Supplemental Text

Comparison of PCR-based techniques to identify specific centromeres to results of PERCON analysis

We found a positive correlation ($r = 0.7345$, $p < 0.0001$, Fig S8a) with the PCR data and estimates of centromere array size generated by the PERCON analysis, the most recent analysis of the human assembly (Shepelev 2015, Miga 2014). As expected, the correlation of the PERCON bioinformatics analysis and our bioinformatics analysis on the 1000 genomes data set showed a very strong correlation ($r= 0.926$, $p < 0.00001$, Fig S8b). We compared eighteen of the arrays studied by the PERCON investigators, and for which we have PCR data, to determine what percentage of each estimated array our primers can detect. Half of these arrays give concordant estimates of size by PCR and bioinformatics analysis (Table 1). However, the size of arrays D7Z1, D13Z1/D21Z1, and D20Z2 was estimated to be 3-to 8-fold higher, and the arrays D4Z1, D12Z3, D15Z3 and D17Z2 were estimated to be 10-50 fold lower, by PCR. These discrepancies might be the result of differences in analytical parameters with these two techniques. For example, optimizing PCR conditions, such as elevating the annealing temperature to increase specificity for a given centromere, leads to detection of lower levels of the array.

Distinguishing Centromeres 13 and 21

Due to the importance of Chromosome 21 in Trisomy 21, being able to distinguish its centromere would be very valuable for diagnostics and for studying the pathogenesis of this syndrome. However, the centromere α-repeat arrays of Chromosome 13 D13Z1 and

Chromosome 21 D21Z1 are nearly ~100% similar: they differ by only two single nucleotides (Pellestor et al. 1994; Nilsson et al. 1997). We developed PCR assays using modified primers that recognize the nucleotide variations in D13Z1 or D21Z1 (Fig. S4). To detect these mutations, we designed primers containing a locked nucleic acid (LNA) modification at the base targeting the nucleotide change and optimized the PCR conditions in order to detect either D13Z1 or D21Z1 nucleotide substitution (Ballantyne et al. 2008) (Supplemental Tables 1 and 2, Fig. S4). The PCR reaction to detect D13Z1 has an LNA primer that specifically binds to a D13Z1 base substitution and another primer that binds to both arrays (Fig. S5a). The PCR assay for D21Z1 contains primers with LNA modifications that detect a nucleotide substitution of D21Z1, but also include a primer clamp that recognizes a base substitution of D13Z1 and prevents its amplification (Supplemental Table 1 and 2). This combination of primers and a clamp substantially reduced the amplification of D13Z1 while allowing the amplification of D21Z1 (Fig. S5b). PCR reactions designed to detect either D13Z1 or D21Z1 were assessed in the context of all human chromosomes, verifying their specificity (Fig. S4b).

We measured the levels of D13Z1 and D21Z1 in healthy individuals of different ethnicity, including 28 Caucasians, 2 African Americans, 5 people from sub-saharan Africa, 5 Chinese, 5 Jewish Ashkenazi individuals, 5 Italians, 5 Pakistanis, 5 Middle Eastern people, 5 Japanese, and 5 Siberian Russian individuals. While we identify variation between individuals, no significant differences were found between populations.

Centromere PCR assays to detect normal chromosomal ploidy and nondisjunction defects

Our PCR assays are able to infer the number of chromosomes by determining the abundance of α-repeats in the centromere of each chromosome. Thus, we can screen DNA samples to determine the number of sex and somatic chromosomes with the purpose of distinguishing genetic gender as well as identifying chromosome aneuploidy in nondisjunction congenital defects. We quantitated the number of α-repeats in the arrays of chromosomes X and Y in order to identify gender. We were able to detect the DYZ3 array from chromosome Y in the male population but not in the female population (Fig. S9). Centromere X DXZ1, while varying among individuals, was consistently found at about 2 times the levels in females as in males. DXZ1 was found at different levels in one individual with trisomy X as well as in one phenotypically female with an XY karyotype, whose content of α-repeats was either three times higher or similar to a male, respectively. In the latter sample, we also detected DYZ3. Therefore, our PCR assays can determine the sex chromosome ploidy of an individual and can be used to detail nondisjunction genetic disorders of sex chromosomes, such as triple or quadruple X syndromes, Klinefelter XXY syndrome, Turner syndrome (monosomy X), super male XYY syndrome, and other disorders associated with sex chromosome ploidy.

We also quantitated the number of α-repeats in arrays of chromosomes 8 and 18 in karyotypically normal individuals and individuals with trisomy 8 (Warkany syndrome) or 18 (Edwards syndrome). We were able to detect a ~1.5 fold increase in the α-repeat content of D8Z2 in individuals with trisomy 8 compared to the karyotypically normal population (Fig. S9). A similar increase was detected in D18Z1 and D18Z2 of chromosome 18 in individuals with trisomy 18. In contrast to the centromere reduction of Chr 21 in trisomy 21 individuals, we did not detect instability in the centromere of Chr 8 or 18 in trisomy 8 or 18 individuals,

respectively, suggesting that centromere 21 is more unstable than other centromeres as previously reported (Lo et al. 1999). Our PCR assay can thus reliably distinguish between those with trisomy 8 or 18 and those without the defect. We did not find evidence of centromere instability in trisomy 8, 13, or 18 (Fig. 4, Fig. S9). These findings are consistent with the possibility that centromeric instability is not a feature of all trisomies, but only of trisomy 21. However, sequencing of the pericentric areas of the chromosomes involved in other trisomies will be necessary to develop specific markers and understand whether pericentromere instability exists in trisomies other than 21.

Quantifying Centromeric Transcript Abundance

Human centromere sequences have been thought to be tightly packaged into heterochromatin structures and transcriptionally inert. Recent evidence suggests that these areas of the genome are not only transcriptionally active, but that their transcripts regulate chromosome stability and cell division (Wong et al. 2007). Studies have shown the expression of transcripts from centromere arrays in chromosomes 4, 9, and 13/21, yet a more complete map of centromere α-repeat array transcripts has not yet been generated. We have used our centromere PCR assays to measure the levels of transcription of individual human α-repeat arrays and observed that Centromere α-repeat Satellite RNAs (CeASaRs) originating from all centromere arrays are found in benign epithelial prostate and prostate cancer cell lines (Fig. S10a). These CeASaRs are genuine RNA sequences, as we detect these transcripts in samples carefully treated with DNAse that do not produce amplification products without a reverse transcription step in the PCR reactions. Therefore, our PCR assays can be used to study transcription levels of essentially all of the

1 distinct human α-repeat arrays arising from centromeres. Furthermore, expression analysis of
2 these arrays revealed that the levels of RNA correlate to the size of the arrays (Fig. S10b).

3