Sequence assembly and validation of chromosome 2p11

Our initial characterization within human 2p11 began with a pericentromeric reference BAC 101B6 (AC002038) sequence which mapped in close proximity to the centromere of human chromosome 2 (Horvath et al. 2000). Using this anchor clone as a reference, we generated a physical map in both directions in an effort to bridge into higher-order alpha satellite sequence. Overlaps were confirmed using a previously established strategy of paralogy sequence variant (PSV) tagging of BACs (Horvath et al. 2000). A minimal tiling path was eventually generated that extended over 700kb (Figure 1a, Supplemental Table 1). Due to the highly duplicated nature of this region, an additional 892 BACs were excluded during this exercise where the STS showed 94-99.8% sequence identity. In addition, structural variation within the pericentromeric region required single haplotype contiguity within RPCI-11. All BACs selected as part of this process were sequenced in collaboration with Washington University in St. Louis. The results from this study have now been incorporated into the chromosome 2 build 34 as NT_034508 (Hillier 2005).

Construction of the 2p11 physical map was challenging for two reasons. First, centromeric and pericentromeric regions are typically devoid of useful STS (sequence tagged site) markers. STS markers have been widely used in the past to aid in approximate clone placement in the genome, but become ambiguous in regions containing duplications. Second, multiple copies of some duplicons exist on chromosome 2 and elsewhere in the human genome, and therefore made correct

construction of this contig particularly challenging. When obtaining putative overlapping BACs, it was necessary to use high standards (end sequence placement >99.95% sequence identity over 400bp, and matching chromosome 2 monochromosomal hybrid DNA sequence over one kb or more) to confirm overlaps. As expected, extension of the 2p11 sequence proximally into the centromere proved difficult, yet possible. This was accessible only after screening end-sequences of over 445 BACs containing alpha satellite to recover BAC clones that extended in the correct direction.

Due to the highly duplicated nature of this particular portion of chromosome 2p11, the existence of multiple highly identical (>99%) segmental duplications and the potential for structural rearrangement near the centromere its assembly has changed in recent genome builds. We therefore chose a series of independent methods to validate the BAC assembly of this region. First, paralogous sequence variants were designed within the sequence overlaps of BACs (see numbered boxes in Figure 1). These paralogous sequence variants (or PSVs) were primer pairs designed within "unique" regions of a BAC that could be PCR amplified and sequenced (Horvath et al. 2000). Based on the sequence analysis of these 31 STS and comparison with a monochromosomal chromosome 2 source, we confirmed map location to chromosome 2 as well as overlaps among these clones. Second, we confirmed clone integrity and sequence assembly by comparison of the assembled sequence with clone insert size (as determined by pulsed-field gel electrophoresis). Third, we compared the assembly of the region to that of the genome by performing genomic Southern hybridizations between BAC clones and chromosome 2 monochromosomal hybrid DNAs (Figure 1B). Finally, fiber FISH

experiments validated sequence overlaps within monochromosomal somatic cell hybrid cell lines (Figure 1C).

Alpha satellite analysis

Two-color FISH analyses of many of the above 2p11 BACs in conjunction with higher-order alpha satellite probes showed close proximity to the centromere of chromosome 2. To confirm the distance between pericentromeric duplicons and the centromere, we conducted sequence analyses of the targeted region that indicated we had transitioned into ~170kb of alpha-satellite DNA. Two previous studies of chromosome 2 higher-order alpha satellite indicated it belonged to suprachromosomal family 2 satellite because of its dimeric monomer organization (Haaf and Willard 1992; Rocchi et al. 1990). Although neither study was able to identify a characteristic higher-order periodicity upon restriction enzyme digestion, XbaI digestion of chromosome 2 alpha satellite DNA resulted in a prominent 680bp fragment. A ladder of fragments was observed in 342bp intervals (the size of a dimer), signifying loss of restriction sites in adjacent regions (Haaf and Willard 1992). Sequence analysis of three chromosome 2 alpha satellite clones suggested that a structure of at least two tetrameric repeats made up one higher-order repeat unit (Haaf and Willard 1992). To confirm that both alpha satellite clones in this contig contained chromosome 2 higher-order satellite sequences, we subjected them to XbaI digestion followed by a Southern blot using known higher-order 2 satellite DNA as a probe (Supplemental Figure 1A). Both the agarose gel and Southern blot revealed a 680bp fragment as well as a ladder of fragments every 342bp in both 2p11 clones, which

was absent from a chromosome 16 alpha satellite containing BAC (lanes 3 and 4 versus lane 1, Supplemental Figure 1A).

To assess the relationship between all known alpha satellite monomers and the satellite monomers in this 2p11 contig, neighbor-joining phylograms were constructed using a database of higher-order alpha satellite repeat structures. A representative subset of monomers from the finished clones AC025223 and AC144896 were compared with all previously characterized higher order satellite monomers including known higher-order alpha satellite monomers from chromosome 2 (J04773 and M81229) (Haaf and Willard 1992; Rocchi et al. 1990) and New World and Old World monkey sequences (obtained by BAC end sequence analysis (She and Eichler, unpublished)). Multiple clades of monomers are evident in this tree (Supplemental Figure 1B). Old World monkey sequences form a bifurcating clade with high bootstrap support which separate it from all other human sequences. Known chromosome 2 higher-order sequences (D2Z1) are intermixed with monomers extracted from the 2p11 sequence. Multiple 2p11 monomers are highly identical to known higher-order chromosome 2 satellite monomers ranging between 98.7 and 99.4% identity, consistent with higher order alpha-satellite sequences for this chromosome.

SUPPLEMENTAL METHODS

Hybridization

Human BAC RPCI-11 (segments 1, 2, 4, 5), human cosmid libraries LLNL-01AH, LLNL-02AE, LLNL-07NC01Y, LANL-11CO1, and LLNL-22N, orangutan CHORI-253 (segments 1 and 2), and baboon RPCI-41 BAC libraries (segments 1 and 2), were hybridized with PCR-generated probes amplified from human genomic DNA as previously described (Horvath et al. 2003) (Table 4, see Supplemental Table 2 for the probes used). All previously identified PUC false positives were eliminated from the BAC positive lists before PCR analysis. Since no false positive lists exist for the cosmid libraries, all positives were used in PCR assays and only those that PCR amplified were used in further analyses. Probes made for radioactive hybridization were purified from pooled PCR products using Qiagen's QIAquick® PCR purification kit (250) according to the manufacturer's recommendations (Qiagen, Valencia, CA). Approximately fifty nanograms of purified product was random-hexamer labeled with [α-32P] dCTP using Amersham's Ready-To-Go® DNA labeling beads according to the manufacturer's recommendations (Amersham, Piscataway, NJ). One mg of sonicated salmon sperm DNA was used in all hybridization experiments to block the membranes (Stratagene, La Jolla, CA). Genomic Southern blots were performed using 1µg BAC DNA, 5µg total genomic DNA or 10µg monochromosomal 2 hybrid DNA. Human genomic lymphoblastoid lines (GM-15036 and GM-15038) were obtained from the human Polymorphism Discovery Resource (PDR) panel and monochromosomal 2 hybrid fibroblast lines (GM-11686 and GM-11712) were obtained from Coriell Cell Repositories (Camden, NJ). All cell line DNA was isolated using the PUREGENE® DNA isolation

kit (Gentra systems, Minneapolis, MN) according to the manufacturer's recommendations. All genomic, monochromosomal and BAC DNA was transferred to Zeta-Probe® membranes (BIO-RAD, Hercules, CA). Genomic blots were hybridized at least 16 hours in 20mL hybridization solution (1% glycine, 10% dextran sulfate, 3X SSC, 0.2% PVP-Ficoll, 50mM $NaPO_4$ pH 6.8, 50% deionized formamide, 0.2% BSA) at 55ºC. Following hybridization, each blot was rinsed twice in 2X SSC and then washed twice for 5 minutes each in 50mM Tris pH 8.6, 2mM EDTA pH 8.0, 1% SDS, 1M NaCl heated to 65ºC. The next two washes were for 10 minutes each at 65ºC (50mM Tris pH 8.6, 0.5% SDS, 0.5M NaCl) followed by two washes for 10 minutes each at 65 ºC (0.1% SDS, 0.1X SSC). The final two washes were for 10 minutes each at 68ºC in 0.1% SDS, 0.1X SSC.

Fluorescent *in situ* Hybridization

Human and primate metaphase chromosomes (Figure 2) from *H. sapiens*, *P. troglodytes*, *G. gorilla*, *P. pygmaeus*, and *H. lar* were prepared from lymphoblastoid lines as previously described (Horvath et al. 2000) and hybridized using human cosmid or BAC DNA. Extended fiber FISH was conducted using the DIRVISH protocol (Parra and Windle 1993).

REFERENCES

Haaf, T. and H. Willard. 1992. Organization, polymorphism and molecular cytogenetics of chromosome-specific alpha-satellite DNA from the centromere of chromosome 2. *Genomics* **13:** 122-128.

Hillier, L.W.e.a. 2005. Generation and annotation of the DNA sequences of human chromosomes 2 and 4. *Nature*.

Horvath, J., S. Schwartz, and E. Eichler. 2000. The mosaic structure of a 2p11 pericentromeric segment: A strategy for characterizing complex regions of the human genome. *Genome Res* **10:** 839-852.

Horvath, J.E., C.L. Gulden, J.A. Bailey, C. Yohn, J.D. McPherson, A. Prescott, B.A. Roe, P.J. De Jong, M. Ventura, D. Misceo et al. 2003. Using a pericentromeric interspersed repeat to recapitulate the phylogeny and expansion of human centromeric segmental duplications. *Mol Biol Evol* **20:** 1463-1479.

Parra, I. and B. Windle. 1993. High resolution visual mapping of stretched DNA by fluorescent hybridization. *Nat Genet* **5:** 17-21.

Rocchi, M., A. Baldini, N. Archidiacono, S. Lainwala, O.J. Miller, and D.A. Miller. 1990. Chromosome-specific subsets of human alphoid DNA identified by a chromosome 2-derived clone. *Genomics* **8:** 705-709.