Supplemental Methods

DNA samples

DNA samples from Human/Rodent somatic hybrid cell lines (each containing a single human chromosome) and their parental rodent cells were obtained from a human Chromosomal DNA mapping panel (NIGMS Human/Rodent Somatic Cell Hybrid Mini Mapping Panel # 2 DNA, Coriell Cell Repositories).

DNA samples from people of different genders and ethnic origins were obtained from the HRC2 Human Random Control DNA panel 2 (Sigma Aldrich, 96 Caucasian people) and from Human variation panels HD03 (Indo Pakistani), HD05 (Middle Eastern), HD07 (Japanese), HD12 (Africans South of the Sahara), HD20 (Russian Kransnodar), HD21 (Italian), HD22 (Ashkenazi Jewish), HD32 (Chinese), and samples from Mbpygmy (NA10492, NA10493, NA10494, NA10495, NA10496) from Coriell Cell Repositories.

DNA samples from individuals with trisomy 13 (Catalog numbers: AG12070, NA00526, NA02948, NA03330, and NA00503), trisomy 21 (AGPDOWN Aging DNA panel-down syndrome), and an individual with an XY genotype who is phenotypically female (Catalog number: NA02598) were obtained from the Coriell Cell Repositories. Human tissue from trisomy 21 individuals was also obtained from the NICHD Brain and Tissue Bank for Developmental Disorders at the University of Maryland, Baltimore, MD (Catalog numbers: UMB4904, UMB1258, UMB0707, and UMB5277). To assure that the trisomy 21 DNA samples used in our study have a trisomy 21 karyotype, we performed a PCR genetic test for the Chromosome 21 short tandem repeat D21S167 and the S100B gene (Yang et al. 2005) and verified that the trisomy 21 DNA samples had extra copies on Chromosome 21 in contrast to samples with a normal karyotype. The karyotypically normal, immortalized hTERT cell line CHON-002 was purchased from ATCC.

RNA samples

RNA was isolated from the prostate epithelial cell lines RWPE-1, PNT2, 957E-hTERT and the prostate cancer cell lines PC3, VCaP, LNCaP, and DU145 using TriZol reagent (ThermoFischer Scientific) as recommended by the manufacturer. The authenticity of these cells lines was verified by genotyping. Isolated RNA was treated with the Turbo DNA-free kit (Ambion) to remove contaminant traces of DNA, which was subsequently verified by PCR.

Search for centromere specific markers

We performed a systematic and comprehensive analysis of the literature and sequence databases to annotate centromere sequences, and have identified unique centromere markers for 23 of the 24 human chromosomes (the exception is Chromosome 19). Several of these sequences, identified in previous studies by Southern blotting analysis, have been recently annotated and expanded to the most recent assembly of the human genome project (hg38) by Karen Miga (Miga et al. 2014; Miga 2015). Our laboratory identified markers for pericentromeric sequences previously. The accession numbers of the sequences analyzed in this study are indicated in the Supplemental Table 2. Due to the sequence similarity between α-repeats, primers were designed and validated to specifically detect each array. We previously designed primers and probes to specifically detect proviruses K111 and K222 at pericentromeric loci (Contreras-Galindo et al.

2011, 2013; Zahn et al. 2015). All the sequences reported in this manuscript were previously deposited in the NCBI database, nucleotide section (http://www.ncbi.nlm.nih.gov/nuccore). Sequences for alpha repeats are reported in the supplemental material. The Accession numbers are: BX248407.26, BX248407.26, AJ295044.1, J04773.1, Z12006.1, Z12011.1, M26920.1, GJ211907.1, AC142529.3, M16037.1, M64779.1, M64320.1, X63622.1, M21452.1, M28221.1, D29750.1, D29750.1, GJ211955.2, GJ211961.2, GJ211965.2, GJ211967.2, GJ211968.2, GJ211969.2, M22273.1, GJ211972.2, GJ211986.2, AF237720.1, M58446.1, M13882.1, GJ212053.1, M65181.1, M65182.1, AC004164, AC006504, X58269.1, D29750.1, D29750.1, GJ212162.2, GJ212162.2, GJ212163.2, X02418.1, and GJ212193.1

Real-Time qRT-PCR

Real Time RT-PCR was carried out using the same qPCR conditions described in the main manuscript to quantitate human centromere arrays with the addition of an RT step. The PCR reaction contained 0.2 microliter of the Murine Leukemia Virus Reverse Transcriptase (MLV-RT) and the PCR reaction was preceded by an RT step of 30 min at 50 °C. DNA contamination was ruled out by conducting PCR on samples that were not treated with MLV-RT. We also used water as a negative control.

Chromatin immunoprecipitation (ChIP)

ChIP assays were performed using the iDeal ChIP-seq kit for Histones (Diagenode) following the procedures described by the manufacturer to assess the association of centromeric proteins

CENPA and CENPB with centromere repeats. Briefly, approximately 70% to 80% confluent LnCaP cells grown in 75 cm^2 flasks were fixed with 0.1% formaldehyde for 8 minutes to cross-link protein to DNA. The fixation step was stopped by adding Glycine solution. Cells were lysed with the Lysis buffers iL1 and iL2 included in the kit and the chromatin was sheared to 100 to 600 bp with a sonicator in the presence of a protease inhibitor cocktail. Chromatin was immunoprecipitated overnight using a specific monoclonal antibody to CENPA (Abcam, ab13939), a polyclonal antibody to CENPB (Abcam, ab134144, clone EPR6930), or with non-specific IgG antibodies. ChIP was performed on 100 μL of shredded chromatin using 20 μL of DiaMag Protein A-coated magnetic beads and 2 μg of target antibody or control rabbit IgG antibody per reaction in the presence of a cocktail of protease inhibitors overnight at 4 °C. After several washes, the chromatin was eluted in buffer iE1 and iE2 as described by the manufacturer. DNA was separated from the protein fraction by isopropanol precipitation. Centromere protein occupancy on target arrays, meaning the number of α-repeats in each centromere array sequence bound to the centromeric proteins CENPA or CENPB, was measured by qPCR with our qPCR assays for centromeric repeats, using the primers and optimized conditions shown in supplementary Tables 1 and 2. The relative amount of immunoprecipitated DNA compared to input DNA was calculated using the following equation: % recovery = $2^\wedge(Ct_{input} - Ct_{sample})$, which assumes that the efficiency of the PCR reaction is 100%. The relative percentage of each array occupied by CENPA and CENPB was calculated by dividing the number of α-repeats in each centromere array precipitated with centromeric proteins to the total number of α-repeats in the array determined in the input DNA. The fold enrichment was also determined based on the cycle differences (ΔCt) between the sample Vs. control (IgG). The genes *TOP3A*, *DEK*, and beta-actin (*ACTNB*) served as negative controls in the centromere ChIP studies, as these genes localize to chromosomal arms. Water was used as a negative control for the qPCR reactions.

Amplification and Sequencing of K111 LTR insertions

K111 insertions were amplified by PCR using the Expand Long Range dNTPack PCR kit (Roche Applied Science, Indianapolis, IN) as previously described (Contreras-Galindo et al. 2011, 2013). The amplification products were cloned into the topo TA vector (Invitrogen, Carlsbad, CA) and sequenced. Sequences of K111-related insertions amplified from DNA of human/rodent cell hybrids containing a single human chromosome shown in the supplementary Figure S11 are deposited in the NCBI database with Accession Numbers (JQ790790 - JQ790967). The primers P1/P4 amplify K111 insertions in several human chromosomes. The sequences amplified from trisomy 21 patients are deposited in the NCBI database with the Accession Numbers MF624880 - MF625017

Immunofluorescence-Fluorescent In Situ Hybridization (IF-FISH)

The extent of CENPB binding to the centromere of Chr 21 in karyotypically normal CHON-002 cells and cells from trisomy 21 Subject A 1258 and Subject B 5277 was determined by IF-FISH. Cells were synchronized in metaphase for 16 h in 10 µg/mL Colchicine solution (KaryoMAX® Colcemid™ Solution, Thermo Fischer Scientific). Cells were tripsynized, suspended in 0.56 % KCl hypotonic solution for 15 min at 37 °C, and then suspended in hypotonic solution with 0.05% Tween. To make chromosome spreads, cells were spotted onto slides by centrifugation at

1600 rpm for 5 min. IF was done by fixing the chromosome spreads with 100% MeOH at -20 °C for 10 min. Chromosome spreads were washed with PBS, permeabilized 2 times in PBST (PBS + 0.05% Triton X) for 5 min, and incubated with blocking solution (PBST + 5% Bovine Serum Albumin (BSA)) for 30 min. Chromosome spreads were incubated with mouse anti-CENPB primary antibody for 1 h (Santa Cruz Biotechnology, CENPB Antibody (C-10): sc-376392), washed three times in PBST for 5 min, and then incubated with a red-fluorescent secondary antibody for 45 min (Alexa Flour 594 rabbit anti-mouse IgG 1:1000, Thermo Fischer Scientific, Cat No A27027). Spreads were washed three times with PBST, and then two times with PBS. Chr 1 was identified by its large size. For the FISH analysis to identify Chr 21, chromosome spreads were denatured with 100% MeOH for 10 min at -20 °C, incubated with XCP 21 green chromosome paint (Meta Systems probes, D-0321-100-FI) at 75 °C for 2 min and then incubated at 37 °C overnight. Chromosome spreads were washed in 0.4X SSC (Saline Sodium Citrate) at 72 °C for 2 min and 2X SSC for 30 sec, and then rinsed with PBS. They were then counterstained with DAPI solution (Thermo Fischer Scientific, ProLong® Gold Antifade Mountant with DAPI, Invitrogen) and visualized with a NIKON Eclipse Ti-S Inverted fluorescent microscope. The specificity of CENPB staining was assessed by its punctated staining on the centromere of each chromatid (a pair of dots in each chromosome at metaphase) and the lack of this staining when incubating only with the secondary antibody. The diameter of CENPB binding along the centromeres of Chr 1 and Chr 21 was measured using the NIS-Elements Software.

Real-Time qPCR for centromeres 13 and 21 using LNA primers and clamps

The homolog arrays D13Z1/D21Z1 present in the centromere of chromosomes 13 and 21 are almost ~100% identical except for the presence of two nucleotide substitutions, T/C and A/G (Pellestor et al. 1994; Nilsson et al. 1997). A qPCR was developed to detect these substitutions using primers that have a locked nucleic acid (LNA) modification targeting these nucleotide variations (Ballantyne et al. 2008). The primers also contained LNA modification at the bases just before and after the substitution. The description of the primers is provided in the Supplemental Table 1. LNA primers offer markedly increased affinity for the complementary strand, compared to traditional DNA primers. We used these modified primers in a PCR reaction to specifically detect either D13Z1 or D21Z1. We used DNA isolated from human/rodent somatic cell hybrids that contain either Chromosome 13 or 21 to optimize the assay. We used mouse and/or hamster DNA and water as negative controls. The PCR assay was carried out as described above and consisted of an enzyme activation step at 95 °C for 10 minutes and 20 cycles of 15 s of denaturation at 95 °C and 30 sec of annealing/extension at the temperature reported in Supplemental Table 2. A temperature gradient showed that at increasing annealing temperatures, the LNA primers were able to uniquely amplify the centromere repeat containing the nucleotide variation. For centromere 13 D13Z1, a forward LNA primer and a regular reverse oligonucleotide (that binds to either D13Z1 or D21Z1; see Supplementary Figure S4) exclusively detected the centromere of Chromosome 13 but not 21 at an annealing/extension temperature of 68 °C (See Supplementary Figure S5a). For the detection and quantitation of the centromere 21 D21Z1, we used forward and reverse LNA primers that recognize the D21Z1 nucleotide variations together with an LNA primer clamp that recognizes the substitution of D13Z1. This clamp consisted of the same forward LNA primer that detects D13Z1 but this time phosphorylated at the 3’ end to inhibit the amplification of D13Z1 sequences. The LNA PCR reaction was shown to specifically amplify D21Z1 and exhibited substantially reduced amplification of D13Z1 when an annealing/extension temperature of 64 °C was used (See Supplementary Figure S5b). The specificity of the LNA primers was evaluated in human/rodent hybrid cell lines that contain a single human chromosome as shown in Figure 1. An LNA primer was also designed to detect specific mutations in the pericentromeric K111 *gag* and not other HERV-Ks (Figure 4d, Supplementary Table 1). We used a clone that contains a K111 fragment that extends from the LTR region to the *pol* gene as a positive control. As a negative control, we used DNA from the Hut 78 cell line, which lacks K111 but has all other HERV-Ks (Zahn et al. 2015). Water was used as a further negative control. The specific detection of K111 was verified by sequencing analysis of the PCR products.

Next generation sequencing (NGS) data analysis

The number of copies of α-repeats in each centromeric array, the numbers of pericentromeric K111 and K222, and the number of single copy genes *TOP3A*, *CCR5*, *DEK*, *RENIN*, *GAPDH*, and *ACTNB*, were analyzed in sequence libraries from diverse populations generated in the 1000 genomes project. We screened the 101 bp read libraries for sequences that match each query sequence, allowing no more than 10 bp mismatches and/or indels. For pericentromere K111 proviruses, we screened for sequences that match to the 5’ and 3’ integration sites and adjacent flanking sequence. Using this parameter enabled specific detection of K111 as opposed to other HERV-K proviruses that exhibit alternative integration sites. Detection criteria were constrained by reads containing at least 20 bp of the flanking sequence (CER: centromere repeat elements), 20 bp of the K111 LTR, and the GAATTC target site duplication as previously analyzed in our laboratory (Contreras-Galindo et al. 2011, 2013). For the K222 proviruses, we screened for reads that have 20 bp of the CER flanking sequence and 20 of the adjacent *prt* unique to K222 proviruses (Zahn et al. 2015). We additionally screened these libraries to detect single copy genes using the same criteria used to identify centromeric arrays. We normalized these values to the total number of sequence reads in each group by dividing the number of sequence reads that hit the query sequences to the total number of reads and multiplied these values by one million bp of sequence. The calculated data were compared to the number of reads that match to the genes *TOP3*A, *DEK*, *CCR5*, *RENIN*, *GAPDH*, and *ACTNB* using the same exercise in order to contrast the abundance of centromere arrays to that of single copy or small-number-of-copy genes.

In silico sequence analysis

The K111-related LTR sequences obtained in the DNA of human cells, trisomy 21 samples, and DNA from human/rodent chromosomal cell hybrids were BLASTed to the NCBI database. The sequences were aligned in BioEdit and exported to the MEGA 7 matrix. The K111 tree was generated using Bayesian inference- MrBayes v 3.2 (Huelsenbeck et al. 2001; Ronquist and Huelsenbeck 2003) with four independent chains run for at least 10,000,000 generations until sufficient trees were sampled to generate more than 99% credibility.

Statistical analysis

Statistical analysis was performed using GraphPad Prism 7. Statistically significant differences between the numbers of α-repeats in each centromere array and K111 and K222 pericentromeric proviruses between samples were calculated using the student's t-test. The relative enrichment of centromeric array DNA associated with the centromere proteins CENPA and CENPB in ChIP experiments (using IgG as a control antibody); the size of arrays in chromosome 13 and 21 (including pericentromere proviruses K111 and K222) in patients with trisomy 13 or 21 and healthy individuals; the size of the arrays in Chr X and Chr Y in male and female populations; and the sizes of CENPB spots in trisomy and healthy cells were compared using an unpaired t-test. Statistical differences between the size of centromere arrays in chr 8 and 18 between patients with trisomy 8, trisomy 18, and healthy individuals was calculated using the ANOVA test and significant differences between the groups estimated using the Dunnett's multiple comparisons test. Correlations of values obtained by qPCR, qRT-PCR, *in silico* analysis of the 1000 Genomes Project or the PERCON analysis, and by the estimated average number reported in the literature were measured using the Pearson correlation test. Two-tailed p values were considered significant at $p < 0.05$. The statistical analysis is reported in Table S4.