals, and the sequences we have added are from four additional individual genomes (Table 3.1, see description of the mapped telomere fosmid resource in Supplementary Information); it is important to consider these limitations in the interpretation of read-mapping results (see Discussion).

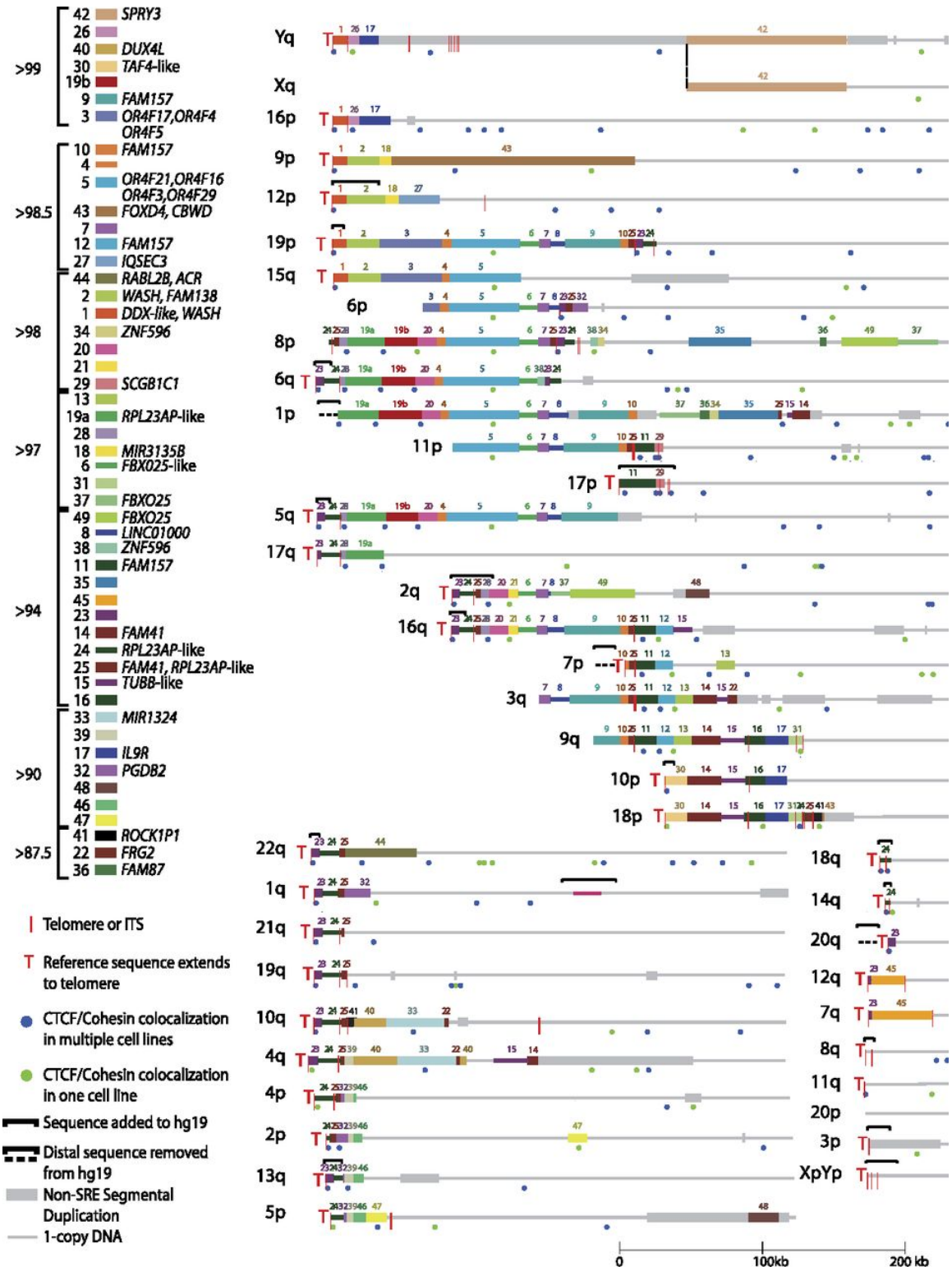
>99
42 SPRY3
26
40 DUX4L
30 TAF4-like
19b
9 FAM157
3 OR4F17,OR4F4
OR4F5
>98.5
10 FAM157
4
5 OR4F21,OR4F16
OR4F3,OR4F29
43 FOXD4, CBWD
7
12 FAM157
27 IQSEC3
>98
44 RABL2B, ACR
2 WASH, FAM138
1 DDX-like, WASH
34 ZNF596
20
21
29 SCGB1C1
13
19a RPL23AP-like
28
>97
18 MIR3135B
6 FBX025-like
31
37 FBXO25
49 FBXO25
8 LINC01000
38 ZNF596
11 FAM157
35
45
>94
23
14 FAM41
24 RPL23AP-like
25 FAM41, RPL23AP-like
15 TUBB-like
16
33 MIR1324
39
>90
17 IL9R
32 PGDB2
48
46
47
41 ROCK1P1
>87.5
22 FRG2
36 FAM87
Telomere or ITS
Reference sequence extends to telomere
CTCF/Cohesin colocalization in multiple cell lines
CTCF/Cohesin colocalization in one cell line
Sequence added to hg19
Distal sequence removed from hg19
Non-SRE Segmental Duplication
1-copy DNA
Yq
Xq
16p
9p
12p
19p
15q
6p
8p
6q
1p
11p
17p
5q
17q
2q
16q
7p
3q
9q
10p
18p
22q
1q
21q
19q
10q
4q
4p
2p
13q
5p
18q
14q
20q
12q
7q
8q
11q
20p
3p
XpYp
0
100kb
200 kb

Figure 3.1 – Sequence organization of updated subtelomere sequence assemblies. The assemblies are oriented with the telomere on the left and aligned to maximize paralogous blocks of SREs following the methods described in Linardopoulou et al. (2005). Regions of the assemblies differing from hg19 are indicated by the black brackets above the altered region of the assembly. An internal gap in the 1q assembly is indicated by the magenta line segment. The pseudoautosomal region of Xq and Yq shares the same reference sequence and is indicated by the thick gray line distal to the dotted line. Blocks 43 and 44 are shown as subtelomere paralogs because they are duplicated at the 2q site of an ancestral telomere fusion; other internal paralogies are not shown or analyzed here. A selection of named transcripts mapping primarily to the indicated blocks is listed; a much larger number of uncharacterized transcripts and ncRNAs is not shown here but is annotated on the subtelomere browser. The average percentage of identity shared by copies of paralogous blocks is indicated by the groupings to the left of the color key. The positions of telomeres, ITSs, and CTCF/cohesion colocalization sites in the three cell types examined in detail are as indicated in the figure.

3.4.3 Subtelomere Annotation

The hybrid genome was used to annotate subtelomeric sequence features as described in Ambrosini et al (2007), and to map several ChIP-seq datasets of particular interest to subtelomere function [19]. Figure 3.2 illustrates these annotations for the first 250 kb of the 19p subtelomere. Both coding and non-coding transcripts are abundant in SRE regions; while some are clearly functional, most are not well-characterized [21,40,51,52].

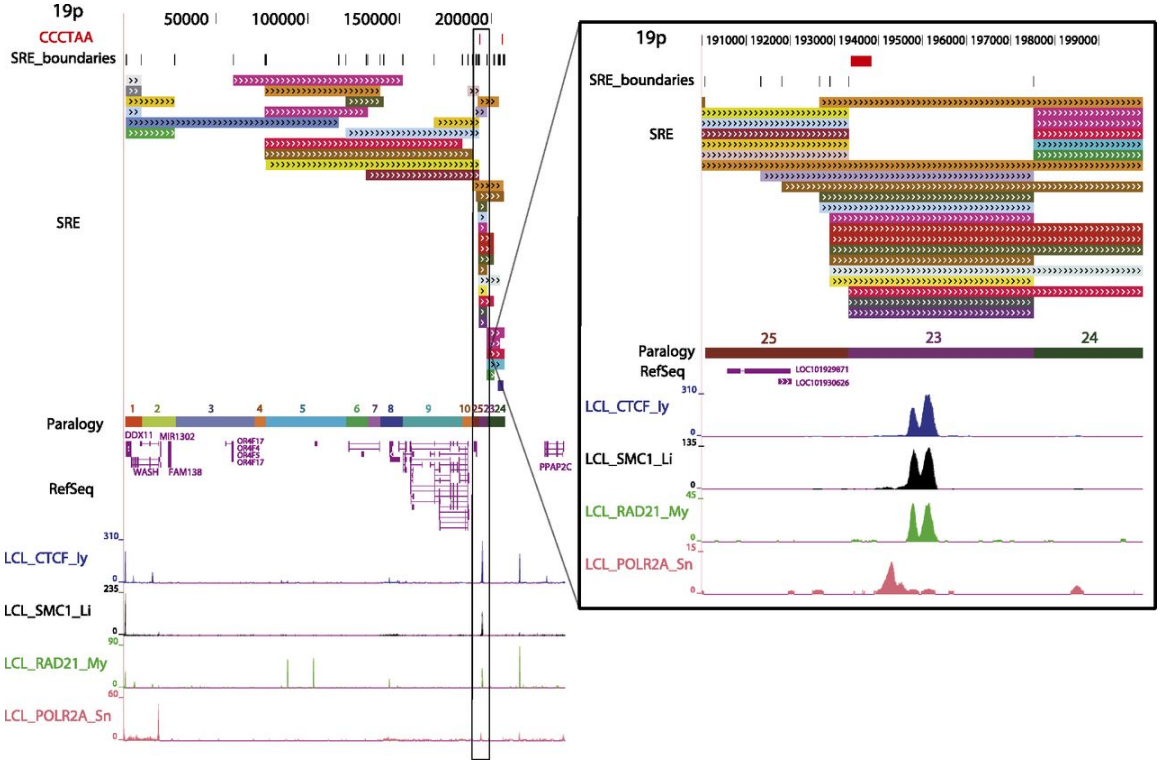

Figure 3.2 – Subtelomere annotation features. The first 250 kb of the 19p subtelomere assembly is shown to illustrate key features of subtelomere sequence organization annotated on our browser. Coordinate 1 on the browser corresponds to the centromeric end of the terminal repeat tract [i.e., the last (CCCTAA)n repeat unit before subtelomere DNA starts]. The 207-kb-long SRE region on 19p is subdivided into duplication modules ("duplicons") defined by segments of similarity (>90% nucleotide identity, >1 kb in length) between 19p and other subtelomeres (Ambrosini et al. 2007). Each rectangle represents a separate duplicon. Duplicated segments are identified by chromosome (color) as described previously (Ambrosini et al. 2007); additional details included on the live browser but omitted for the sake of clarity include the subject subtelomere identity, starting and ending coordinates of the duplicon in the subject subtelomere sequence, and the percentage of nucleotide sequence similarity of non-RepeatMasked sequences from the duplicon segment of the subject subtelomere to 19p (vader.wistar.upenn.edu/humansubtel). Each SRE boundary is indicated on a single track (SRE_boundaries), as are the internal telomere-like sequence (ITS) islands as defined in Methods (red ticks in the CCCTAA track). Gene models for transcripts included in the RefSeq (shown) (Pruitt et al. 2012) and Ensembl (hidden in this figure) (Flicek et al. 2012) transcript databases were mapped using Spidey (Wheelan et al. 2001). The paralogy track corresponds to the blocks, as shown in Figure 3.1. Enrichment profiles for four ChIP-seq data sets originally mapped only to subterminal DNA sequences (Deng et al. 2012) are displayed. (Inset) Close-up view of an internal SRE boundary region showing the association of the boundaries with an ITS (red rectangle on top line) and enrichment peaks for CTCF, cohesin subunits SMC1A and RAD21, and RNA polymerase II large subunit (POLR2A).

A paralogy map for SRE regions was prepared based upon the paralogy blocks defined previously [44] to facilitate graphic visualization of similar sequence segments occurring in multiple telomeres (see Figure 3.1, and methods). Previously defined paralogy blocks covered most SRE regions, but we identified 5 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 Linnardopoulou (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 % 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 [53,54]. 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 multi-read 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).

Using this pipeline, enrichment profiles for four of the ChIP-seq datasets originally mapped only to telomere-adjacent DNA sequences [19] are displayed in Fig. 3.2 on the subtelomere browser after mapping to the entire hybrid genome using the multi-read mapping approach, then displaying the distal 500 kb on the subtelomere browser (see Methods). The same subterminal binding enrichments for CTCF, SMC1A, RAD21, and RNA polymerase II large subunit (POLRA2) which were found and validated by ChIP-qPCR in our previous work [19] are

evident in the current annotation (less than 3 kb from the telomere tract at 19p in Fig. 3.2; see Supplementary Fig. 3.8 for other telomeres). In addition, enrichment peaks for these proteins throughout the 19p subtelomere region are shown in Fig. 3.2 (see Supplementary Fig. 3.8 for other subtelomeres). The inset highlights an internal SRE boundary region shared by many duplicons, showing the proximity of these boundaries with an ITS (red rectangle on top line) and enrichment peaks for CTCF, cohesin subunits SMC1 and RAD21, and POLR2A. Interestingly, the sequences adjacent to this ITS share similar but non-identical 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 [19]; this ITS-adjacent binding site corresponds to a 23-mer element which, 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 2 and 4 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 [19];, 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 %).

3.4.4 SRE Boundary Enrichments

Publicly available CTCF and cohesin subunit ChIP-seq datasets from human ES cells and primary diploid fibroblasts (IMR90) were mapped in the same fashion and compared with the LCL data. All of the datasets used in this study and their mapping characteristics are summarized in Supplementary Table 3.9. 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. 3.3; Supplementary Fig. 3.8). For example terminal boundary RAD21 enrichment peaks were almost always visible at some level in the bedGraphs of the expected subtelomeres, but for some datasets many of the peaks did not

reach the MACS significance threshold set for peak-calling subtelomere-wide ($p < 1.0E-4$; Supplementary Table 3.10). 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. 3.3; Supplementary Fig. 3.8). Differences in library quality and depth, as well as differences in the antibodies used for ChIP-seq (Supplementary Table 3.9) 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 (e.g., compare relative CTCF peak heights and relative RAD21 peak heights on distal 6q in Fig. 3.3, and in all subtelomeres in Supplementary Fig. 3.8 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 [55]. Similarly, a subset of cohesin binding sites are known to display cell-type specificity, co-localizing with tissue-specific transcription factors [56]. 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.

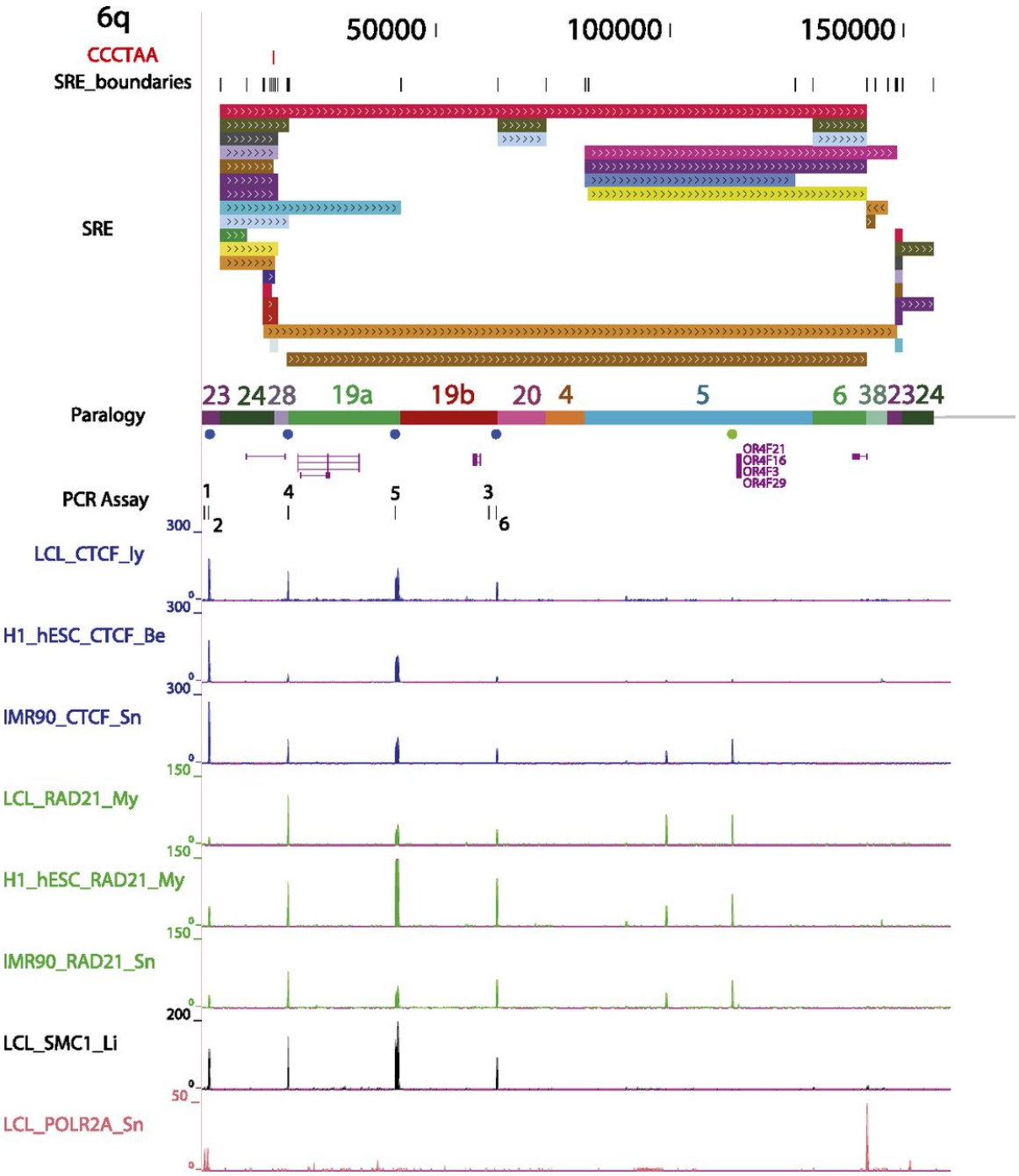

Figure 3.3 – Example of an annotated subtelomere with CTCF and cohesin binding enrichment peaks from multiple cell types. The first 160 kb of 6q is shown in our browser. The PCR assay track marks the primer sites used for ChIP-qPCR (see Fig. 3.4). In addition to the ChIP-seq data sets shown in Figure **3.**2 for LCLs (Deng et al. 2012), enrichment profiles for CTCF and RAD21 are shown following mapping of the ENCODE Project ChIP-seq data sets from the pluripotent human embryonic stem cell line H1-hESC and the primary fibroblast cell line IMR90.

Visually noted apparent association of CTCF and cohesin peaks with some SRE boundaries was analyzed systematically using only significant peaks called for each dataset by MACS [57]. 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 less than 3 kb from this boundary [19], 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.

Table 2. SRE boundary enrichments

3-kb window						
Cell line_ChIP-seq data set	SRE/SRE enrichment	*P*-value	Terminal enrichment	*P*-value	ITS enrichment	*P*-value
CTCF						
LCL_CTCF_Iy	1.200	0.07314714	10.789	5.05593×10^{-19}	2.012	0.00019951
LCL_CTCF_W_Li	1.192	0.0762587	10.260	1.8887×10^{-18}	2.033	0.00013915
H1-hESC_CTCF_Be	1.189	0.1219057	13.011	1.19914×10^{-19}	2.308	4.3071×10^{-5}
H1-hESC_CTCF_My	1.419	0.00707443	15.530	1.15952×10^{-21}	1.771	0.00366826
IMR90_CTCF_Sn	0.967	0.6399912	8.880	2.33908×10^{-14}	1.725	0.00419184
Cohesin						
LCL_RAD21_My	1.200	0.3232894	4.247	0.001311649	2.298	0.00136006
LCL_SMC1_Li	1.192	0.006035461	15.391	2.66361×10^{-23}	1.794	0.0021072
H1-hESC_RAD21_My	1.189	0.009682766	8.535	5.39758×10^{-14}	2.369	7.83×10^{-6}
H1-hESC_RAD21_Sn	1.617	0.000260144	5.980	5.41325×10^{-5}	2.614	0.00038859
IMR90_RAD21_Sn	0.967	0.2708427	4.407	0.012587796	0.963	0.53558534
Colocalized CTCF & cohesin						
LCL_CTCF_Iy & LCL_RAD21_My	1.478	0.00664919	5.980	0.000158034	3.236	1.9703×10^{-5}
LCL_CTCF_W_Li & LCL_SMC1_Li	1.450	0.00309262	15.857	1.10817×10^{-23}	2.587	2.5738×10^{-6}
H1-hESC_CTCF_Be & H1-hESC_RAD21_My	1.564	0.003089455	7.788	7.27985×10^{-6}	2.837	0.00036953
H1-hESC_CTCF_My & H1-hESC_RAD21_Sn	1.485	0.004011074	15.505	5.16979×10^{-19}	1.581	0.01862117
IMR90_CTCF_Sn & IMR90_RAD21_Sn	1.157	0.2708427	4.407	0.012587796	1.926	0.05584247
Tel repeat	1.545	0.000145563	All	NA	All	NA

Table 3.2 – SRE boundary enrichments

This analysis confirmed the strong association of CTCF and cohesin sites with the terminal boundaries in the cell types examined, and also revealed a strong association of CTCF and cohesin sites with ITSs in all datasets except for IMR90 RAD21 (Table 3.2). 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 3.2). However, boundary analysis of just the strictly co-localized 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 3.2). The positions of all co-localized CTCF and cohesin peaks occurring in at least one of these three cell types are shown relative to SRE organization in Figure 3.1.

3.4.5 Experimental validation of ChIP-seq peaks by ChIP-qPCR

Several recent reports have suggested that some human ITSs bind shelterin components TERF1 and TERF2 [58,59], which seems plausible given the demonstrated ability of TERF1 and TERF2 to interact with the very short TTAGGGTT motif in some contexts [60,61]. This could have important functional implications and suggest potential long-range interaction of ITSs with telomeres. We therefore mapped TERF1 and TERF2 ChIP-seq datasets we prepared from LCLs, as well as publically available TERF1 and TERF2 ChIP-seq datasets from a transformed BJ fibroblast cell line [59]. 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 (Supplementary Figure 3.7A), 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 comprised 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 include proximal regions of terminal repeat tracts, known to contain extended regions of telomere-like sequences interspersed with pure (TTAGGG)n repeats [62]. When telomere and telomere-like sequences were specifically removed from the datasets, peaks at all subtelomeric ITSs disappeared (Supplementary Figure 3.7A). Examination of read orientations underneath a typical ITS peak compared to 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 towards the peak of the enrichment (Supplementary Figure 3.7B and 3.7C). 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 co-localization 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 3.4 and in Supplementary Figure 3.9. In Figure 3.4A, the co-localized CTCF and RAD21 sites in segments of the 6q and 16q SRE regions were examined; each of these sites were 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 2,4,5,6,8,9,10) 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 [19], whereas more distant subtelomeric sites at positions 3-10 show only background TERF1 and TERF2 levels. In Figure 3.4B, the expected CTCF and RAD21 enrichments are also seen in assays corresponding to co-localized CTCF and RAD21 sites predicted by ChIP-seq (positions 2,4,5,6,8). The telomere-adjacent Xq sites at positions 1 and 2 detect the expected TERF1 and TERF2 enrichments [19], 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 datasets, 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 datasets and binding enrichment by ChIP-qPCR, while showing anticipated ChIP-qPCR enrichments at CTCF and RAD21 co-localization sites predicted by ChIP-seq (Supplementary Figure 3.9). 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 datasets cannot be used to predict true TERF1 and TERF2 binding.

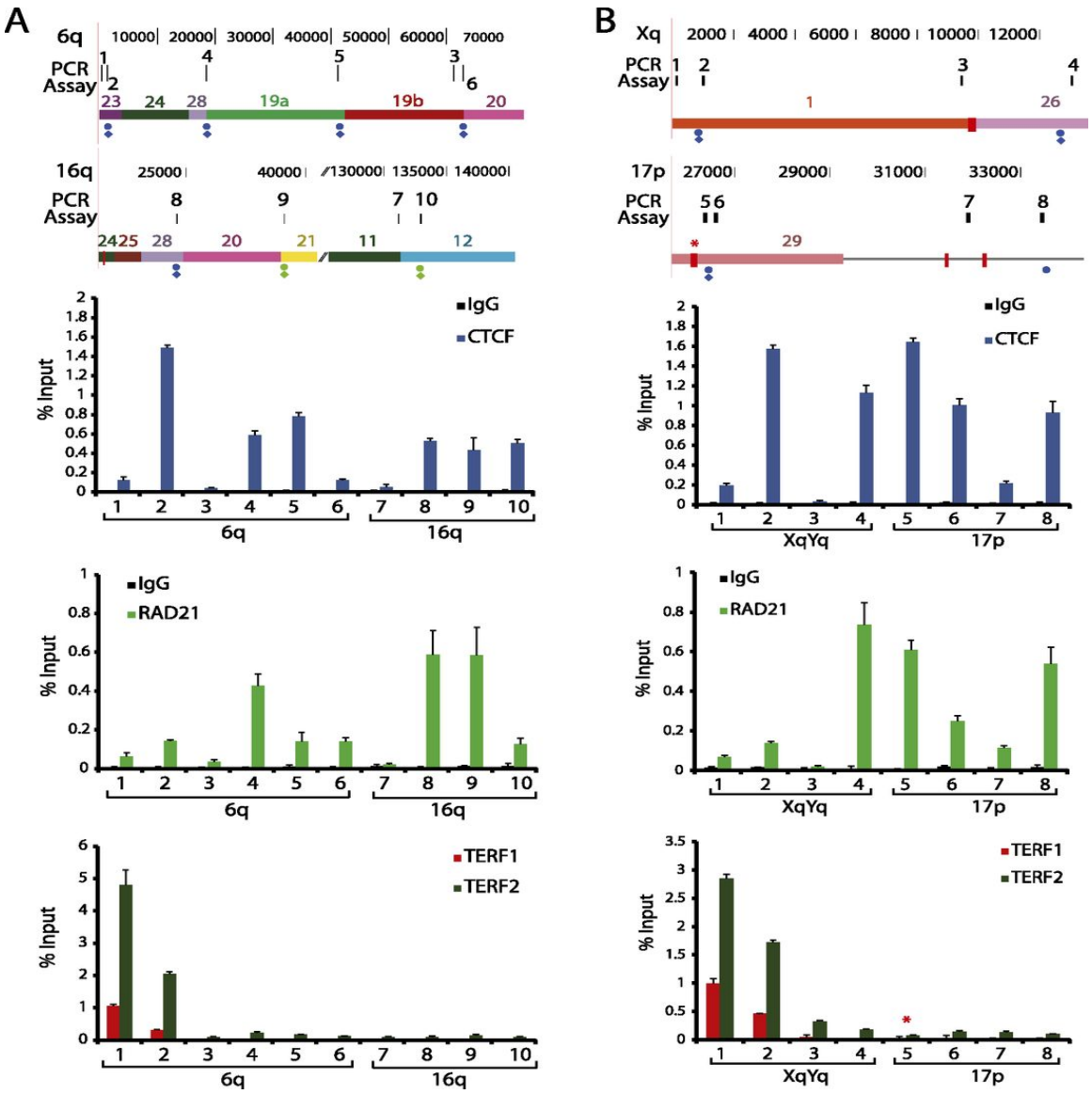

Figure 3.4 – ChIP-qPCR analysis of subtelomeric DNA protein binding sites predicted by ChIP-seq data set mappings. Candidate sites of CTCF, cohesin, TERF1, and TERF2 binding were analyzed by ChIP-qPCR. Segments of the 6q and 16q (A) and the Xq and 17p (B) subtelomeres are shown, with the coordinates (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/or 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).

3.4.6 CTCF datasets 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 datasets from four primary cell lines (HMEC, human mammany epithelial cells; SAEC, small airway epithelial cells; HRE, human renal cortical epithelium; 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 (Supplementary Table 3.11) as were found in ChIP-seq datasets for LCLs, human ES cells, and IMR90 (Supplementary Table 3.10). As with the individual CTCF datasets for LCLs, ES cells, and IMR90, non-terminal SRE/SRE boundary associations with just CTCF peaks were usually not significant in the cell lines. Cohesin ChIP-seq datasets were not available for most of these cell lines, so we could not determine co-localized 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 datasets as well as experimental validation.

3.5 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. 3.1). This represents a significant advance over the previous human subtelomere assemblies [40,43]. We also provide a multi-read mapping pipeline that enables the systematic analysis of distal chromosome regions using short-read sequencing based methods, leveraging

the wealth of public genome-wide datasets 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.

The updated subtelomere reference assemblies are subject to caveats as are all regions of the reference human genome sequence; they are comprised 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[21,44], and this may complicate interpretation of ChIP-seq signal strengths at specific high-similarity SRE sites when comparing datasets from non-isogenic 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 dataset from a genome with more copies of the SRE segment than a dataset 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 2-fold [44,52]; 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 dataset. 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 [19], 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 multi-read mapping approach for ChIP-seq short-read datasets 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 dataset) when only uniquely mapping reads were considered. This is illustrated dramatically in Figure 3.4, where all of the sites predicted by the multi-read mapping in the SRE regions were missed in the analysis considering only uniquely mapping reads. Comparison of the multi-read 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 multi-read peak for SRE sites, indicating that some fraction of the reads were mapping uniquely to the site but the unique enrichment peak was too weak to be called statistically significant. However, as we showed previously [19], in SRE regions with very high sequence similarity to paralogs there was no detectable enrichment in the uniquely mapping datasets.

Because peaks detected using the multi-read 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 predicted CTCF and RAD21 peaks shown in Fig 3.4 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 Supplementary Table 3.12).

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 3.2; [19]). When we considered only the CTCF and cohesin subunit peaks that co-localized exactly (see Figure 3.1), the significance of association with telomere-adjacent DNA and ITSs typically increased, while the co-localized peak association with SRE/SRE boundaries reached significance for the ES and LCL lines but not for IMR90 (Table 3.2). Strong cohesin sites co-localizing with CTCF have been implicated in long-range chromosomal interactions [56], suggesting co-localized cohesin/CTCF sites may mediate DNA looping and long-range DNA interactions as well as regulate transcription [56,63,64]. 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 [19] 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 [56]. Using our subtelomere browser and bioinformatics pipeline to leverage the rich public resource of additional short-read datasets 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.

3.6 DATA ACCESS

DNA sequence for gap-filling clones and clone fragments were submitted to the NCBI GenBank (https://www.ncbi.nlm.nih.gov/genbank/) with the following accessions:: AC215217, AC215219, AC213859, AC213860, AC215218,

AC213861, AC215220, AC215221, AC225782, AC225782, AC215524, AC215522, AC226150, KF477190, KF477189, KF477188, KF477185, KF477187, KF477186, KF477184 (see Table 3.1). The TERF1 and TERF2 ChIP-seq datasets generated as part of this study were submitted to the NCBI GEO (http://www.ncbi.nlm.nih.gov/gds) with the following accessions: GSM1328844 and GSM1328845. Each of the 500 kb subtelomere reference assemblies are available as a concatenated FASTA file in Supplementary File 1. New SRE paralogy blocks 45-49,19a, and 19b are available as a concatenated FASTA file in Supplementary File 2. The subtelomere browser link is (vader.wistar.upenn.edu/humansubtel).

3.7 ACKNOWLEDGEMENTS

We acknowledge contributions from the Wistar Cancer Center Core facilities in Bioinformatics and Genomics, and especially Priyankara Wickramasinghe for maintainence and updating of the Subtelomere Browser in the Wistar Bioinformatics Core facility and Brett Taylor for assistance with the computational resources in the Wistar Center for Systems and Computational Biology. This work was supported by the Wistar Cancer Center core grant (P30 CA10815) and the Commonwealth Universal Research Enhancement Program, PA Department of Health. Additional support was provided by the Philadelphia Health Care Trust and a predoctoral NRSA F31 Diversity award (N.S.), and NIH grants to HR (R21CA143349 and R21HG007205) and PL (RO1CA140652). ZD was supported by an American Heart Association Grant (11SDG5330017), and RD by R01LM011297. EE was supported by HG004120 and gratefully acknowledges the technical/bioinformatics assistance of Maika Malig and Jeff Kidd. Work done by TG, CF, LC, and RW was supported by U54 HG003079.

Author Contributions: NS and HR designed the experiments for the project, and SH and SP carried out the mapping and sequencing experiments in HR's Lab. Fosmid clones were provided by EE, and full Sanger sequencing of selected clones was done by CF, LC, TG, and RW. RG, AW, and RD assisted with the initial analysis of ChIP-seq data. ZD and PL carried out the ChIP-qPCR experiments, and assisted with interpretation of the data. NS developed the bionformatic

pipeline for multimapping ChIP-Seq analysis. NS and HR led the analysis and interpretation of the data, assembled the figures and wrote the manuscript.

3.8 Supplementary Information

3.8.1 Supplementary Figures

Fosmid End Sequence (FES) mapping, gap-filling, and detection of telomeric structural variants

A set of end-sequenced fosmid libraries derived from sheared human genomic DNA were screened for clones containing the telomere terminal repeat sequence (TTAGGG)n. Because of the orientation of this repeat at all terminal repeat tracts, the distal end-sequence from a telomere-terminal fosmid will always contain a (CCCTAA)n pattern. We initially computationally screened the G248 and the ABC7 fosmid libraries for the presence of (CCCTAA)n in their end-sequences. The G248 library was prepared to validate the original human reference genome assembly [279], then was used to detect genomic structural variation [294]. The ABC7 library was the first structural variation fosmid library for which complete paired end-sequence data were available [295,296].

Analyses of (CCCTAA)n-containing end-sequence reads and mapping of their mate-pair end sequences to subtelomeric DNA showed that requiring a perfect (CCCTAA)4 match reliably identified authentic telomere-containing fosmid clones (Supplementary Table 3.3; Supplementary Table 3.4). Both of these libraries contained fewer (CCCTAA)n sequences than expected from their 12x clone coverage (Supplementary Table 3.3), and both were clearly skewed towards loss of (CCCTAA)n sequences upon quality processing of sequence traces to remove low-quality bases; inspection of individual traces showed that a high fraction of CCCTAA -containing sequence reads were poor quality and short, but many of the poor-quality reads clearly contained terminal (CCCTAA)n by visual inspection of the sequence trace patterns. Matches corresponding to internal (CCCTAA)n-like islands were very rare, probably because these islands are known to be quite small (less than 250 bp) and somewhat degenerate [165], and the expected overall sequence coverage by the sequenced fosmid ends is low (about 0.5x). By comparison, the target

size of terminal repeat tracts expected to contain mostly perfect (CCCTAA)n in lymphoblastoid cell line DNA is roughly 3-12 kb [297].

The mate pairs of (CCCTAA)-containing fosmid end sequences from 183 fosmids from the G248 library and 353 fosmids from the ABC7 library were mapped back to reference subtelomere assemblies [162]; all but a few mapped either uniquely to a known subtelomere assembly or to a known SRE. These mappings identified some fosmids that should bridge existing subterminal gaps in the reference sequence and additional fosmids that appeared to represent structural variants of SRE regions (Supplementary Table 3.4); each of these clones were fully sequenced (Table 3.1). In addition, a 12q half-YAC-derived cosmid that bridged a gap in the distal subtelomeric 12q assembly was sequenced (Table 3.1).

Additional experiments confirmed that underrepresentation of telomere regions in the roughly 12-fold coverage G248 and ABC7 libraries was mainly due to the relatively poor quality and short length of the (CCCTAA)n-containing sequence reads (see Supplementary Table 3.3 and Supplementary Fig 3.5). Restriction mapping of multiple clones from each telomere (Supplementary Fig 3.6) and analysis of sequence reads from subterminal duplicon/terminal repeat junctions showed that (CCCTAA)n tract deletions were confined entirely to regions within the terminal repeat tract itself, and subterminal sequences were not affected by the (CCCTAA) sequence deletion (Supplementary Fig. 3.6). This observation is consistent with earlier observations in yeast [298] and *E. coli* (Riethman, unpublished) that (CCCTAA)n tracts longer than about a 1000 bp are not maintained in either host.

An additional seven structural variation libraries were screened computationally for (CCCTAA)n in end sequences, identifying telomere clones (Supplementary Table 3.5), and the mate-pair mappings were analyzed relative to the available human reference assemblies [162]. Mate-pair mappings from the three libraries with the highest coverages in identified (CCCTAA)n sequence (ABC7, ABC8 and ABC14) were characterized in detail (Supplementary Table 3.6). In addition to the variants described above for ABC7, potential truncation alleles for XpYp were identified in the

ABC8 and ABC14 libraries, and a potential new allele for the 17p telomere was identified in ABC14. No additional structural variants could be identified on the basis of unique mate-pair mappings. However, differential clustering of mate-pair reads with SRE regions of the previous reference assembly [162] in these libraries (Supplementary Table 3.6) suggest large SRE-associated structural variation amongst these genomes [266] that will require long-range analytical methods to characterize further. While the exact localization of many of the SRE-mapping telomeric fosmids is not possible using mate-pair mappings, this information in combination with known similarities amongst subtelomere duplicon families and the depth of clone coverage indicates that all or nearly all telomere-terminal fragments are represented amongst the (CCCTAA)4 –selected clones from each library. In addition, these mapped telomere fosmid resources can be used to refine sequences and further explore allelic variation near specific telomeres. For example, the 8q and 18q telomeres both retain sequencing ambiguities and potential mis-assemblies immediately adjacent to the (TTAGGG)n tract in the current version of the reference sequence (hg19). High-resolution mapping and sequencing of the distal portions of telomere fosmid clones (Supplementary Figure 3.6 legend) from the ABC8 and ABC14 libraries identified several related but distinct alleles corresponding to each of these telomeres (Table 3.1).

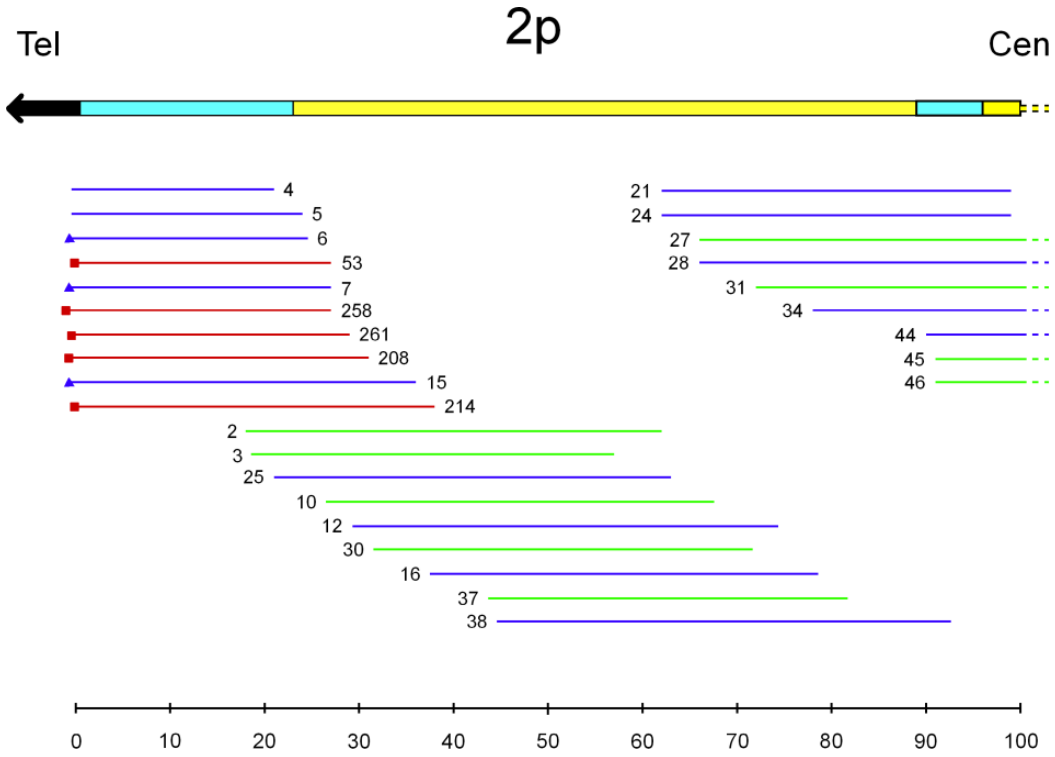

Figure 3.5 – G248 fosmid coverage of 2p subtelomere. Experiments showed that the apparent differences and underrepresentation of telomere regions in the roughly 12- fold coverage G248 and ABC7 libraries was mainly due to the criteria used to declare a (CCCTAA)n hit in the initial computational screens and the often relatively poor quality and short length of the (CCCTAA)n-containing sequence reads.

To illustrate this, the terminal 100 kb of 2p reference sequence was used to query the G248 end-sequence library (after processing reads from the library to remove low-quality regions of traces). Megablast parameters used to match the reads were (-D 3 –p 95 –W 12 –t 21), and the BLAST output results were stringently refined so that only hits with a % identity greater than or equal to 98 % and alignment length greater than or equal to 100 bases were retained. The raw sequence reads of the mate pairs of all of these telomerically oriented near-perfect single matches mapping to within 40 kb of the 2p telomere were examined; 5 corresponded to the original (CCCTAA)n hits from the initial screen using Quality-filtered reads (red), 3 had recognizable (paired or greater) (CCCTAA)n motifs but the reads were removed from the library during the trace quality trimming procedure (blue diamonds) and two lacked any mate pair in the database, suggesting failed sequencing reads So the actual depth of clone coverage is similar to what one would expect for a 12x , 40 kb random shear library close to an absolute end of a source DNA fragment. The positions of end-sequence matches for the non-(CCCTAA)n ends of the terminal fosmids (from 27 kb to 40 kb from the start of the (TTAGGG)n tract) likely reflects the variable stretches of (TTAGGG)n sequence originally present in the size-selected fosmid clones; the fosmids with an end-sequence mapping closest to the telomere tract carried the longest (TTAGGG)n stretch, those the farthest from the telomere carried the shortest, but in every case all but the most proximal 300 – 800 bp of the telomere tract was deleted. Our paired end mappings (blue) also revealed additional fosmid coverage throughout the region in addition to that found in the UCSC browser (green line segments), perhaps because we did not mask interspersed repeats in our end-sequence mapping procedure. These experiments showed that, for both the G248 and ABC7 libraries the relatively stringent criteria used for declaring a (CCCTAA)n hit resulted in roughly 5-6 fold coverage of terminal fosmids, and by relaxing these criteria slightly and making use of end-pairs mapping to distal subtelomere regions we could increase the coverage to 8-10 fold.

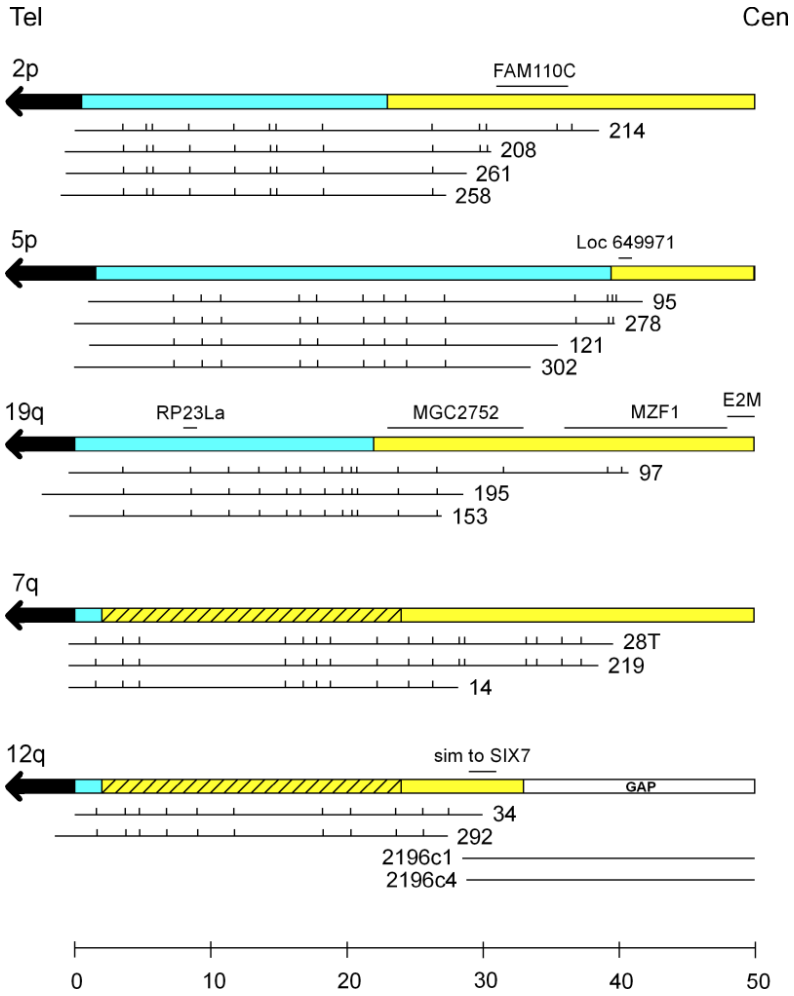

Figure 3.6 – Stability of Subterminal DNA in Fosmids

Each of the mapped terminal G248 fosmids and a selection of the terminal ABC7 fosmids (provided by Evan Eichler) were fingerprint-mapped. Those mapping to a single telomere yielded overlapping fingerprints, with the exception of 4q which yielded two sets of fingerprints and 4p which yielded several sets of related overlapping fingerprints (perhaps due to the contribution of acrocentric short-arm telomeres, which are known to have sequences highly similar to 4p; Youngman et al., 1992). The fingerprint contig maps obtained for G248 fosmids mapping to several completed subtelomere assembly ends agree with each other and with the mapped

position of the non-(CCCTAA)n mate-pair reads on the subtelomere reference assembly (Suppplementary Fig 3.6). In each case, the telomeric end of the fosmid insert contains a short (< 1 kb) stretch of mostly (CCCTAA)n sequence (usually with some non-canonical hexamer repeats as well; Baird et al., 1995) immediately adjacent to the fosmid cloning site. Since the libraries were constructed from sheared DNA that was size-selected, most of the terminal fosmid clones must have lost some fraction of their initial (CCCTAA)n tract to a length that could be stably maintained in the fosmid, typically 300 bp to 800 bp. Both the restriction mapping and the sequence reads of the subterminal duplicon-terminal repeat junction from sets of independently isolated fosmids mapping to single loci indicate that (CCCTAA)n tract deletions were confined entirely to regions greater than 300 bp telomeric of the subterminal duplicon-terminal repeat boundary, and subterminal sequences were not affected by the (CCCTAA) sequence deletion. This observation is consistent with the size of remaining human telomere tract lengths seen on terminal telomere fragments cloned in yeast, and indicate that (CCCTAA)n tracts longer than about a kb are not maintained in either cloning system.

Sequencing of distal ends of Terminal Fosmids

Directed sequencing was used to obtain data on subterminal sequences for selected fosmid clones. We followed a protocol similar to that of Raymond et al. (2005) using purified fosmid DNA and BigDye Terminator sequencing using custom primers corresponding to known human subterminal sequences. Initial reactions were primed from sites across the subterminal 5 kb of DNA immediately adjacent to the start of the telomere repeat tract. Over most of this region, high-quality reads > 600 bp were obtained. However, in regions immediately adjacent to the start of the terminal repeat tract, including a very CG-rich region and the beginning of the hexamer repeat tract itself, the read lengths were well below this average, often in the 200 – 250 base range. We found that a simple modification of the sequencing protocol to include a 5-min controlled-heat denaturation step of the template prior to addition of cycle sequencing reagents (Kieleczawa, 2006) doubled the read lengths in most cases. We were able to obtain reads extending about 300 bases into the hexamer repeat tract from an adjacent subterminal priming

site for multiple fosmids mapping to the same telomeres, and from this sequence could distinguish not more than two sets of closely related sequences from these fosmids for a given source genome (i.e., either G248 or ABC7). This gives us further confidence that, while the fosmids clones lose the distal part of the initially ligated telomere tract in the cloning and propagation of the fosmid in bacteria, the subterminal and immediately adjacent beginning of the terminal repeat tract are not affected by this deletion and carry an accurate copy of the subterminal genomic DNA.

This general strategy of directed sequencing off of these terminal fosmid templates was used to acquire high-quality sequence from fosmids containing alleles of the 8q and 18q subterminal sequence. Custom primers made according to the sequence of a reference allele were used to generate the first round of sequence reads from both strands and, following assembly, gaps and low-quality regions were filled by a second round of directed sequencing based on the assembly of the first round of reads. In cases where the gaps were too large to be filled by single reads, PCR amplicons spanning the predicted gap were prepared from the variant fosmid and sequenced. This method can be made quite efficient, especially since custom primers corresponding to high sequence identify regions of paralogous subterminal repeats can be used for multiple clones carrying similar subterminal sequences (Riethman 2008b).

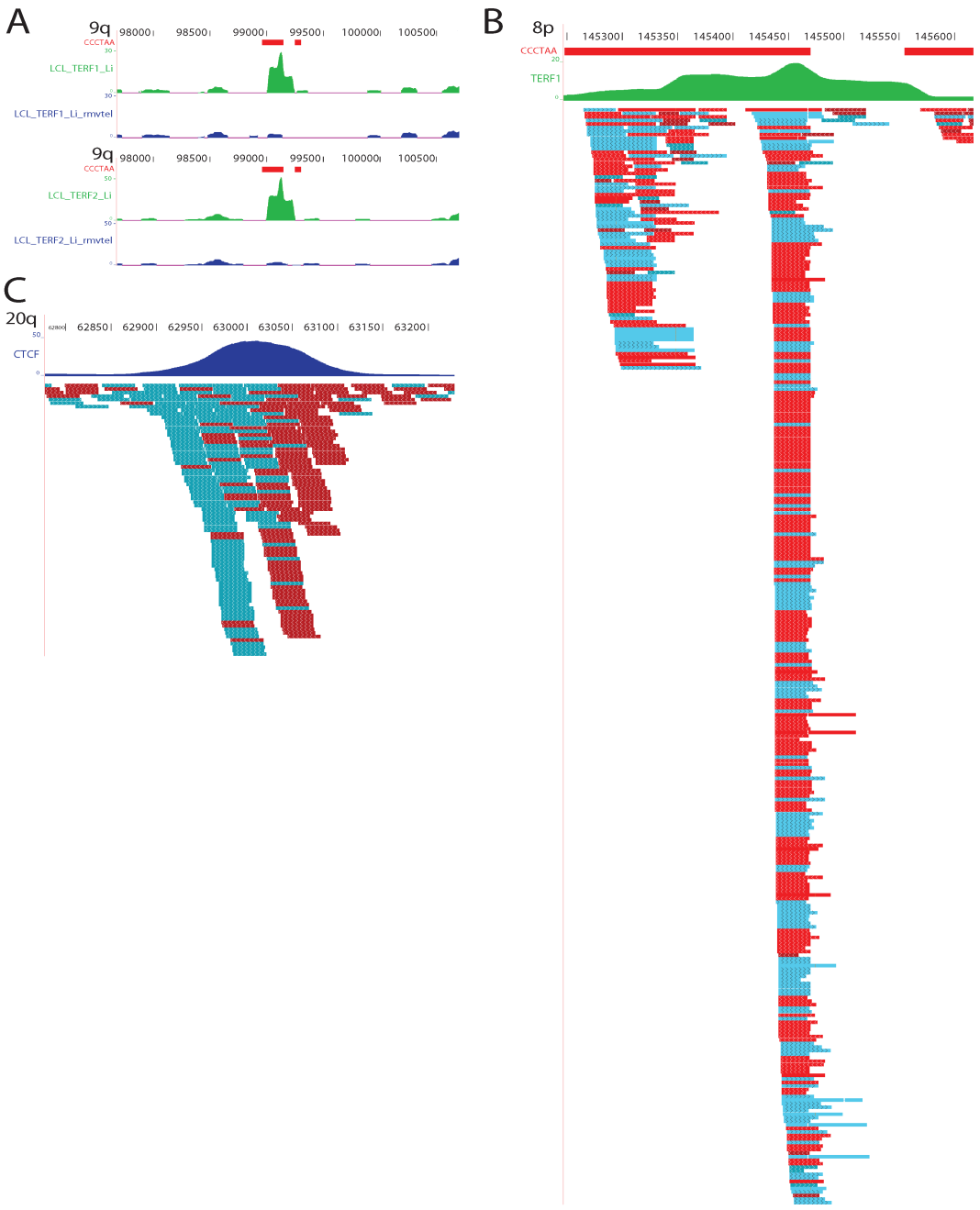

Figure 3.7 – TERF1 and TERF2 ChIP-seq peak analysis. 3A. Artifactual enrichment peaks at an Internal Telomere-like Sequence (ITS). Enrichment tracks from TERF1 and TREF2 ChIP-seq of DNA from LCLs are shown. Green: enrichment peaks for TRF1 (top) and TRF2 (bottom) based upon positions of uniquely mapped reads in the sample vs the control datasets. Blue: enrichment profile for same datasets following removal of telomere-like reads. 3B. TERF1 ChIP-seq read pile-ups at an ITS. TERF1 ChIP-seq reads mapped to an ITS prior to removal of telomere-like sequences. Note the random orientation of reads in the pile-ups and the abnormal peak shape. 3C. CTCF ChIP-seq reads mapping to a true binding site. Note the strand-specificity of the reads contributing to the central peak.

Large multipage figure available as figure S4 at
http://genome.cshlp.org/content/24/6/1039/suppl/DC1

Figure 3.8 – Annotated Subtelomeres (screen shots of all subtelomeres)

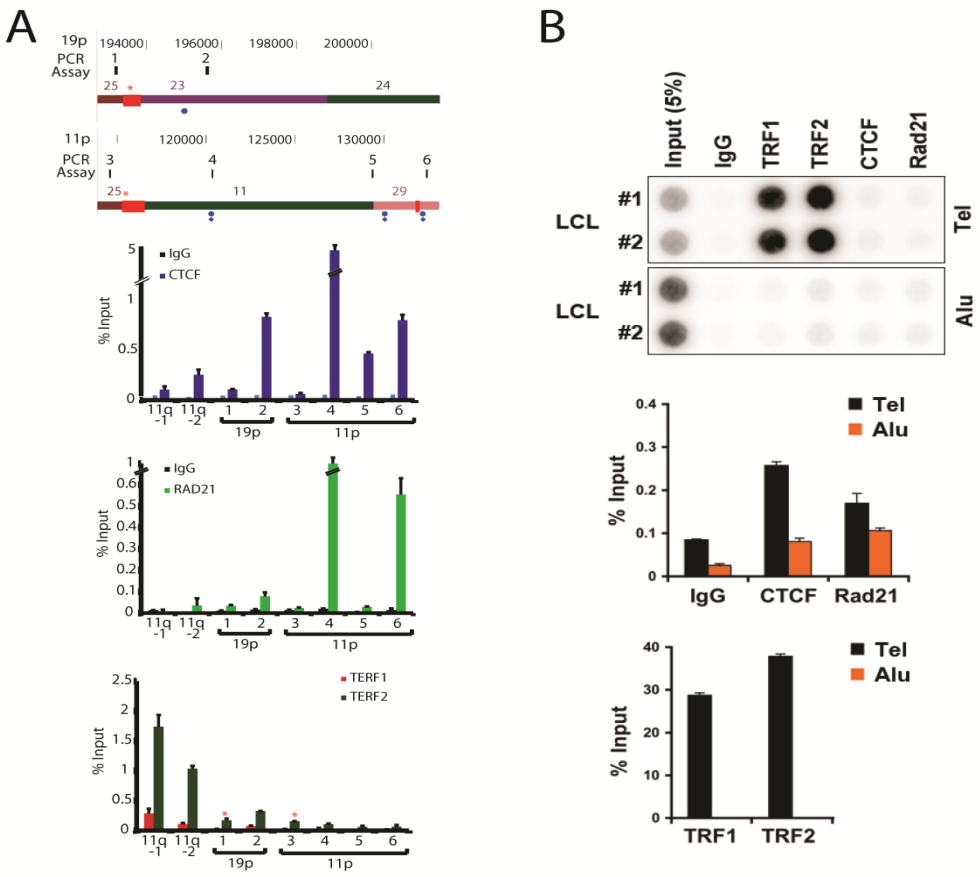

Figure 3.9 – ChIP analysis of CTCF, RAD21, TERF1, and TERF2 binding at subtelomeric candidate sites predicted by ChIP-seq dataset mappings. A) ChIP-qPCR analysis of factors binding at 19p and 11p subtelomeres in LCLs. Segments of the 19p and 11p subtelomeres are shown, with the coordinates (in bp) shown 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 asterix. The positions of co-localized CTCF and Cohesin (RAD21) peaks called in LCLs, ES, or IMR90 cells are shown as green (LCL only) or blue dots (all three cells), and 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 the bar graphs represent the average of % input (mean ± SD) for each ChIP from three independent ChIP experiments. qPCR assays for DNA immediately adjacent to the 11q telomere (primer sets 11q-1 and 11q-2) were used here as positive controls for TERF1 and TERF2 binding and a positive control for a previously validated subtelomeric CTCF /RAD21 co-localization site (11q-2). B) Dot-blotting was used as a control to validate the efficiency of TERF1 and TERF2 ChIP. ChIP DNA were dot-blotted, and assayed by hybridization with either ^{32}P-labeled $(TTAGGG)_4$ or ^{32}P-labeled *Alu* probe. Upper panels: a representative dot-blots was shown in duplicates; Lower panels: Quantification of dot-blots for indicated antibodies. Bar graph represents average values of % input for each ChIP (Mean ± SD) from three independent ChIP experiments

Library	Library Size	(CCCTAA)1	(CCCTAA)2	(CCCTAA)3	(CCCTAA)4	(CCCTAA)6
G248 raw	2,300,845	454,900	483	213	183	151
G248 q-processed	1,737,926	266,435	328	179	157	133
ABC7 raw	2,152,783	480,751	732	397	353	305
ABC7 q-processed	1,539,295	233,354	311	159	138	114

Table 3.3 – Telomere sequence screen of end-sequences from the G248 and ABC7 fosmid libraries.The number of reads with perfect matches identified for each (CCCTAA)n multimer from 1 to 6 is shown for the raw trace reads and for traces processed to remove low-quality bases. The end-sequence reads for fosmids fom the G248 and ABC7 libraries were downloaded from the NCBI Sequence Trace Archive. The sequence traces were quality-trimmed both from the 5' end and the 3' end , until a 30-base window contained at least 28 bases with a Phred Q value greater than or equal to 30. A simple pattern-match algorithm was used to identify perfect (CCCTAA)n stretches of at last the length shown in each column, in either the raw read or the quality-processed read. Both libraries contained fewer (CCCTAA)n sequences than expected from 12x clone coverage, and both were skewed towards loss of (CCCTAA)-containing sequences upon quality processing of sequence traces. The green boxes indicate the number of (CCCTAA)n-containing reads which matched at least (CCCTAA)4; nearly all of the mate-pairs of these reads mapped back to subtelomeric sequence in reference assemblies.

Tel	G248 fosmids mapped uniquely *	G248 Fosmids Span a terminal gap in Reference Sequence	New Structural Variant	ABC7 fosmids mapped uniquely *	ABC7 Fosmids Span a terminal gap in Reference Sequence	New Structural Variant
1p	-			-		
1q	6	-		6	-	
2p	5	-		3	-	
2q	0	-		2	+	+
3p	0	-		5	+	+
3q	0	-		0	-	
4p	19	-		33	-	

4q	3	+		5	+	
5p	4	-		5	-	
5q	0	-		0	-	
6p	0	-		0	-	
6q	0	-		0	-	
7p	0	-		2	Truncation	+
7q	3	-		9	-	
8p	0	-		0	-	
8q	4	-		5	-	
9p	0	-		0	-	
9q	0	-		0	-	
10p	0	-		8	+	
10q	2	-		3	-	
11p	0	-		0	-	
11q	3	-		5	-	
12p	0	-		4	+	+
12q	2	+		0		
13q	6	+		8	+	
14q	2	+		5	+	
15q	0	-		0	-	
16p	4	-		0	-	
16q	0	-		0	-	
17p	0	-		0	-	
17q	0	-		0	-	
18p	1	-		7	-	
18q	3	-		3	-	

19p	0	-		0	-	
19q	4	-		6	-	
20p	0	-		0	-	
20q	0	-		3	Truncation	+
21q	3	-		3	-	
22q	3	+		8	+	
Xp/Yp	3	-		4	-	
Xq/Yq	5	-		3	-	
Block 1	8			18		
Block 2	6			7		
Srpt	49			127		
Total mapped Reads	148			306		
No mate pair seq	13			33		

Table 3.4 – Mate-pair mappings of telomere fosmid end sequences from G248 and ABC7 to the human reference assembly. The mate pairs of (CCCTAA)-containing fosmid end sequences from 183 fosmids from the G248 library and 353 fosmids from the ABC7 library were mapped back to the reference subtelomere assemblies; all but a few mapped either uniquely to a known subtelomere assembly or to a known subtelomeric duplicon. Subtelomeric Blocks 1 and 2 correspond to specific classes of SREs characterized previously which contain DNA < 40 kb from the (TTAGGG)n tract at many subtelomeres (Ambrosini et al.,2007). Srpt identifies all other classes of SRE; telomeres in Supplementary Table 3.4 with 0 mate-pair matches have SREs that extend beyond 40 kb from the (TTAGGG)n tract and are thus only represented by fosmid ends in SREs. Comparison of these mate-pair mappings with the reference assemblies identified fosmids that should bridge existing subterminal gaps in the reference sequence and additional fosmids that appeared to represent structural variants of subterminal regions. The roughly 6-fold coverage increase of mappings for 4p is likely due to the contribution of acrocentric short-arm telomeres, which are known to have sequences highly similar to 4p (Youngman et al., 1992). One fosmid clone from clearly variant and/or gap-containing locus was sequenced.

Library	Library Size	(CCCTAA)4	Clones per allele
G248	2,300,845	183	1.9
G248 q-processed	1,737,926	157	1.7
ABC7	2,152,783	353	3.8

ABC7 q-processed	1,539,295	138	1.5
ABC8	3,888,476	784	8.5
ABC9	2,084,892	112	1.2
ABC10	2,121,489	238	2.5
ABC11	1,966,644	110	1.2
ABC12	2,366,708	215	2.3
ABC13	2,057,345	132	1.4
ABC14	2,089,193	619	6.7

Table 3.5 – (CCCTAA)4 - containing end sequences in Structural Variation Fosmid Libraries. The number of reads with perfect matches identified for (CCCTAA)4 in the raw trace reads is shown. Dividing the number of (CCCTAA)4-containing reads by 92 telomere alleles per genome gives the estimated clone coverage per telomere allele from this screen, assuming random coverage of telomere alleles. The relative efficiency of each screen in recovering (CCCTAA)n-containing clones depended primarily upon the quality of telomeric end-sequence available for that library, except for ABC8 which contained twice the number of clones as the other structural variation libraries.

Tel	ABC7	ABC7_clustering	Comments	ABC8	ABC8_Clustering	Comments	ABC14	ABC14_clustering	Comments
	# mate-pair reads	some reads binned by BLAST but not mapped within subtel (poor seq)		# mate-pair reads			# mate-pair reads		
1p	0						0		
1q	6	2_15-16K, 4_21-24K		9	8_30-40K, no clustering		15	27-38K, no clustering	
2p	7	1_24K, 6_33-38K		13	12_30-35K, 1_38K		17	32-41K, no clustering	
2q	0						0		
3p	5	4_16-24K, 1_33K		10	10_14-30K, no clustering		7	26-33K, no clustering	
3q	0						0		
4p	28	5_18-20K, 22_26-36K		42	6_18-20K, 34_25-41K		46	27-38K, no clustering	
4p4q	20	15--26K, no clustering		23	13-26K, no clustering		2		
4q	0			1			1	14K	
4q10q	7	3_16-18K, 1_34K		28	15_14-25K, 3_36K, 2_82K		25	3_18K, 10_28-29K, 9_34-37K	
5p	10	31-37K, no clustering		9	9_20-39K, no clustering		7	31-43K, no clustering	

5q	0						0		
6p	0						0		
6q	0						0		
7p	3	3_96-105	truncated allele(s)				0		
7q	4	1_20K, 3_35-40K		9	8_21-35K, no clustering		0		
7q_12q	6	4_27-30K, 1_40K		10	10_23-38K, no clustering		5	33-38, no clustering	
8p	0						0		
8q	9	2_24-25K, 7_31-41K		18	14_29-40K, 2_42-48K		10	39-43K, no clustering	
9p	0						0		
9q	0						0		
10p	1			1			1	35K	
10p18p	14	20-32K, no clustering		22	22_25-38K, no clustering		7	28-38K, no clustering	
10q	3	3_20-25K		5	4_24-25K		0		
11p	0						0		
11q	10	9_27-36K, 1_41K		13	9_25-40K, no clustering		20	32-40K, no clustering	
12p	2	2_8K					0		
12q	1			8	3_28-31K		1		
13q	8	1_15K, 7_20-28K		23	20_14-29K, no clustering		11	23-29K, no clustering	
14q	7	24-30K, no clustering		15	14_22-29K, no clustering		6	30-37K, no clustering	
15q	0						0		
16p	2	32-33K		10	1_13K, 9_22-35K		26	31-41K, no clustering	
16q	0						0		
17p	0						4	46-53K, no clustering	New 17p allele
17q	0						0		
18p	0			0			1	39K	
18q	5	1_19, 4_27-32		7	7_24-31K, no		20	22-34K, no	

					clustering			clustering	
19p	0			0			0		
19q	9	27-36K, no clustering		8	1_18K, 7_29-40K		12	31-41K, no clustering	
20p	0			0			0		
20q	5	72-85K, no clustering	truncated allele(s)	0			0		
21q	4	3_28-31K, 1_40K		21	20_23-38K, no clustering		11	35-40K, no clustering	
22q	9	26-37K, no clustering		17	17_23-36K, no clustering		9	27-39K, no clustering	
XpYp	6	2_119-120K_trunc, 4_DXYS20region	unreliable FES mappings s DXYS20 minisatellite	20	18_104-109K, 1_113K	truncation alleles	18	108-119K, no clustering	Truncation alleles
XqYq	3	32-42K		19	19_25-42K, no clustering		28	30-44K, no clustering	
Srpt_Block1	28	3_1K, 25_7-23K		76	16_4-11K, 61_14-25K		58	1_12K, 57_17-25K	
Srpt_Block2	16	16_1-14K		58	58_1-14K		44	44_3-14K	
Srpt_9p	18	24-36K, not clustered		34	32_26-39		49	30-40K, unclustered	
Srpt_6q	35	7_35-39K, 20_49-68K, 8_122-132K		79	1_24K, 15_34-42K, 39_55-67K, 5_122-124K		29	11_34-39K, 9_63-66K, 9_124-129K	
Srpt_2q	15	10-19K, unclustered		21	21_14-19K		50	13-21K, unclustered	
Srpt_8p	2	16K		2	1_8K, 1_36K		23	20_24-27K, 3_36K	
Srpt_12				31	5_1-3K,		2	9K	

p					24_8-17K				
Srpt_1p				0			18	26-28K, unclustered	
Srpt_7p				6	6_30-41K				
Srpt_17q				4	4_30-42K				
Srpt_unloc	10			6			1		
Srpt_total	124			317			274		
Acro1	3			5			3		
L1-nohit				3					
No_MP	33			63			21		
Total	354			743			608		

Table 3.6 – Subtelomere mapping distribution of mate-pairs of (CCCTAA)n reads from ABC7, ABC8, and ABC14 libraries. Mate-pair sequences of telomere-containing end sequences from these three libraries, which had the highest depth of coverage in identified telomere clones (Supplementary Table 3.5) were mapped to the human reference subtelomere assembly [5,7]. For the ABC7 library, a small set of additional reads identified by matches to (CCCTAA)2 and (CCCTAA)3 were added to those identified by (CCCTAA)4 to try and increase the depth of telomere clone coverage; but because of the high background of non-telomere clones identified with this strategy, it was abandoned after identifying about 50 additional bona fide telomere clones. Sequences from each library mapping uniquely were identified first, followed by those mapping to subtelomere duplicons known to exist on discrete pairs of telomeres (4p/4q, 4q/10q, 7q/12q, 10p/18p). The 4p and the 4p/4q reads are also shared by the acrocentric short-arm subtelomeres [10].

The remaining mate pair reads all contained SRE elements that cannot be mapped uniquely or to small discrete subtelomere subsets. These were characterized by mapping to SRE regions in the reference assembly in the following order, with matching mate pair reads removed from the remaining pool after each step: Srpt Blocks 1 and 2 [7]; then SRE regions of 9p; 6q; 2q; 8p; 12p; 1p; 7p; and 17q. The intent of the sequential mapping strategy was to compare the mapping patterns from each of the libraries and look for evidence of structural variation within the SRE regions. In addition, clustering of mate-pair mapping sites within the subtelomere assemblies was evaluated, again to gain insight into possible structural variation. While the clustering of mate-pair reads in SRE regions cannot define specific discrete structural variants (because the limited fosmid length does not permit mate-pair anchoring to single-copy DNA), the variable patterns of mate-pair mappings between genomes does provide confirmation of the high level of structural variation involving SRE regions [9] in these genomes.

Excel spreadsheet table available as table S5 at
http://genome.cshlp.org/content/24/6/1039/suppl/DC1

Table 3.7 – Clone-based Subtelomere Assemblies. For each subtelomere, the clone-based subtelomere assembly we used as well as its relationship to current GRC tpfs, and to additional clones mapping to subtelomere regions, are presented. See Separate Table 3.7 Excel Spreadsheet.

Chr	Start (bp)	End (bp)
chr1	1021369	248615650

chr2	510263	242625701
chr3	553282	197462429
chr4	510158	190543738
chr5	511808	180408075
chr6	560001	170563944
chr7	510233	158628563
chr8	510001	145798747
chr9	510354	140653430
chr10	556360	135024680
chr11	560001	134446463
chr12	619742	133341530
chr13	1	114600843
chr14	1	106792067
chr15	1	102021022
chr16	560028	89805666
chr17	508889	80695004
chr18	510616	77520492
chr19	552164	58618863
chr20	560001	62417956
chr21	1	47619786
chr22	1	50746323
chrX	639412	154759556
chrY	589412	58862562

Table 3.8 – Hybrid genome joining coordinates of hg19. Coordinates of hg19 to which the updated 500 kb subtelomeric assemblies were added prior to the annotation described here. Coordinates were identified by BLASTing the last (most centromeric) 10kb of the subtelomere sequence to hg19. The most internal coordinate plus one were the coordinates of hg19 sequence that was joined to the 500kb subtelomere seqeuence. The p-arm sequence of each 500 kb subtelomere assembly as given was attached at the p-arm coordinate, and the reverse complement of the 500 kb q-arm sequences were attached at the indicated q-arm coordinates.

Excel spreadsheet table available as table S7 at
http://genome.cshlp.org/content/24/6/1039/suppl/DC1

Table 3.9 – Datasets used in this study and quality metrics. Table includes all data set tracks and information on data set origins, their matched control, and the specific companies and product numbers for antibodies used. The abbreviations for the antibody-providing companies are: MIL, Millipore; sc, Santa Cruz Biotechnology; ab, Abcam; and BL, Bethyl Laboratories. Metrics include number of peaks called in hybrid genome using all reads, and only uniquely mapping reads. FRiP (Fraction of Reads in Peak), for all reads: partial reads (mapping likelihood) was counted in peaks called using all reads. FRiP for only unique reads is reads in peaks called using only uniquely mapping reads. PBC (PCR Bottleneck Coefficient) is the number of genomic positions with one read mapping to it (uniquely mapping or partial mapping), divided by the total number of genomic positions with at least one read mapping (uniquely mapping or partial mapping). NSC (Normalized Strand Cross-correlation coefficient) is the ratio of maximal cross-correlation value over the background cross-correlation. RSC (Relative Strand Cross-correlation coefficient) is the maximal cross correlation value minus the background cross-correlation, divided by the cross-correlation at the read length minus the background cross-correlation. For detail see **[292]** and http://genome.ucsc.edu/ENCODE/qualityMetrics.html.

Excel spreadsheet table available as table S8 at
http://genome.cshlp.org/content/24/6/1039/suppl/DC1

Table 3.10 – SRE boundary enrichment statistics. Raw Peak Counts of Subtelomere Boundary Enrichments. All Tables have columns corresponding to different boundary categories, and total for all boundaries in the SRE region. Boundary categories are 1copy/SRE (duplicon ends at unique subtelomere sequence), Gap (duplicon ends at most terminal complete sequence but not telomere), SRE/SRE (duplicon ends within SRE region), SRE/SD (duplicon ends at genomic duplicon), Terminal (duplicon ends at telomere), All_Bndries (all boundaries). Rows correspond to datasets. A. Raw Counts - Counts of peaks in association with different boundary categories and total. Additional row, Sequence, is the amount of sequence within the window of the boundary type, or the total SRE region. B. Percent – This table shows the percent of total peaks associated with a boundary type, for the total column this is always 100%. The additional row still corresponds to the amount of sequence within the window for the boundary type, the percent of sequence within the window of the boundary type out of the total SRE. This is the expected percentage of peaks in association, if the peaks are distributed randomly in the SRE. C. Enrichment – The ratio of percentage for a category and dataset over the expected percentage for that category. D. P Value – The p value calculated using a binomial test with the expected percentage as the probability of success, and the associated number of peaks and total number of peaks as success and trials.

3kb

Raw Counts

Cell Line	Cell Description	SRE/SRE (Internal)	Terminal	ITS	Total Peaks
HMEC	Primary	44	24	20	114
MCF-7	Cancer	45	22	16	91
SAEC	Primary	46	25	19	116
A549	Cancer	51	26	20	113
HRE	Primary	51	23	17	124
HEK293	Cancer	46	26	19	109
HRPEpiC	Primary	40	22	18	85
Weri-RB	Cancer	46	28	17	101
Sequence	-	1457453	90060	309076	3770157

Percent

Cell Line	Cell Description	SRE/SRE Internal)	Terminal	ITS	Total Peaks
HMEC	Primary	38.60%	21.05%	17.54%	100.00%
MCF-7	Cancer	49.45%	24.18%	17.58%	100.00%
SAEC	Primary	39.66%	21.55%	16.38%	100.

					00%
A549	Cancer	45.13%	23.01%	17.70%	100.00%
HRE	Primary	41.13%	18.55%	13.71%	100.00%
HEK293	Cancer	42.20%	23.85%	17.43%	100.00%
HRPEpiC	Primary	47.06%	25.88%	21.18%	100.00%
Weri-RB	Cancer	45.54%	27.72%	16.83%	100.00%
Expected	-	38.66%	2.39%	8.20%	100.00%

Enrichment

Cell Line	Cell Description	SRE/SRE Internal)	Terminal	ITS	Total Peaks
HMEC	Primary	1.00	8.81	2.14	1.00
MCF-7	Cancer	1.28	10.12	2.14	1.00
SAEC	Primary	1.03	9.02	2.00	1.00
A549	Cancer	1.17	9.63	2.16	1.00
HRE	Primary	1.06	7.76	1.67	1.00
HEK293	Cancer	1.09	9.99	2.13	1.00
HRPEpiC	Primary	1.22	10.84	2.58	1.00
Weri-RB	Cancer	1.18	11.61	2.05	1.00

P_value

Cell Line	Cell Description	SRE/SRE (Internal)	Terminal	ITS
HMEC	Primary	0.540728025	3.7319E-16	3.39208E-05
MCF-7	Cancer	0.023394174	2.7681E-16	0.000117401
SAEC	Primary	0.44736285	5.1024E-17	0.000128178
A549	Cancer	0.09468902	2.1781E-18	2.00843E-05
HRE	Primary	0.316283786	2.9931E-14	0.003450501
HEK293	Cancer	0.252883536	8.2942E-19	3.41008E-05
HRPEpiC	Primary	0.070614004	5.7555E-17	1.94656E-06
Weri-RB	Cancer	0.094350627	4.5939E-22	6.87747E-05

Table 3.11 – CTCF boundary analysis for 4 primary and 4 immortalized cell lines.
Boundary analysis for CTCF in additional cell lines.

Primer Name	Sequence	PCR Assay	Number of copies with amplicon match	Known Loci
		Figure 4A		
6q-1-For	GGCAGCAAACGGGAAAGA	1	10	6q,1q,2q,5q,10q,13q,16q,21q,22q (Tel)
6q-1-Rev	TGCCTGCCTTTGGGATAACT	1		
6q-2-For	CAGAGACGAGTGGAACCTGAGTAAT	2	14	6q,1q,2q,4q,5q,8p,10q,13q,16q,19p,19q,21q,22q,(Tel); 2qfus (internal)
6q-2-Rev	TGGGCAAGCTGGTCCTGTAG	2		
6q-4-For	GGCAGCTACGTCCTCTCTTGA	4	7	6q,5q,1p,8p,17q,2q,16q (Tel)
6q-4-Rev	GCAAACTAAGCAACAATGAAACAGA	4		
6q-5-For	CCTGATGGAGTCTAAATGCAGTGA	5	5	6q,5q,1p,8p,17q (Tel)
6q-5-Rev	TCCATCCACCCCCTCCTT	5		
6q-3-For	TTCTTACTTATCAGGGTGCTCATCTACT	3	6	6q,5q,1p,8p (Tel); chr1,chrY (internal)
6q-3-Rev	GTCCCTCCAAGGAAAATTCCA	3		
6q-6-For	CCCTGGGTGCTTCACCATT	6	6	6q,5q,1p,8p (Tel) chr1,chrY (internal)
6q-6-Rev	GAAGAATTTAGTGAAGGGTCAGTTTACA	6		
16q-8-For	GGCAGCTACGTCCTCTCTTGA	8	3	16q,2q,8p (Tel)
16q-8-Rev	GCAAACTAAACAACAACAATGAAACA	8		
16q-9-For	GGCTGCCACCTGCTGTTG	9	6	16q,2q,1p (Tel); chr1,chrY, chr10 (internal)
16q-9-Rev	TGCTCTCCAGTCCAGTGTTCTG	9		
16q-10-For	CGGTGGATCTCCGAAGTTCA	10	4	16q,7p,9q,3q (Tel)
16q-10-Rev	GGCTTCAGCTGGTTTTTCAAA	10		
16q-7-For	CTGGAAGCACCCCCACTTC	7	4	16q,17p,11p,7p (Tel)
16q-7-Rev	CAGTCATTTGGCCCCGTAA	7		
		Figure 4B		
Xyq-1-For	CCCCTTGCCTTGGGAGAA	1	3	XqYq,9p,19p (Tel)

Xyq-1-Rev	GAAAGCAAAAGCCCCTCTGA	1		
XqYq-2-For	GGTGGAACTTCAGTAATCCGAAA	2	5	XqYq,9p,15q,16p,19p (Tel)
XqYq-2-Rev	AGCAAGCGGGTCCTGTAGTG	2		
Xq-3-For	TCCCCGTGCCCTAATGG	3	7	XqYq,9p,12p,15q,16p,19p (Tel); 2qfus (internal).WASH gene
Xq-3-Rev	TGAGCCCCCTGCACACA	3		
Xq-4-For	CACGCACCGCGTCTCA	4	2	XqYq,16p (Tel)
Xq-4-Rev	TCCTCATAGTGGCCGCAAA	4		
17p-5-For	CAAGGATCTTGGTCTTCACAGAGA	5	2	17p,11p (Tel)
17p-5-Rev	GCTGATGGCATCCACATGAC	5		
17p-6-For	CCCCCAGGGCCTTCAAC	6	2	17p,11p (Tel)
17p-6-Rev	GGCTTGAGGTACATCTTCCATCA	6		
17p-7-For	CGGAGAAGGCTGCTATTGGA	7	1	17p (Tel)
17p-7-Rev	GAGCTCGCCACCTTCTTGTT	7		
17p-8-For	GGGTTAAGCAGTGCACGAGAGT	8	1	17p (Tel)
17p-8-Rev	AACCTCCCGATGCATGGA	8		
		Supplementary Figure 5		
11q-1-For	TGCGGCCCCGAATTG	11q-1	1	11q (Tel)
11q-1-Rev	GTTTCTCAGCACAGACCTTGGA	11q-1		
11q-2-For	GAGACCTGATGTCCCAATTCTTAAC	11q-2	1	11q (Tel)
11q-2-Rev	CTGTGTTCTTAGAGAGTGTTTCTTGGA	11q-2		
19p-1-For	TGAGGCAGTGAAGGACGTAGAG	1	11	19p,1p,3q,7p,8p, 9q,11p,16q,19q,22q (Tel), 2qfus (internal).
19p-1-Rev	AACGGGCTTCCAGGAGCTA	1		
19p-2-For	TCCTGCCTCTGTCTCAAGTCTATG	2	2	19p,18p (Tel)
19p-2-Rev	CAGGGACCTAAGGCAGTAGCA	2		
11p-5-For	CACCAGCATTGTCCCCACTA	5	12	11p,1p,3q,4p,4q,7p,8p,9q, 16q,17p,19p,19q,22q (Tel),2qfus (internal)
11p-5-Rev	CCACAACCCCGAGCATACTG	5		
11p-6-For	GTGGCGCCATGGTTCAG	6	6	11p,17p,7p,16q,3q,9q (Tel)
11p-6-Rev	CACATCCGCCGAGAAACTG	6		
11p-7-For	TACCCTGGCCACCTTGGA	7	6	11p,17p,7p,16q,3q,9q (Tel)

11p-7-Rev	AAGATTAAGGACACGACCATGACA	7		
11p-8-For	CAACCCCAAGCCCTCCTT	8	2	11p,17p (Tel)
11p-8-Rev	GGGACAGGCTTGAGGTACATCT	8		

Table 3.12 – ChIP-qPCR primers used. PCR primers used for ChIP-qPCR assays corresponding to the PCR fragments shown in Fig. **3.5** and Supp. Fig. **3.6**. Each primer is shown next to its mate with sequence and the name of the PCR product as indicated on the figure. The number of distinct genomic loci matching the primer set computationally is given. The genomic locations of each of these loci are indicated; "Tel" indicates the SRE regions shown in Figure **3.**1. The number of non-subtelomeric duplicated sites (internal), where they exist for an assay, are as indicated.

CHAPTER 4: MOUSE SUBTELOMERE ANALYSIS

4.1 Abstract

Human subtelomeres contain sequences important for telomere length regulation and stability, including transcriptional regulatory sequences for telomeric repeat containing RNAs (TERRA). However, relatively little is known about the sequence organization and function of DNA adjacent to mouse telomeres. Here, we analyze mouse subtelomeric sequence features in the context of segmental duplications, annotated genes, and position relative to the mouse terminal telomere tract. We adapt our recently-described multi-read mapping pipeline for analyzing massively parallel short-read datasets to the mouse genome, and display the results of these annotations relative to the start of mouse telomeres in a custom mouse subtelomere browser (http://vader.wistar.upenn.edu/mousesubtel). We find very little similarity in mouse subtelomere sequence organization as compared to human. Mouse lacks the enrichment for telomere-adjacent co-localized CTCF/cohesin sites shown to be important for TERRA regulation and telomere protection in human, and also lacks internal telomere-like sequence (ITS)-associated and subtelomeric repeat element (SRE) boundary –associated CTCF/cohesin sites found in human subtelomeres. Instead of the large, recently duplicated and telomerically oriented SREs typical of human subtelomeres, we found smaller, more ancient and more randomly oriented SREs at mouse subtelomeres. We found very high enrichment of MurSatRep1and MM4SAT interspersed repeat families at mouse subtelomeres, and mapping of ribosome-depleted RNAseq datasets from the mouse ENCODE project revealed clustering of MurSatRep1-containing transcripts near mouse telomeres. The MurSatRep1 repeat element had previously been shown to be enormously enriched in mouse lincRNAs (Kelley and Rinn, 2012); interestingly, the recently-described mouse TERRA locus found in the 18q subtelomere [299] contains one of the MurSatRep1-containing transcripts we detect here. Our results suggest little conservation between mouse and human subtelomeres and widely disparate mechanisms of cis-regulation of telomere length and stability.

4.2 Introduction

Telomeres are extraordinarily dynamic chromosomal structures. They are essential for genome stability and faithful chromosome replication, and mediate a host of key biological activities including cell cycle regulation, cellular aging, movements and localization of chromosomes within the nucleus, and transcriptional regulation of subtelomeric genes [300,301]. Specialized functions involving telomeric and subtelomeric DNA have evolved in a wide range of eukaryotes: for example, frequent subtelomeric gene conversion provides diversity for surface antigens in Trypanosomes [302], and rapidly-evolving subtelomeric gene families confer selective advantages for closely-related yeast strains [303].

A conserved, (TTAGGG)n tract forms the DNA component of each chromosome terminus in humans and other vertebrates [6,8]. Telomerase-associated and telomerase-independent pathways for maintaining (TTAGGG)n repeats exist; the major telomerase-independent pathways are recombination-based, sometimes involve co-amplification of subtelomeric sequences along with the simple repeat tracts found at chromosome termini (Bryan et al., 1995; Henson et al., 2002; Lundblad and Wright, 1996; Marciniak et al., 2005) and can generate very long and heterogeneous stretches of (TTAGGG)n-containing repeats [305,308]. Transcription of subtelomeric genes can be regulated by (TTAGGG)n tract length [309,310] and by subtelomeric repeat content and abundance, possibly by contributing specific sequence elements necessary for local silencing[311,312] or by providing extended homology regions required for somatic pairing and heterochromatin formation [313].

Subtelomeric DNA, along with pericentromeric chromosome regions, are preferential sites of segmentally duplicated DNA. Estimated to comprise approximately 5% of the human and mouse genomes [314,315], this class of low-copy repeat DNA is characterized by very high sequence similarity (90 % to >99.5 %) between homology tracts, and variable but sometimes very large tract lengths (1 kb to > 200 kb). Segmental duplications can pre-dispose associated chromosome segments to genetic instability and have been connected with many genetic

diseases [316,317]. Evolutionarily recent duplicative transposition of these large DNA tracts has led to the generation of new gene families and to the formation of fusion transcripts with potentially new functions [166].

In the human genome subtelomeric DNA is highly enriched in segmental duplications [315,318]. These regions of human chromosomes contain mosaic patchworks of duplicons [162,319,320] apparently generated by translocations involving the tips of chromosomes, followed by transmission of unbalanced chromosomal complements to offspring [321]. Along with highly elevated sister chromatid exchange (SCE) rates in subtelomeres [322], these studies indicate that human subtelomeres are duplication-rich hotspots of DNA breakage and repair as well as rapidly-evolving regions of the genome.

A telomeric repeat-containing family of RNAs (TERRA) is transcribed from subtelomeres into the (TTAGGG)n tracts [323–325] and, along with telomere-associated shelterin proteins, is believed to form an integral component of a functional telomere; perturbation of its abundance and/or localization causes telomere dysfunction and genome instability [323,326]. Subtelomeric DNA elements regulate both TERRA levels and haplotype-specific (TTAGGG)n tract length and stability [326–330], with accumulating evidence for specific epigenetic modulation of these effects [326,327,331–333]. Heterogeneously-sized TERRA transcripts with as yet ill-defined transcription start sites and potential splice patterns originate in many, perhaps all human subtelomere regions [323,325,334,335], with the sizes of the larger transcripts (greater than 15 kb) suggesting structural overlap with some transcribed subtelomeric gene families [336,337]. 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 chromatin organizing factors CTCF and cohesin are integral components of most human subtelomeres and important for the regulation of TERRA transcription and telomere end protection [334].

In contrast to human subtelomeres, relatively little is known about the DNA sequences adjacent to mouse telomeres. Here, we analyze mouse subtelomeric sequence features in the context of segmental duplications, annotated genes, and position relative to the mouse terminal telomere tract. We find very little similarity between mouse and human subtelomere sequence organization. Mouse lacks the enrichment for telomere-adjacent co-localized CTCF/cohesin sites shown to be important for TERRA regulation and telomere protection in human. We found very high enrichment of MurSatRep1and MM4SAT interspersed repeat families at mouse subtelomeres, and mapping of ribosome-depleted RNAseq datasets from the mouse ENCODE project revealed clustering of MurSatRep1-containing transcripts near mouse telomeres. The MurSatRep1 repeat element had previously been shown to be enormously enriched in mouse lincRNAs [338]; interestingly, the recently-described mouse TERRA locus found in the 18q subtelomere [299] contains a MurSatRep1-containing transcript we detect here. Our results suggest little conservation between mouse and human subtelomeres and widely disparate mechanisms of cis-regulation of telomere length and stability.

4.3 Methods

4.3.1 Sequence Feature Annotation

SRE and SD annotation were carried out as described previously [162] Duplicon boundaries were defined as the end positions of duplicon 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 declared a single boundary for analysis purposes. Subtelomere sequence assemblies were analyzed with RepeatMasker [284] and Tandem Repeats Finder [285]. The ITS track was built from repeatmasker results, showing both (CCCTAA)n and (TTAGGG)n hits. The MurSatRep1 and MMSAT4 track was also built from repeatmasker results. Ensembl transcripts [339], and RefSeq genes [340] were aligned to subtelomeres using Spidey [288]. Refseq genes require a 99% identity match to be shown on the mm9 assembly, but when transferred using lift over coordinates to mm10 they did not always retain the same

matching percentage with the new reference sequence. For this reason we re-mapped the ref seq genes to mouse subtelomeres and provide a toggle for displaying those that match with different levels of sequence identity, with 98 % used as the default.

4.3.2 Subtelomere Sequence Characterization

Repeatmasker [284] results were used to characterize the repeat content of the mouse subtelomere and compare with the rest of the mouse genome. Repeatmasker was run on the subtelomere sequences and combined with results for the remainder of the genome which is available from the UCSC browser for mm10. Sizes of repeat regions were calculated from repeatmasker bed tracks, and the amount of segmentally duplicated DNA was calculated as the size of our defined duplicon blocks. Total duplicated sequence counts each pairwise alignment relative to each query subtelomere sequence and calculates the % divergence from each, so this analysis permits some duplicons to be counted multiple times.

Short-read-based Annotation. Datasets analyzed in this study are listed with their specific sample and control GEO accessions, as well as the specific antibodies used and their sources as listed in their GEO entries, in Supplementary Table 2. Data sets were downloaded as raw data FASTQ files from the ENCODE project [289] through the UCSC portal or the GEO database through the NCBI short read archive. The majority of the data sets are from the ENCODE labs of Snyder (Sydh/Stanford) and Ren (LICR), . Our previously described multimapping ChIP-Seq pipeline (Stong et al., 2014) was employed. Briefly reads were aligned to the hybrid genome using BWA 0.6.2 [239], allowing multimapping up to 101 locations (-n 101). A mapping likelihood (ml) weight was used to uniformly distribute a read to all possible mapping locations. The mappings are then used to build a genome wide fragment density. Signal tracks were built as the fold enrichment of signal over a control. Peak calls were made using MACS 2.0.10 using the sample and control bedgraphs. First bdgcmp –m ppois, was called setting ppois as the method, calculating p value tracks. Peaks were called using bdgpeakcall –l 50 –c 4, setting minimum peak length to 50, and a p value significance cut of 4 (10-4) [291]. Overall quality and mapping metrics for the datasets were determined as described [341] and are

included in Supplementary Table 2. RNA-seq datasets were mapped using TopHat(Trapnell et al., 2009) which allows for multimapping reads. Tracks were then built using the same uniform weighting of mapping postions and the ChIP-seq tracks and were normalized to reads per million.

4.3.3 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/mousesubtel). The entire subtelomere region of interest is displayed by typing it in the format chrNq:1-500000. The subtelomere browser has similar navigation and mapped dataset selection functionalities as the UCSC Genome Browser [293].

4.3.4 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.

4.4 Results

4.4.1 Telomeres and subtelomeres in the mouse genome

The current mouse genome assembly (Build 38/mm10) includes subtelomere DNA immediately adjacent to the (TTAGGG)n tract for all q-arm telomeres except 4q. Prior to the completion and release of build 38, we had isolated multiple telomere sequence-containing

clones for each of the 1q, 2q, 3q, 5q, 6q, 7q, 10q, 11q, 12q, 13q, 14q, 16q, 17q, 18q, 19q, and XqYq telomeres using either a novel telomere-junction-specific screen of the CH25 BAC library or a computational screen of end-sequences from two fosmid libraries (see Supplementary Figure 4.7 and Supplementary Table 4.2). The same screen yielded singleton telomere-containing clones that contained telomere-adjacent DNA mapping to 8q, 9q, and 15q. Single subtelomere localization sites were confirmed by FISH analysis of screened BAC clones from the 1q, 5q, 6q, 7q, 13q, 14q, 18q and XqYq subtelomeres, and multiple localization sites including the expected subtelomere were found for clones from the 10q, 11q, 16q, 17q and 19q subtelomeres. Representative, telomere-adjacent clones from this screen were fully sequenced and incorporated into build 38 (1q, 3q, 6q,7q, 8q, 9q, 10q, 11q, 12q, 13q, 14q, 15q, 16q, 17, 18q, 19q, XqYq; Supplementary Figure 4.7 and Supplementary Table 4.2). Representative telomere-adjacent DNA from the p-arm telomeres (all of which share a short subtelomeric sequence bridging from the telomere to the centromere) was taken from Choo and co-workers [342]. Finally, three overlapping fosmid clones containing a telomere and members of the Vmn2r gene family were assigned to 4q (it being the only remaining telomere), and the longest of the three fosmid sequences was appended to the distal end of the build 38 4q assembly.

For all of the q-arm telomeres, the string of NNNs following the last clone sequence in Build 38 was removed, as was the distal telomere tract from the sequenced telomere clone, so that the last base at these telomeres corresponds to the start of the terminal repeat tract. Thus, upon display of the reverse complement of the last 500 kb of each chromosome q-arm, coordinate 1 corresponds to the start of the terminal (TTAGGG)n tract on the strand oriented towards the centromere (to maintain a consistent starting coordinate for subtelomere annotation). Except for the addition of the p-arm subtelomere (represented by a single 13 kb sequence appended to build 38, where the telomere sequence was trimmed to coordinate 1 of the centromerically oriented strand); and the above-mentioned addition of 29 kb of subtelomere sequence to 4q, the modified mouse genome analyzed here is identical to build 38 with the

telomere sequence trimmed to the start of the terminal repeat tract. It incorporates all mouse telomere-adjacent DNA .

4.4.2 Subtelomere Annotation

The modified mouse build 38 genome was used to annotate subtelomeric sequence features as described in Ambrosini et al (2007), and to map available mouse next-gen sequence datasets of potential interest for mouse subtelomere function based upon our previous studies of human subtelomeres [334]. Figures 4.1 and 4.2 illustrate these annotations for the first 175 kb of the mouse 8q, 13q, 2q and 17q telomeres; Supplementary Figure 4.8 displays these annotations for all of the mouse subtelomeres, and the mouse subtelomere browser permits viewing and selection of tracks for these and additional datasets (http://vader.wistar.upenn.edu/mousesubtel).

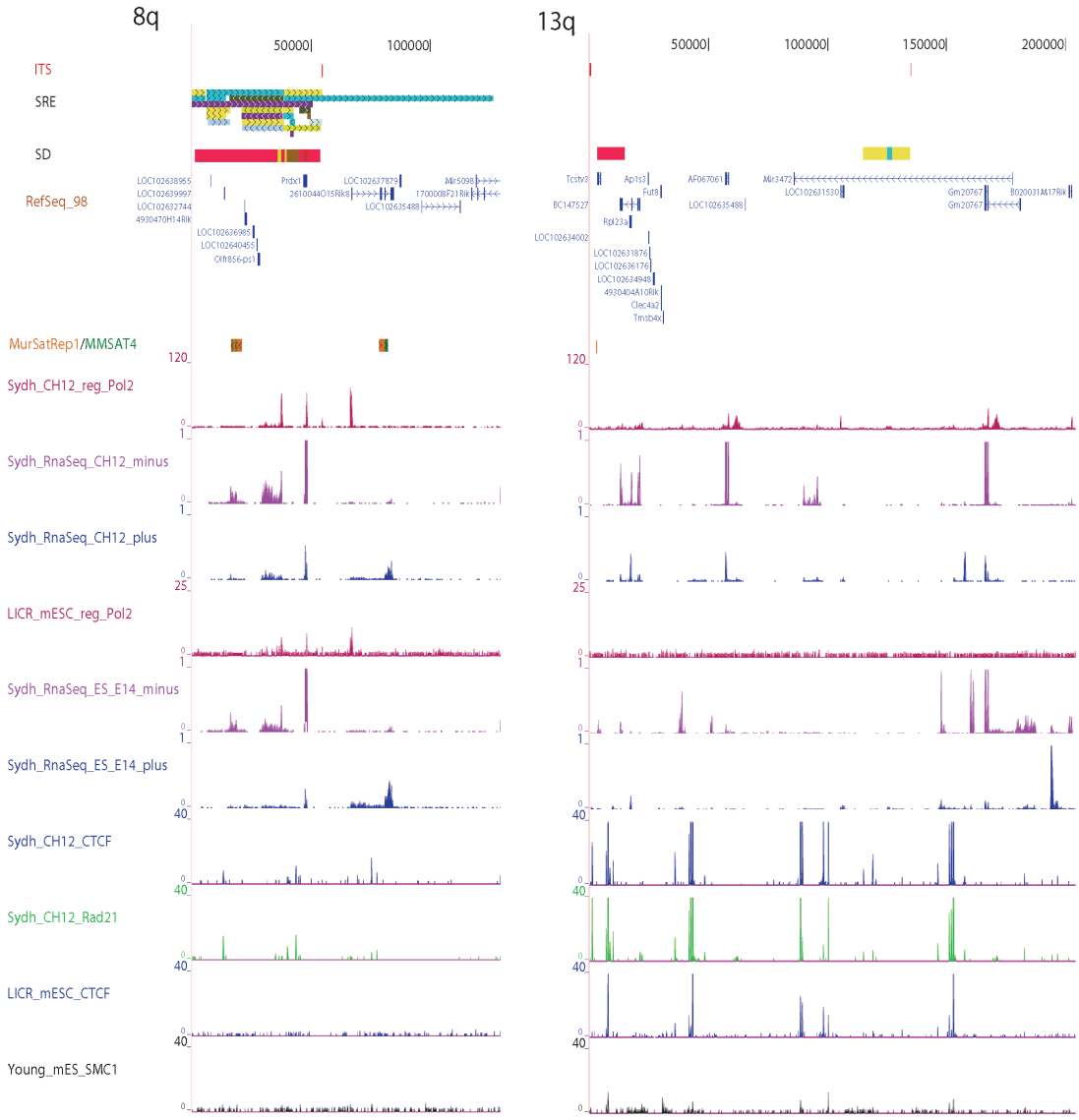

Figure 4.1 – Subtelomere Annotation Features. The first 175 kb of the 8q and 13q mouse subtelomeres are shown to illustrate key features of subtelomere sequence organization annotated on our browser. Coordinate 1 on the browser corresponds to the centromeric end of the terminal repeat tract (i.e., the last (CCCTAA)n repeat unit before subtelomere DNA starts). The 140 kb long SRE region on 8q is subdivided into duplication modules ("duplicons") defined by segments of similarity (> 90% nucleotide identity, > 1kb in length) between 8q and other subtelomeres (Ambrosini et al., 2007). Each rectangle represents a separate duplicon. Duplicated segments are identified by chromosome (color) as described previously (Ambrosini et al., 2007); additional detail included on the live browser but omitted in Fig. 1 for the sake of clarity include the subject subtelomere identity, starting and ending coordinates of the duplicon in the subject subtelomere sequence, and % nucleotide sequence similarity of non-RepeatMasked sequences from the duplicon segment of the subject subtelomere to 8q (http://vader.wistar.upenn.edu/mousesubtel). The 13q subtelomere lacks SREs. Segmental duplications (SD track; shown as compressed into single rectangle in Fig 4.2, can be separated into single duplicons in 'full" mode on the browser) correspond to duplicons with non-subtelomeric localization sites. These can be in addition to SRE copies or exclusively non-subtelomeric copies (e.g., the SDs on 13q). The Internal Telomere-like sequence (ITS) islands as defined in Methods are indicated as red tics on the ITS track. Gene models for transcripts included in the RefSeq (Pruitt et al., 2012)

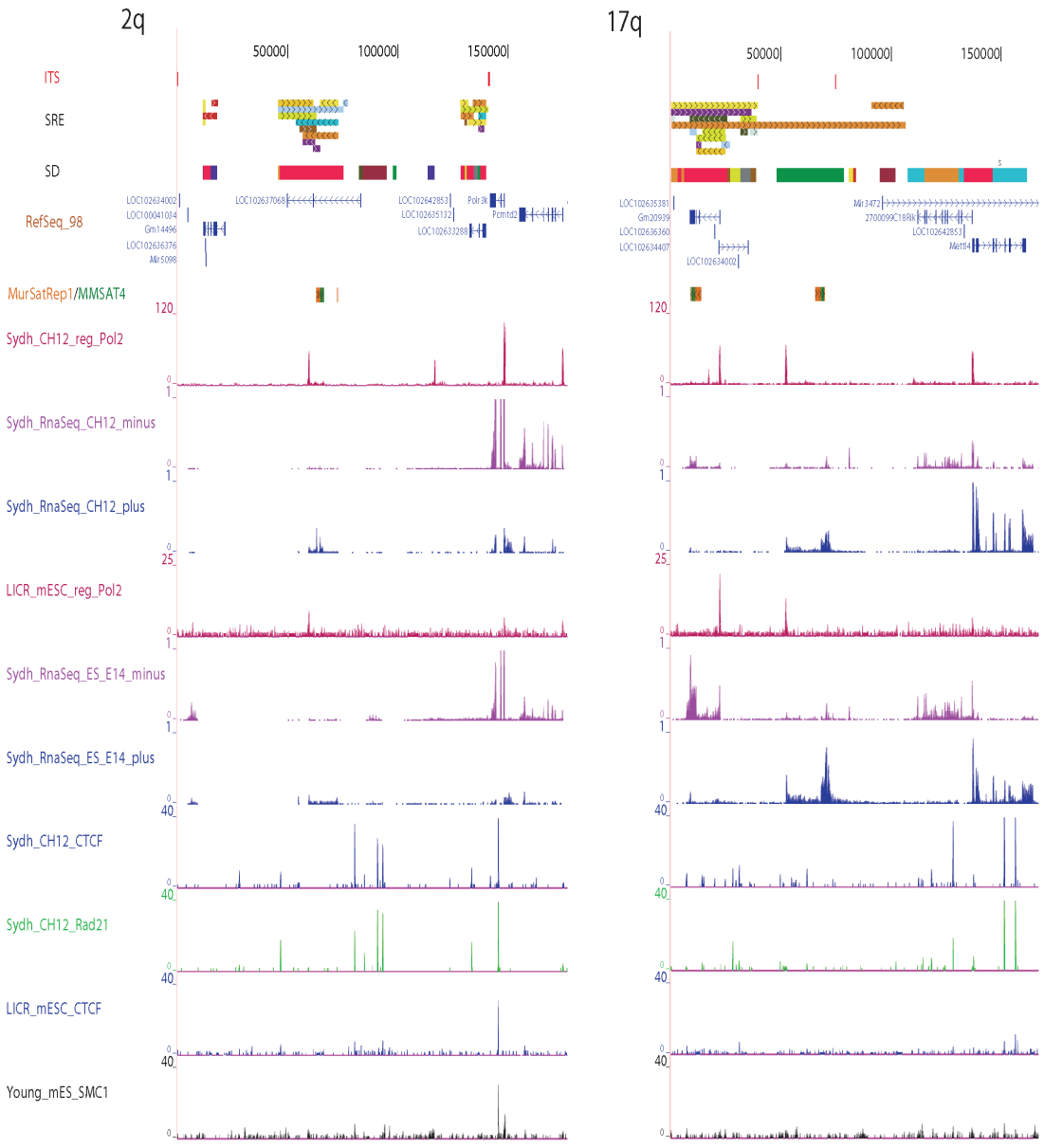

Figure 4.2 – Examples of annotated subtelomeres 2q and 17q. Examples of annotated subtelomeres 2q and 17q. Track descriptions are the same as for

Subtelomeric duplications were defined and displayed according to chromosome color as described in Ambrosini et al. 2007. The sequence of the most distal 500 kb of each mouse subtelomere region in the modified Build 38 genome was used to query the the entire mouse genome (see Methods). Adjacent and properly oriented BLAST matches with ≥ 90 % nucleotide

sequence identity and ≥ 1kb in size were assembled into chains as described Ambrosini et al. 2007. The query sequence and each aligned region identified in this manner were termed "duplicons" defined by that query, and this set of homologous sequences is a single "module". Each module was thus defined by a set of pairwise alignments with the query subtelomere sequence, and a % nucleotide sequence identity of each chained pairwise alignment was derived from the BLAST alignments. If a duplicon identifies more than one subtelomere region these are displayed in the subtelomeric repeat element (SRE) track, and if it identifies non-subtelomeric copies these are displayed in the segmental duplication (SD) track.

	100kb		500kb	
	mSubTel	hSubTel	mSubTel	hSubTel
SRE only	8.35%	27.54%	2.40%	9.50%
SD only	7.32%	5.28%	5.45%	5.27%
SRE/SD	19.80%	24.63%	5.01%	7.08%
Total Duplicated	35.47%	57.45%	12.86%	21.85%

Table 4.1 – Segmental Duplication Content of the Mouse Subtelomere. Percent of most distal 100 and 500kb that are either exclusily SRE, SD, or both, and the total segemental duplication amount.

A comparison of mouse and human subtelomeric duplications reveals several striking differences. First, mouse SRE regions are about half as abundant as human (7.41 % of subtelomere vs 16.58 % of subtelomere), whereas mouse SD regions cover about the same fraction of subtelomeres as do human SDs (10.45 % vs 12.34 %). (Table I). Second, whereas human SREs are nearly universally oriented in the same direction relative to the telomere (Linnardopoulou et al., 2005; Ambrosini et al., 2007), mouse SREs have a much higher fraction of duplicons oriented in the opposite direction as well (Figures 4.1 and 4.2, Supplementary Figure 4.8; http://vader.wistar.upenn.edu/mousesubtel), which is inconsistent with the model proposed for generation of human SREs [321]. Most striking, however, is the much greater part of the subtelomere occupied by very high similarity (less than 3 % divergence in pairwise alignments) SREs in human relative to mouse (Figure 4.3), indicating that most human SREs arose much more recently than mouse SREs.

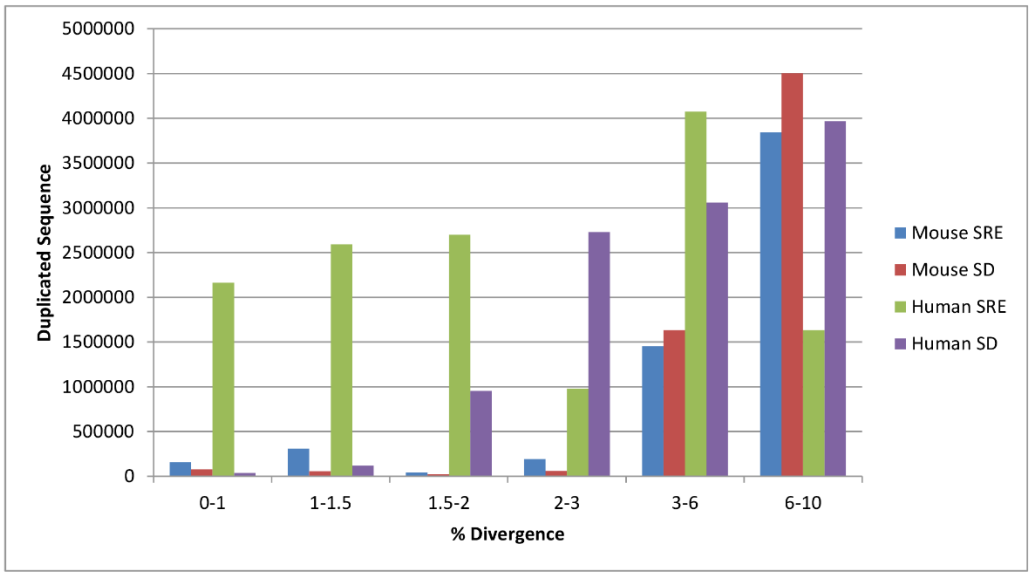

Figure 4.3 – Comparison of duplicated sequence in the mouse and human subtelomere. Total pairwise alignments of sequences involved in segmental duplications within the subtelomere (SRE) or internally in the rest of the genome (SD), in both mouse and human.. Sequence count is grouped by percent divergence between aligned pairs of duplicated sequences, and total mouse subtelomeric sequence (in bp) is normalized to total human subtelomere sequence.

Similarly, there are dramatic differences in the internal telomere-like sequence (ITS) distribution between mouse and human subtelomeres (Figure 4.4). Whereas in human subtelomeres most ITSs are in SRE regions, in mouse subtelomeres the vast majority of ITSs are in either 1-copy regions or in SDs, with only a handful localized to SREs. Human subtelomeric ITSs are distributed over a broader range of lengths and are highly enriched at subtelomeres relative to the rest of the genome, whereas mouse ITSs tend to be much shorter (especially those localizing to SREs) and not as enriched at subtelomeres relative to internal genomic sites.

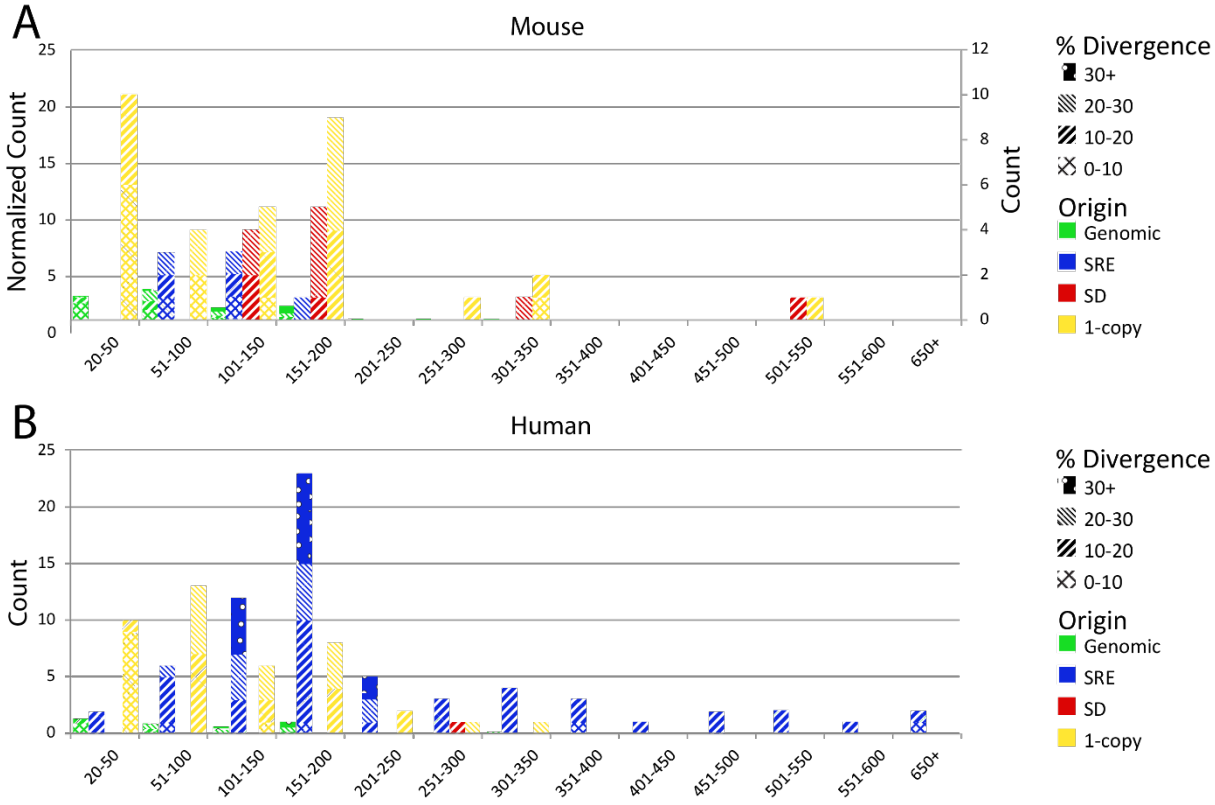

Figure 4.4 – Characteristics of interstitial telomere sequences in the mouse (A) and human (B) genome. ITS counts are shown corresponding to their length (x axis), percent divergence (pattern), and location in the genome (color). Sequences are identified as occurring within subtelomere repeat elements (SRE) subtelomere segmental duplication (SD), single copy subtelomere (1-copy), or internal genomic sequence. Counts of internal genomic occurrences were normalized to the size of the subtelomere, and mouse subtelomere content was normalized to human. Percent divergence shows the similarity of the ITS sequence to a perfect (TTAGGG)n sequence.

Similar to human subtelomeres [318], the overall GC content, SINE, LINE, and DNA transposon composition of murine subtelomeres was similar to the genome-wide average and not heavily skewed between SREs, SDs, and 1-copy DNA (Figure 4.5). LTR content was somewhat enriched (about double the genome average) in mouse SREs and SDs. Strikingly, there were very high enrichments of the frequently co-occuring MMSAT4 and MurSatRep1 repeat elements in both the SRE and SD regions (e.g., greater than 200-fold enrichment for MurSatRep1 in SREs; Figure 4.5).

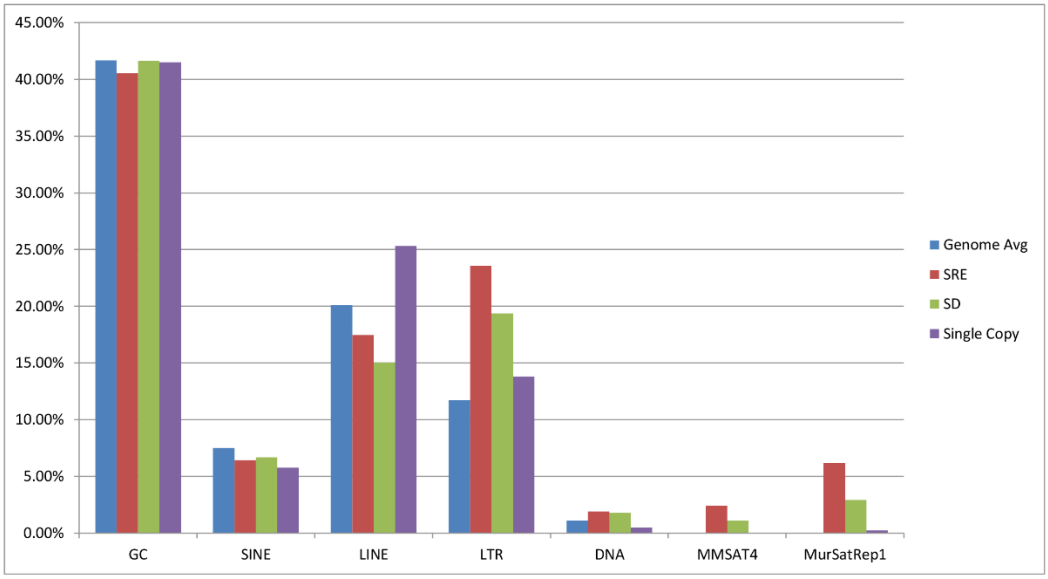

Figure 4.5 – Sequence composition of murine subtelomeres. The percent of sequence defined as GC content or an identified interspersed repeat for the entire genome (blue), and the specified subtelomere regions, Subtelomeric Repeat Elements (red), Segment Duplication (green), Single Copy (purple). Genome wide averages were calculated from mm10.

4.4.3 Annotation of subtelomeric CTCF and cohesin binding sites using ChIPseq datasets.

Recent work from our group indicates that chromatin organizing factors CTCF and cohesin are integral components of most human subtelomeres and important for the regulation of TERRA transcription and telomere end protection [334]. Co-localized CTCF/cohesin sites map within 3 kb of the start of most human terminal repeat tracts, and associate with internal telomere-like sequences and with SRE boundaries in at least some human cell types[334,343]. We therefore wished to determine whether CTCF/cohesin localized to similar features in mouse subtelomeres. We identified appropriate, existing CTCF and cohesin ChIP-seq datasets from mouse immortalized white blood cells (CH12), mouse erythroid leukemia cells (MEL), mouse

embryonic stem cells (ES_E14), and mouse embryonic fibroblast cells (MEF) [344]; see Supplementary Table 2).As was the case with human subtelomeres, the mapping of short-read data sets to mouse subtelomere regions requires special consideration because of the 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 [345,346]. 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 multi-read track. The multiread tracks are shown in the figures; tracks for uniquely mapping reads can be found in the mouse subtelomere browser (http://vader.wistar.upenn.edu/mousesubtel).

Immediately clear from the mapping data is the relative lack of CTCF and cohesin binding sites adjacent to mouse telomere tracts (Figs 4.1 and 4.2, Supplementary Figure 4.8; http://vader.wistar.upenn.edu/mousesubtel).; only the 12q and 13q telomeres had CTCF and cohesin peaks within 3 kb of the telomere tracts, in stark contrast to most human telomeres [334,343]. Likewise, there was no association of CTCF and cohesin peaks with mouse ITSs or mouse SRE boundaries in any of the cell types (Supplementary Table 4.4), in contrast to what was observed in most human cell types [334,343]. Since the human telomere-adjacent subtelomeric CTCF/cohesin sites were associated with TERRA regulation and telomere protection, this suggests very different mechanisms for these functions have evolved at mouse telomeres.

We also investigated mouse subtelomeric transcription by mapping RNA PolII binding sites using ChIPseq datasets and RNAseq reads in datasets from these same cell types, and displaying the subtelomeric regions on the mouse browser (Figs 4.1 and 4.2, Supplementary Figure 4.8; (http://vader.wistar.upenn.edu/mousesubtel). The RNA PolII peaks indicate the positions of possible promoters and transcription pause sites, with smaller enrichments often visible along the gene bodies of actively transcribed loci. The strand-specific enrichments on the

RNAseq tracks show where reads from the sequenced RNA population accumulated. The minus strand RNAseq reads (Purple) are oriented towards the telomere, and the plus strand RNAseq reads are oriented away from the telomere (Blue). The RNAseq reads in these datasets were derived from ribosomal RNA-depleted, polyA+ and polyA- RNA molecules from total RNA that yielded 300 to 350 bp cDNAs for sequencing (from GEO record associated with the RNAseq Datasets; see Supplementary Table 4.3), and so include most mRNA and lincRNA species but exclude small RNAs.

As expected, some but not all of the RNA PolII peaks localize to known promoters (based upon the Refseq gene models), and the position and intensity of the peaks depend in many cases upon cell type (Figures 4.1 and 4.2, Supplementary Figure 4.8; mouse subtelomere browser (http://vader.wistar.upenn.edu/mousesubtel). In contrast to human subtelomeres, there are no RNA PolII enrichments immediately adjacent to telomere repeat tracts. RNAseq enrichments can be seen for many annotated genes, but also for many unannotated stretches. Strikingly, the MurSatRep1 repeat element frequently overlaps with RNAseq enrichments in at least one of the 3 cell types analyzed (see Supplementary Figure 4.8), and when it overlaps it is always in the sense orientation. Of the 30 MurSatRep1 clusters found in mouse subtelomeres, 19 overlap with RNAseq enrichments in the sense orientation. Three sites which do not overlap with RNAseq enrichments correspond to (TTAGGG)n –adjacent MurSatRep1 clusters oriented away from the telomere (9q, 16q, and 19q); a cluster immediately adjacent to the 6q telomere and oriented towards the telomere does overlap with a detectable RNAseq enrichment (Supplementary Figure 4.8; also see 6q tracks on web browser (http://vader.wistar.upenn.edu/mousesubtel).

de Silanes et al. (2014) recently described a mouse TERRA locus near the 18q telomere, and presented data suggesting that it was the major source of TERRA RNA in the mouse, with possible additional minor contribution from the 9q subtelomere. Figure 4.6 shows the 18q locus with our annotation, along with the positions of the putative TERRA promoter regions described by de Silanes et al. (2014); Supplementary Figures 4.9 and 4.10 show the same two subtelomeres with several additional CTCF, Cohesin, and RNA PolII ChIPseq datasets as well as

additional RNAseq datasets. RNA PolII peaks localize to the A3 and A2 promoters defined by de Silanes et al. (2014), and RNAseq enrichments on the strand oriented towards the telomere are clearly visible in both the CH12 and the ES datasets (Figure 4.6) as well as the MEL dataset (Supplementary Figures 4.9 and 4.10), although only the ES track shows additional enrichment downstream of that found in CH12 and MEL. The putative TERRA-enriched datasets described by de Silanes et al. (2014) show only slight enrichments or no detectable enrichment over their control RNAseq datasets (Supplementary Figures 4.9 and 4.10). MurSatRep1 overlaps with the 18q TERRA and is oriented on its sense strand (Figure 4.6). However, in contrast to human TERRA (Deng et al., 2012), there is no evidence for CTCF/cohesin colocalization with the 18q TERRA transcript promoters.

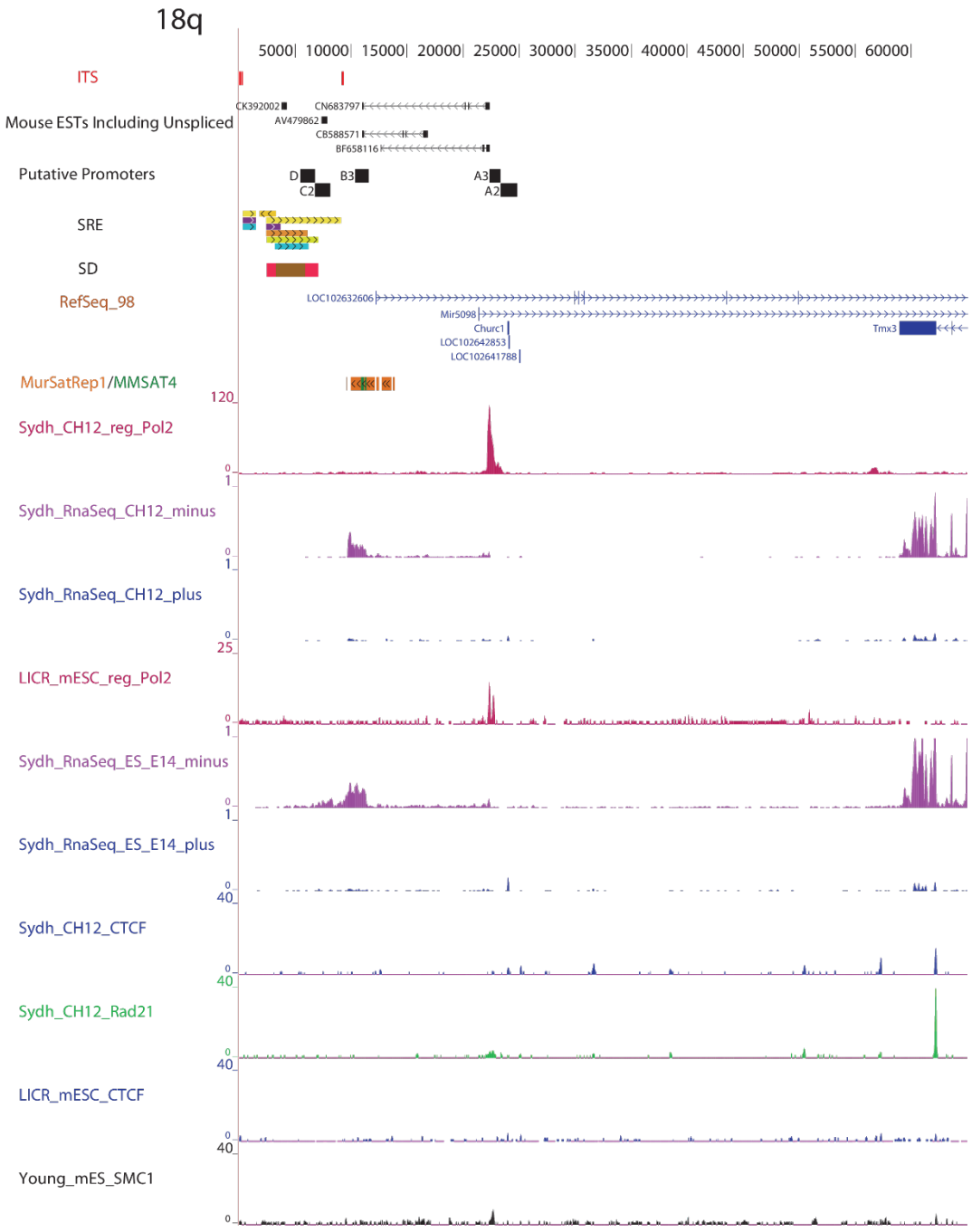

Figure – 4.6 Subtelomere features of mouse 18q relative to TERRA-associated features as described by de Silanes et al. (2014). Track descriptions are as in Figure 1. The positions of TERRA promoter regions as defined in de Salines et al. (2014) are shown for reference points.

4.5 Discussion

Since finished sequences from the telomere-containing BAC and fosmid clones we identified, isolated, and initially characterized (see Supplementary Figure 4.7 and Supplementary Table 4.2) have already been incorporated into the current mouse assembly (Build 38/mm10), we used a mouse assembly very similar to that build as our reference genome. We appended two segments of subtelomeric DNA, one at 4q (for which we had identified the remaining q-arm subtelomere sequence) and the other a representative sequence from the highly similar p-arm subtelomeres (Kalitsis et al., 2006) in order to include all telomere-adjacent sequence for our analysis. By trimming the sequenced distal telomere repeats and the placeholder NNNNs representing the unsequenced telomere tracts at mouse telomeres, we provided a consistent coordinate system for mouse subtelomere annotations, which were thus set relative to the base at the start of each terminal repeat tract and oriented from the most distal base towards the centromere. This allowed subtelomere sequence features near individual telomeres to be analyzed and visualized on the mouse browser relative to the same position at the start of each terminal repeat tract.

Our analysis indicates that human and mouse subtelomeres evolved via distinct mechanisms. Large, highly similar and identically oriented SREs that predominate at human subtelomeres are absent in the mouse. Instead, we observe smaller, more ancient and more randomly oriented SREs in the mouse. The current model for explaining human SRE structure - translocations involving the tips of chromosomes, followed by transmission of unbalanced chromosomal complements to offspring [321] – cannot explain SRE structure in mouse. While human SRE structure and chromosomal distribution evolved very recently and in fact is highly polymorphic in the population, mouse SREs are more divergent and, if they originally evolved in a manner similar to human SREs, their sequence organization has subsequently been disrupted by recombination or repair events resulting in locally inverted duplications. Since subtelomere regions are poorly assembled or absent in WGS assemblies, there is no reliable information

available on the subtelomere structure of multiple mouse species; this information would be quite valuable in understanding how these regions evolved and whether the unusually long telomere repeat tracts in mice played a role in suppressing recent subtelomeric duplication events or perhaps promoting inversions.

A relatively large fraction of mouse subtelomeres lack SREs entirely (3q, 5q, 6q, 7q, 12q, 13q, 14q, 15q), and for some of these subtelomeres known genes extend to very near the start of the terminal (TTAGGG) tracts (e.g., wntless homolog (Wls) at 3q, bicaudal D homolog 1 (Bicd1) at 6q, MAS-related GPR, member D (Mrgprd) at 7q, transmembrane protein 196 (Tmem196) at 12q, and 2-cell-stage, variable group, member 3 (Tcstv3) at 13q). It is notable that each of these telomere-adjacent transcripts are oriented telomerically. 18q, which contains a relatively small SRE region, has recently been described as a major site for TERRA transcription in mouse (de Silanes et al., 2014). The position of the putative upstream promoter sites (A2 and A3, Figure 4.6) correspond to RNAPolII peaks in our analysis, whereas the downstream B3, C2, and D promoter regions overlap with properly oriented RNAseq read enrichments in the ES datasets but not the immortalized lymphocyte CH12 dataset (Figure 4.6). While TERRA molecules derived from non-18q telomeres were not found (with the possible exception of 9q; de Silanes et al, 2014), other subtelomeres were neither exhaustively sampled, nor sampled in multiple cell types by de Salines et al. (2014); it is therefore possible that some of the existing telomerically oriented transcripts could read through subtelomeres into telomeres to form TERRAs, as was suggested by recent TERRA-enriched RNAseq mappings to human subtelomeres that overlap with the WASH transcript family [335].

Strikingly, a mouse repeat element of unknown origin (MurSatRep1) which is known to be enormously enriched in mouse lincRNAs where it is almost always transcribed in the sense orientation [338] is very highly enriched in mouse SREs (>200 fold) as well as in subtelomeric SDs (98-fold). Of the 30 MurSatRep1 clusters we found in mouse subtelomeres, 19 overlap with RNAseq enrichments in the sense orientation. The 18q TERRA locus includes one of these overlap sites; additional telomerically oriented sites within 20 kb of the (TTAGGG)n tract are

found at 8q, 13q, 17q (Figures 4.1 and 4.2), as well as at 10q and 6q (Supplementary Figure 4.8, (http://vader.wistar.upenn.edu/mousesubtel). Each of these subtelomeric sites corresponds to a lincRNA potentially capable of extending into the telomere (TTAGGG)n tract, (thus giving rise to TERRAs), and merits further analysis.

Waves of retrotransposon expansions have remodelled genome organization and CTCF binding in murine lineages [347]; these dramatic evolutionary changes, along with the very different segmental duplication evolution trajectories taken at mouse vs human subtelomeres, may together account for the completely different CTCF and cohesin binding site organization at mouse and human subtelomeres. Our results showing a nearly complete lack of CTCF/cohesin near mouse telomeres and no association with SRE or ITS boundaries indicate radically different mechanisms for TERRA regulation, telomere protection and potential cis-regulation of telomere lengths in mouse relative to human. Given the differences in telomere lengths, overall lifespan, susceptibility to cancer, and other telomere-associated properties in the two species, this is perhaps not surprising. It may be one of many functional consequences of the accelerated evolution of subtelomeric genome regions, where a wide variety of lineage-specific and species-specific functionalities have arisen [302,303,348,349].

4.6 Supplemental Figures

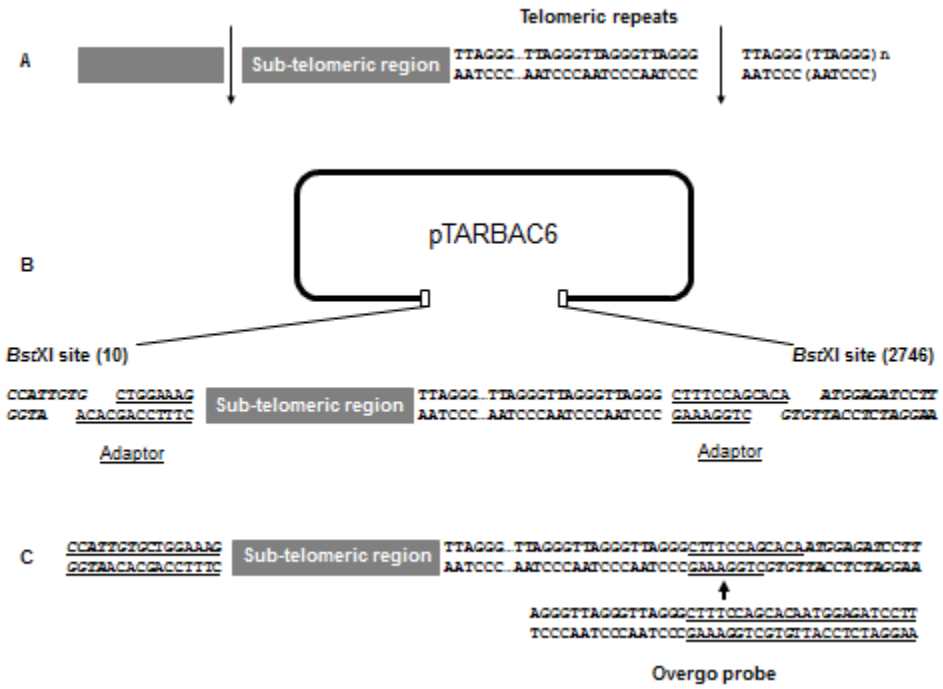

Figure 4.7 – Telomeric BAC Isolation. The CH25 BAC library was prepared using sheared DNA from strain c57bl/6j and cloning into the vector pTARBAC6 (Osoegawa et al., 2007). The library was screened using labeled overgo probes (Vollrath, 1999) specific for the junction of vector and telomere repeat sequence in order to specifically identify clones which contain an insert fragment with the telomere repeat at one end. BAC clones thus identified were colony-purified, end-sequenced, and localized to metaphase chromosomes using FISH. The non-telomere BAC end sequences were mapped to the assembled mouse genome using BLAST (Altschul et al., 1990). The combination of FISH localizations (or multi-site localization) and end sequence match were used to select candidate telomere BACs for full sequencing. The telomeric BACs thus identified and the sequenced clones incorporated into Build38/mm10 are listed in Supplementary Table 1.

Large multipage figure available at https://shiek-db.wistar.upenn.edu/riethman/suppfig48.pdf

Figure 4.8 – Snapshots of Annotated Mouse Subtelomeres. Screenshots of the mouse subtelomere browser showing static versions of tracks specifically discussed in this paper are shown. See the live browser (http://vader.wistar.upenn.edu/mousesubtel) for custom track selection, zooming, and track organization.

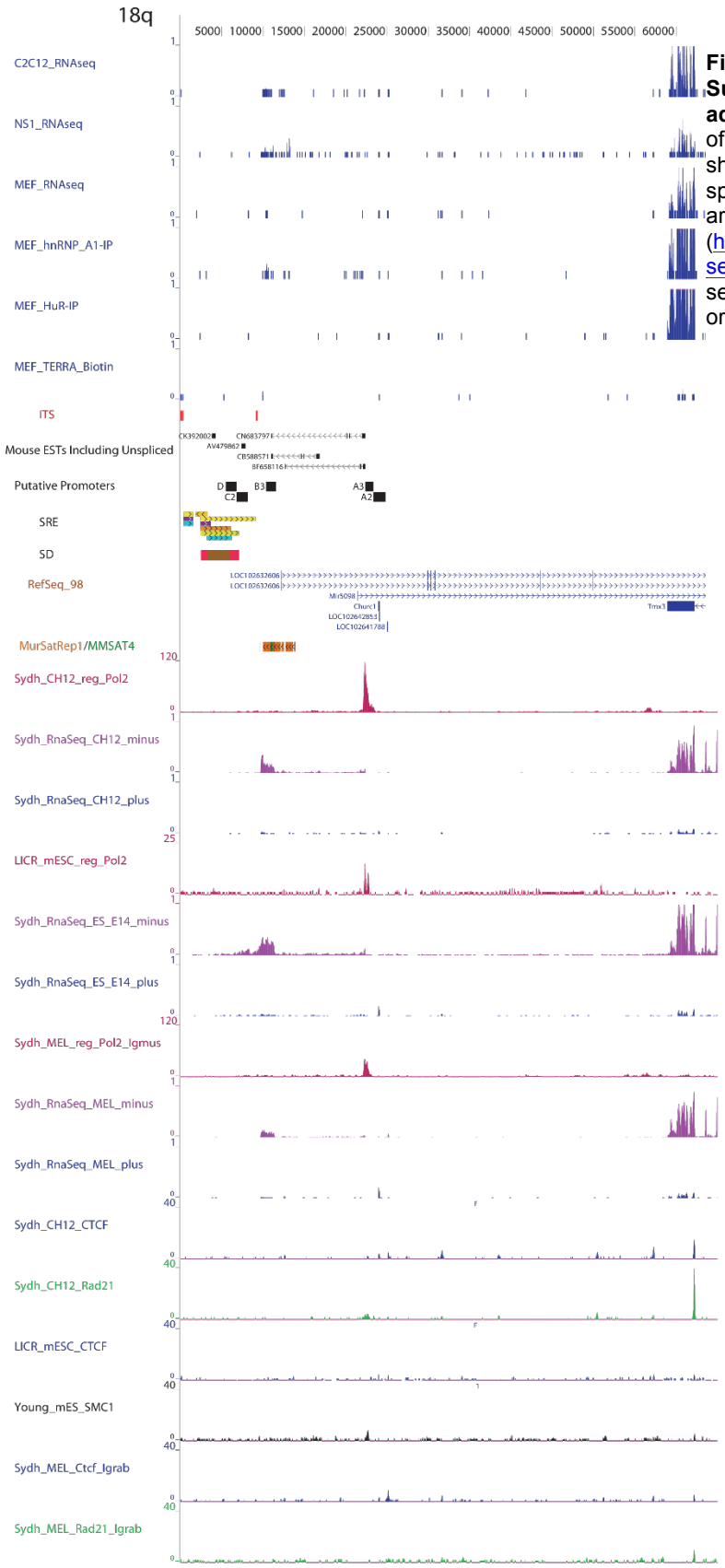

Figure 4.9 – Mouse 18q Subtelomere Annotated using additional datasets. Screenshots of the mouse subtelomere browser showing static versions of tracks specifically discussed in this paper are shown. See the live browser (http://vader.wistar.upenn.edu/mousesubtel) for custom track selection, zooming, and track organization.

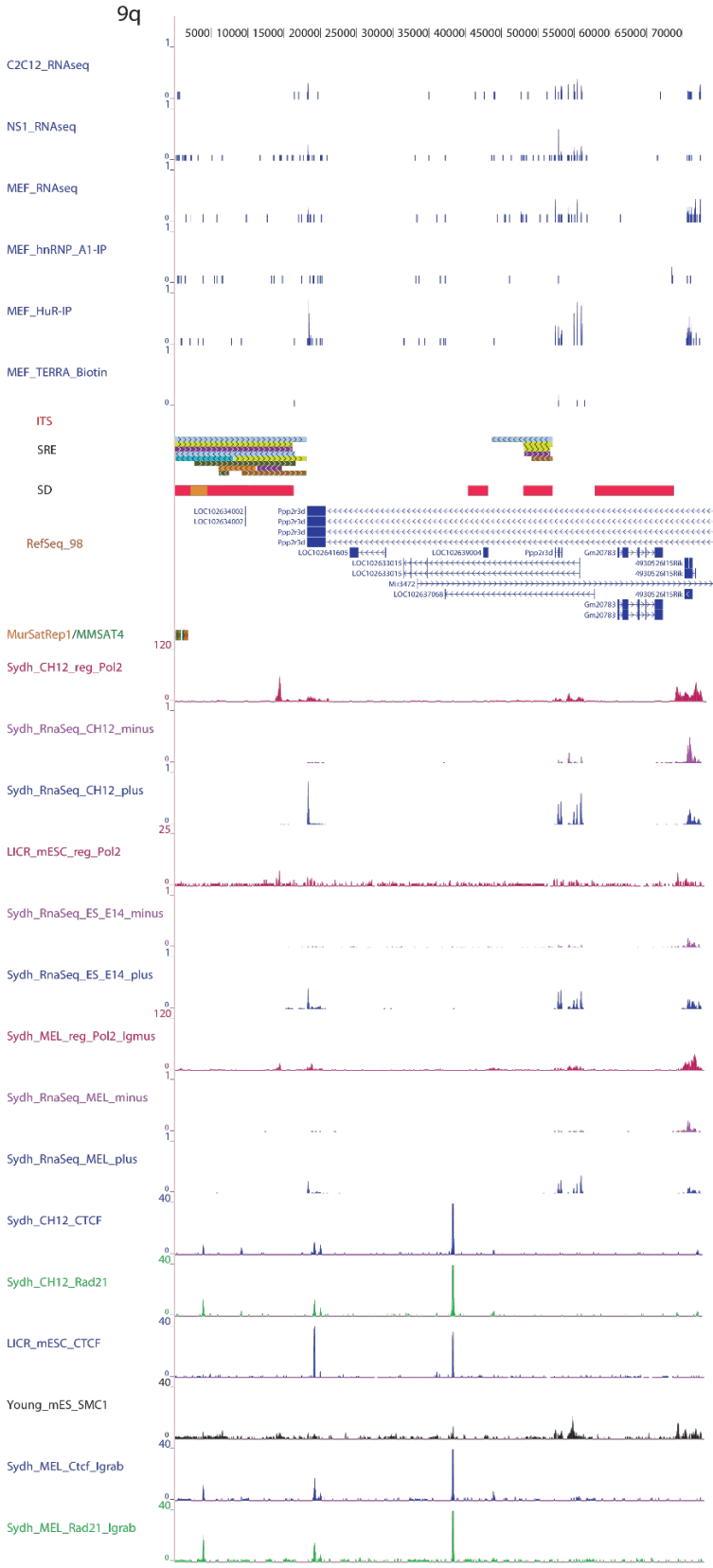

Figure 4.10 – Mouse 9q Subtelomere Annotated using additional datasets. The same tracks as in Supplementary Figure 3, as shown for 9q.

Table 4.2 – Mouse Subtelomeric Clones. Telomeric BACs were identified as described in the legend to Supplemenary Figure 4.7, and are highlighted in Supplementary Table 4.2. Fosmid End Sequence (FES) mapping: An end-sequenced fosmid library derived from sheared mouse genomic DNA (WI-2; Church et al., 2009) was screened for clones containing the telomere terminal repeat sequence (TTAGGG)n. Because of the orientation of this repeat at all terminal repeat tracts, the distal end-sequence from a telomere-terminal fosmid will always contain a (CCCTAA)n pattern. As for a similar computational screen of human fosmid libraries (Stong et al., 2014), requiring a perfect (CCCTAA)4 match reliably identified authentic telomere-containing fosmid clones. The mate pairs of (CCCTAA)-containing fosmid end sequences from the WI-2 library were mapped back to reference subtelomere assemblies; all but a few mapped either uniquely to a known subtelomere assembly or to a known SRE. These mappings identified fosmids that should bridge existing subterminal gaps in the reference sequence, and are identified in Supplementary Table 4.2.

Excel spreadsheet available from https://shiek-db.wistar.upenn.edu/riethman/supptab3.xls

Table 4.3 – Datasets used in this study and quality metrics. Table includes all data set tracks and information on data set origins, their matched control, and the specific companies and product numbers for antibodies used. The abbreviations for the antibody-providing companies are: MIL, Millipore; sc, Santa Cruz Biotechnology; ab, Abcam; and BL, Bethyl Laboratories. Metrics include number of peaks called in hybrid genome using all reads, and only uniquely mapping reads. FRiP (Fraction of Reads in Peak), for all reads: partial reads (mapping likelihood) was counted in peaks called using all reads. FRiP for only unique reads is reads in peaks called using only uniquely mapping reads. PBC (PCR Bottleneck Coefficient) is the number of genomic positions with one read mapping to it (uniquely mapping or partial mapping), divided by the total number of genomic positions with at least one read mapping (uniquely mapping or partial mapping). NSC (Normalized Strand Cross-correlation coefficient) is the ratio of maximal cross-correlation value over the background cross-correlation. RSC (Relative Strand Cross-correlation coefficient) is the maximal cross correlation value minus the background cross-correlation, divided by the cross-correlation at the read length minus the background cross-correlation. For detail see (Landt et al., 2012) and http://genome.ucsc.edu/ENCODE/qualityMetrics.html.

Excel spreadsheet available from https://shiek-db.wistar.upenn.edu/riethman/supptab4.xls

Table 4.4 – SRE boundary enrichment statistics. Raw Peak Counts of Subtelomere Boundary Enrichments. All Tables have columns corresponding to different boundary categories, and total for all boundaries in the SRE region. Boundary categories are 1copy/SRE (duplicon ends at unique subtelomere sequence), Gap (duplicon ends at most terminal complete sequence but not telomere), SRE/SRE (duplicon ends within SRE region), SRE/SD (duplicon ends at genomic duplicon), Terminal (duplicon ends at telomere), All_Bndries (all boundaries). Rows correspond to datasets. A. Raw Counts - Counts of peaks in association with different boundary categories and total. Additional row, Sequence, is the amount of sequence within the window of the boundary type, or the total SRE region. B. Percent – This table shows the percent of total peaks associated with a boundary type, for the total column this is always 100%. The additional row still corresponds to the amount of sequence within the window for the boundary type, the percent of sequence within the window of the boundary type out of the total SRE. This is the expected percentage of peaks in association, if the peaks are distributed randomly in the SRE. C. Enrichment – The ratio of percentage for a category and dataset over the expected percentage for that category. D. P Value – The p value calculated using a binomial test with the expected percentage as the probability of success, and the associated number of peaks and total number of peaks as success and trials.

CHAPTER 5: TELOMERE ANALYSIS USING TASER

5.1 Abstract

In order for a cell to gain limitless replicative potential it must have an activated telomere maintenance pathway that compensates for the telomere shortening induced by DNA replication attrition or by DNA damage induced rapid telomere deletion, either of which can create a dysfunctional telomere. Telomere maintenance pathway reactivation occurs in developing cancer cells through activation of telomerase expression or by the recombination based ALT pathway. Both mechanisms cause changes to the telomere sequence in a cell. The dynamics of these changes over time in a developing cancer are not well understood, in part because studying these changes requires focused assays to measure telomere length. A number of cancer genome sequencing projects have extensively sampled and sequenced tumors and corresponding adjacent normal tissue. These whole genome sequencing datasets contain telomere sequence reads that can be leveraged to study the changes in telomeres that occur in cancer. We have developed a pipeline, TASeR, to capture this information and used it to identify short and potentially dysfunctional telomeres in cancer samples. TASeR was used to analyze samples from the prostate cancer sequencing genome project. We found prostate cancers have shortened telomeres relative to paired normal samples, specifically a loss of perfect telomere sequence. These changes in telomere sequence content were used to build a classifier to separate samples as either cancer or normal.

5.2 Introduction

Carcinogenesis is a multistep process in which the accumulation of mutations in a cell allows it to evade a number of checks and deficiencies to gain the ability and drive to proliferate uncontrollably[128,134,135]. Telomere maintenance plays a central role in this development. In a precancerous but DNA damage repair deficient context dysfunctional telomeres contribute to a mutator phenotype[350,351]. Telomeres too short to bind a sufficient density of shelterin components can no longer form a functional protective structure. In a normal context DNA damage response elements lead the cell to senescence or apoptosis. In the absence of a functional allele of these elements the perceived damage of a DNA double strand break will be "repaired"[74]. This creates telomere telomere fusions or an unbalanced translocation between chromosomes. When these cells divide it leads to aneuploidy. These changes in copy number can disrupt repressive mechanisms or amplify growth pathways progressing the cell towards malignancy. Short dysfunctional telomeres that accumulate with age can contribute to a mutator phenotype enabling cancer[352].

Telomeres also play a vital role in enabling the unlimited proliferative potential of cancer cells. As precancerous cells begin to divide their telomeres shorten with each cell division due to the end replication problem, exonucleic activity, and DNA breaks in telomeres caused by replicative stress. Once the telomere is insufficiently long the cell will enter crisis and some small subset of cells in crisis will emerge with an activated telomere maintenance mechanism[123,124]. Approximately 85% of cancer types have active telomerase expression and the remaining 15% activate the ALT pathway[125,126]. In cells that have activated the ALT pathway the recombinatory mechanism of elongating telomere results in transfer of large sections of sequence between telomeres resulting in heterogeneous telomere lengths including extremely long telomeres[82]. In sharp contrast telomeres in telomerase activated cancer cells remain extremely short, bordering on dysfunctional[129].

The dynamics of telomere length through the initiation and progression of cancer is unclear. At which point in the development of cancer telomeres need to be elongated can lead to important understanding of how telomeres contribute or are swept into carcinogenesis. In order to study these dynamics two major methods to measure telomere length are used telomere repeat fragment (TRF)[142,143] analysis or quantitative PCR (qPCR)[150]. Telomere Repeat Fragment analysis depends on gel electrophoresis of genomically digested DNA, allowing the large telomere fragment to be shown on a gel. The highest throughput telomere length measurement is a quantitative PCR method. By comparing the fluorescence signal from telomere primers to those from a single copy locus provides a ratio that is highly proportional to the telomere length. These methods are relatively easy to perform, however have the limitation of measuring the average telomere length in a sample. Chromosome specific telomere length measurements require metaphase arrest, or can only be performed on a select number of telomeres where unique subtelomere sequence exists[147,148,151]. In order to study telomere dynamics these specific methods are employed, using samples to focus on this limited aspect of cancer development.

Cancer sequencing genome projects are making available large numbers of whole genome sequencing (WGS) data sets for many different cancer types. These datasets have initially been used to investigate the single nucleotide and copy number mutations to identify changes that are consistently found in a certain cancer type. Prostate cancer is a highly occurring heterogeneous disease where the majority of patients respond well to treatment, but those with advanced disease at the time of diagnosis have a poor prognosis[353]. Prostate cancers are known to have a characteristic genome instability with a recurrent TMPRSS2-ERG fusion, with the second partner being a member of the ETS family[354]. A prostate cancer genome effort has focused on this genome instability, finding coordinated rearrangement leading to large scale genome restructuring, which they term chromoplexy[355].

The library preparation techniques used to generate the data sets from these sequencing efforts captures telomeric sequence that was inevitably screened out from downstream analyses.

It is impossible to map these sequences, as there is no useful reference sequence for the telomere, and if it did exist a telomere read would map to thousands of positions throughout the telomere sequences. Instead of attempting to align a read to a specific genomic index, we took the approach of "aligning" a read to a portion of the telomere based on its sequence content. Here we present a pipeline Telomere Analysis from Sequencing Reads (TASER) to capture the telomere content of a WGS sample allowing for comparison between samples.

5.3 Methods

5.3.1 Use and Optimization of RepeatMasker

The problem of repeat identification was solved early on for genome masking. Repetitive sequence was masked in sequences to simplify the alignment of matches by avoiding searching through the many possible matches in repetitive sequence. RepeatMasker[284] is the most widely used tool for this task however it was not made with the short sequences generated by high throughput sequencing in mind. Repeatmasker was run with options to only search for simple repeats, and alignments were reported in the orientation of the repeat to mark at which base in the telomere repeat possible mutations occur (-nocut -no_is -noint -a –inv). To further optimize the use of RepeatMasker samples were parallelized using GNU Parallel[356]. Samples were split into 1 million read portions and run simultaneously up to the number of cores available. RepeatMasker results were parsed to categorize reads. Reads with an identified telomere repeat were identified as TR (telomere repeat). A telomere repeat with a point mutation was classified as TRPM (telomere repeat point mutation). Point mutations with a PHRED score of less than 10 were ignored as these were low quality base calls that were likely to actually conform to the telomere repeat. After initial tests analyzing the number of mutations per read it appeared that reads with up to 7 mutations included reads which were actually perfect telomere reads (SupplementalFigure 5.17). These reads with up to 7 mutations were reclassified as perfect telomere reads. Reads with repeat patterns CCCGAA, CCCCAA, CCCTAG, TTAGGC (each of which had been characterized previously in the proximal regions of terminal repeat tracts;

[336,357]for review) were classified as TLSR (telomere like simple repeat). Reads containing more than one identified simple repeat were classified as mixed. Reads with some other simple repeat identified were classified as other. Reads with no simple repeat identified by RepeatMasker were mapped to the 15kb telomere reference[265] with bowtie[238]. Reads with an alignment were classified as mapped, and the remaining reads were classified as unknown. Mate pairs are classified separately before being re-paired with their corresponding mate. Total read counts for all paired categories were counted. The counts are then normalized to account for differences in sequencing depth by dividing the count by the total library size. They are further normalized to account for PCR duplicates. The PCR duplicate rate of mapped reads were calculated by finding,counting, and determining the average rate of read pairs with the exact same mapping positions reported within the 15kb subtelomere regions. The telomere read counts were then normalized by this duplicate rate.

5.3.2 Summary Statistics

Summary statistics for telomere reads were calculated based upon the known structure of human telomere terminal repeat tracts (Fig 5.1) and our biological expectation for how that might change in cancer (Fig 5.2). Total telomere is the sum of all TR, TRPM, and TLSR reads without regard to mate pair information. The boundary statistic is the count of reads where one read is classified as mapped, in pairing with a TR, TRPM, or TLSR read. Percent perfect telomere is the TR reads divided by the total telomere reads. The mutation interspersion ratio is the ratio of TR reads in pairing with TR reads to TR reads in pairing with either TRPM or TLSR reads. R was used to calculate summary statistics for the data analyzed and to generate the box plot figures.

5.3.3 Dataset

The prostate cancer genome sequencing project data is available from dbGaP under accession number phs000447 and its corresponding SRA entry SRP011021. Samples were

sequenced on Illumina HiSeq 2000[358,359].

Classifier

A binary classifier was built to classify a sample as tumor or normal based on it's telomere characteristcics catptured by TASER. Logistic regression was performed using generalized linear models tool in R[360] setting the family to binomial. The only significant factor was found to be the TR_TR measurement. A conditional logistic regression was also performed but resulted in similar coefficients. For simplicity the logistic regression was used to measure

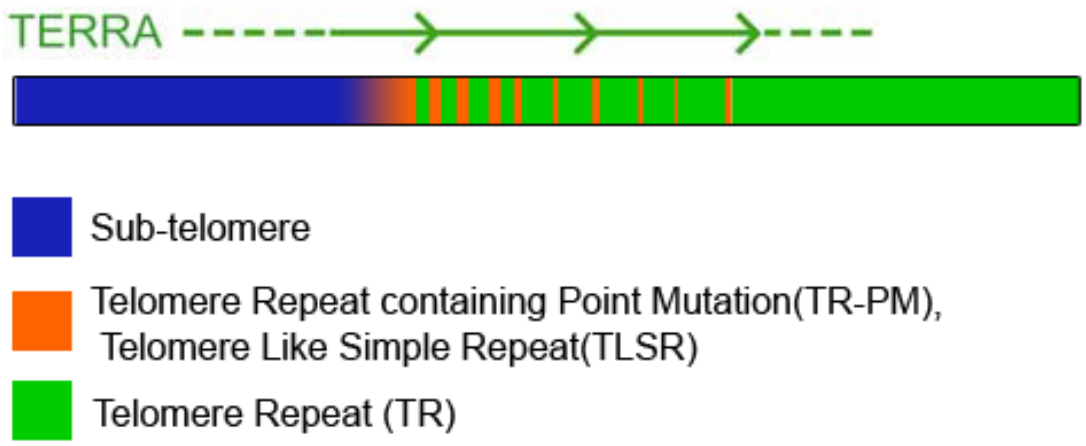

Figure 5.1 – Idealized telomere structure.. The telomere structure is known to end in perfect telomere repeat (TTAGGG)n (green), while closer to the subtelomere mutations in the telomere repeat and simple repeats similar to the telomere are found interspersed with the canonical perfect telomere sequence (orange). The beginning of the telomere is adjacent to the subtelomere sequence, which includes the subtelomere repeat elements and the telomere associated repeats (blue). The subtelomere also contains the transcription start sites of TERRA which is transcribed through the telomere.

cross validation using the boot package in R[361,362]. Analysis of variance was used to look for significant differences in the features measured between different tumor stages for both direct tumor measurements and the change (log ratio) between normal and cancer of a sample.

5.4 Results

5.4.1 Motivation

In the absence of a reference sequence we consider an ideogram of telomere terminal repeat sequence based upon experimental analysis (Baird et al., 1995; summarized in Riethman, 2008; Figure 1). The telomere sequence is known to end in perfect repetition of the canonical telomere sequence, TTAGGG. However all sequencing efforts extending from the subtelomere into the telomere tract include mutation events and telomere like simple repeats (i.e. TTAGGGG). Instead of aligning reads to a reference and obtaining a mapping position we instead classify reads as containing a perfect telomere repeat, or a mutated telomere sequence. In this way we capture some positional information, as a mutated sequence is more likely to have originated closer to the subtelomere. TASER was also developed for use on paired sequencing data. By considering the categorization of both mate pairs, you gain further insight to the sequence structure of the telomeres in the sample. In comparing TASER results from different samples changes in telomere sequence content and structure can be observed. For example comparing TASER results between a normal and tumor sample you would expect (in a telomerase positive tumor) for there to be a decrease in total telomere content, and an increase in the proportion of mutated telomere reads (Figure 2).

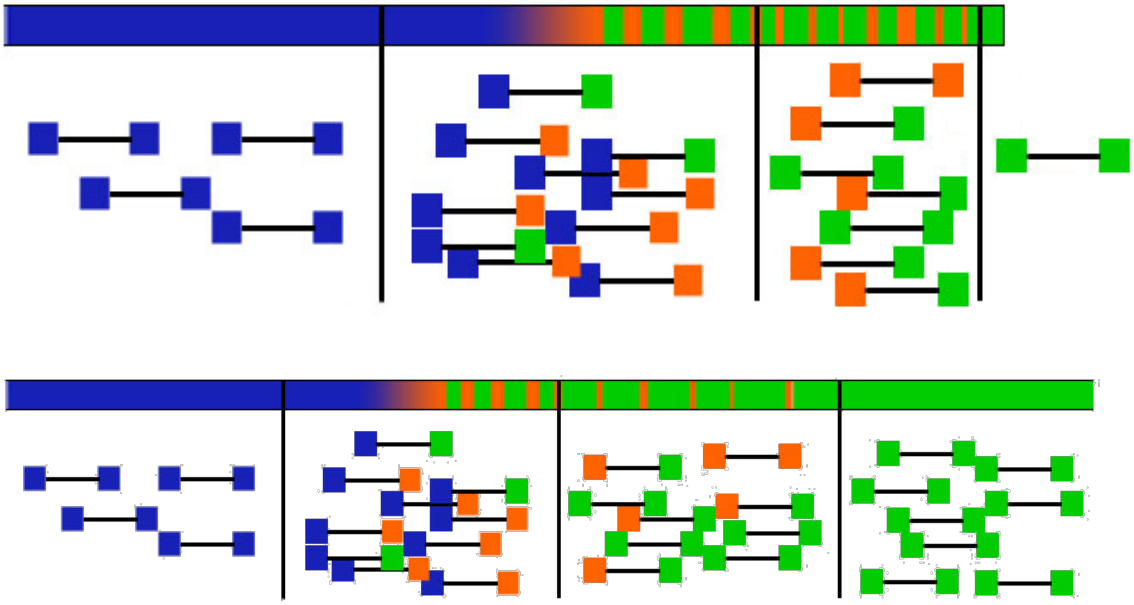

Figure 5.2 – Changes in TASER measurement distribution due to telomere shortening. TASER classifies reads as containing some type of telomere sequence which is paired with a read that is also classified. In the case of a shortened telomere the distribution of these categorized reads changes.

In order to evaluate whether the reported telomere state was consistent with our biological expectations we first calculated intuitive summary statistics which summarize the telomere state captured in a way that we have some expectation for the difference between a normal sample and a cancer sample. We calculated four intuitive summary statistics, total telomere, boundary, percent perfect telomere, and interspersion ratio. Total telomere is the sum of all telomere reads, ignoring mate pair relationships. This is proportional to the total telomere sequence in the cell. The boundary statistic is the number of reads where a read that maps to the proximal 15kb of the subtelomere is in pairing with a telomere read. The number of these reads will vary with changes in the number of telomeres in a cell, i.e. cells with aneuploidy and an abnormal karyotype will have changes in this metric in comparison to a normal cell. Percent

perfect telomere is the amount of telomere reads that are perfect out of the total telomere reads, which includes perfect and mutated telomere reads. In samples with shortened telomeres the mutated telomere repeats will make up a larger portion of the total. The mutation interspersion ratio measures the overall integrity of telomeres. This is expected to degrade with telomere shortening, but it is not clear how this metric would be affected by ALT, which could potentially disperse proximal telomere sequences throughout telomeres by recombination.

5.4.2 Prostate Cancer has short dysfunctional telomeres

TASER was run on 53 paired tumor normal prostate cancer WGS data generated at the Broad (106 samples total). After normalization for dataset size and PCR duplicate rate, the total telomere content, while a small portion of the total sequenced library, shows an overall difference between the normal and cancer samples (Figure 3). The cancer samples have less total telomere content indicating they have shorter telomeres, which should be reflected in the standard measurement techniques using TRF or a qPCR measurement. This overall result is very consistent with the individual pairwise analysis of normal and tumor samples (Sup Fig 5.9), where only 2/53 normal-cancer pairs show an increase in telomere length (Sup Fig 5.13).

There is no significant difference in the boundary reads between normal and cancer samples overall (Figure 4), reflecting no globally consistent changes in telomere number in prostate cancer. However, an increase in the spread of the box plot indicates increased variation in the sample set which could be due to increased genomic instability in the cancer cells or heterogeneity in the tumor samples sequenced. When individual normal-cancer pairs are examined using this metric (Sup Fig 5.10), there are subsets of tumors where this metric increases, decreases, and stays roughly the same/increases slightly (Sup Fig 5.14). These results thus suggest distinct subclasses of tumors with increased telomere number, decreased telomere number, and slight to no difference in telomere number. This more granular analysis of the boundary metric might therefore be useful in distinguishing telomere number change subsets within tumor types and thus could be a useful new parameter for tumor classification.

There is also a significant decrease in both the percent perfect reads (Figure 5) and mutation interspersion ratio (Figure 6). This indicates that the cancer samples have more mutated telomere sequence relative to the total telomere, and that it is more interspersed in the remaining perfect telomere sequence. This is most simply explained by the preferential loss of distally located perfect telomere repeat tracts in cancer cells. These global results are generally supported by the individual pairwise analyses of these metrics (Sup Fig 5.11 and Sup Fig 5.12), but there are a substantial number of outliers where these two metrics are increased in cancer relative to normal (SupFig 5.15 and Sup Fig 5.16), which merits more detailed experimental analyses.

These changes we see in telomere sequence content can play an important role in the alteration of telomere function in cancer. The loss of perfect telomere sequence, and an increased proportion and interspersion of telomere like sequence, might be expected to destabilize the binding of TRF1 and TRF2 to the telomere. This could affect the overall stability of the shelterin complex, with consequent loss of telomere integrity and increased genome instability. Thus the quality as well as the quantity of telomere repeat tract may be important for telomere integrity in cancer.

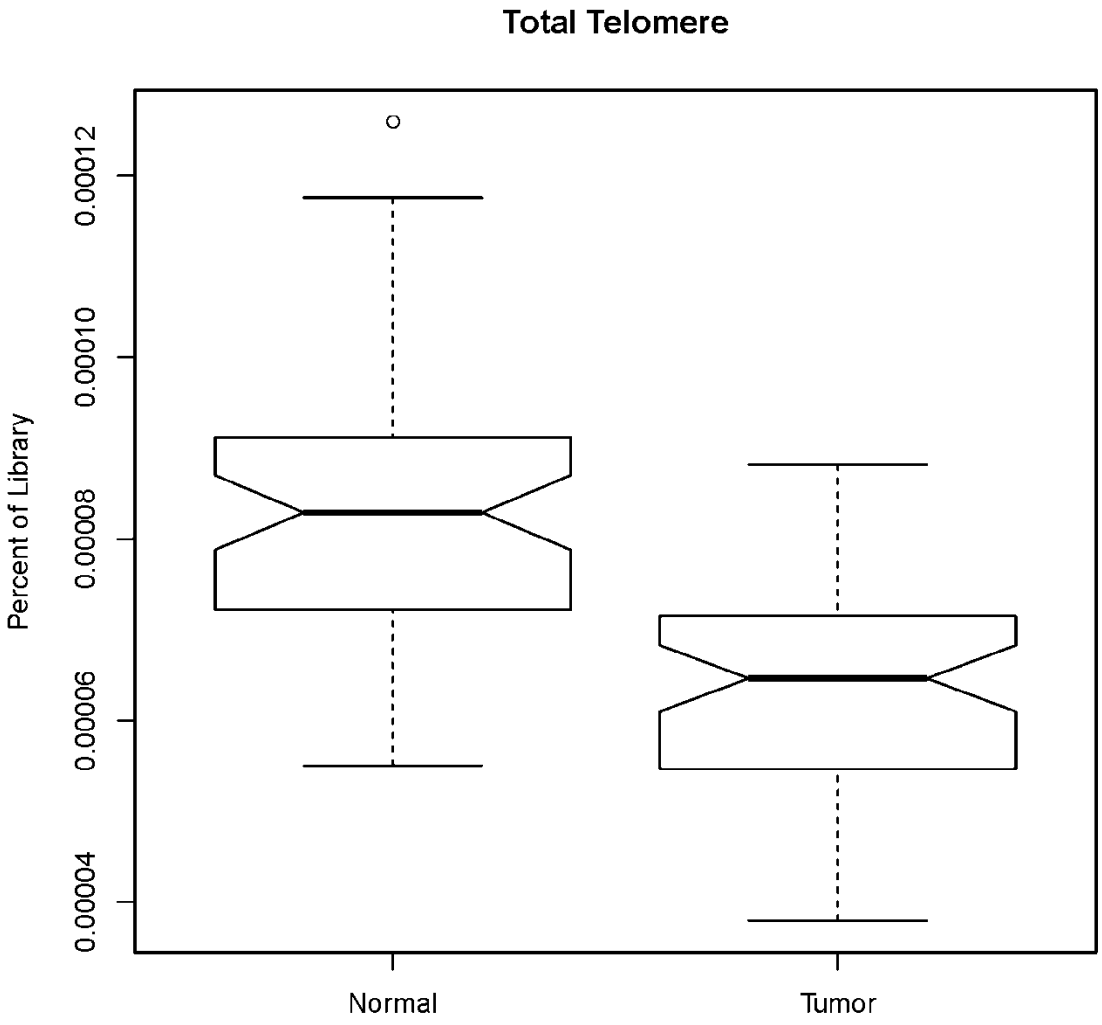

Figure 5.3 – Box plot of total telomere measurement for normal and cancer samples. Box plots show the distribution of values in tumor and normal samples. Values are percent of library, number of reads normalized by library size. The number of reads for total telomere is the sum of all telomere categories, TR, TRPM, and TLSR.

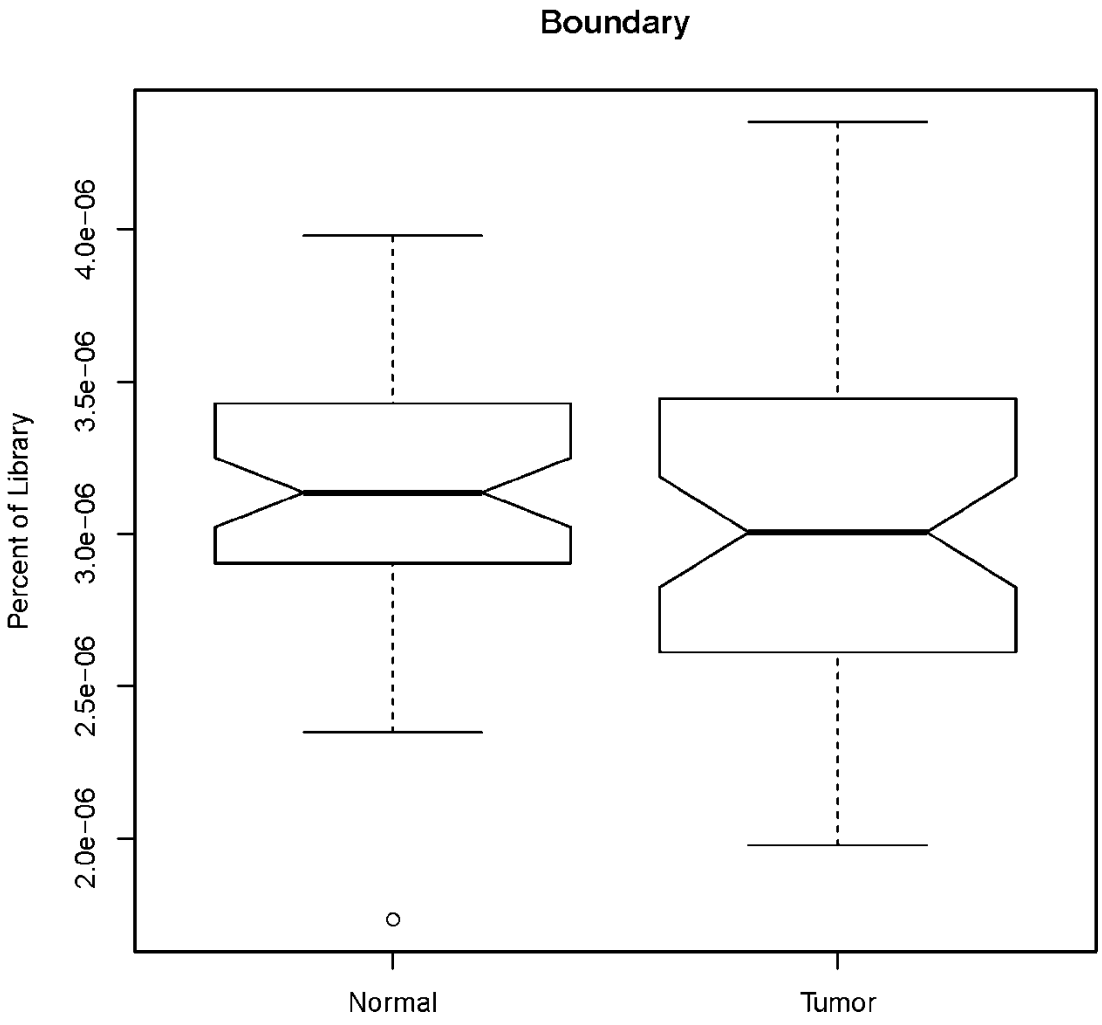

Figure 5.4 – Box plot of boundary measurement for normal and cancer samples. Box plots show the distribution of values in tumor and normal samples. Values are percent of library, number of reads normalized by library size. Boundary reads are any category of telomere reads, TR, TRPM, or TLSR, in pairing with a mapped read.

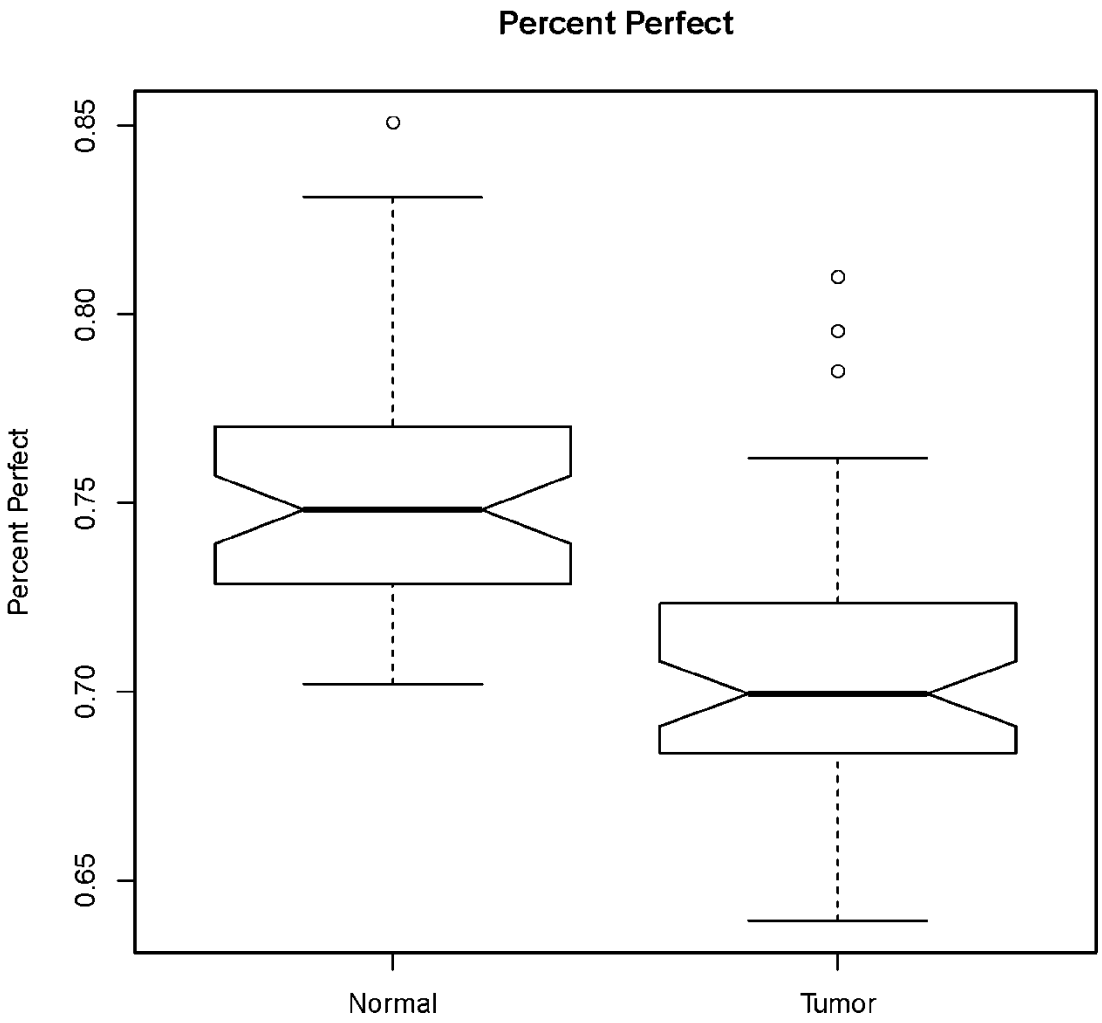

Figure 5.5 – Box plot of percent perfect measurement for normal and cancer samples. Box plots show the distribution of values in tumor and normal samples. Values are the percent of the telomere reads that perfect, TR, out of total telomere reads TR, TRPM, and TLSR.

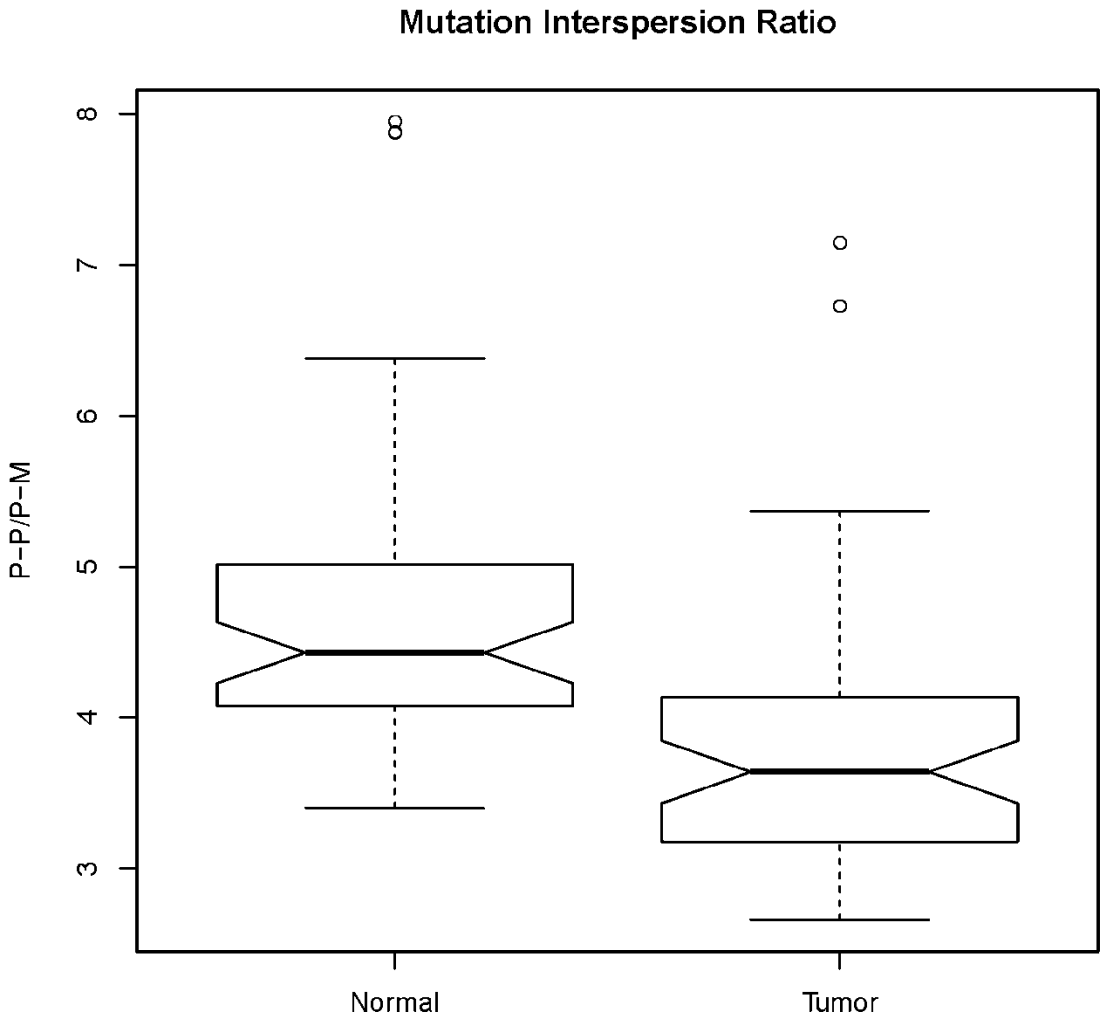

Figure 5.6 – Box plot of mutation interspersion ratio measurement for normal and cancer samples. Box plots show the distribution of values in tumor and normal samples. Values are the calculated ratio. A higher value indicated more perfect/perfect telomere read pairs or less perfect/mutant telomere read pairs.

5.4.3 Telomere metrics can be used to identify cancer

The normalized read amounts for paired read categories were used as features in a logistic regression to classify cancer and normal samples based on the telomere information contained in their WGS data. Pruning of the full feature set resulted in one feature being responsible for the separation of tumor and normal samples, the perfect telomere reads in pairing with perfect telomere reads (Figure 7). This is consistent with the expected changes between normal and cancer samples. Cancer cells after dividing hundreds of times would have shortened telomeres, losing the perfect sequence known to be found at the end of telomeres. For the comparison between tumor and normal samples this single feature is quite powerful at separating tumor and normal samples, properly classifying samples with 87% accuracy (Figure 8). Changes in the presence of mutated telomere reads could play an important role in separating different cancer clonal populations or stages of disease. A logistic regression was also performed on tumor samples considering the TMPRSS2-ERG fusion status (with fusion, or without), as the two groups. An ANOVA was also performed considering the various pathological stage classifications and grouping samples by the age of the patient. The change in telomere state for samples (normal minus tumor) was also considered for both these analyses. These analyses were unable to separate different groups within the tumor samples sequenced.

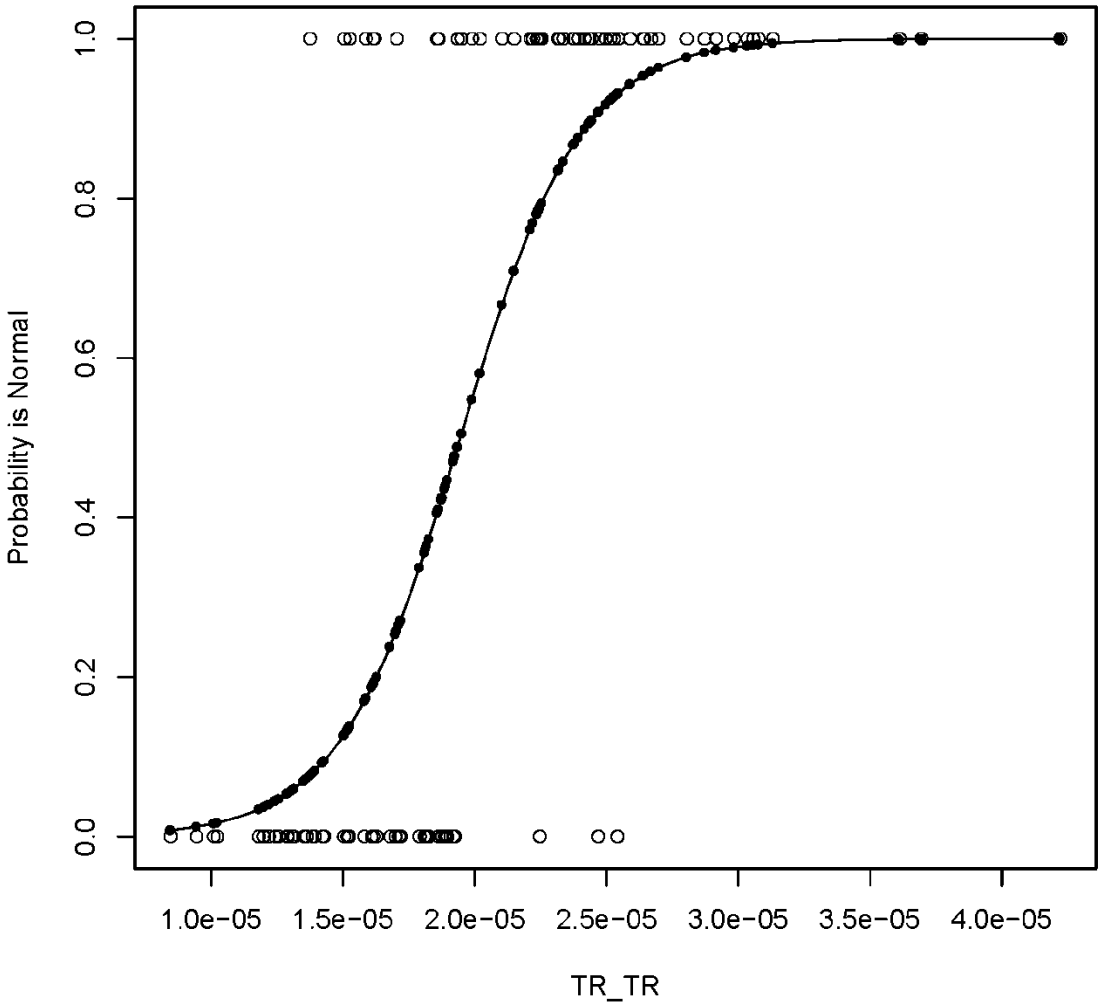

Figure 5.7 – Fitted logistic regression model for tumor classifier. Fitted logistic function used to separate tumor and normal samples bases on the amount of TR_TR reads in the sample. The model line is shown with samples falling on it indicating the likelihood that a sample is normal. Along the bottom tumor samples are shown, along the top normal samples are shown both at the position of their TR_TR content.

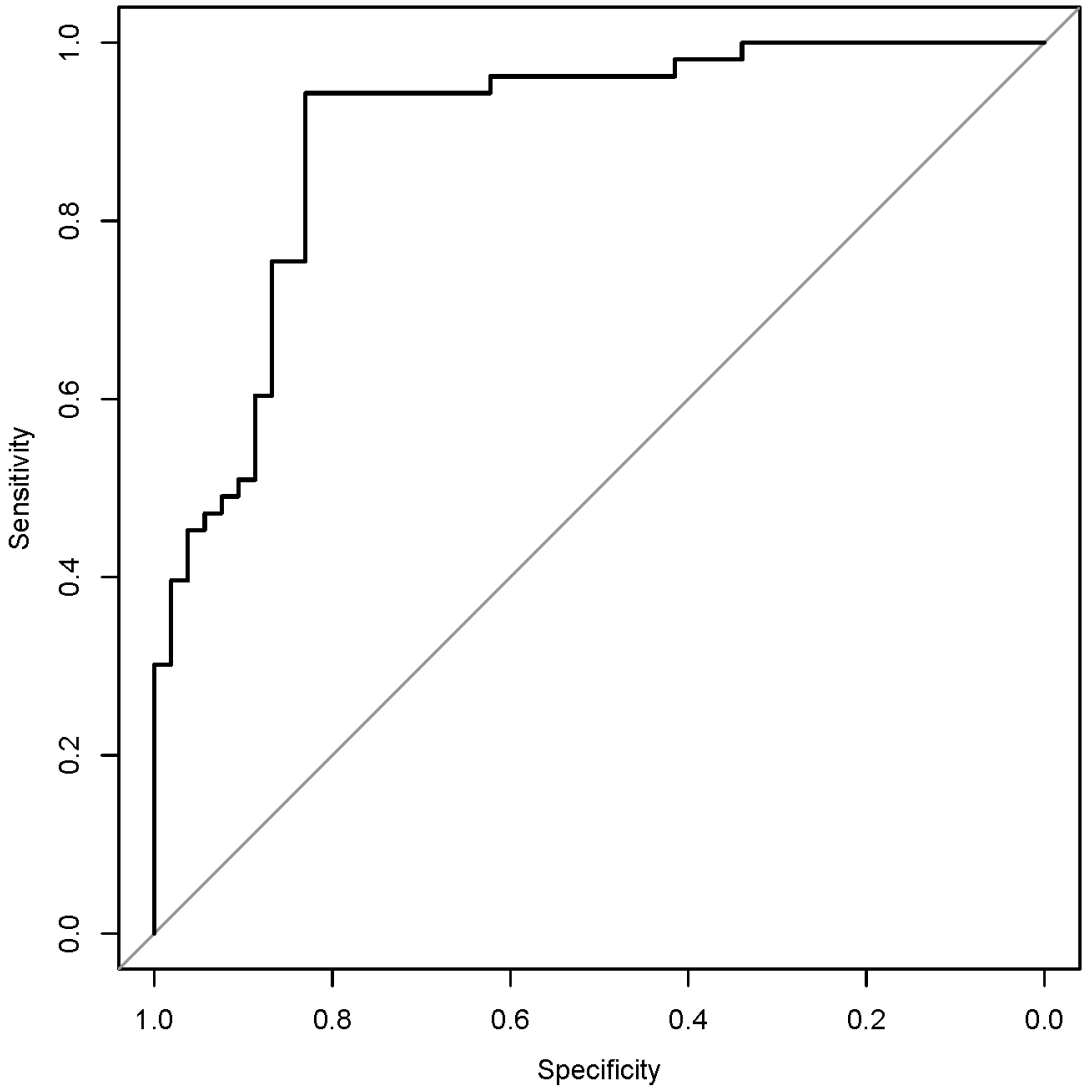

Figure 5.8 – ROC curve of tumor classifier. Receiver operator characteristic curve for tumor classifier. Area under the curve is 0.87.

5.5 Discussion

TASER is able to capture the information about both telomere sequence and interspersion of telomere like sequences in the (TTAGGG)n tract from WGS data. Here we applied it to 53 paired normal prostate cancer data sets and were able to recapitulate expected telomere differences between the normal and cancer samples. We were unable to find significant differences between the telomere state of cancer samples that correlated with reported stage or age of the patient. The samples analyzed were collected during surgery, requiring that they were advanced to a certain stage for them to be included in this data set. Additionally they are paired with limited phenotypic data that is available publically. The use of WGS on diverse types of datasets, including studies looking at clonal evolution, and in cancers where a broader range of stages of disease can be sampled, will allow for extensive investigation of the dynamics of how telomeres change in the progression of a cancer and between clones in a tumor. As WGS cost continues to decline it will become an integral part of cancer treatment. We show here the potential of telomere information in WGS to be used to distinguish cancer from normal samples. It could also be a useful feature in further sub-classifying tumor types and aggressiveness; for example, more extensive sampling of tumors in different stages of progression, or isolation of cellular subsets from tumors by microdissection or sorting of tumor cell subpopulations followed by WGS could reveal telomere state distinctions detectable by TASER. These studies will be greatly facilitated by an increasing ability to inexpensively acquire WGS data from small cell numbers.

5.6 Supplemental Figures

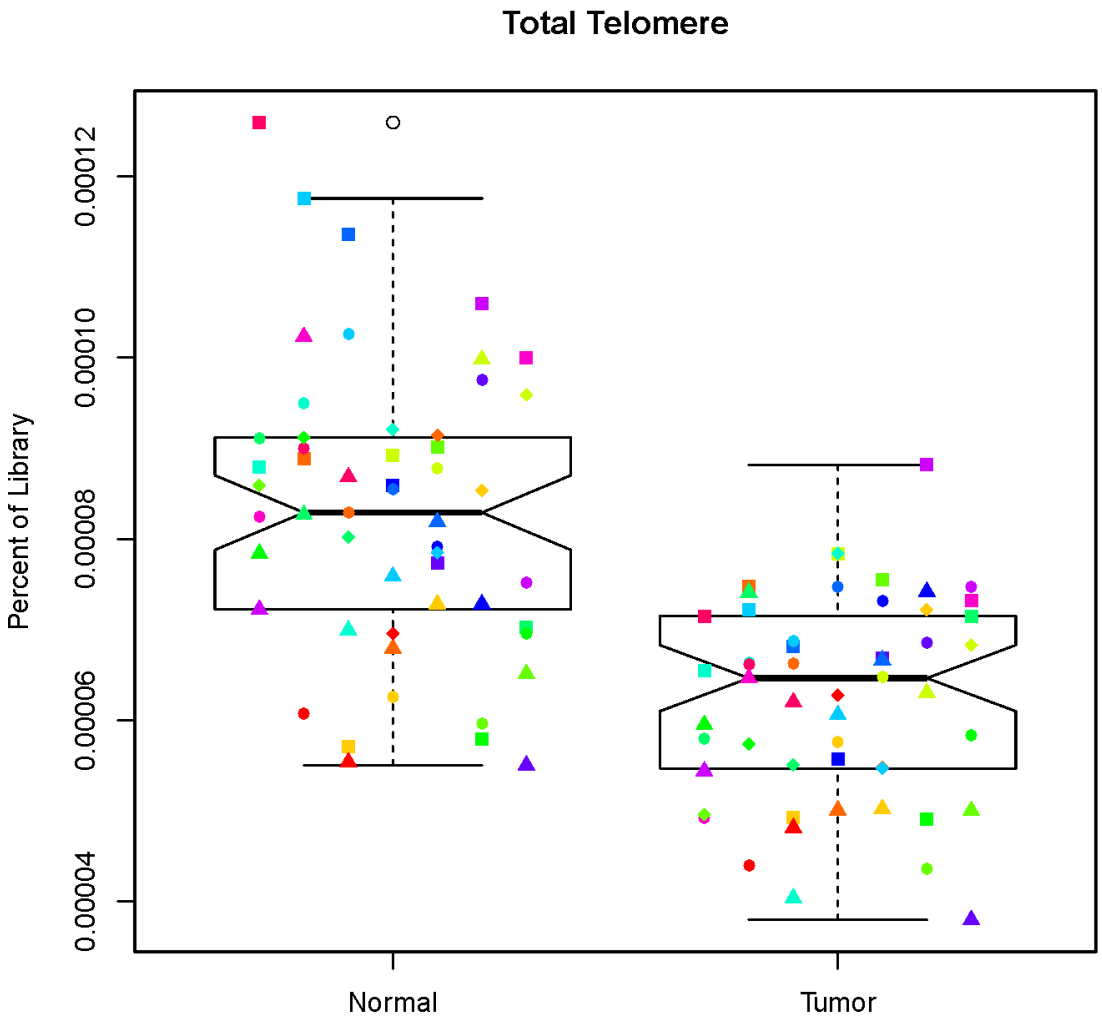

Figure 5.9 – Box plot of total telomere measurement for normal and cancer samples with individual sample points. Box plots show the distribution of values in tumor and normal samples. Values are percent of library, number of reads normalized by library size. The number of reads for total telomere is the sum of all telomere categories, TR, TRPM, and TLSR. Individual samples are shown as a uniquely colored and shaped point.

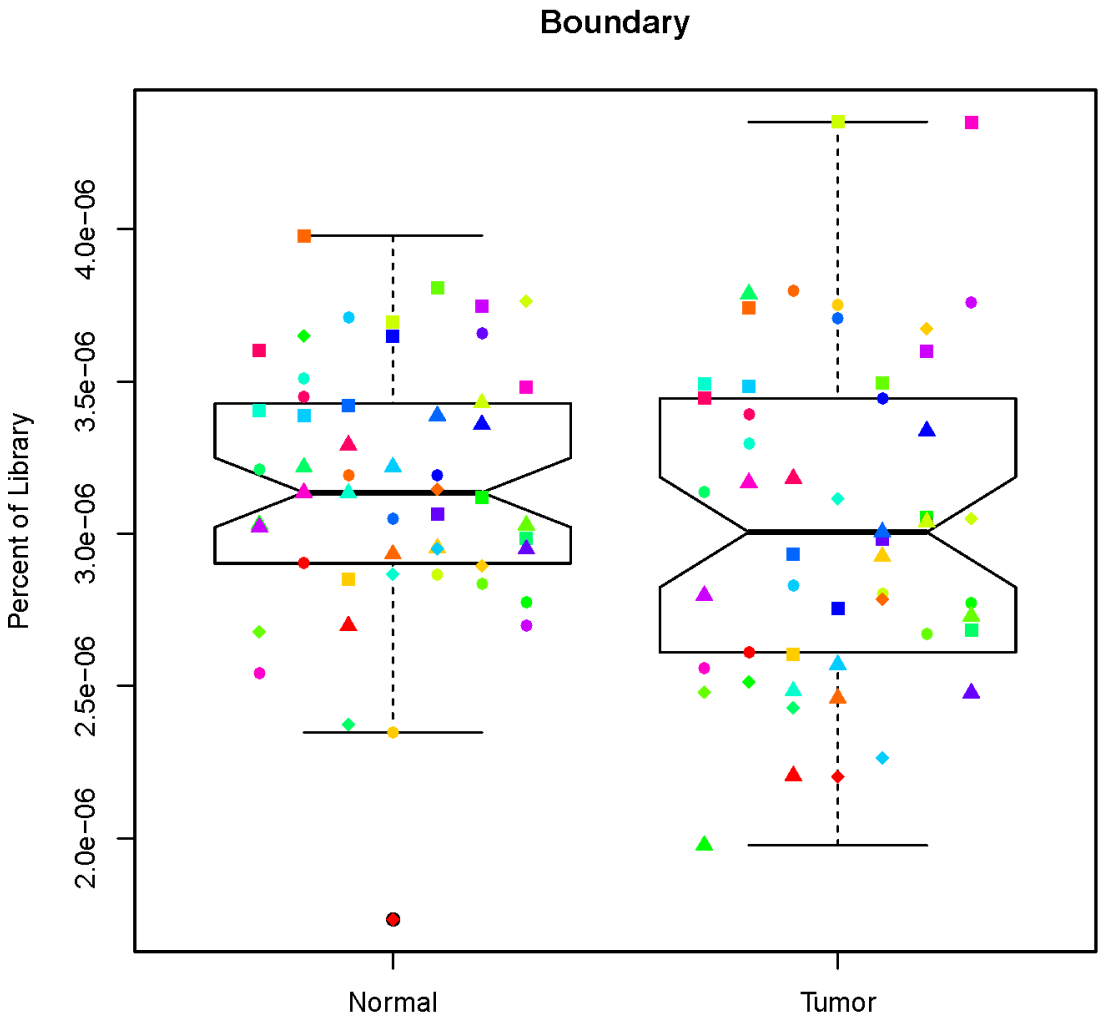

Figure 5.10 – Box plot of boundary measurement for normal and cancer samples with individual sample points. Box plots show the distribution of values in tumor and normal samples. Values are percent of library, number of reads normalized by library size. Boundary reads are any category of telomere reads, TR, TRPM, or TLSR, in pairing with a mapped read. Individual samples are shown as a uniquely colored and shaped point.

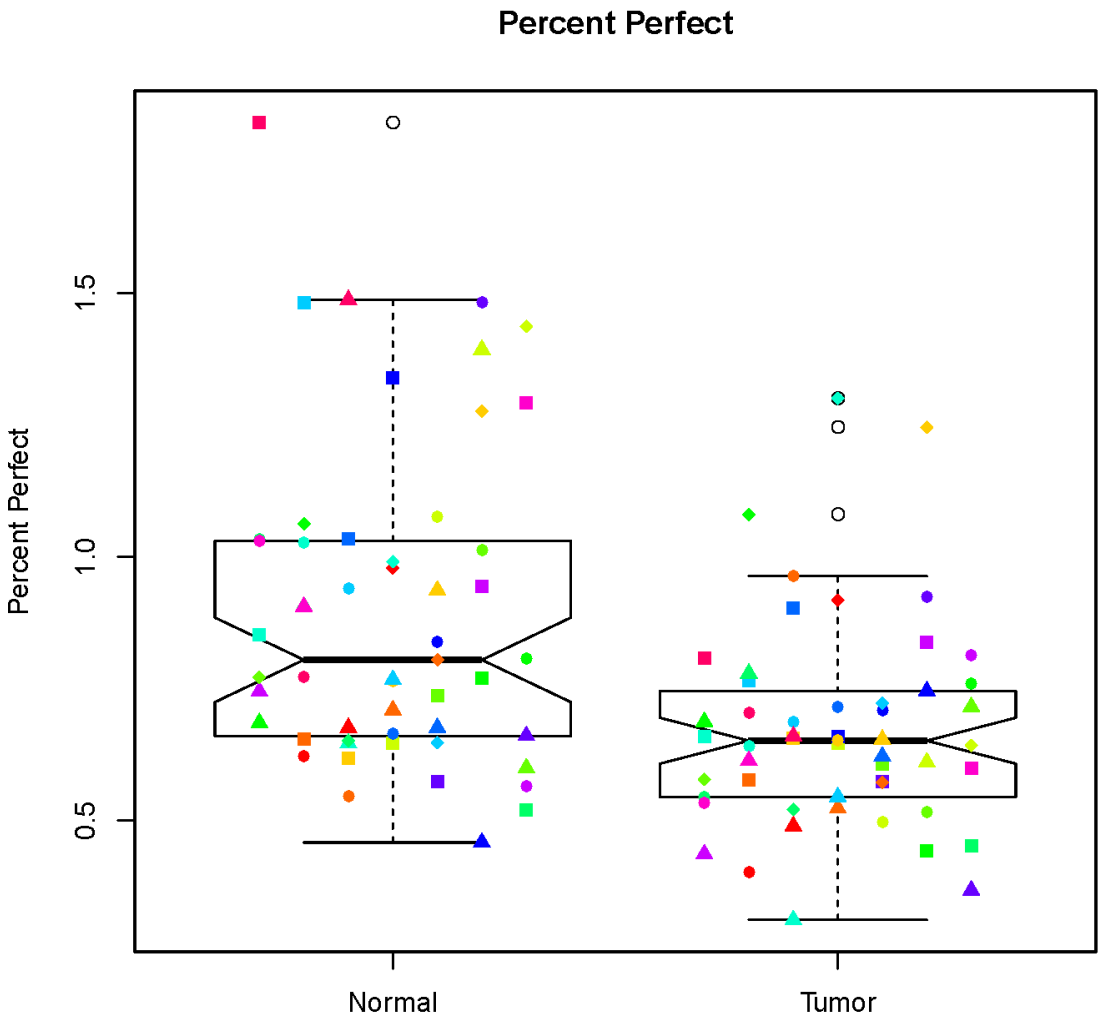

Figure 5.11 – Box plot of percent perfect measurement for normal and cancer samples with Individual sample points. Box plots show the distribution of values in tumor and normal samples. Values are the percent of the telomere reads that perfect, TR, out of total telomere reads TR, TRPM, and TLSR.

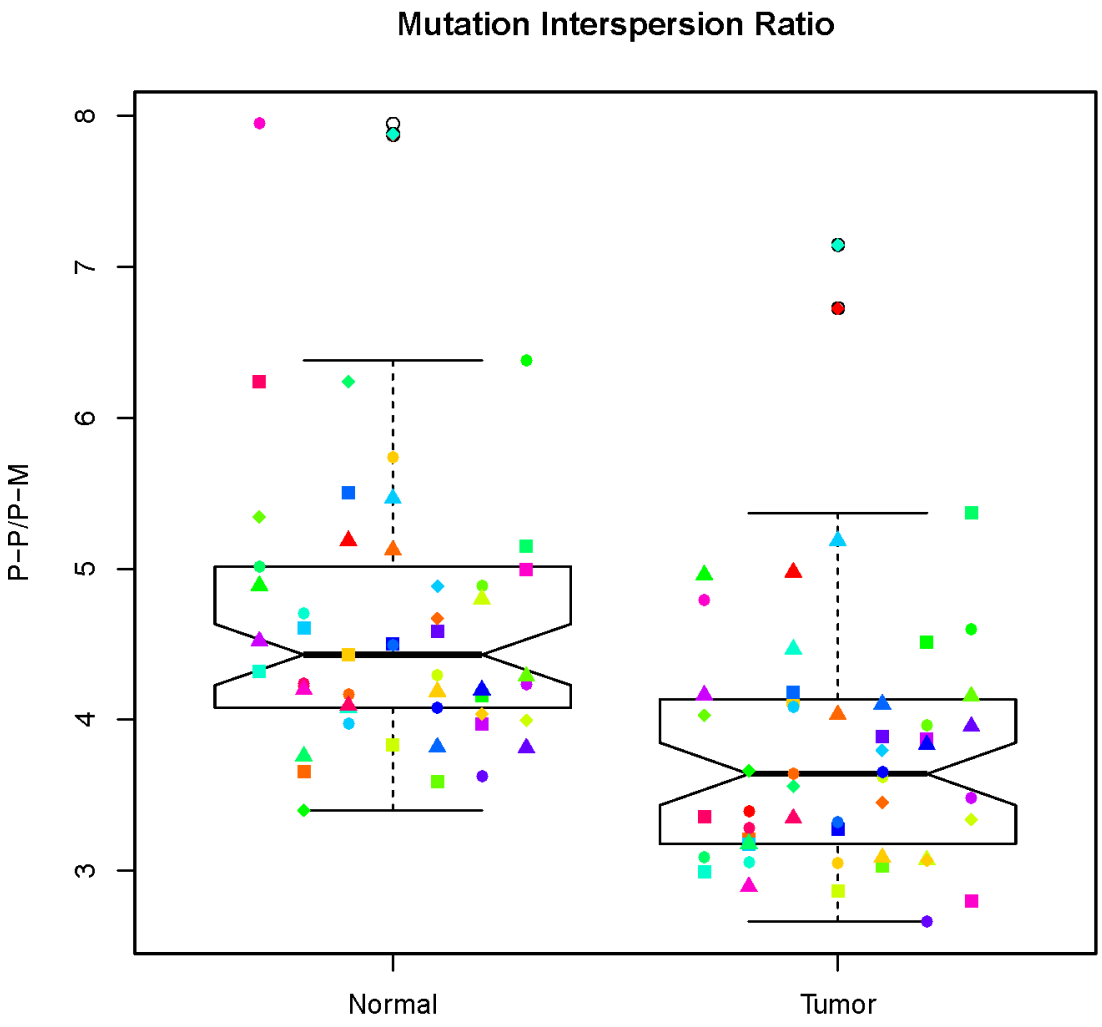

Figure 5.12 – Box plot of mutation interspersion ratio measurement for normal and cancer samples with individual sample points. Box plots show the distribution of values in tumor and normal samples. Values are the calculated ratio. A higher value indicated more perfect/perfect telomere read pairs or less perfect/mutant telomere read pairs. Individual samples are shown as a uniquely colored and shaped point.

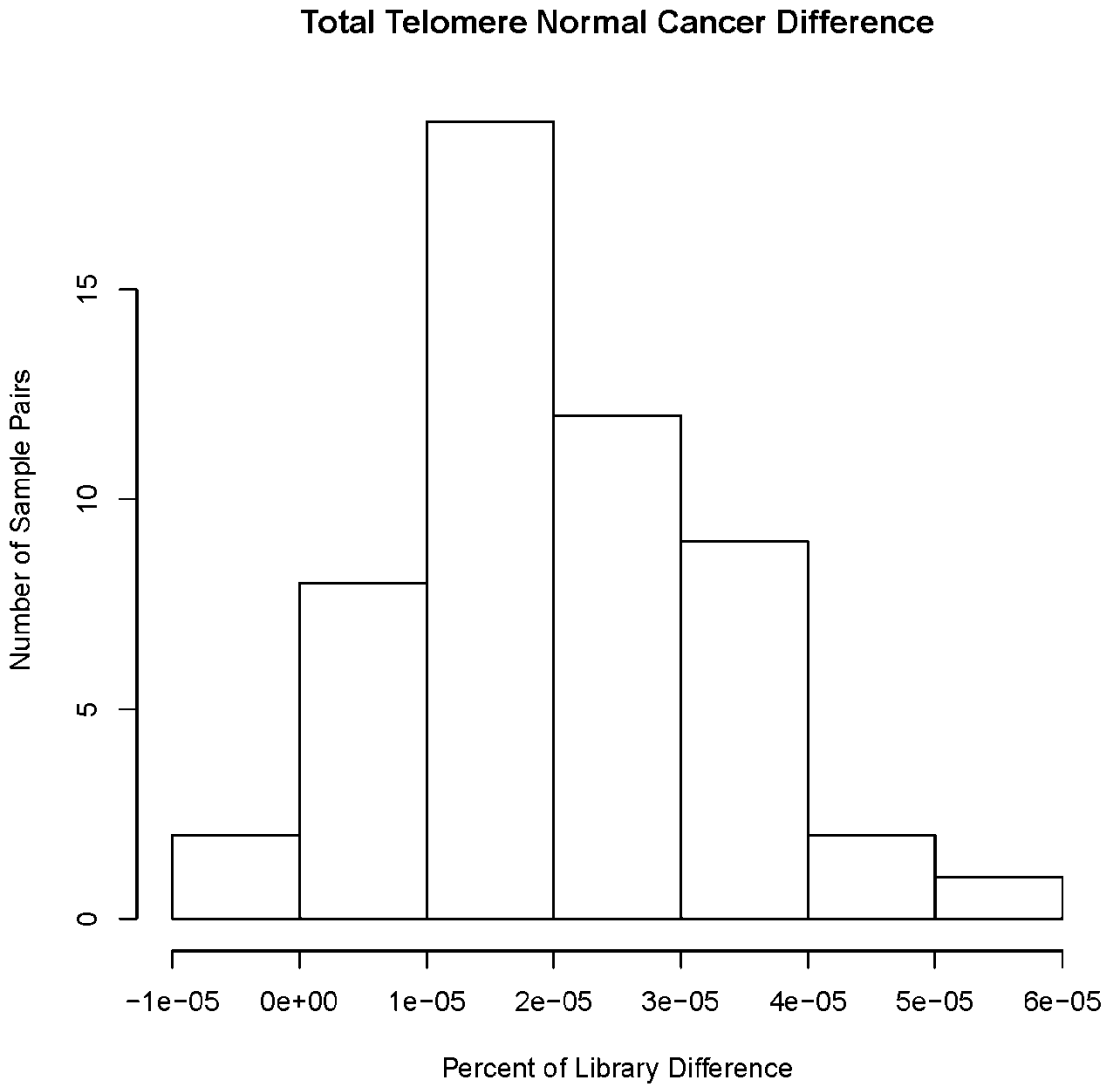

Figure 5.13 – Histogram of difference in total telomere measurement between normal and cancer samples. Box plots show the distribution of values in tumor and normal samples.

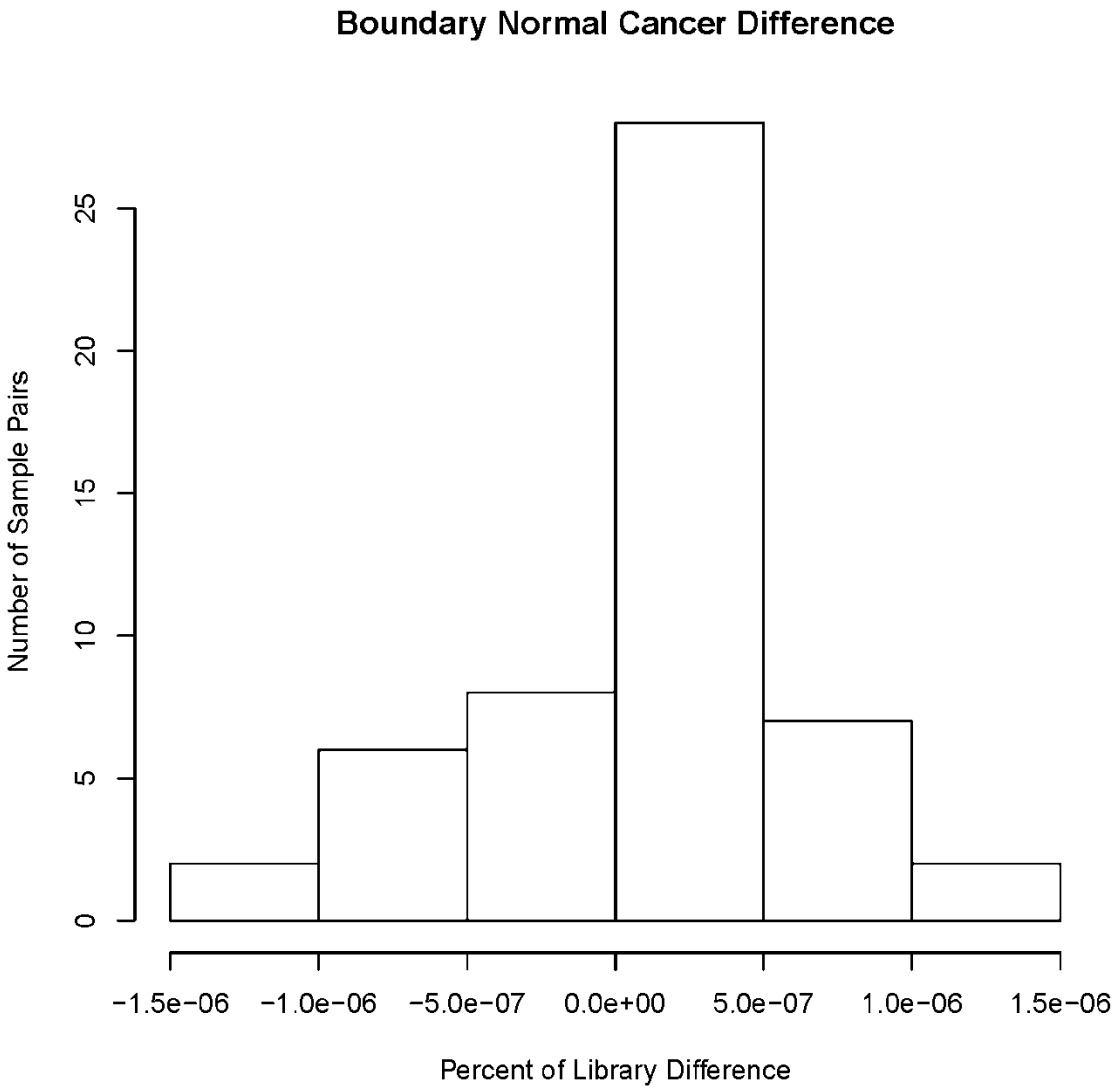

Figure 5.14 – Histogram of difference in boundary measurement between normal and cancer samples. Box plots show the distribution of values in tumor and normal samples.

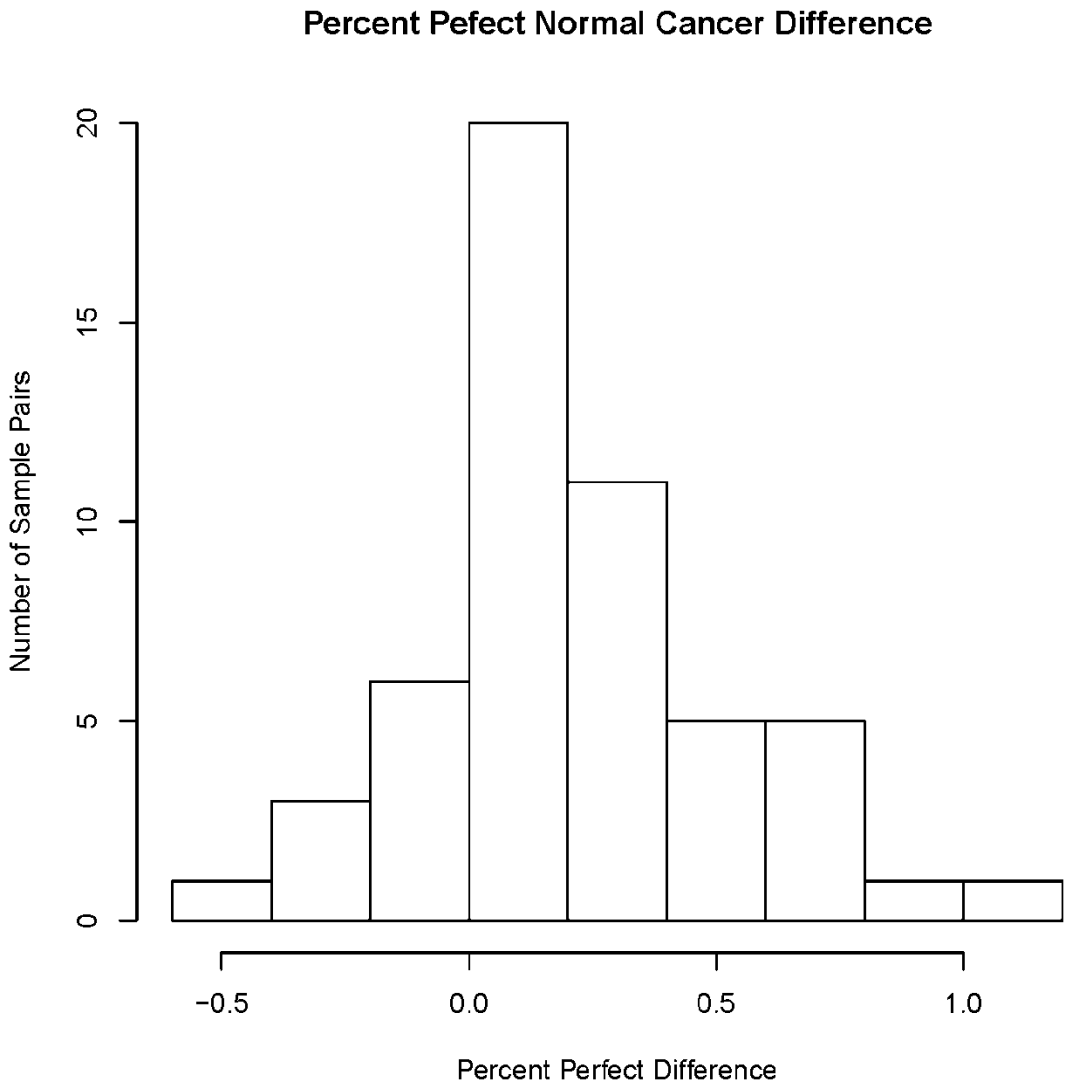

Figure 5.15 – Histogram of difference in percent perfect measurement between normal and cancer samples. Box plots show the distribution of values in tumor and normal samples.

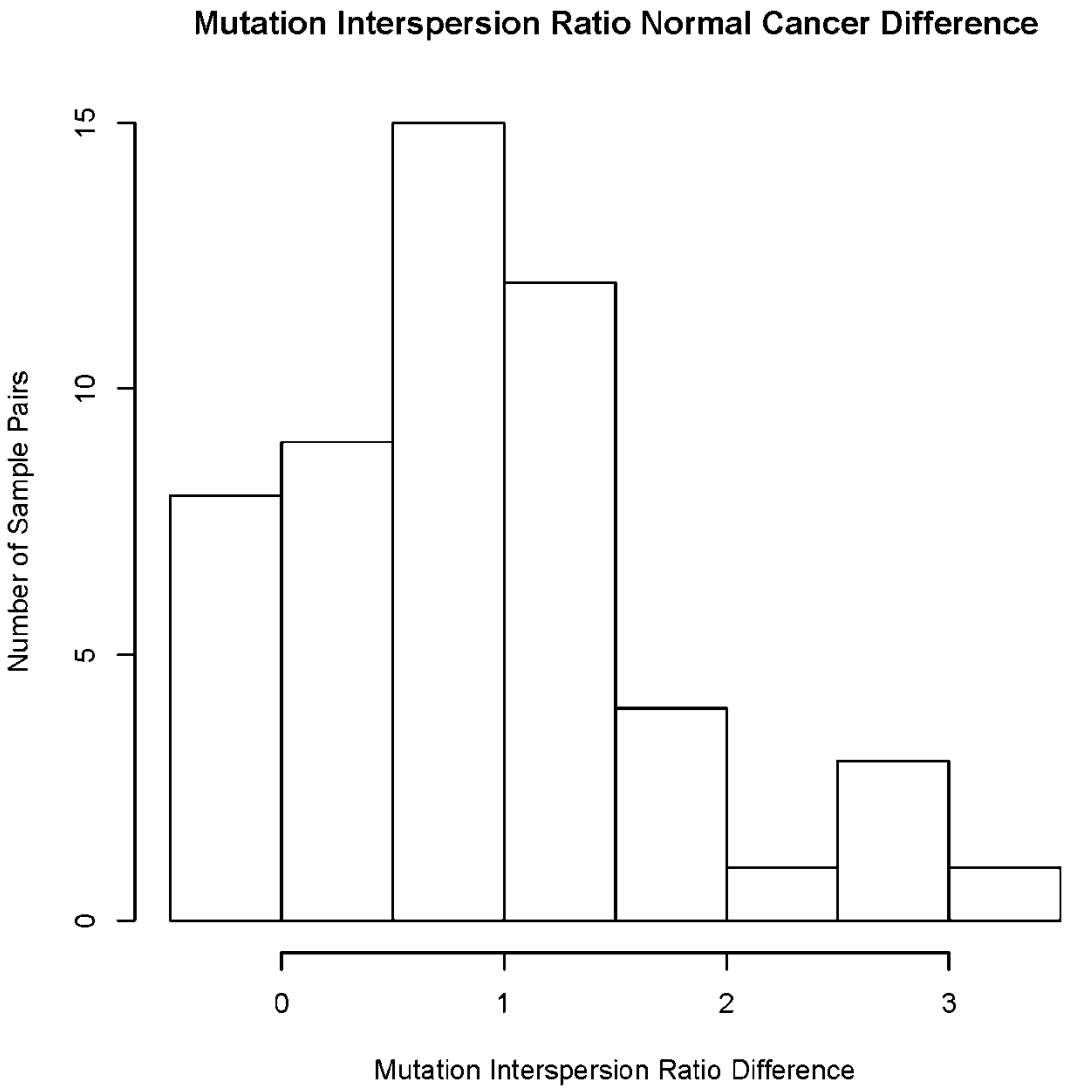

Figure 5.16 – Histogram of difference in mutation interspersion ratio measurement between normal and cancer samples. Box plots show the distribution of values in tumor and normal samples.

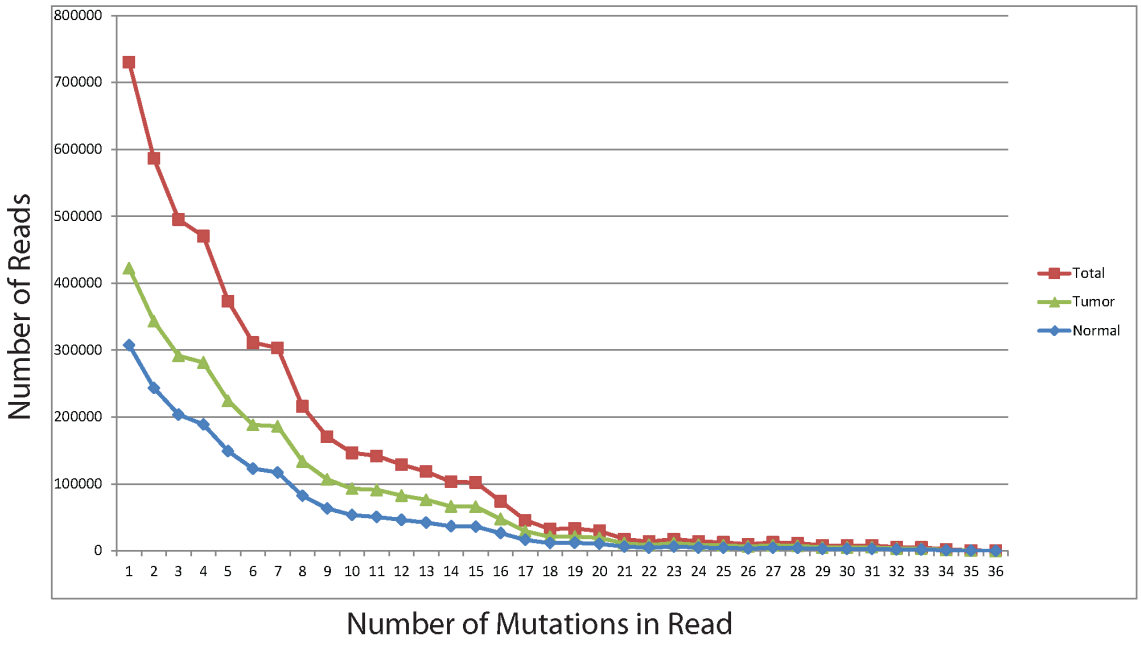

Figure 5.17 – Plots showing the number of reads for increasing numbers of point mutations. The number of reads with increasing number of point mutations for tumor samples, normal samples and combined in total. Because we cannot distinguish a point mutation from a sequencing error in single reads, and there are a potentially large number of Illumina reads with multiple sequencing errors, we elected to categorize otherwise perfect telomere reads with up to a few point mutations/sequencing errors as "perfect" telomere reads. To estimate a cut-off threshold for this category we plotted the number of reads for increasing numbers of point mutations. There is an inflection point in the points of the curves at seven point mutations. This suggests that the counts for reads with less than 7 point mutations might include reads that have been misclassified and called point mutations but are actually sequencing errors, although this cannot be determined for a given read. .

CHAPTER 6: Conclusion

In this dissertation I have presented the novel computational methods I have developed which are necessary to analyze the repeat rich telomere and subtelomere using HTS data. While different strategies are needed to deal with differing numbers of copies in the genome, incorporating multimapping reads allows the use of HTS methods to study telomere biology. Without these methods multimapping reads generated from the telomere and subtelomere are filtered out of the analysis, leaving this important region of the genome unanalyzed.

In chapter 2 I presented the findings of our initial mappings to the 15kb adjacent to the telomere in the human genome. This analysis showed that CTCF and cohesin colocalized immediately adjacent to most human telomeres, and layed the groundwork for subsequent functional studies by our collaborators, who demonstrated the importance of CTCF and cohesin binding for TERRA transcription and telomere stability. In chapter 3 I presented my completed ChIP-seq analysis pipeline and the results of extending the mapping of ChIP-seq data sets to the entire 500kb subtelomere. We analyzed a number of additional data sets, finding stable patterns of regulation in cell types of differing developmental stages. Using significance calls from the completed ChIP-seq pipeline we were able to find an association of CTCF and cohesin binding with boundaries within the subtelomere repeats, and with ITS sites throughout the genome. Both CTCF and cohesin are known to reorganize the 3D structure of the genome. In the subtelomere CTCF and cohesin colocalization in the subtelomere repeats may be responsible for DNA looping to bring distal enhancer elements into proximity with TERRA promoters. ITS sites are hotspots of recombination, which may also be mediated by the structural changes induced by CTCF and cohesin. Investigation of long range interaction in the subtelomere and at

ITS sites will enhance understanding of the role of CTCF and cohesin at these loci. The use of Hi-C and Chia-pet libraries in combination with strategies incorporating multimapping reads will enable this investigation.

In chapter 4 I describe the sequence characteristics of the mouse subtelomere. I show the mouse subtelomere has less duplicated sequence, and the sequence that exists appears to have arisen by a mechanism different than that in human subtelomeres. Parallel to the human subtelomere analysis CTCF and cohesin subunit ChIP-seq data sets were mapped using my novel pipeline for datasets from a number of cell types representing different developmental stages. There is little subtelomere CTCF and cohesin colocalization immediately adjacent to the mouse telomere, showing TERRA transcription is regulated through different mechanisms in the mouse genome. The expanded use of directional RNA-seq data revealed possible evidence of TERRA transcription as it allows observation of transcripts being transcribed towards the telomere, overlapping with known lincRNA sequence elements. Further investigation of TERRA transcription is necessary to find factors regulating its transcription.

In chapter 5 I present the TASER pipeline which captures telomere sequence content from WGS data sets. In TASER I have optimized existing tools for their use in studying changes in telomere read distribution between samples. TASER was used on 53 paired tumor normal pairs from the prostate cancer genome sequencing project. I was able to show expected changes in the prostate cancer, less telomere sequence, particularly the perfect telomere sequence found at the end of the telomere. The information captured by TASER was also used to distinguish normal and cancer samples. While it was not possible with the samples analyzed, the telomere state captured by TASER could be used to better stratify tumors into types of disease or treatment groups; this will be greatly facilitated by sampling tumors at different stages of

progression, and by an increasing ability to inexpensively acquire WGS data from small cell numbers.

This dissertation brings the advances in HTS to the study of telomere biology in the application of data sets generated by WGS, ChIP-seq, and RNA-seq. While other telomere studies have incorporated computational work, here I comprehensively approach the problems that arise in studying the telomere and subtelomere, developing two novel methods to enable proper consideration of the data being analyzed. This has led to important findings in the regulation of TERRA transcription, and telomere dynamics in cancer. The subtelomere analysis has also created a valuable resource as all subtelomere analysis and other characterization is available online, for both the human and mouse subtelomere.

BIBLIOGRAPHY

[1] Muller H. The remaking of chromosomes. Collect Net 1938;13:181–95.

[2] McClintock B. The Stability of Broken Ends of Chromosomes in Zea Mays. Genetics 1941;26:234–82.

[3] Blackburn EH, Gall JG. A tandemly repeated sequence at the termini of the extrachromosomal ribosomal RNA genes in Tetrahymena. J Mol Biol 1978;120:33–53.

[4] Oka Y, Shiota S, Nakai S, Nishida Y, Okubo S. Inverted terminal repeat sequence in the macronuclear DNA of Stylonychia pustulata. Gene 1980;10:301–6.

[5] Klobutcher LA, Swanton MT, Donini P, Prescott DM. All gene-sized DNA molecules in four species of hypotrichs have the same terminal sequence and an unusual 3' terminus. Proc Natl Acad Sci U S A 1981;78:3015–9.

[6] Moyzis RK, Buckingham JM, Cram LS, Dani M, Deaven LL, Jones MD, et al. A highly conserved repetitive DNA sequence, (TTAGGG)n, present at the telomeres of human chromosomes. Proc Natl Acad Sci U S A 1988;85:6622–6.

[7] Allshire RC, Gosden JR, Cross SH, Cranston G, Rout D, Sugawara N, et al. Telomeric repeat from T. thermophila cross hybridizes with human telomeres. Nature 1988;332:656–9. doi:10.1038/332656a0.

[8] Meyne J, Ratliff RL, Moyzis RK. Conservation of the human telomere sequence (TTAGGG)n among vertebrates. Proc Natl Acad Sci U S A 1989;86:7049–53.

[9] Lejnine S, Makarov VL, Langmore JP. Conserved nucleoprotein structure at the ends of vertebrate and invertebrate chromosomes. Proc Natl Acad Sci U S A 1995;92:2393–7.

[10] Baird DM, Britt-Compton B, Rowson J, Amso NN, Gregory L, Kipling D. Telomere instability in the male germline. Hum Mol Genet 2006;15:45–51. doi:10.1093/hmg/ddi424.

[11] Henderson ER, Blackburn EH. An overhanging 3' terminus is a conserved feature of telomeres. Mol Cell Biol 1989;9:345–8.

[12] Wright WE, Tesmer VM, Huffman KE, Levene SD, Shay JW. Normal human chromosomes have long G-rich telomeric overhangs at one end. Genes Dev 1997;11:2801–9.

[13] Watson JD. Origin of concatemeric T7 DNA. Nature New Biol 1972;239:197–201.

[14] Olovnikov AM. A theory of marginotomy. The incomplete copying of template margin in enzymic synthesis of polynucleotides and biological significance of the phenomenon. J Theor Biol 1973;41:181–90.

[15] Olovnikov AM. [Principle of marginotomy in template synthesis of polynucleotides]. Dokl Akad Nauk SSSR 1971;201:1496–9.

[16] Makarov VL, Hirose Y, Langmore JP. Long G Tails at Both Ends of Human Chromosomes Suggest a C Strand Degradation Mechanism for Telomere Shortening. Cell 1997;88:657–66. doi:10.1016/S0092-8674(00)81908-X.

[17] Sfeir AJ, Chai W, Shay JW, Wright WE. Telomere-End Processing. Mol Cell 2005;18:131–8. doi:10.1016/j.molcel.2005.02.035.

[18] Harley CB, Futcher AB, Greider CW. Telomeres shorten during ageing of human fibroblasts. Nature 1990;345:458–60.

[19] Allsopp RC, Chang E, Kashefi-Aazam M, Rogaev EI, Piatyszek MA, Shay JW, et al. Telomere shortening is associated with cell division in vitro and in vivo. Exp Cell Res 1995;220:194–200. doi:10.1006/excr.1995.1306.

[20] Parkinson GN, Lee MPH, Neidle S. Crystal structure of parallel quadruplexes from human telomeric DNA. Nature 2002;417:876–80. doi:10.1038/nature755.
[21] Lipps HJ, Rhodes D. G-quadruplex structures: in vivo evidence and function. Trends Cell Biol 2009;19:414–22. doi:10.1016/j.tcb.2009.05.002.
[22] Biffi G, Tannahill D, McCafferty J, Balasubramanian S. Quantitative visualization of DNA G-quadruplex structures in human cells. Nat Chem 2013;5:182–6. doi:10.1038/nchem.1548.
[23] Chang C-C, Kuo I-C, Ling I-F, Chen C-T, Chen H-C, Lou P-J, et al. Detection of quadruplex DNA structures in human telomeres by a fluorescent carbazole derivative. Anal Chem 2004;76:4490–4. doi:10.1021/ac049510s.
[24] Griffith JD, Comeau L, Rosenfield S, Stansel RM, Bianchi A, Moss H, et al. Mammalian telomeres end in a large duplex loop. Cell 1999;97:503–14.
[25] Stansel RM, de Lange T, Griffith JD. T-loop assembly in vitro involves binding of TRF2 near the 3' telomeric overhang. EMBO J 2001;20:5532–40. doi:10.1093/emboj/20.19.5532.
[26] Zhong Z, Shiue L, Kaplan S, de Lange T. A mammalian factor that binds telomeric TTAGGG repeats in vitro. Mol Cell Biol 1992;12:4834–43.
[27] Chong L, van Steensel B, Broccoli D, Erdjument-Bromage H, Hanish J, Tempst P, et al. A human telomeric protein. Science 1995;270:1663–7.
[28] Bilaud T, Brun C, Ancelin K, Koering CE, Laroche T, Gilson E. Telomeric localization of TRF2, a novel human telobox protein. Nat Genet 1997;17:236–9. doi:10.1038/ng1097-236.
[29] Broccoli D, Smogorzewska A, Chong L, de Lange T. Human telomeres contain two distinct Myb-related proteins, TRF1 and TRF2. Nat Genet 1997;17:231–5. doi:10.1038/ng1097-231.
[30] Baumann P, Cech T, FREE. Pot the putative telomere end-binding protein in fission yeast and humans. Sci 292 Abstr 2001;1 SRC - GoogleScholar:1171–5.
[31] Baumann P, Podell E, Cech TR. Human Pot1 (protection of telomeres) protein: cytolocalization, gene structure, and alternative splicing. Mol Cell Biol 2002;22:8079–87.
[32] Loayza D, Parsons H, Donigian J, Hoke K, de Lange T. DNA binding features of human POT1: a nonamer 5'-TAGGGTTAG-3' minimal binding site, sequence specificity, and internal binding to multimeric sites. J Biol Chem 2004;279:13241–8. doi:10.1074/jbc.M312309200.
[33] Yang Q, Zheng Y-L, Harris CC. POT1 and TRF2 cooperate to maintain telomeric integrity. Mol Cell Biol 2005;25:1070–80. doi:10.1128/MCB.25.3.1070-1080.2005.
[34] Kim S, Kaminker P, Campisi J. TIN2, a new regulator of telomere length in human cells. Nat Genet 1999;23:405–12. doi:10.1038/70508.
[35] Li B, Oestreich S, de Lange T. Identification of Human Rap1: Implications for Telomere Evolution. Cell 2000;101:471–83. doi:10.1016/S0092-8674(00)80858-2.
[36] Houghtaling BR, Cuttonaro L, Chang W, Smith S. A Dynamic Molecular Link between the Telomere Length Regulator TRF1 and the Chromosome End Protector TRF2. Curr Biol 2004;14:1621–31. doi:10.1016/j.cub.2004.08.052.
[37] Liu D, Safari A, O'Connor MS, Chan DW, Laegeler A, Qin J, et al. PTOP interacts with POT1 and regulates its localization to telomeres. Nat Cell Biol 2004;6:673–80. doi:10.1038/ncb1142.

[38] Ye JZ-S, Hockemeyer D, Krutchinsky AN, Loayza D, Hooper SM, Chait BT, et al. POT1-interacting protein PIP1: a telomere length regulator that recruits POT1 to the TIN2/TRF1 complex. Genes Dev 2004;18:1649–54. doi:10.1101/gad.1215404.
[39] Liu D, O'Connor MS, Qin J, Songyang Z. Telosome, a Mammalian Telomere-associated Complex Formed by Multiple Telomeric Proteins. J Biol Chem 2004;279:51338–42. doi:10.1074/jbc.M409293200.
[40] De Lange T. Shelterin: the protein complex that shapes and safeguards human telomeres. Genes Dev 2005;19:2100–10.
[41] Surovtseva YV, Churikov D, Boltz KA, Song X, Lamb JC, Warrington R, et al. Conserved telomere maintenance component 1 interacts with STN1 and maintains chromosome ends in higher eukaryotes. Mol Cell 2009;36:207–18. doi:10.1016/j.molcel.2009.09.017.
[42] Miyake Y, Nakamura M, Nabetani A, Shimamura S, Tamura M, Yonehara S, et al. RPA-like mammalian Ctc1-Stn1-Ten1 complex binds to single-stranded DNA and protects telomeres independently of the Pot1 pathway. Mol Cell 2009;36:193–206. doi:10.1016/j.molcel.2009.08.009.
[43] Price CM, Boltz KA, Chaiken MF, Stewart JA, Beilstein MA, Shippen DE. Evolution of CST function in telomere maintenance. Cell Cycle Georget Tex 2010;9:3157–65. doi:10.4161/cc.9.16.12547.
[44] Gu P, Min J-N, Wang Y, Huang C, Peng T, Chai W, et al. CTC1 deletion results in defective telomere replication, leading to catastrophic telomere loss and stem cell exhaustion. EMBO J 2012;31:2309–21. doi:10.1038/emboj.2012.96.
[45] Von Zglinicki T. Oxidative stress shortens telomeres. Trends Biochem Sci 2002;27:339–44.
[46] Von Zglinicki T, Pilger R, Sitte N. Accumulation of single-strand breaks is the major cause of telomere shortening in human fibroblasts. Free Radic Biol Med 2000;28:64–74.
[47] Greider CW, Blackburn EH. Identification of a specific telomere terminal transferase activity in Tetrahymena extracts. Cell 1985;43:405–13.
[48] Greider CW, Blackburn EH. A telomeric sequence in the RNA of Tetrahymena telomerase required for telomere repeat synthesis. Nature 1989;337:331–7. doi:10.1038/337331a0.
[49] Morin GB. The human telomere terminal transferase enzyme is a ribonucleoprotein that synthesizes TTAGGG repeats. Cell 1989;59:521–9.
[50] Feng J, Funk WD, Wang SS, Weinrich SL, Avilion AA, Chiu CP, et al. The RNA component of human telomerase. Science 1995;269:1236–41.
[51] Meyerson M, Counter CM, Eaton EN, Ellisen LW, Steiner P, Caddle SD, et al. hEST2, the putative human telomerase catalytic subunit gene, is up-regulated in tumor cells and during immortalization. Cell 1997;90:785–95.
[52] Nakamura TM, Morin GB, Chapman KB, Weinrich SL, Andrews WH, Lingner J, et al. Telomerase catalytic subunit homologs from fission yeast and human. Science 1997;277:955–9.
[53] Avilion AA, Piatyszek MA, Gupta J, Shay JW, Bacchetti S, Greider CW. Human Telomerase RNA and Telomerase Activity in Immortal Cell Lines and Tumor Tissues. Cancer Res 1996;56:645–50.
[54] Yi X, Tesmer VM, Savre-Train I, Shay JW, Wright WE. Both Transcriptional and Posttranscriptional Mechanisms Regulate Human Telomerase Template RNA Levels. Mol Cell Biol 1999;19:3989–97.
[55] Poole JC, Andrews LG, Tollefsbol TO. Activity, function, and gene regulation of the catalytic subunit of telomerase (hTERT). Gene 2001;269:1–12.

[56] Hayflick L, Moorhead P. The serial cultivation of human diploid cell strains. Exp Cell Res CrossRefMedlineWeb Sci 1961;25 SRC - GoogleScholar:585–621.
[57] Hayflick L. The limited in vitro lifetime of human diploid cell strains. Exp Cell Res CrossRefMedlineWeb Sci 1965;37 SRC - GoogleScholar:614–36.
[58] Allsopp RC, Vaziri H, Patterson C, Goldstein S, Younglai EV, Futcher AB, et al. Telomere length predicts replicative capacity of human fibroblasts. Proc Natl Acad Sci U S A 1992;89:10114–8.
[59] Karlseder J, Smogorzewska A, de Lange T. Senescence induced by altered telomere state, not telomere loss. Science 2002;295:2446–9. doi:10.1126/science.1069523.
[60] Bodnar AG, Ouellette M, Frolkis M, Holt SE, Chiu CP, Morin GB, et al. Extension of life-span by introduction of telomerase into normal human cells. Science 1998;279:349–52.
[61] Atadja P, Wong H, Garkavtsev I, Veillette C, Riabowol K. Increased activity of p53 in senescing fibroblasts. Proc Natl Acad Sci U S A 1995;92:8348–52.
[62] Shiloh Y. ATM and related protein kinases: safeguarding genome integrity. Nat Rev Cancer 2003;3:155–68. doi:10.1038/nrc1011.
[63] Burma S, Chen BP, Murphy M, Kurimasa A, Chen DJ. ATM Phosphorylates Histone H2AX in Response to DNA Double-strand Breaks. J Biol Chem 2001;276:42462–7. doi:10.1074/jbc.C100466200.
[64] Ward IM, Chen J. Histone H2AX is phosphorylated in an ATR-dependent manner in response to replicational stress. J Biol Chem 2001;276:47759–62. doi:10.1074/jbc.C100569200.
[65] D' Adda di Fagagna F, Reaper PM, Clay-Farrace L, Fiegler H, Carr P, Von Zglinicki T, et al. A DNA damage checkpoint response in telomere-initiated senescence. Nature 2003;426:194–8. doi:10.1038/nature02118.
[66] Zou Y, Sfeir A, Gryaznov SM, Shay JW, Wright WE. Does a sentinel or a subset of short telomeres determine replicative senescence? Mol Biol Cell 2004;15:3709–18. doi:10.1091/mbc.E04-03-0207.
[67] D' Adda di Fagagna F, Teo S-H, Jackson SP. Functional links between telomeres and proteins of the DNA-damage response. Genes Dev 2004;18:1781–99. doi:10.1101/gad.1214504.
[68] Herbig U, Jobling WA, Chen BPC, Chen DJ, Sedivy JM. Telomere shortening triggers senescence of human cells through a pathway involving ATM, p53, and p21(CIP1), but not p16(INK4a). Mol Cell 2004;14:501–13.
[69] Karlseder J, Broccoli D, Dai Y, Hardy S, de Lange T. p53- and ATM-dependent apoptosis induced by telomeres lacking TRF2. Science 1999;283:1321–5.
[70] Smogorzewska A, Lange T de. Different telomere damage signaling pathways in human and mouse cells. EMBO J 2002;21:4338–48. doi:10.1093/emboj/cdf433.
[71] Celli GB, de Lange T. DNA processing is not required for ATM-mediated telomere damage response after TRF2 deletion. Nat Cell Biol 2005;7:712–8. doi:10.1038/ncb1275.
[72] Celli GB, Denchi EL, de Lange T. Ku70 stimulates fusion of dysfunctional telomeres yet protects chromosome ends from homologous recombination. Nat Cell Biol 2006;8:885–90. doi:10.1038/ncb1444.
[73] Wang Y, Ghosh G, Hendrickson EA. Ku86 represses lethal telomere deletion events in human somatic cells. Proc Natl Acad Sci 2009;106:12430–5. doi:10.1073/pnas.0903362106.
[74] Sfeir A, Lange T de. Removal of Shelterin Reveals the Telomere End-Protection Problem. Science 2012;336:593–7. doi:10.1126/science.1218498.

[75] Konishi A, Lange T de. Cell cycle control of telomere protection and NHEJ revealed by a ts mutation in the DNA-binding domain of TRF2. Genes Dev 2008;22:1221–30. doi:10.1101/gad.1634008.
[76] Gisselsson D, Jonson T, Petersén A, Strömbeck B, Dal Cin P, Höglund M, et al. Telomere dysfunction triggers extensive DNA fragmentation and evolution of complex chromosome abnormalities in human malignant tumors. Proc Natl Acad Sci U S A 2001;98:12683–8. doi:10.1073/pnas.211357798.
[77] Gisselsson D, Lv M, Tsao S-W, Man C, Jin C, Höglund M, et al. Telomere-mediated mitotic disturbances in immortalized ovarian epithelial cells reproduce chromosomal losses and breakpoints from ovarian carcinoma. Genes Chromosomes Cancer 2005;42:22–33. doi:10.1002/gcc.20094.
[78] Plug-DeMaggio AW, Sundsvold T, Wurscher MA, Koop JI, Klingelhutz AJ, McDougall JK. Telomere erosion and chromosomal instability in cells expressing the HPV oncogene 16E6. Oncogene 2004;23:3561–71. doi:10.1038/sj.onc.1207388.
[79] Compton SA, Choi J-H, Cesare AJ, Özgür S, Griffith JD. Xrcc3 and Nbs1 Are Required for the Production of Extrachromosomal Telomeric Circles in Human Alternative Lengthening of Telomere Cells. Cancer Res 2007;67:1513–9. doi:10.1158/0008-5472.CAN-06-3672.
[80] Wang RC, Smogorzewska A, de Lange T. Homologous Recombination Generates T-Loop-Sized Deletions at Human Telomeres. Cell 2004;119:355–68. doi:10.1016/j.cell.2004.10.011.
[81] Bryan TM, Englezou A, Dalla-Pozza L, Dunham MA, Reddel RR. Evidence for an alternative mechanism for maintaining telomere length in human tumors and tumor-derived cell lines. Nat Med 1997;3:1271–4.
[82] Reddel RR. Alternative lengthening of telomeres, telomerase, and cancer. Cancer Lett 2003;194:155–62.
[83] Londoño-Vallejo JA, Der-Sarkissian H, Cazes L, Bacchetti S, Reddel RR. Alternative Lengthening of Telomeres Is Characterized by High Rates of Telomeric Exchange. Cancer Res 2004;64:2324–7. doi:10.1158/0008-5472.CAN-03-4035.
[84] Cesare AJ, Griffith JD. Telomeric DNA in ALT Cells Is Characterized by Free Telomeric Circles and Heterogeneous t-Loops. Mol Cell Biol 2004;24:9948–57. doi:10.1128/MCB.24.22.9948-9957.2004.
[85] Heiss NS, Knight SW, Vulliamy TJ, Klauck SM, Wiemann S, Mason PJ, et al. X-linked dyskeratosis congenita is caused by mutations in a highly conserved gene with putative nucleolar functions. Nat Genet 1998;19:32–8. doi:10.1038/ng0598-32.
[86] Vulliamy TJ, Knight SW, Mason PJ, Dokal I. Very short telomeres in the peripheral blood of patients with X-linked and autosomal dyskeratosis congenita. Blood Cells Mol Dis 2001;27:353–7. doi:10.1006/bcmd.2001.0389.
[87] Mitchell JR, Wood E, Collins K. A telomerase component is defective in the human disease dyskeratosis congenita. Nature 1999;402:551–5. doi:10.1038/990141.
[88] Vulliamy T. The RNA component of telomerase is mutated in autosomal dominant dyskeratosis congenita. Nature 2001;413:432–5. doi:10.1038/35096585.
[89] Armanios M, Chen J-L, Chang Y-PC, Brodsky RA, Hawkins A, Griffin CA, et al. Haploinsufficiency of telomerase reverse transcriptase leads to anticipation in autosomal dominant dyskeratosis congenita. Proc Natl Acad Sci U S A 2005;102:15960–4. doi:10.1073/pnas.0508124102.

[90] Savage SA, Giri N, Baerlocher GM, Orr N, Lansdorp PM, Alter BP. TINF2, a component of the shelterin telomere protection complex, is mutated in dyskeratosis congenita. Am J Hum Genet 2008;82:501–9. doi:10.1016/j.ajhg.2007.10.004.
[91] Armanios M, Blackburn EH. The telomere syndromes. Nat Rev Genet 2012;13:693–704. doi:10.1038/nrg3246.
[92] De la Fuente J, Dokal I. Dyskeratosis congenita: advances in the understanding of the telomerase defect and the role of stem cell transplantation. Pediatr Transplant 2007;11:584–94. doi:10.1111/j.1399-3046.2007.00721.x.
[93] Polvi A, Linnankivi T, Kivelä T, Herva R, Keating JP, Mäkitie O, et al. Mutations in CTC1, encoding the CTS telomere maintenance complex component 1, cause cerebroretinal microangiopathy with calcifications and cysts. Am J Hum Genet 2012;90:540–9. doi:10.1016/j.ajhg.2012.02.002.
[94] Anderson BH, Kasher PR, Mayer J, Szynkiewicz M, Jenkinson EM, Bhaskar SS, et al. Mutations in CTC1, encoding conserved telomere maintenance component 1, cause Coats plus. Nat Genet 2012;44:338–42. doi:10.1038/ng.1084.
[95] Alder JK. Short telomeres are a risk factor for idiopathic pulmonary fibrosis. Proc Natl Acad Sci USA 2008;105:13051–6. doi:10.1073/pnas.0804280105.
[96] Armanios MY, Chen JJ-L, Cogan JD, Alder JK, Ingersoll RG, Markin C, et al. Telomerase mutations in families with idiopathic pulmonary fibrosis. N Engl J Med 2007;356:1317–26. doi:10.1056/NEJMoa066157.
[97] Yamaguchi H, Calado RT, Ly H, Kajigaya S, Baerlocher GM, Chanock SJ, et al. Mutations in TERT, the gene for telomerase reverse transcriptase, in aplastic anemia. N Engl J Med 2005;352:1413–24. doi:10.1056/NEJMoa042980.
[98] Yamaguchi H. Mutations of the human telomerase RNA gene (TERC) in aplastic anemia and myelodysplastic syndrome. Blood 2003;102:916–8. doi:10.1182/blood-2003-01-0335.
[99] Du H-Y, Pumbo E, Ivanovich J, An P, Maziarz RT, Reiss UM, et al. TERC and TERT gene mutations in patients with bone marrow failure and the significance of telomere length measurements. Blood 2009;113:309–16. doi:10.1182/blood-2008-07-166421.
[100] Calado RT, Regal JA, Kleiner DE, Schrump DS, Peterson NR, Pons V, et al. A spectrum of severe familial liver disorders associate with telomerase mutations. PloS One 2009;4:e7926. doi:10.1371/journal.pone.0007926.
[101] Calado RT, Young NS. Telomere Diseases. N Engl J Med 2009;361:2353–65. doi:10.1056/NEJMra0903373.
[102] Marrone A, Stevens D, Vulliamy T, Dokal I, Mason PJ. Heterozygous telomerase RNA mutations found in dyskeratosis congenita and aplastic anemia reduce telomerase activity via haploinsufficiency. Blood 2004;104:3936–42. doi:10.1182/blood-2004-05-1829.
[103] Parry EM, Alder JK, Qi X, Chen JJ-L, Armanios M. Syndrome complex of bone marrow failure and pulmonary fibrosis predicts germline defects in telomerase. Blood 2011;117:5607–11. doi:10.1182/blood-2010-11-322149.
[104] Blasco MA, Lee HW, Hande MP, Samper E, Lansdorp PM, DePinho RA, et al. Telomere shortening and tumor formation by mouse cells lacking telomerase RNA. Cell 1997;91:25–34.
[105] Lee HW, Blasco MA, Gottlieb GJ, Horner JW, Greider CW, DePinho RA. Essential role of mouse telomerase in highly proliferative organs. Nature 1998;392:569–74.

[106] Herrera E, Samper E, Martín-Caballero J, Flores JM, Lee HW, Blasco MA. Disease states associated with telomerase deficiency appear earlier in mice with short telomeres. EMBO J 1999;18:2950–60. doi:10.1093/emboj/18.11.2950.
[107] Armanios M. Short telomeres are sufficient to cause the degenerative defects associated with aging. Am J Hum Genet 2009;85:823–32. doi:10.1016/j.ajhg.2009.10.028.
[108] Vaziri H, Dragowska W, Allsopp RC, Thomas TE, Harley CB, Lansdorp PM. Evidence for a mitotic clock in human hematopoietic stem cells: loss of telomeric DNA with age. Proc Natl Acad Sci U S A 1994;91:9857–60.
[109] Cherif H, Tarry JL, Ozanne SE, Hales CN. Ageing and telomeres: a study into organ- and gender-specific telomere shortening. Nucleic Acids Res 2003;31:1576–83. doi:10.1093/nar/gkg208.
[110] Rufer N, Brümmendorf TH, Kolvraa S, Bischoff C, Christensen K, Wadsworth L, et al. Telomere Fluorescence Measurements in Granulocytes and T Lymphocyte Subsets Point to a High Turnover of Hematopoietic Stem Cells and Memory T Cells in Early Childhood. J Exp Med 1999;190:157–68. doi:10.1084/jem.190.2.157.
[111] Allsopp RC, Morin GB, DePinho R, Harley CB, Weissman IL. Telomerase is required to slow telomere shortening and extend replicative lifespan of HSCs during serial transplantation. Blood 2003;102:517–20. doi:10.1182/blood-2002-07-2334.
[112] Rossi DJ, Bryder D, Seita J, Nussenzweig A, Hoeijmakers J, Weissman IL. Deficiencies in DNA damage repair limit the function of haematopoietic stem cells with age. Nature 2007;447:725–9. doi:10.1038/nature05862.
[113] Choudhury AR, Ju Z, Djojosubroto MW, Schienke A, Lechel A, Schaetzlein S, et al. Cdkn1a deletion improves stem cell function and lifespan of mice with dysfunctional telomeres without accelerating cancer formation. Nat Genet 2007;39:99–105. doi:10.1038/ng1937.
[114] Rudolph KL, Chang S, Lee HW, Blasco M, Gottlieb GJ, Greider C, et al. Longevity, stress response, and cancer in aging telomerase-deficient mice. Cell 1999;96:701–12.
[115] Satoh M, Minami Y, Takahashi Y, Tabuchi T, Itoh T, Nakamura M. Effect of intensive lipid-lowering therapy on telomere erosion in endothelial progenitor cells obtained from patients with coronary artery disease. Clin Sci 2009;116:827–35. doi:10.1042/CS20080404.
[116] Brouilette SW, Moore JS, McMahon AD, Thompson JR, Ford I, Shepherd J, et al. Telomere length, risk of coronary heart disease, and statin treatment in the West of Scotland Primary Prevention Study: a nested case-control study. The Lancet 2007;369:107–14. doi:10.1016/S0140-6736(07)60071-3.
[117] O'Donnell CJ, Demissie S, Kimura M, Levy D, Gardner JP, White C, et al. Leukocyte Telomere Length and Carotid Artery Intimal Medial Thickness The Framingham Heart Study. Arterioscler Thromb Vasc Biol 2008;28:1165–71. doi:10.1161/ATVBAHA.107.154849.
[118] Fitzpatrick AL, Kronmal RA, Gardner JP, Psaty BM, Jenny NS, Tracy RP, et al. Leukocyte Telomere Length and Cardiovascular Disease in the Cardiovascular Health Study. Am J Epidemiol 2007;165:14–21. doi:10.1093/aje/kwj346.
[119] Brouilette S, Singh RK, Thompson JR, Goodall AH, Samani NJ. White Cell Telomere Length and Risk of Premature Myocardial Infarction. Arterioscler Thromb Vasc Biol 2003;23:842–6. doi:10.1161/01.ATV.0000067426.96344.32.
[120] Jiang H, Schiffer E, Song Z, Wang J, Zürbig P, Thedieck K, et al. Proteins induced by telomere dysfunction and DNA damage represent biomarkers of human aging and disease. Proc Natl Acad Sci U S A 2008;105:11299–304. doi:10.1073/pnas.0801457105.

[121] Campisi J. Senescent Cells, Tumor Suppression, and Organismal Aging: Good Citizens, Bad Neighbors. Cell 2005;120:513–22. doi:10.1016/j.cell.2005.02.003.
[122] Chin L, Artandi SE, Shen Q, Tam A, Lee SL, Gottlieb GJ, et al. p53 deficiency rescues the adverse effects of telomere loss and cooperates with telomere dysfunction to accelerate carcinogenesis. Cell 1999;97:527–38.
[123] Hara E, Tsurui H, Shinozaki A, Nakada S, Oda K. Cooperative effect of antisense-Rb and antisense-p53 oligomers on the extension of life span in human diploid fibroblasts, TIG-1. Biochem Biophys Res Commun 1991;179:528–34. doi:10.1016/0006-291X(91)91403-Y.
[124] Counter CM, Avilion AA, LeFeuvre CE, Stewart NG, Greider CW, Harley CB, et al. Telomere shortening associated with chromosome instability is arrested in immortal cells which express telomerase activity. EMBO J 1992;11:1921–9.
[125] Kim NW, Piatyszek MA, Prowse KR, Harley CB, West MD, Ho PL, et al. Specific association of human telomerase activity with immortal cells and cancer. Science 1994;266:2011–5. doi:10.1126/science.7605428.
[126] Shay JW, van der Haegen BA, Ying Y, Wright WE. The Frequency of Immortalization of Human Fibroblasts and Mammary Epithelial Cells Transfected with SV40 Large T-Antigen. Exp Cell Res 1993;209:45–52. doi:10.1006/excr.1993.1283.
[127] Henson JD, Hannay JA, McCarthy SW, Royds JA, Yeager TR, Robinson RA, et al. A robust assay for alternative lengthening of telomeres in tumors shows the significance of alternative lengthening of telomeres in sarcomas and astrocytomas. Clin Cancer Res Off J Am Assoc Cancer Res 2005;11:217–25.
[128] Hanahan D, Weinberg RA. The Hallmarks of Cancer. Cell 2000;100:57–70. doi:10.1016/S0092-8674(00)81683-9.
[129] Ducray C, Pommier J-P, Martins L, Boussin FD, Sabatier L. Telomere dynamics, end-to-end fusions and telomerase activation during the human fibroblast immortalization process. Oncogene 1999;18:4211–23. doi:10.1038/sj.onc.1202797.
[130] Deng W, Tsao SW, Guan X-Y, Lucas JN, Si HX, Leung CS, et al. Distinct profiles of critically short telomeres are a key determinant of different chromosome aberrations in immortalized human cells: whole-genome evidence from multiple cell lines. Oncogene 2004;23:9090–101. doi:10.1038/sj.onc.1208119.
[131] Britt-Compton B, Capper R, Rowson J, Baird DM. Short telomeres are preferentially elongated by telomerase in human cells. FEBS Lett 2009;583:3076–80. doi:10.1016/j.febslet.2009.08.029.
[132] Hemann MT, Strong MA, Hao LY, Greider CW. The shortest telomere, not average telomere length, is critical for cell viability and chromosome stability. Cell 2001;107:67–77.
[133] Zhao Y, Sfeir AJ, Zou Y, Buseman CM, Chow TT, Shay JW, et al. Telomere Extension Occurs at Most Chromosome Ends and Is Uncoupled from Fill-In in Human Cancer Cells. Cell 2009;138:463–75. doi:10.1016/j.cell.2009.05.026.
[134] Fusenig NE, Boukamp P. Multiple stages and genetic alterations in immortalization, malignant transformation, and tumor progression of human skin keratinocytes. Mol Carcinog 1998;23:144–58.
[135] Fearon ER, Vogelstein B. A genetic model for colorectal tumorigenesis. Cell 1990;61:759–67. doi:10.1016/0092-8674(90)90186-I.
[136] Lengauer C, Kinzler KW, Vogelstein B. Genetic instabilities in human cancers. Nature 1998;396:643–9. doi:10.1038/25292.

[137] Artandi SE, Chang S, Lee SL, Alson S, Gottlieb GJ, Chin L, et al. Telomere dysfunction promotes non-reciprocal translocations and epithelial cancers in mice. Nature 2000;406:641–5. doi:10.1038/35020592.
[138] Windle B, Draper BW, Yin YX, O'Gorman S, Wahl GM. A central role for chromosome breakage in gene amplification, deletion formation, and amplicon integration. Genes Dev 1991;5:160–74. doi:10.1101/gad.5.2.160.
[139] Maser RS, Choudhury B, Campbell PJ, Feng B, Wong K-K, Protopopov A, et al. Chromosomally unstable mouse tumours have genomic alterations similar to diverse human cancers. Nature 2007;447:966–71. doi:10.1038/nature05886.
[140] O'Hagan RC, Chang S, Maser RS, Mohan R, Artandi SE, Chin L, et al. Telomere dysfunction provokes regional amplification and deletion in cancer genomes. Cancer Cell 2002;2:149–55. doi:10.1016/S1535-6108(02)00094-6.
[141] Smogorzewska A, Karlseder J, Holtgreve-Grez H, Jauch A, de Lange T. DNA Ligase IV-Dependent NHEJ of Deprotected Mammalian Telomeres in G1 and G2. Curr Biol 2002;12:1635–44. doi:10.1016/S0960-9822(02)01179-X.
[142] Chadeneau C, Hay K, Hirte HW, Gallinger S, Bacchetti S. Telomerase Activity Associated with Acquisition of Malignancy in Human Colorectal Cancer. Cancer Res 1995;55:2533–6.
[143] Engelhardt M, Drullinsky P, Guillem J, Moore MA. Telomerase and telomere length in the development and progression of premalignant lesions to colorectal cancer. Clin Cancer Res 1997;3:1931–41.
[144] Odagiri E, Kanda N, Jibiki K, Demura R, Aikawa E, Demura H. Reduction of telomeric length and c-erbb-2 gene amplification in human breast cancer, fibroadenoma, and gynecomastia. Relationship to histologic grade and clinical parameters. Cancer 1994;73:2978–84. doi:10.1002/1097-0142(19940615)73:12<2978::AID-CNCR2820731215>3.0.CO;2-5.
[145] Meeker AK, Gage WR, Hicks JL, Simon I, Coffman JR, Platz EA, et al. Telomere length assessment in human archival tissues: combined telomere fluorescence in situ hybridization and immunostaining. Am J Pathol 2002;160:1259–68. doi:10.1016/S0002-9440(10)62553-9.
[146] Meeker AK, Hicks JL, Gabrielson E, Strauss WM, De Marzo AM, Argani P. Telomere shortening occurs in subsets of normal breast epithelium as well as in situ and invasive carcinoma. Am J Pathol 2004;164:925–35. doi:10.1016/S0002-9440(10)63180-X.
[147] Van Heek NT, Meeker AK, Kern SE, Yeo CJ, Lillemoe KD, Cameron JL, et al. Telomere shortening is nearly universal in pancreatic intraepithelial neoplasia. Am J Pathol 2002;161:1541–7. doi:10.1016/S0002-9440(10)64432-X.
[148] Chin K, de Solorzano CO, Knowles D, Jones A, Chou W, Rodriguez EG, et al. In situ analyses of genome instability in breast cancer. Nat Genet 2004;36:984–8. doi:10.1038/ng1409.
[149] Meeker AK, Hicks JL, Iacobuzio-Donahue CA, Montgomery EA, Westra WH, Chan TY, et al. Telomere length abnormalities occur early in the initiation of epithelial carcinogenesis. Clin Cancer Res Off J Am Assoc Cancer Res 2004;10:3317–26. doi:10.1158/1078-0432.CCR-0984-03.
[150] Cawthon RM. Telomere measurement by quantitative PCR. Nucleic Acids Res 2002;30:e47–e47. doi:10.1093/nar/30.10.e47.
[151] Baird DM, Rowson J, Wynford-Thomas D, Kipling D. Extensive allelic variation and ultrashort telomeres in senescent human cells. Nat Genet 2003;33:203–7. doi:10.1038/ng1084.

[152] Lansdorp PM, Verwoerd NP, van de Rijke FM, Dragowska V, Little MT, Dirks RW, et al. Heterogeneity in telomere length of human chromosomes. Hum Mol Genet 1996;5:685–91.
[153] Zijlmans JMJM, Martens UM, Poon SSS, Raap AK, Tanke HJ, Ward RK, et al. Telomeres in the mouse have large inter-chromosomal variations in the number of T2AG3 repeats. Proc Natl Acad Sci 1997;94:7423–8.
[154] Londoño-Vallejo JA, DerSarkissian H, Cazes L, Thomas G. Differences in telomere length between homologous chromosomes in humans. Nucleic Acids Res 2001;29:3164–71.
[155] Nordfjäll K, Larefalk A, Lindgren P, Holmberg D, Roos G. Telomere length and heredity: Indications of paternal inheritance. Proc Natl Acad Sci U S A 2005;102:16374–8. doi:10.1073/pnas.0501724102.
[156] Graakjaer J, Pascoe L, Der-Sarkissian H, Thomas G, Kolvraa S, Christensen K, et al. The relative lengths of individual telomeres are defined in the zygote and strictly maintained during life. Aging Cell 2004;3:97–102. doi:10.1111/j.1474-9728.2004.00093.x.
[157] Graakjaer J, Bischoff C, Korsholm L, Holstebroe S, Vach W, Bohr VA, 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 2003;124:629–40.
[158] Graakjaer J, Der-Sarkissian H, Schmitz A, Bayer J, Thomas G, Kolvraa S, et al. Allele-specific relative telomere lengths are inherited. Hum Genet 2006;119:344–50. doi:10.1007/s00439-006-0137-x.
[159] 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 2006;15:725–33. doi:10.1093/hmg/ddi486.
[160] Royle NJ, Hill MC, Jeffreys AJ. Isolation of Telomere Junction Fragments by Anchored Polymerase Chain Reaction. Proc R Soc Lond B Biol Sci 1992;247:57–67. doi:10.1098/rspb.1992.0009.
[161] Brown WRA, MacKinnon PJ, Villasanté A, Spurr N, Buckle VJ, Dobson MJ. Structure and polymorphism of human telomere-associated DNA. Cell 1990;63:119–32. doi:10.1016/0092-8674(90)90293-N.
[162] Ambrosini A, Paul S, Hu S, Riethman H. Human subtelomeric duplicon structure and organization. Genome Biol 2007;8:R151. doi:10.1186/gb-2007-8-7-r151.
[163] Rudd MK, Friedman C, Parghi SS, Linardopoulou EV, Hsu L, Trask BJ. Elevated rates of sister chromatid exchange at chromosome ends. PLoS Genet 2007;3:e32. doi:10.1371/journal.pgen.0030032.
[164] Linardopoulou EV, Williams EM, Fan Y, Friedman C, Young JM, Trask BJ. Human subtelomeres are hot spots of interchromosomal recombination and segmental duplication. Nature 2005;437:94–100. doi:10.1038/nature04029.
[165] Riethman H, Ambrosini A, Castaneda C, Finklestein J, Hu X-L, Mudunuri U, et al. Mapping and initial analysis of human subtelomeric sequence assemblies. Genome Res 2004;14:18–28. doi:10.1101/gr.1245004.
[166] Bailey JA, Gu Z, Clark RA, Reinert K, Samonte RV, Schwartz S, et al. Recent Segmental Duplications in the Human Genome. Science 2002;297:1003–7. doi:10.1126/science.1072047.
[167] Riethman H. Human telomere structure and biology. Annu Rev Genomics Hum Genet 2008;9:1–19. doi:10.1146/annurev.genom.8.021506.172017.

[168] Stavenhagen JB, Zakian VA. Internal tracts of telomeric DNA act as silencers in Saccharomyces cerevisiae. Genes Dev 1994;8:1411–22. doi:10.1101/gad.8.12.1411.
[169] Lundblad V. Telomere maintenance without telomerase. Oncogene 2002;21:522–31. doi:10.1038/sj.onc.1205079.
[170] Mondello C, Pirzio L, Azzalin CM, Giulotto E. Instability of Interstitial Telomeric Sequences in the Human Genome. Genomics 2000;68:111–7. doi:10.1006/geno.2000.6280.
[171] Azzalin CM, Nergadze SG, Giulotto E. Human intrachromosomal telomeric-like repeats: sequence organization and mechanisms of origin. Chromosoma 2001;110:75–82.
[172] Iafrate AJ, Feuk L, Rivera MN, Listewnik ML, Donahoe PK, Qi Y, et al. Detection of large-scale variation in the human genome. Nat Genet 2004;36:949–51. doi:10.1038/ng1416.
[173] Sebat J, Lakshmi B, Troge J, Alexander J, Young J, Lundin P, et al. Large-Scale Copy Number Polymorphism in the Human Genome. Science 2004;305:525–8. doi:10.1126/science.1098918.
[174] Redon R, Ishikawa S, Fitch KR, Feuk L, Perry GH, Andrews TD, et al. Global variation in copy number in the human genome. Nature 2006;444:444–54. doi:10.1038/nature05329.
[175] Hazelrigg T, Levis R, Rubin GM. Transformation of white locus DNA in Drosophila: Dosage compensation, zeste interaction, and position effects. Cell 1984;36:469–81. doi:10.1016/0092-8674(84)90240-X.
[176] Gottschling DE, Aparicio OM, Billington BL, Zakian VA. Position effect at S. cerevisiae telomeres: Reversible repression of Pol II transcription. Cell 1990;63:751–62. doi:10.1016/0092-8674(90)90141-Z.
[177] Linardopoulou 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 2007;3:e237. doi:10.1371/journal.pgen.0030237.
[178] Trask BJ, Friedman C, Martin-Gallardo A, Rowen L, Akinbami C, Blankenship J, 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 1998;7:13–26. doi:10.1093/hmg/7.1.13.
[179] Luke B, Panza A, Redon S, Iglesias N, Li Z, Lingner J. The Rat1p 5′ to 3′ Exonuclease Degrades Telomeric Repeat-Containing RNA and Promotes Telomere Elongation in Saccharomyces cerevisiae. Mol Cell 2008;32:465–77. doi:10.1016/j.molcel.2008.10.019.
[180] Vrbsky J, Akimcheva S, Watson JM, Turner TL, Daxinger L, Vyskot B, et al. siRNA–Mediated Methylation of Arabidopsis Telomeres. PLoS Genet 2010;6:e1000986. doi:10.1371/journal.pgen.1000986.
[181] Azzalin CM, Reichenbach P, Khoriauli L, Giulotto E, Lingner J. Telomeric Repeat–Containing RNA and RNA Surveillance Factors at Mammalian Chromosome Ends. Science 2007;318:798–801. doi:10.1126/science.1147182.
[182] Schoeftner S, Blasco MA. Developmentally regulated transcription of mammalian telomeres by DNA-dependent RNA polymerase II. Nat Cell Biol 2008;10:228–36. doi:10.1038/ncb1685.
[183] Porro A, Feuerhahn S, Reichenbach P, Lingner J. Molecular Dissection of Telomeric Repeat-Containing RNA Biogenesis Unveils the Presence of Distinct and Multiple Regulatory Pathways. Mol Cell Biol 2010;30:4808–17. doi:10.1128/MCB.00460-10.
[184] Nergadze SG, Farnung BO, Wischnewski H, Khoriauli L, Vitelli V, Chawla R, et al. CpG-island promoters drive transcription of human telomeres. RNA N Y N 2009;15:2186–94.

[185] Deng Z, Norseen J, Wiedmer A, Riethman H, Lieberman PM. TERRA RNA Binding to TRF2 Facilitates Heterochromatin Formation and ORC Recruitment at Telomeres. Mol Cell 2009;35:403–13. doi:10.1016/j.molcel.2009.06.025.
[186] Scheibe M, Arnoult N, Kappei D, Buchholz F, Decottignies A, Butter F, et al. Quantitative interaction screen of telomeric repeat-containing RNA reveals novel TERRA regulators. Genome Res 2013;23:2149–57. doi:10.1101/gr.151878.112.
[187] Pfeiffer V, Lingner J. TERRA Promotes Telomere Shortening through Exonuclease 1–Mediated Resection of Chromosome Ends. PLoS Genet 2012;8:e1002747. doi:10.1371/journal.pgen.1002747.
[188] Maicher A, Kastner L, Dees M, Luke B. Deregulated telomere transcription causes replication-dependent telomere shortening and promotes cellular senescence. Nucleic Acids Res 2012;40:6649–59. doi:10.1093/nar/gks358.
[189] Flynn RL, Centore RC, O'Sullivan RJ, Rai R, Tse A, Songyang Z, et al. TERRA and hnRNPA1 orchestrate an RPA-to-POT1 switch on telomeric single-stranded DNA. Nature 2011;471:532–6. doi:10.1038/nature09772.
[190] De Silanes IL, d' Alcontres MS, Blasco MA. TERRA transcripts are bound by a complex array of RNA-binding proteins. Nat Commun 2010;1:33. doi:10.1038/ncomms1032.
[191] Le PN, Maranon DG, Altina NH, Battaglia CLR, Bailey SM. TERRA, hnRNP A1, and DNA-PKcs Interactions at Human Telomeres. Front Oncol 2013;3. doi:10.3389/fonc.2013.00091.
[192] Sampl S, Pramhas S, Stern C, Preusser M, Marosi C, Holzmann K. Expression of Telomeres in Astrocytoma WHO Grade 2 to 4: TERRA Level Correlates with Telomere Length, Telomerase Activity, and Advanced Clinical Grade. Transl Oncol 2012;5:56–IN4. doi:10.1593/tlo.11202.
[193] Ng LJ, Cropley JE, Pickett HA, Reddel RR, Suter CM. Telomerase activity is associated with an increase in DNA methylation at the proximal subtelomere and a reduction in telomeric transcription. Nucleic Acids Res 2009;37:1152–9. doi:10.1093/nar/gkn1030.
[194] Deng Z, Wang Z, Xiang C, Molczan A, Baubet V, Conejo-Garcia J, et al. Formation of telomeric repeat-containing RNA (TERRA) foci in highly proliferating mouse cerebellar neuronal progenitors and medulloblastoma. J Cell Sci 2012;125:4383–94. doi:10.1242/jcs.108118.
[195] Anderson DE, Losada A, Erickson HP, Hirano T. Condensin and cohesin display different arm conformations with characteristic hinge angles. J Cell Biol 2002;156:419–24. doi:10.1083/jcb.200111002.
[196] Haering CH, Löwe J, Hochwagen A, Nasmyth K. Molecular Architecture of SMC Proteins and the Yeast Cohesin Complex. Mol Cell 2002;9:773–88. doi:10.1016/S1097-2765(02)00515-4.
[197] How cohesin and CTCF cooperate in regulating gene expression - Springer n.d. doi:10.1007/s10577-008-9017-7.
[198] Donze D, Adams CR, Rine J, Kamakaka RT. The boundaries of the silenced HMR domain in Saccharomyces cerevisiae. Genes Dev 1999;13:698–708.
[199] Losada A, Hirano M, Hirano T. Identification of Xenopus SMC protein complexes required for sister chromatid cohesion. Genes Dev 1998;12:1986–97. doi:10.1101/gad.12.13.1986.
[200] Sumara I, Vorlaufer E, Gieffers C, Peters BH, Peters J-M. Characterization of Vertebrate Cohesin Complexes and Their Regulation in Prophase. J Cell Biol 2000;151:749–62. doi:10.1083/jcb.151.4.749.

[201] Waizenegger IC, Hauf S, Meinke A, Peters J-M. Two Distinct Pathways Remove Mammalian Cohesin from Chromosome Arms in Prophase and from Centromeres in Anaphase. Cell 2000;103:399–410. doi:10.1016/S0092-8674(00)00132-X.
[202] Wendt KS, Yoshida K, Itoh T, Bando M, Koch B, Schirghuber E, et al. Cohesin mediates transcriptional insulation by CCCTC-binding factor. Nature 2008;451:796–801. doi:10.1038/nature06634.
[203] Zhang B, Jain S, Song H, Fu M, Heuckeroth RO, Erlich JM, et al. Mice lacking sister chromatid cohesion protein PDS5B exhibit developmental abnormalities reminiscent of Cornelia de Lange syndrome. Dev Camb Engl 2007;134:3191–201. doi:10.1242/dev.005884.
[204] Parelho V, Hadjur S, Spivakov M, Leleu M, Sauer S, Gregson HC, et al. Cohesins Functionally Associate with CTCF on Mammalian Chromosome Arms. Cell 2008;132:422–33. doi:10.1016/j.cell.2008.01.011.
[205] Stedman W, Kang H, Lin S, Kissil JL, Bartolomei MS, Lieberman PM. Cohesins localize with CTCF at the KSHV latency control region and at cellular c-myc and H19/Igf2 insulators. EMBO J 2008;27:654–66. doi:10.1038/emboj.2008.1.
[206] Rubio ED, Reiss DJ, Welcsh PL, Disteche CM, Filippova GN, Baliga NS, et al. CTCF physically links cohesin to chromatin. Proc Natl Acad Sci 2008;105:8309–14. doi:10.1073/pnas.0801273105.
[207] Lobanenkov VV, Nicolas RH, Adler VV, Paterson H, Klenova EM, Polotskaja AV, et al. A novel sequence-specific DNA binding protein which interacts with three regularly spaced direct repeats of the CCCTC-motif in the 5'-flanking sequence of the chicken c-myc gene. Oncogene 1990;5:1743–53.
[208] Hark AT, Schoenherr CJ, Katz DJ, Ingram RS, Levorse JM, Tilghman SM. CTCF mediates methylation-sensitive enhancer-blocking activity at the H19/Igf2 locus. Nature 2000;405:486–9. doi:10.1038/35013106.
[209] Kanduri C, Pant V, Loukinov D, Pugacheva E, Qi C-F, Wolffe A, et al. Functional association of CTCF with the insulator upstream of the H19 gene is parent of origin-specific and methylation-sensitive. Curr Biol 2000;10:853–6. doi:10.1016/S0960-9822(00)00597-2.
[210] Hikichi T, Kohda T, Kaneko-Ishino T, Ishino F. Imprinting regulation of the murine Meg1/Grb10 and human GRB10 genes; roles of brain-specific promoters and mouse-specific CTCF-binding sites. Nucleic Acids Res 2003;31:1398–406. doi:10.1093/nar/gkg232.
[211] Bell AC, Felsenfeld G. Methylation of a CTCF-dependent boundary controls imprinted expression of the Igf2 gene. Nature 2000;405:482–5. doi:10.1038/35013100.
[212] Pant V, Kurukuti S, Pugacheva E, Shamsuddin S, Mariano P, Renkawitz R, et al. Mutation of a single CTCF target site within the H19 imprinting control region leads to loss of Igf2 imprinting and complex patterns of de novo methylation upon maternal inheritance. Mol Cell Biol 2004;24:3497–504.
[213] Phillips JE, Corces VG. CTCF: master weaver of the genome. Cell 2009;137:1194–211. doi:10.1016/j.cell.2009.06.001.
[214] Fuss SH, Omura M, Mombaerts P. Local and cis Effects of the H Element on Expression of Odorant Receptor Genes in Mouse. Cell 2007;130:373–84. doi:10.1016/j.cell.2007.06.023.

[215] Lomvardas S, Barnea G, Pisapia DJ, Mendelsohn M, Kirkland J, Axel R. Interchromosomal Interactions and Olfactory Receptor Choice. Cell 2006;126:403–13. doi:10.1016/j.cell.2006.06.035.
[216] Murrell A, Heeson S, Reik W. Interaction between differentially methylated regions partitions the imprinted genes Igf2 and H19 into parent-specific chromatin loops. Nat Genet 2004;36:889–93. doi:10.1038/ng1402.
[217] Kurukuti S, Tiwari VK, Tavoosidana G, Pugacheva E, Murrell A, Zhao Z, et al. CTCF binding at the H19 imprinting control region mediates maternally inherited higher-order chromatin conformation to restrict enhancer access to Igf2. Proc Natl Acad Sci 2006;103:10684–9. doi:10.1073/pnas.0600326103.
[218] Engel N, Raval AK, Thorvaldsen JL, Bartolomei SM. Three-dimensional conformation at the H19/Igf2 locus supports a model of enhancer tracking. Hum Mol Genet 2008;17:3021–9. doi:10.1093/hmg/ddn200.
[219] Dressman D, Yan H, Traverso G, Kinzler KW, Vogelstein B. Transforming single DNA molecules into fluorescent magnetic particles for detection and enumeration of genetic variations. Proc Natl Acad Sci 2003;100:8817–22. doi:10.1073/pnas.1133470100.
[220] Ronaghi M, Karamohamed S, Pettersson B, Uhlén M, Nyrén P. Real-Time DNA Sequencing Using Detection of Pyrophosphate Release. Anal Biochem 1996;242:84–9. doi:10.1006/abio.1996.0432.
[221] Ronaghi M, Uhlén M, Nyrén P. A Sequencing Method Based on Real-Time Pyrophosphate. Science 1998;281:363–5. doi:10.1126/science.281.5375.363.
[222] Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, et al. Genome sequencing in microfabricated high-density picolitre reactors. Nature 2005. doi:10.1038/nature03959.
[223] Landegren U, Kaiser R, Sanders J, Hood L. A ligase-mediated gene detection technique. Science 1988;241:1077–80. doi:10.1126/science.3413476.
[224] Valouev A, Ichikawa J, Tonthat T, Stuart J, Ranade S, Peckham H, et al. A high-resolution, nucleosome position map of C. elegans reveals a lack of universal sequence-dictated positioning. Genome Res 2008;18:1051–63. doi:10.1101/gr.076463.108.
[225] Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, et al. Accurate whole human genome sequencing using reversible terminator chemistry. Nature 2008;456:53–9. doi:10.1038/nature07517.
[226] Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol 1990;215:403–10. doi:10.1016/S0022-2836(05)80360-2.
[227] Smith TF, Waterman MS. Identification of common molecular subsequences. J Mol Biol 1981;147:195–7. doi:10.1016/0022-2836(81)90087-5.
[228] Gotoh O. An improved algorithm for matching biological sequences. J Mol Biol 1982;162:705–8. doi:10.1016/0022-2836(82)90398-9.
[229] Ma B, Tromp J, Li M. PatternHunter: faster and more sensitive homology search. Bioinformatics 2002;18:440–5. doi:10.1093/bioinformatics/18.3.440.
[230] Li R, Li Y, Kristiansen K, Wang J. SOAP: short oligonucleotide alignment program. Bioinformatics 2008;24:713–4. doi:10.1093/bioinformatics/btn025.
[231] Jiang H, Wong WH. SeqMap: mapping massive amount of oligonucleotides to the genome. Bioinformatics 2008;24:2395–6. doi:10.1093/bioinformatics/btn429.
[232] Li H, Ruan J, Durbin R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome Res 2008;18:1851–8. doi:10.1101/gr.078212.108.

[233] Rumble SM, Lacroute P, Dalca AV, Fiume M, Sidow A, Brudno M. SHRiMP: Accurate Mapping of Short Color-space Reads. PLoS Comput Biol 2009;5:e1000386. doi:10.1371/journal.pcbi.1000386.
[234] Abouelhoda MI, Kurtz S, Ohlebusch E. Replacing suffix trees with enhanced suffix arrays. J Discrete Algorithms 2004;2:53–86. doi:10.1016/S1570-8667(03)00065-0.
[235] Burrows M, Wheeler DJ, Burrows M, Wheeler DJ. A block-sorting lossless data compression algorithm. 1994.
[236] Ferragina P, Manzini G. Opportunistic data structures with applications. 41st Annu. Symp. Found. Comput. Sci. 2000 Proc., 2000, p. 390–8. doi:10.1109/SFCS.2000.892127.
[237] Li R, Yu C, Li Y, Lam T-W, Yiu S-M, Kristiansen K, et al. SOAP2: an improved ultrafast tool for short read alignment. Bioinformatics 2009;25:1966–7. doi:10.1093/bioinformatics/btp336.
[238] Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol 2009;10:R25. doi:10.1186/gb-2009-10-3-r25.
[239] Li H, Durbin R. Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics 2009;25:1754–60. doi:10.1093/bioinformatics/btp324.
[240] Batzer MA, Deininger PL. ALU REPEATS AND HUMAN GENOMIC DIVERSITY. Nat Rev Genet 2002;3:370–9. doi:10.1038/nrg798.
[241] Schmid CW, Deininger PL. Sequence organization of the human genome. Cell 1975;6:345–58. doi:10.1016/0092-8674(75)90184-1.
[242] McKenna AH. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res 2010;20:1297–303. doi:10.1101/gr.107524.110.
[243] Ye K, Schulz MH, Long Q, Apweiler R, Ning Z. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. Bioinformatics 2009;25:2865–71. doi:10.1093/bioinformatics/btp394.
[244] Chen K, Wallis JW, McLellan MD, Larson DE, Kalicki JM, Pohl CS, et al. BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. Nat Methods 2009;6:677–81. doi:10.1038/nmeth.1363.
[245] Ding L, Ley TJ, Larson DE, Miller CA, Koboldt DC, Welch JS, et al. Clonal evolution in relapsed acute myeloid leukaemia revealed by whole-genome sequencing. Nature 2012;481:506–10. doi:10.1038/nature10738.
[246] Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods 2008;5:621–8. doi:10.1038/nmeth.1226.
[247] Li B, Ruotti V, Stewart RM, Thomson JA, Dewey CN. RNA-Seq gene expression estimation with read mapping uncertainty. Bioinformatics 2010;26:493–500. doi:10.1093/bioinformatics/btp692.
[248] Jiang H, Wong WH. Statistical inferences for isoform expression in RNA-Seq. Bioinformatics 2009;25:1026–32. doi:10.1093/bioinformatics/btp113.
[249] Lu F, Tsai K, Chen H-S, Wikramasinghe P, Davuluri RV, Showe L, et al. Identification of host-chromosome binding sites and candidate gene targets for Kaposi's sarcoma-associated herpesvirus LANA. J Virol 2012;86:5752–62. doi:10.1128/JVI.07216-11.
[250] Lee B-K, Bhinge AA, Battenhouse A, McDaniell RM, Liu Z, Song L, et al. Cell-type specific and combinatorial usage of diverse transcription factors revealed by genome-wide

binding studies in multiple human cells. Genome Res 2012;22:9–24. doi:10.1101/gr.127597.111.
[251] Kasowski M, Grubert F, Heffelfinger C, Hariharan M, Asabere A, Waszak SM, et al. Variation in transcription factor binding among humans. Science 2010;328:232–5. doi:10.1126/science.1183621.
[252] Riethman H, Ambrosini A, Paul S. Human subtelomere structure and variation. Chromosome Res Int J Mol Supramol Evol Asp Chromosome Biol 2005;13:505–15. doi:10.1007/s10577-005-0998-1.
[253] DeScipio C, Spinner NB, Kaur M, Yaeger D, Conlin LK, Ambrosini A, et al. Fine-mapping subtelomeric deletions and duplications by comparative genomic hybridization in 42 individuals. Am J Med Genet A 2008;146A:730–9. doi:10.1002/ajmg.a.32216.
[254] Faulkner GJ, Forrest ARR, Chalk AM, Schroder K, Hayashizaki Y, Carninci P, et al. A rescue strategy for multimapping short sequence tags refines surveys of transcriptional activity by CAGE. Genomics 2008;91:281–8. doi:10.1016/j.ygeno.2007.11.003.
[255] Kim TH, Abdullaev ZK, Smith AD, Ching KA, Loukinov DI, Green RD, et al. Analysis of the vertebrate insulator protein CTCF-binding sites in the human genome. Cell 2007;128:1231–45. doi:10.1016/j.cell.2006.12.048.
[256] Palm W, de Lange T. How shelterin protects mammalian telomeres. Annu Rev Genet 2008;42:301–34. doi:10.1146/annurev.genet.41.110306.130350.
[257] Meier A, Fiegler H, Muñoz P, Ellis P, Rigler D, Langford C, et al. Spreading of mammalian DNA-damage response factors studied by ChIP-chip at damaged telomeres. EMBO J 2007;26:2707–18. doi:10.1038/sj.emboj.7601719.
[258] Coppé J-P, Desprez P-Y, Krtolica A, Campisi J. The senescence-associated secretory phenotype: the dark side of tumor suppression. Annu Rev Pathol 2010;5:99–118. doi:10.1146/annurev-pathol-121808-102144.
[259] Davalos AR, Coppe J-P, Campisi J, Desprez P-Y. Senescent cells as a source of inflammatory factors for tumor progression. Cancer Metastasis Rev 2010;29:273–83. doi:10.1007/s10555-010-9220-9.
[260] Jaskelioff M, Muller FL, Paik J-H, Thomas E, Jiang S, Adams A, et al. Telomerase reactivation reverses tissue degeneration in aged telomerase deficient mice. Nature 2011;469:102–6. doi:10.1038/nature09603.
[261] Sahin E, Depinho RA. Linking functional decline of telomeres, mitochondria and stem cells during ageing. Nature 2010;464:520–8. doi:10.1038/nature08982.
[262] Yehezkel S, Segev Y, Viegas-Péquignot E, Skorecki K, Selig S. Hypomethylation of subtelomeric regions in ICF syndrome is associated with abnormally short telomeres and enhanced transcription from telomeric regions. Hum Mol Genet 2008;17:2776–89. doi:10.1093/hmg/ddn177.
[263] Caslini C, Connelly JA, Serna A, Broccoli D, Hess JL. MLL associates with telomeres and regulates telomeric repeat-containing RNA transcription. Mol Cell Biol 2009;29:4519–26. doi:10.1128/MCB.00195-09.
[264] Arnoult N, Van Beneden A, Decottignies A. Telomere length regulates TERRA levels through increased trimethylation of telomeric H3K9 and HP1α. Nat Struct Mol Biol 2012;19:948–56. doi:10.1038/nsmb.2364.
[265] Deng Z, Wang Z, Stong N, Plasschaert R, Moczan A, Chen H-S, et al. A role for CTCF and cohesin in subtelomere chromatin organization, TERRA transcription, and telomere end protection. EMBO J 2012;31:4165–78. doi:10.1038/emboj.2012.266.

[266] Riethman H. Human subtelomeric copy number variations. Cytogenet Genome Res 2008;123:244–52. doi:10.1159/000184714.
[267] Bushey AM, Dorman ER, Corces VG. Chromatin insulators: regulatory mechanisms and epigenetic inheritance. Mol Cell 2008;32:1–9. doi:10.1016/j.molcel.2008.08.017.
[268] Ohlsson R, Lobanenkov V, Klenova E. Does CTCF mediate between nuclear organization and gene expression? BioEssays News Rev Mol Cell Dev Biol 2010;32:37–50. doi:10.1002/bies.200900118.
[269] Ottaviani A, Schluth-Bolard C, Gilson E, Magdinier F. D4Z4 as a prototype of CTCF and lamins-dependent insulator in human cells. Nucl Austin Tex 2011;1:30–6. doi:10.4161/nucl.1.1.10799.
[270] Ottaviani A, Rival-Gervier S, Boussouar A, Foerster AM, Rondier D, Sacconi S, et al. The D4Z4 macrosatellite repeat acts as a CTCF and A-type lamins-dependent insulator in facio-scapulo-humeral dystrophy. PLoS Genet 2009;5:e1000394. doi:10.1371/journal.pgen.1000394.
[271] Ottaviani A, Schluth-Bolard C, Rival-Gervier S, Boussouar A, Rondier D, Foerster AM, et al. Identification of a perinuclear positioning element in human subtelomeres that requires A-type lamins and CTCF. EMBO J 2009;28:2428–36. doi:10.1038/emboj.2009.201.
[272] Hirano T. At the heart of the chromosome: SMC proteins in action. Nat Rev Mol Cell Biol 2006;7:311–22. doi:10.1038/nrm1909.
[273] Nasmyth K, Haering CH. The structure and function of SMC and kleisin complexes. Annu Rev Biochem 2005;74:595–648. doi:10.1146/annurev.biochem.74.082803.133219.
[274] Dorsett D. Cohesin: genomic insights into controlling gene transcription and development. Curr Opin Genet Dev 2011;21:199–206. doi:10.1016/j.gde.2011.01.018.
[275] Kagey MH, Newman JJ, Bilodeau S, Zhan Y, Orlando DA, van Berkum NL, et al. Mediator and cohesin connect gene expression and chromatin architecture. Nature 2010;467:430–5. doi:10.1038/nature09380.
[276] Canudas S, Smith S. Differential regulation of telomere and centromere cohesion by the Scc3 homologues SA1 and SA2, respectively, in human cells. J Cell Biol 2009;187:165–73. doi:10.1083/jcb.200903096.
[277] Remeseiro S, Cuadrado A, Carretero M, Martínez P, Drosopoulos WC, Cañamero M, et al. Cohesin-SA1 deficiency drives aneuploidy and tumourigenesis in mice due to impaired replication of telomeres. EMBO J 2012;31:2076–89. doi:10.1038/emboj.2012.11.
[278] Bisht KK, Daniloski Z, Smith S. SA1 binds directly to DNA through its unique AT-hook to promote sister chromatid cohesion at telomeres. J Cell Sci 2013;126:3493–503. doi:10.1242/jcs.130872.
[279] International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. Nature 2004;431:931–45. doi:10.1038/nature03001.
[280] Der-Sarkissian H, Vergnaud G, Borde Y-M, Thomas G, Londoño-Vallejo J-A. Segmental polymorphisms in the proterminal regions of a subset of human chromosomes. Genome Res 2002;12:1673–8. doi:10.1101/gr.322802.
[281] Mefford HC, Trask BJ. The complex structure and dynamic evolution of human subtelomeres. Nat Rev Genet 2002;3:91–102. doi:10.1038/nrg727.
[282] Altschul S. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997;25:3389–402. doi:10.1093/nar/25.17.3389.
[283] Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinforma Oxf Engl 2010;26:841–2. doi:10.1093/bioinformatics/btq033.

[284] Smit AFA, Hubley R, Green P. RepeatMasker Open-3.0. 1996.
[285] Benson G. Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res 1999;27:573–80.
[286] Flicek P, Amode MR, Barrell D, Beal K, Brent S, Carvalho-Silva D, et al. Ensembl 2012. Nucleic Acids Res 2012;40:D84–90. doi:10.1093/nar/gkr991.
[287] Pruitt KD, Tatusova T, Brown GR, Maglott DR. NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy. Nucleic Acids Res 2012;40:D130–5. doi:10.1093/nar/gkr1079.
[288] Wheelan SJ, Church DM, Ostell JM. Spidey: a tool for mRNA-to-genomic alignments. Genome Res 2001;11:1952–7. doi:10.1101/gr.195301.
[289] ENCODE Project Consortium, RM M, J S, M S, I D, RC H, et al. A user's guide to the encyclopedia of DNA elements (ENCODE). PLoS Biol 2011;9:e1001046. doi:10.1371/journal.pbio.1001046.
[290] Kharchenko PV, Tolstorukov MY, Park PJ. Design and analysis of ChIP-seq experiments for DNA-binding proteins. Nat Biotechnol 2008;26:1351–9. doi:10.1038/nbt.1508.
[291] Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, et al. Model-based analysis of ChIP-Seq (MACS). Genome Biol 2008;9:R137. doi:10.1186/gb-2008-9-9-r137.
[292] Landt SG, Marinov GK, Kundaje A, Kheradpour P, Pauli F, Batzoglou S, et al. ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. Genome Res 2012;22:1813–31. doi:10.1101/gr.136184.111.
[293] Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, et al. The Human Genome Browser at UCSC. Genome Res 2002;12:996–1006. doi:10.1101/gr.229102.
[294] Tuzun E, Sharp AJ, Bailey JA, Kaul R, Morrison VA, Pertz LM, et al. Fine-scale structural variation of the human genome. Nat Genet 2005;37:727–32. doi:10.1038/ng1562.
[295] Kidd JM, Cooper GM, Donahue WF, Hayden HS, Sampas N, Graves T, et al. Mapping and sequencing of structural variation from eight human genomes. Nature 2008;453:56–64. doi:10.1038/nature06862.
[296] Kidd JM, Graves T, Newman TL, Fulton R, Hayden HS, Malig M, et al. A human genome structural variation sequencing resource reveals insights into mutational mechanisms. Cell 2010;143:837–47. doi:10.1016/j.cell.2010.10.027.
[297] Sprung CN, Davey DSP, Withana NP, Distel LV, McKay MJ. Telomere length in lymphoblast cell lines derived from clinically radiosensitive cancer patients. Cancer Biol Ther 2008;7:638–44. doi:10.4161/cbt.7.5.5762.
[298] Riethman HC, Moyzis RK, Meyne J, Burke DT, Olson MV. Cloning human telomeric DNA fragments into Saccharomyces cerevisiae using a yeast-artificial-chromosome vector. Proc Natl Acad Sci 1989;86:6240–4. doi:10.1073/pnas.86.16.6240.
[299] De Silanes IL, Graña O, De Bonis ML, Dominguez O, Pisano DG, Blasco MA. Identification of TERRA locus unveils a telomere protection role through association to nearly all chromosomes. Nat Commun 2014;5. doi:10.1038/ncomms5723.
[300] Feuerbach F, Galy V, Trelles-Sticken E, Fromont-Racine M, Jacquier A, Gilson E, et al. Nuclear architecture and spatial positioning help establish transcriptional states of telomeres in yeast. Nat Cell Biol 2002;4:214–21. doi:10.1038/ncb756.
[301] Blasco MA, Gasser SM, Lingner J. Telomeres and telomerase. Genes Dev 1999;13:2353–9.
[302] McCulloch R, Rudenko G, Borst P. Gene conversions mediating antigenic variation in Trypanosoma brucei can occur in variant surface glycoprotein expression sites lacking 70-base-pair repeat sequences. Mol Cell Biol 1997;17:833–43.

[303] Carlson M, Celenza JL, Eng FJ. Evolution of the dispersed SUC gene family of Saccharomyces by rearrangements of chromosome telomeres. Mol Cell Biol 1985;5:2894–902. doi:10.1128/MCB.5.11.2894.
[304] Bryan TM, Englezou A, Gupta J, Bacchetti S, Reddel RR. Telomere elongation in immortal human cells without detectable telomerase activity. EMBO J 1995;14:4240–8.
[305] Henson JD, Neumann AA, Yeager TR, Reddel RR. Alternative lengthening of telomeres in mammalian cells. Oncogene 2002;21:598–610. doi:10.1038/sj.onc.1205058.
[306] Lundblad V, Wright WE. Telomeres and Telomerase: A Simple Picture Becomes Complex. Cell 1996;87:369–75. doi:10.1016/S0092-8674(00)81358-6.
[307] Marciniak RA, Cavazos D, Montellano R, Chen Q, Guarente L, Johnson FB. A novel telomere structure in a human alternative lengthening of telomeres cell line. Cancer Res 2005;65:2730–7. doi:10.1158/0008-5472.CAN-04-2888.
[308] Rizki A, Lundblad V. Defects in mismatch repair promote telomerase-independent proliferation. Nature 2001;411:713–6.
[309] Baur JA, Zou Y, Shay JW, Wright WE. Telomere Position Effect in Human Cells. Science 2001;292:2075–7. doi:10.1126/science.1062329.
[310] Lou Z, Wei J, Riethman H, Baur JA, Voglauer R, Shay JW, et al. Telomere length regulates ISG15 expression in human cells. Aging 2009;1:608–21.
[311] Fourel G, Revardel E, Koering CE, Gilson É. Cohabitation of insulators and silencing elements in yeast subtelomeric regions. EMBO J 1999;18:2522–37. doi:10.1093/emboj/18.9.2522.
[312] Pryde FE, Louis EJ. Limitations of silencing at native yeast telomeres. EMBO J 1999;18:2538–50. doi:10.1093/emboj/18.9.2538.
[313] Donaldson KM, Karpen GH. Trans-Suppression of Terminal Deficiency-Associated Position Effect Variegation in a Drosophila Minichromosome. Genetics 1997;145:325–37.
[314] Church DM, Goodstadt L, Hillier LW, Zody MC, Goldstein S, She X, et al. Lineage-specific biology revealed by a finished genome assembly of the mouse. PLoS Biol 2009;7:e1000112. doi:10.1371/journal.pbio.1000112.
[315] International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. Nature 2004;431:931–45. doi:10.1038/nature03001.
[316] Bailey SM, Cornforth MN, Kurimasa A, Chen DJ, Goodwin EH. Strand-Specific Postreplicative Processing of Mammalian Telomeres. Science 2001;293:2462–5. doi:10.1126/science.1062560.
[317] Stankiewicz P, Lupski JR. Genome architecture, rearrangements and genomic disorders. Trends Genet TIG 2002;18:74–82.
[318] Riethman H, Ambrosini A, Castaneda C, Finklestein J, Hu X-L, Mudunuri U, et al. Mapping and initial analysis of human subtelomeric sequence assemblies. Genome Res 2004;14:18–28. doi:10.1101/gr.1245004.
[319] Der-Sarkissian H, Vergnaud G, Borde Y-M, Thomas G, Londoño-Vallejo J-A. Segmental polymorphisms in the proterminal regions of a subset of human chromosomes. Genome Res 2002;12:1673–8. doi:10.1101/gr.322802.
[320] Mefford HC, Trask BJ. The complex structure and dynamic evolution of human subtelomeres. Nat Rev Genet 2002;3:91–102. doi:10.1038/nrg727.
[321] Linardopoulou E V, Williams EM, Fan Y, Friedman C, Young JM, Trask BJ. Human subtelomeres are hot spots of interchromosomal recombination and segmental duplication. Nature 2005;437:94–100. doi:10.1038/nature04029.

[322] Rudd MK, Friedman C, Parghi SS, Linardopoulou E V, Hsu L, Trask BJ. Elevated rates of sister chromatid exchange at chromosome ends. PLoS Genet 2007;3:e32. doi:10.1371/journal.pgen.0030032.
[323] Azzalin CM, Reichenbach P, Khoriauli L, Giulotto E, Lingner J. Telomeric repeat containing RNA and RNA surveillance factors at mammalian chromosome ends. Science 2007;318:798–801. doi:10.1126/science.1147182.
[324] Schoeftner S, Blasco MA. Developmentally regulated transcription of mammalian telomeres by DNA-dependent RNA polymerase II. Nat Cell Biol 2008;10:228–36. doi:10.1038/ncb1685.
[325] Porro A, Feuerhahn S, Reichenbach P, Lingner J. Molecular dissection of TERRA biogenesis unveils the presence of distinct and multiple regulatory pathways. Mol Cell Biol 2010;30:4808–17. doi:10.1128/MCB.00460-10.
[326] Deng Z, Norseen J, Wiedmer A, Riethman H, Lieberman PM. TERRA RNA binding to TRF2 facilitates heterochromatin formation and ORC recruitment at telomeres. Mol Cell 2009;35:403–13. doi:10.1016/j.molcel.2009.06.025.
[327] Nergadze SG, Farnung BO, Wischnewski H, Khoriauli L, Vitelli V, Chawla R, et al. CpG-island promoters drive transcription of human telomeres. RNA N Y N 2009;15:2186–94. doi:10.1261/rna.1748309.
[328] Graakjaer J, Bischoff C, Korsholm L, Holstebroe S, Vach W, Bohr VA, 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 2003;124:629–40. doi:10.1016/S0047-6374(03)00081-2.
[329] Graakjaer J, Der-Sarkissian H, Schmitz A, Bayer J, Thomas G, Kolvraa S, et al. Allele-specific relative telomere lengths are inherited. Hum Genet 2006;119:344–50. doi:10.1007/s00439-006-0137-x.
[330] 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 2006;15:725–33. doi:10.1093/hmg/ddi486.
[331] Yehezkel S, Segev Y, Viegas-Péquignot E, Skorecki K, Selig S. Hypomethylation of subtelomeric regions in ICF syndrome is associated with abnormally short telomeres and enhanced transcription from telomeric regions. Hum Mol Genet 2008;17:2776–89. doi:10.1093/hmg/ddn177.
[332] Caslini C, Connelly JA, Serna A, Broccoli D, Hess JL. MLL associates with telomeres and regulates telomeric repeat-containing RNA transcription. Mol Cell Biol 2009;29:4519–26. doi:10.1128/MCB.00195-09.
[333] Arnoult N, Van Beneden A, Decottignies A. Telomere length regulates TERRA levels through increased trimethylation of telomeric H3K9 and HP1α. Nat Struct Mol Biol 2012;19:948–56. doi:10.1038/nsmb.2364.
[334] Deng Z, Wang Z, Stong N, Plasschaert R, Moczan A, Chen H-S, et al. A role for CTCF and cohesin in subtelomere chromatin organization, TERRA transcription, and telomere end protection. EMBO J 2012;31:4165–78. doi:10.1038/emboj.2012.266.
[335] Porro A, Feuerhahn S, Delafontaine J, Riethman H, Rougemont J, Linger J. Functional characterization of the TERRA transcriptome at damaged telomere. Nat Commun 2014.
[336] Riethman H. Human telomere structure and biology. Annu Rev Genomics Hum Genet 2008;9:1–19. doi:10.1146/annurev.genom.8.021506.172017.
[337] Riethman H. Human subtelomeric copy number variations. Cytogenet Genome Res 2008;123:244–52. doi:10.1159/000184714.

[338] Kelley D, Rinn J. Transposable elements reveal a stem cell-specific class of long noncoding RNAs. Genome Biol 2012;13:R107. doi:10.1186/gb-2012-13-11-r107.
[339] Flicek P, Amode MR, Barrell D, Beal K, Brent S, Carvalho-Silva D, et al. Ensembl 2012. Nucleic Acids Res 2012;40:D84–90. doi:10.1093/nar/gkr991.
[340] Pruitt KD, Tatusova T, Brown GR, Maglott DR. NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy. Nucleic Acids Res 2012;40:D130–5. doi:10.1093/nar/gkr1079.
[341] Landt SG, Marinov GK, Kundaje A, Kheradpour P, Pauli F, Batzoglou S, et al. ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. Genome Res 2012;22:1813–31. doi:10.1101/gr.136184.111.
[342] Kalitsis P, Griffiths B, Choo KHA. Mouse telocentric sequences reveal a high rate of homogenization and possible role in Robertsonian translocation. Proc Natl Acad Sci 2006;103:8786–91. doi:10.1073/pnas.0600250103.
[343] Stong N, Deng Z, Gupta R, Hu S, Paul S, Weiner AK, et al. Subtelomeric CTCF and cohesin binding site organization using improved subtelomere assemblies and a novel annotation pipeline. Genome Res 2014;24:1039–50. doi:10.1101/gr.166983.113.
[344] Stamatoyannopoulos JA, Snyder M, Hardison R, Ren B, Gingeras T, Gilbert DM, et al. An encyclopedia of mouse DNA elements (Mouse ENCODE). Genome Biol 2012;13:418. doi:10.1186/gb-2012-13-8-418.
[345] Chung D, Kuan PF, Li B, Sanalkumar R, Liang K, Bresnick EH, et al. Discovering transcription factor binding sites in highly repetitive regions of genomes with multi-read analysis of ChIP-Seq data. PLoS Comput Biol 2011;7:e1002111. doi:10.1371/journal.pcbi.1002111.
[346] Wang J, Huda A, Lunyak V V, Jordan IK. A Gibbs sampling strategy applied to the mapping of ambiguous short-sequence tags. Bioinforma Oxf Engl 2010;26:2501–8. doi:10.1093/bioinformatics/btq460.
[347] Schmidt D, Schwalie PC, Wilson MD, Ballester B, Gonçalves A, Kutter C, et al. Waves of retrotransposon expansion remodel genome organization and CTCF binding in multiple mammalian lineages. Cell 2012;148:335–48. doi:10.1016/j.cell.2011.11.058.
[348] Pardue M-L, DeBaryshe PG. Retrotransposons provide an evolutionarily robust non-telomerase mechanism to maintain telomeres. Annu Rev Genet 2003;37:485–511.
[349] Witmer K, Schmid CD, Brancucci NMB, Luah Y-H, Preiser PR, Bozdech Z, et al. Analysis of subtelomeric virulence gene families in Plasmodium falciparum by comparative transcriptional profiling. Mol Microbiol 2012;84:243–59. doi:10.1111/j.1365-2958.2012.08019.x.
[350] Hanahan D, Weinberg RA. Hallmarks of Cancer: The Next Generation. Cell 2011;144:646–74. doi:10.1016/j.cell.2011.02.013.
[351] Feldser DM, Hackett JA, Greider CW. Telomere dysfunction and the initiation of genome instability. Nat Rev Cancer 2003;3:623–7. doi:10.1038/nrc1142.
[352] Artandi SE, DePinho RA. Telomeres and telomerase in cancer. Carcinogenesis 2010;31:9–18. doi:10.1093/carcin/bgp268.
[353] Rubin MA, Maher CA, Chinnaiyan AM. Common Gene Rearrangements in Prostate Cancer. J Clin Oncol 2011;29:3659–68. doi:10.1200/JCO.2011.35.1916.
[354] Tomlins SA, Rhodes DR, Perner S, Dhanasekaran SM, Mehra R, Sun X-W, et al. Recurrent Fusion of TMPRSS2 and ETS Transcription Factor Genes in Prostate Cancer. Science 2005;310:644–8. doi:10.1126/science.1117679.

[355] Baca SC, Prandi D, Lawrence MS, Mosquera JM, Romanel A, Drier Y, et al. Punctuated Evolution of Prostate Cancer Genomes. Cell 2013;153:666–77. doi:10.1016/j.cell.2013.03.021.
[356] Tange O. GNU Parallel - The Command-Line Power Tool. Login USENIX Mag 2011;36:42–7.
[357] 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 1995;14:5433–43.
[358] Berger MF, Lawrence MS, Demichelis F, Drier Y, Cibulskis K, Sivachenko AY, et al. The genomic complexity of primary human prostate cancer. Nature 2011;470:214–20. doi:10.1038/nature09744.
[359] Kumar A, White TA, MacKenzie AP, Clegg N, Lee C, Dumpit RF, et al. Exome sequencing identifies a spectrum of mutation frequencies in advanced and lethal prostate cancers. Proc Natl Acad Sci 2011;108:17087–92. doi:10.1073/pnas.1108745108.
[360] R Development Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing; 2008.
[361] Canty A, Ripley BD. boot: Bootstrap R (S-Plus) Functions. 2014.
[362] Davison AC, Hinkley DV. Bootstrap Methods and Their Applications. Cambridge: Cambridge University Press; 1997.