Genomic and Genetic Definition of a Functional Human Centromere

Mary G. Schueler,[1] Anne W. Higgins,[1]* M. Katharine Rudd,[1] Karen Gustashaw,[1] Huntington F. Willard[1,2]†

The definition of centromeres of human chromosomes requires a complete genomic understanding of these regions. Toward this end, we report integration of physical mapping, genetic, and functional approaches, together with sequencing of selected regions, to define the centromere of the human X chromosome and to explore the evolution of sequences responsible for chromosome segregation. The transitional region between expressed sequences on the short arm of the X and the chromosome-specific alpha satellite array DXZ1 spans about 450 kilobases and is satellite-rich. At the junction between this satellite region and canonical DXZ1 repeats, diverged repeat units provide direct evidence of unequal crossover as the homogenizing force of these arrays. Results from deletion analysis of mitotically stable chromosome rearrangements and from a human artificial chromosome assay demonstrate that DXZ1 DNA is sufficient for centromere function. Evolutionary studies indicate that, while alpha satellite DNA present throughout the pericentromeric region of the X chromosome appears to be a descendant of an ancestral primate centromere, the current functional centromere based on DXZ1 sequences is the product of the much more recent concerted evolution of this satellite DNA.

The centromere is essential for normal segregation of chromosomes in both mitotic and meiotic cells. Paradoxically, although this role is conserved throughout evolution, the sequences that accomplish centromere function in different organisms are not (*1, 2*). These regions of the human genome remain an enigma and have been largely excluded from both the public (*3*) and private (*4*) sequencing efforts due in part to the highly repetitive DNA content of centromere regions. Indeed, reported contigs of pericentromeric regions in the human genome (*3, 4*) terminate at clones containing satellite DNA without reaching the extensive alpha satellite arrays that characterize all normal human centromeres. Alpha satellite DNA, defined by a diverged 171–base pair (bp) motif repeated in a tandem head-to-tail fashion, has been identified at the centromeres of all normal primate chromosomes studied to date (*5*). Stretches of alpha satellite that lack additional sequence structure are termed "monomeric" (*6*) and, where examined, appear to adjoin the euchromatic DNA of human chromosome arms (*7*). Human and great ape chromosomes contain alpha satellite organized hierarchically into higher-order repeat arrays in which a defined number of monomers have been homogenized as a unit to yield large chromosome-specific arrays that span several megabases (Mb) (*5, 8*). Extensive data on the structural and sequence characteristics of these arrays indicate a high degree of sequence homogeneity over regions as large as 3 to 4 Mb (*5*). Where examined, it is these higher-order repeat arrays that colocalize both with the centromere as defined by genetic recombination (*9, 10*) and with the cytologically defined primary constriction and site of a number of centromere and kinetochore proteins that have been implicated in centromere function (*1, 2*). Analysis of sequence variants within these otherwise highly homogeneous arrays supports the hypothesis (*11*) that the occurrence and maintenance of repetitive DNA arrays result from unequal crossover between sister chromatids (*5*). A hallmark of this mechanism is the predicted persistence of diverged repeat units at the edge of the array (*11*). This prediction can be tested at the genomic level only with the availability of contiguous maps and sequence across the transition between such arrays and the largely complete maps of the chromosome arms.

The role of primary DNA sequence in centromere function continues to be a matter of some debate (*1, 2, 12, 13*). Despite great interest in the elements of a functioning centromere, physical or genetic definition of the operational limits of a normal human centromere has not been reported. Complete maps describing the DNA content and organization of pericentromeric regions are essential toward this end, as all sequences present at the primary constriction are potential candidates for functional centromeric DNA, either as a foundation for assembly of a functional kinetochore or as elements involved in sister chromatid cohesion (*2*). Reported here is a map and partial DNA sequence representing a contiguous physical bridge between a functional mammalian centromere and the euchromatin of a chromosome arm, the short arm of the human X chromosome. The map serves not only as a model for mapping comparable regions of other chromosomes (both in the human and other complex genomes), but also as a resource for raising and testing questions about centromere function.

Assembly of a pericentromeric clone contig. To assemble a clone contig that spanned the junction between the termination of the existing contig (*3, 14*) on the X chromosome short arm (Xp) and the ~3 Mb array of DXZ1 alpha satellite sequences known to be present at the primary constriction of the X (*9, 15*), we capitalized on the sequence variation expected from previous analyses of the alpha satellite DNA family (*5*). Despite the often-anticipated difficulties of assembling contigs containing such repetitive DNA, the limited sequence divergence within alpha satellite was sufficient to allow development of specific sequence tagged sites (STSs) (*16*) and thus reliable determination of overlaps between clones. The resulting contig spans more than 500 kb and is characterized at a marker density of 1 STS per 25 kb (Fig. 1). Even though most of these STSs reside in satellite DNA (*17*), the use of high-stringency conditions and rigorous map checking with panels of rodent/human somatic cell hybrids to confirm localization to proximal Xp yielded what appears to be a straightforward clone contig with 5-fold average clone coverage (*16*). Map construction progressed using both in silico strategies and wet bench library screening to walk stepwise from three independent starting points. To walk toward the centromere, the two most proximal STSs from the existing contig in Xp11.21 (*14*) were used to query available databases (*18*). Our second seed position, from which we walked both proximally and distally, was a small localized block of a distinct satellite DNA family, gamma satellite, that is unrelated to alpha satellite and had been previously mapped to this region (*19*). Finally, we walked out from the DXZ1 array itself onto Xp (*18*). Bacterial artificial chromosome (BAC)–end sequencing followed by STS development, map verification, and library screening was completed at each walk step. Throughout the map construction, STS content, sequence analysis, and Southern analyses were used to determine clone integrity, overlap, and orientation.

[1]Department of Genetics, Case Western Reserve University School of Medicine and Center for Human Genetics, and [2]Research Institute, University Hospitals of Cleveland, Cleveland, OH 44106, USA.

*Present address: Brigham and Women's Hospital, Harvard Medical School, Boston, MA 02215 USA.

†To whom correspondence should be addressed. E-mail: willard@uhri.org

To verify that the assembled map is a true representation of X chromosomes, we performed restriction mapping both of the clones themselves (to demonstrate clone consistency) and of human genomic DNA (to demonstrate veracity of the contig). While restriction mapping with rare-cutter enzymes revealed no size or content discrepancies between clones, a number of restriction fragment length polymorphisms were detected in the region, as expected for clones that derive from several different X chromosomes (Fig. 1) (*5, 16*). To test directly whether the map was accurate, pulsed field gel fragment sizes predicted by the clone map were detected within genomic DNA from unrelated human males by Southern analysis (*20*). These data were consistent with expectations from previous long-range restriction mapping of the DXZ1 array (*9*) and suggest that the map derived from the contig is, in fact, representative of X chromosomes in general.

Sequence content of the Xp pericentromeric region. The transition to centromeric satellite sequences on Xp occurs within clone 344I7 (Fig. 1). This 98 kb clone overlaps the most proximal Xp contig (*14*) by 65 kb and contains 12 kb of alpha satellite at its proximal end. Approximately 90% of the clone sequence consists of known repeats, the majority of which are interspersed L1 repeat elements (*21*). This repeat density is over twice the genome average, a result consistent with previous cytogenetic, molecular, and bioinformatic observations on this region of the X (*22*). The transition from this region of dense interspersed repeats to alpha satellite is abrupt, marked by a partial L1 element adjoining an incomplete monomer of alpha satellite. This transition occurs 149 kb centromeric from the most proximal gene on Xp, the *ZXDA* zinc finger gene (*14, 23*). Alpha satellite monomers continue in a head-to-tail orientation for the remaining 12 kb of this clone sequence except for three interspersed repeat elements; each of these elements occurs within an alpha satellite monomer and is flanked by duplication of the site of insertion that is typical of retrotransposition (*21*).

Lying between the satellite junction and the DXZ1 array was an ~450 kb region highly enriched in alpha satellite. This region had not been bridged previously by the public human genome sequencing effort (*3*), while the Celera effort (*4*) includes partial assembly in this region (*24*). By BAC-end sequencing and shotgun sequencing, we further sampled ~33 kb of the region between the satellite junction and DXZ1. This random sample indicated a high percentage of alpha satellite sequences, in addition to gamma satellite, interspersed elements, and a novel 35 bp repeat (*25*); analysis of the partial Celera sequence in the distal portion of this region was largely consistent with this distribution (*24, 25*). Although the total available sequence covers only ~265 kb of the overall ~450 kb region, these analyses yielded no evidence of imbedded genes, other single-copy sequences, or paralogous repetitive sequences (*26*) shared with other chromosomes.

We next performed phylogenetic analysis of approximately 500 alpha satellite monomers sampled from this region to test whether distinct subgroups (clades) of monomers exist within the pericentromeric region, each sharing a common ancestor not shared by the other subgroup. Indeed, two such clades were evident as distinct branches on an evolutionary tree (Fig. 2), as determined by maximum parsimony, likelihood, and neighbor-joining methods (*27*). These data indicate that two evolutionarily distinct classes of alpha satellite are present within the centromeric region of the X chromosome (*27*). All monomers positioned within ~13 kb of the DXZ1 array (Figs. 2 and 3) fall into a clade (the DXZ1 clade) with monomers from the canonical DXZ1 repeat itself (*15*). All other monomers identified in our mapping and sequencing effort fall into a different clade (the monomeric clade) (Fig. 2A). Monomers in this clade were most closely related to a consensus sequence from alpha satellite suprachromosomal family 4 (*6*), consisting of alpha

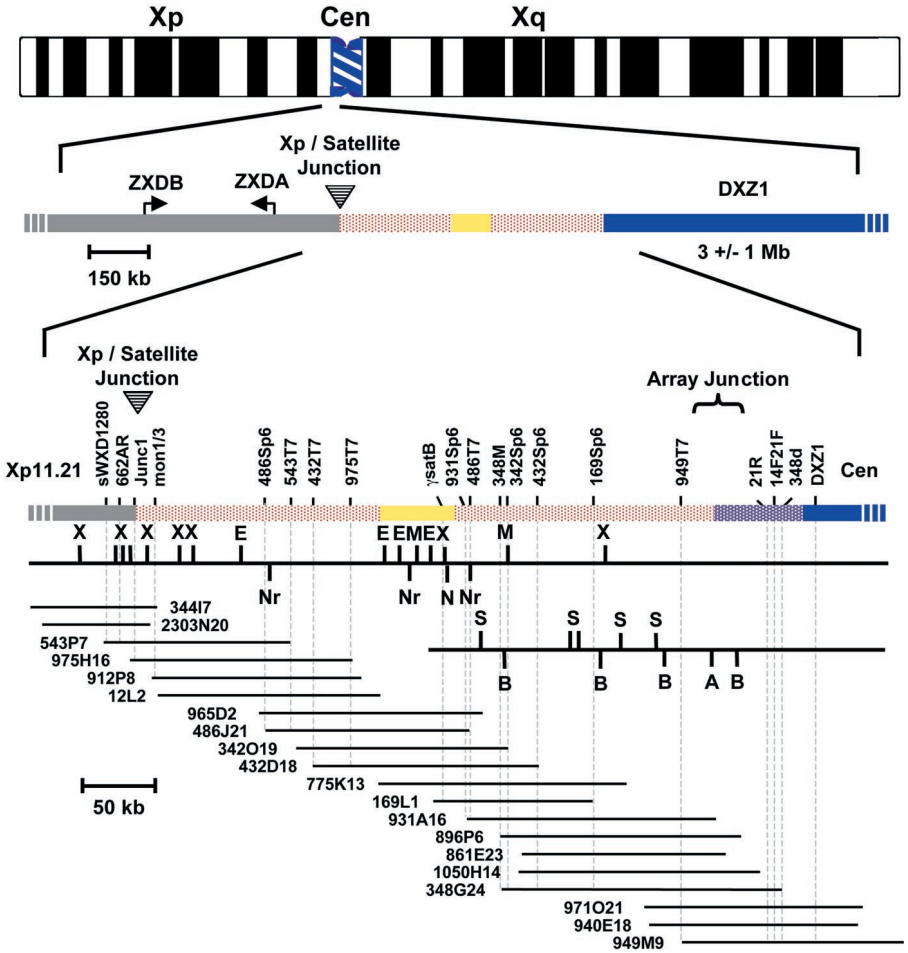

Fig. 1. Physical map of the pericentromeric region of the human X chromosome. The pericentromeric region of Xp is enlarged for orientation, with the large X chromosome-specific alpha satellite array, DXZ1, shown in blue. The junction between Xp euchromatic sequences and centromeric satellites is indicated, as are the nearest expressed genes, *ZXDA* and *ZXDB* (*23*). The 5′ end of *ZXDA* is ~150 kb from the satellite junction and ~600 kb from the DXZ1 array. A contig spanning the region joining the Xp arm with DXZ1 was constructed and is shown at higher resolution in the lower panel. Major genomic sequence features are indicated by color. The solid gray bar indicates non-satellite, euchromatic arm sequences; the red-stippled region contains primarily monomeric alpha satellite (see text); the yellow region contains the ~50 kb gamma satellite array (*19*); the blue-stippled region indicates diverged DXZ1 (see text; Fig. 3). STSs used to construct the contig are listed along the top of the bar and the Xp/satellite and the array junctions are indicated. Dashed lines descending from each STS summarize the STS content of each clone as determined by PCR and, in some cases, Southern analysis. Below the genomic feature bar is the restriction enzyme map determined by Southern analysis using all of the clones displayed. Sites included are those for enzymes XhoI, X; EagI, E; NruI, Nr; MluI, M; NotI, N. A paucity of rare restriction enzyme sites in the proximal region [as predicted for satellite DNA (*8*)] prompted restriction mapping using more frequent cutters; shown are sites for SmaI, S; BglI, B; and ApaI, A. Only clones 775K13 proximal through 949M9 were used to construct this proximal restriction map.

satellite sequences from the human and lower primate genomes that consist of monomeric repeat structure (Fig. 2B) (*27, 28*). While members of this family map to many different centromeric regions in the human genome (*6, 7, 28*), it is important to stress that the members of this clade from the current effort are specific for the X chromosome (*16*).

From this initial sequence analysis, it is evident that monomers within the monomeric clade tend to group together phylogenetically based on their genomic position. Not only are these monomers distinct genomically from members of the DXZ1 clade (Fig. 3), but sampled monomers lying distal to gamma satellite, for example, tend to form a separate group from monomers proximal to gamma satellite. This finding is consistent with the presumed short-range nature of homogenization events involving satellite DNAs (*5, 29*) and suggests that the integration of gamma satellite into an array of monomeric alpha satellite may have served as a punctuating event to drive the subsequently distinct homogenization of sequences proximal and distal to the integration site.

Satellite DNA in transition. Monomeric alpha satellite transitions to DXZ1 alpha satellite at a point ~175 kb proximal to gamma satellite and ~550 kb proximal to the *ZXDA* gene (Fig. 1). To determine the sequence characteristics at the border of the DXZ1 array, we sequenced ~9 kb from this transition region at the proximal end of clones 348G24 and 971O21 (Figs. 1 and 3). To characterize sequence relationships among different members of this highly homogeneous repeated DNA family, we developed an approach to define subgroups of DXZ1 based on analysis of the specific positions within each DXZ1 monomer at which the sequence varies from a consensus DXZ1 monomer. Base changes at each of these sites provide an unambiguous signature for each of the 12 monomers of DXZ1, and the number of such signature sites shared between two monomers is a measure of their relatedness (*30*). Combined analysis of percent sequence identity, shared signature sites, higher order repeat structure, and phylogenetic relationship to canonical DXZ1 identified four types of DXZ1 at the edge of the array (Fig. 3). All four types belong to the DXZ1 clade, as defined earlier (Fig. 2A).

Canonical DXZ1 (type 1) is defined by a 2.0 kb higher-order repeat that dominates the centromeric region of the X chromosome and spans 3 ± 1 Mb (*9, 15*). The canonical repeats show an average of 1 to 2% divergence between copies on the same or different X chromosomes (*29, 31*). Type 2 DXZ1 shares the same higher-order repeat structure, but has substantially greater sequence divergence than documented previously, with only 91 to 97% sequence identity to type 1 units. The degree of sequence identity to type 1 DXZ1 (presumably reflecting the efficiency of sequence homogenization) diminished with increasing distance from the array (Fig. 3).

Immediately distal to type 2 DXZ1 on Xp is type 3 DXZ1 (Fig. 3). These monomers could be assigned to one of the five monomer groups that comprise DXZ1 [thus distinguishing them from monomeric alpha satellite (Fig. 2B)], but were too diverged in sequence to distinguish among different members within a monomer group (*15*). The percent sequence identity between type 3 and type 1 DXZ1 is only 85% (Fig. 3), similar to the percent identity between related alpha satellite families on different chromosomes (*5*). As the physical distance from the type 1 DXZ1 array increased further, the percent identity to DXZ1 dropped (Fig. 3) and the number of shared signature sites diminished until it was not possible to assign a test monomer to any one DXZ1 monomer or group. Nonetheless, phylogenetic analysis clearly indicated that they belong to the DXZ1 clade, thus defining type 4 DXZ1 (Fig. 3). By means of restriction enzyme analysis, sizing, and STS content data, the maximum separation between the monomers operationally defined here as type 4 DXZ1 and the nearest identified monomer from the monomeric clade was determined to be 100 kb (Figs. 1 and 3) (*32*).

Computer modeling of unequal crossover as a mechanism for homogenization of tandem arrays of satellite DNA predicts perseverance of diverged higher-order sequences

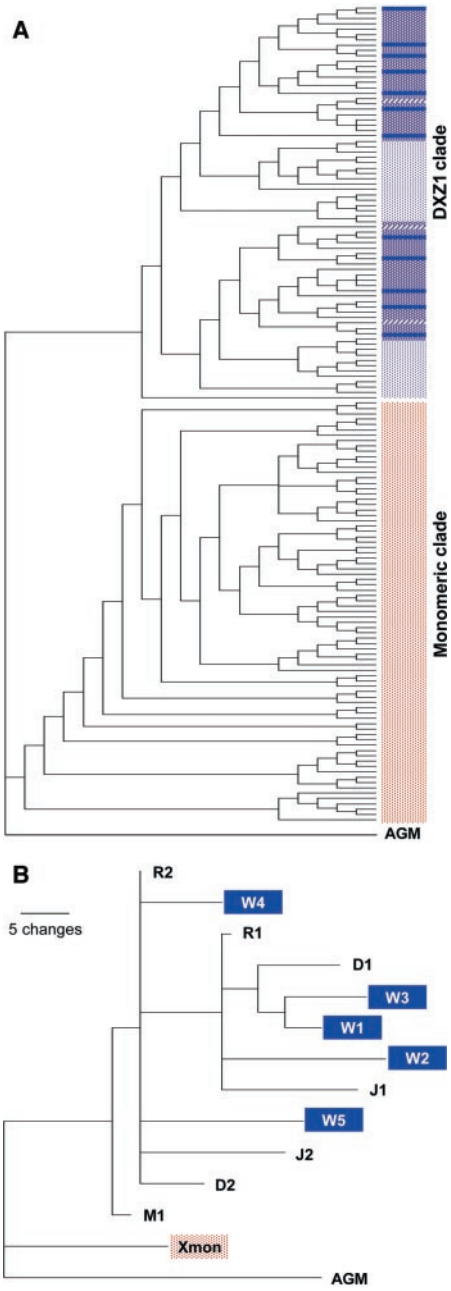

Fig. 2. Phylogenetic analysis of pericentromeric Xp alpha satellite. (**A**) PAUP tree resulting from phylogenetic analysis using maximum parsimony criteria (*27*). The tree is a strict consensus of 1000 trees generated by an Heuristic search. Each taxa is a single 171-bp monomer of alpha satellite. We included 155 taxa in this analysis, including 12 control monomers of type 1 DXZ1 (*9*), 57 monomers distal to gamma satellite, 84 monomers proximal to gamma satellite, and one monomer of African Green Monkey (AGM) alpha satellite as an outgroup. All monomers distal to gamma satellite fall into the monomeric clade, shown in red. The 62 most proximal monomers fall into the DXZ1 clade (Fig. 3). Within the DXZ1 clade, there are two major groups each consisting of a cluster of types 1, 2, and 3 DXZ1 monomers and a cluster of type 4 DXZ1 monomers. The major branch at the DXZ1 clade is well supported by replicate analysis (980 of 1000 bootstraps). Analysis of the full sequence dataset with 474 taxa yields a similar tree with two major clades (*27*). Blue, type 1 DXZ1; dark blue stippled, type 2 DXZ1; blue stripped, type 3 DXZ1; light blue stippled, type 4 DXZ1; red stippled, monomeric alpha satellite. (**B**) The phylogram of consensus monomer sequences resulting from 1000 bootstrap replicates of an Heuristic search using maximum parsimony criteria. Each taxa is an alpha satellite consensus sequence representing monomer types from the major suprachromosomal families (*5, 28*) and a consensus sequence (Xmon) derived from 382 monomers from the satellite region distal to DXZ1 (*27*). W1-W5 represent consensus sequences for the five monomer groups that compose DXZ1. The major branch placing Xmon outside of the clade of human consensus sequences and closest to the AGM sequence is supported by 92% of bootstrap replicates.

at the borders of the array (*11*). These sequences represent members of the array that have been excluded from the most recent and more probable homogenization events within the larger region of homogeneous sequence and thus have undergone random mutation without fixation (*5, 8*). The data presented in Fig. 3 confirm these predictions and thus directly support unequal crossover as a mechanism of homogenization of DXZ1 alpha satellite. While sequence conversion events may play an occasional role in local homogenization, only unequal crossover can explain the range of features that characterizes alpha satellite arrays (*5, 8*): the generation and persistence of a multimeric higher-order repeat length, the extensive spread of sequence variants across millions of basepairs, documented recombination events leading to both deletions and insertions within multimeric repeat units and the rapid fall in sequence identity documented here at the edge of the overall array. None of these observations can be easily accommodated by sequence conversion models.

Our data also provide a genomic and chromosomal context for the unequal crossover model, as the substantially diverged DXZ1 sequences appear to be confined to a relatively small genomic region (<100 kb) that comprises only a few percent of the total length of the overall array. Viewed another way, this means that processes of satellite DNA homogenization are highly efficient on a genomic scale, extending throughout long ranges over the vast majority of the total array length. These considerations refine previous data that documented short-range homogenization within regions of a few hundred kilobases or less (*5, 29*).

Mapping of the functional centromere. To investigate which centromeric sequences are necessary for X chromosome segregation, breakpoints involved in the generation of X chromosome rearrangements were mapped by molecular cytogenetic and genomic methods and sequences implicated in centromere function were inferred by comparative deletion analysis. Isochromosomes of the X chromosome long arm, i(Xq), contain mirror image Xq arms, but lack all or most of Xp (Fig. 4A). Fluorescence in situ hybridization analysis of a series of i(Xq) isolated from Turner syndrome patients (*33, 34*) identified six i(Xq) whose breakpoints occur within the Xp pericentromeric map region described here. Three of these failed to hybridize with even the most proximal probe tested, but still showed clear hybridization signals with DXZ1 (*34*). Thus, elimination of all Xp sequences, including gamma satellite (*19*) and monomeric alpha satellite but not DXZ1 (Fig. 4A), does not effect the mitotic segregation of these naturally occurring chromosomes.

In a complementary manner, we investigated a potential functional requirement for pericentromeric sequences on Xq, using a series of Xp isochromosomes that lack all or most of the Xq arm (*35*). PCR analysis of 11 Xp isochromosomes identified six that contained DXZ1 sequences, but failed to amplify with the most proximal STS on Xq (*36*). For three of the Xp isochromosomes, evidence for breakpoints within the DXZ1 array was obtained by pulsed-field gel analysis (*36*). Thus, analysis of the Xp isochromosomes indicates that deletion of all Xq sequences (including a portion of the DXZ1 array clos-

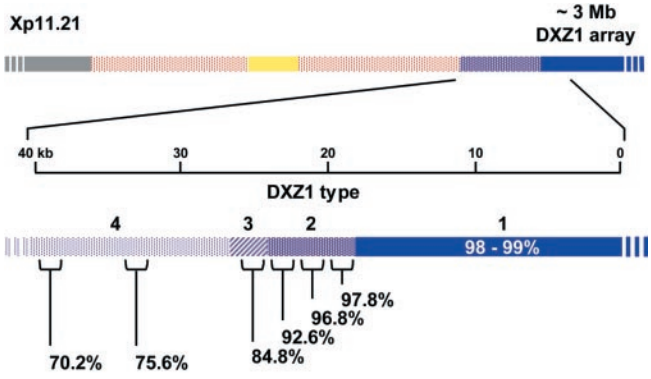

Fig. 3. Expanded region including the edge of the DXZ1 array. Type 1 higher-order repeats are typically 98 to 99% identical to a DXZ1 consensus sequence (*31*). Adjacent higher-order repeat units found at the border of the DXZ1 array were compared with this DXZ1 consensus sequence to determine percent identity. Higher-order repeats of DXZ1 at the border of the array (types 2 and 3) diverge from type 1 as physical distance from the array increases. Monomers of type 4 DXZ1 from two locations were compared with the DXZ1 consensus and are considerably more divergent. Sequence analyzed is continuous for ~9 kb adjacent to type 1 DXZ1. Samples of type 4 repeats are discontinuous. Comparisons based on 4 to 12 monomers for each type 2 sequence and on 3 monomers each for types 3 and 4. Blue, type 1 DXZ1; dark blue stippled, type 2 DXZ1; blue stripped, type 3 DXZ1; light blue stippled, type 4 DXZ1.

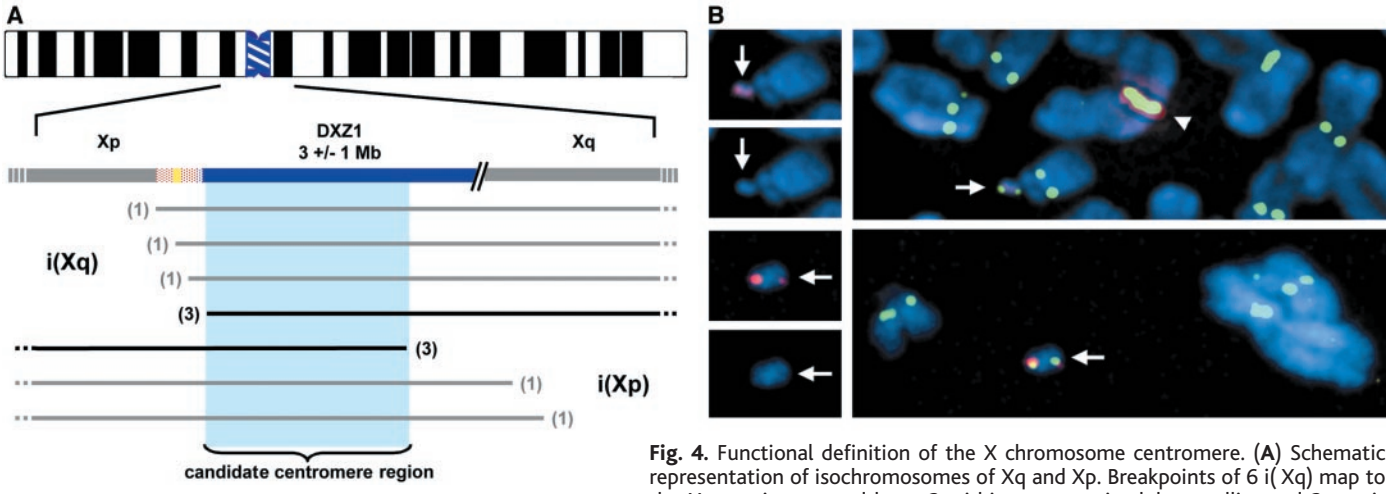

Fig. 4. Functional definition of the X chromosome centromere. (**A**) Schematic representation of isochromosomes of Xq and Xp. Breakpoints of 6 i(Xq) map to the Xp contig reported here, 3 within monomeric alpha satellite and 3 proximally in DXZ1 (*34*). Breakpoints of 5 i(Xp) map within the pericentromeric region, 3 within the DXZ1 array by pulsed-field gel analysis (*36*). Analysis of the 3 most proximal i(Xq) and the 3 i(Xp) with breaks in DXZ1 (shown in bold) define a functional centromere candidate region. (**B**) DXZ1-containing human artificial chromosomes are shown (*38*). Small arrows indicate the artificial chromosomes, identified by FISH using a DXZ1 probe (red). Active centromeres are identified by indirect immunofluorescence using an antibody to CENP-E (green). The overlap between the red and green signals appears yellow. A normal X chromosome is indicated by the arrowhead in the top right panel. Insets at the left show results of FISH alone (top) and DAPI staining (bottom) to identify the artificial chromosomes.

est to the Xq arm) does not effect the mitotic stability of these chromosomes. Together, analysis of the two types of complementary isochromosome establishes DXZ1 sequences as a candidate for functional centromeric sequences on the X (Fig. 4A).

This deletion analysis is predicated on the assumption that the functional centromere in a rearranged chromosome is the same as the functional centromere in the parental chromosome, prior to isochromosome formation. Although this seems intuitively likely and is the accepted basis for similar deletion analyses in a wide range of genetic and genomic studies, it is impossible to exclude epigenetic models (*1, 2, 12*) that invoke movement of the functional centromere from one sequence to an adjacent sequence. However, the frequency of such movement necessary to explain the current data would appear to be inconsistent with the known karyotypic stability of normal human chromosomes. In addition, persistence of DXZ1 sequences in all mitotically stable X isochromosomes would not be a predicted feature of such a "moving target" hypothesis.

To test the potential centromere function of DXZ1 sequences directly, we examined their ability to support formation of a de novo centromere in human cells using an artificial chromosome assay (*37*). The 85-kb insert from a BAC clone containing exclusively type 1 DXZ1 sequences was transferred to a bacteriophage P1–derived artificial chromosome (PAC) vector containing a selectable marker (*38*). This DNA was transfected into human HT1080 cells and artificial chromosome formation was evaluated. Artificial chromosomes containing DXZ1 were generated in ~10% of clones, and four such artificial chromosomes were selected for further characterization. The amount of DXZ1 DNA contained in the artificial chromosomes ranged from an estimated <500 kb to more than 5 Mb. Each was mitotically stable in the absence of drug selection for 30 days and each recruited a centromere protein, the kinesin motor protein CENP-E, that is associated only with active centromeres (*2, 39*), thus indicating that the artificial chromosomes had assembled a functional centromere and kinetochore (Fig. 4B). These data indicate that type 1 DXZ1 sequences are competent to form a de novo centromere, consistent with predictions of the isochromosome deletion mapping data. Whether more diverged members of the DXZ1 clade (or the more distal monomeric alpha satellite) are also capable of artificial chromosome formation is subject to further investigation using the resources described here. However, it is clear from analysis of isochromosomes that such sequences are not required for normal mitotic centromere function (Fig. 4A).

L1 repeats date centromere evolution. We have examined the phylogeny of the L1 family of interspersed repeats (*21*) to explore the evolution of the X pericentromeric region, as L1 family members have been previously described within alpha satellite (*9, 10, 40*). Because alpha satellite is a primate-specific sequence (*5, 28*), one would expect that older, mammalian-wide L1 elements would not be present within alpha satellite regions. Indeed, only those elements mobile since the emergence of prosimians (L1P or primate-specific L1; 60 Ma) (*41*) should be found within alpha satellite, and the presence of L1P elements that were active early in primate evolution (25 to 60 Ma) will indicate regions possibly shared throughout the primate lineage. The collection of L1 elements in Xp euchromatin distal to the satellite junction and within clone 344I7 includes both young (L1P2) and very ancient (L1M4) elements (Fig. 5), demonstrating the likely conservation of this DNA throughout mammalian X chromosome evolution. In contrast, only primate-specific elements mobile near the emergence of Old World monkeys at ~25 Ma (L1PA6 and L1PA7) and more recently mobile elements (L1PA2, L1PA3, and L1PA4) were detected within the X chromosome monomeric alpha satellite region (Fig. 5). Further, a directed effort to identify L1 elements within the DXZ1 array yielded only a single L1, a member of the human-specific L1Hs subfamily (*42*). [The absence of L1Hs and L1P1 elements from more distal portions of the region under study (<0.3% of the X chromosome) presumably reflects their general underrepresentation throughout the genome relative to older L1 elements (*41*).]

These data support an evolutionary scheme in which the currently functional DXZ1 sequences are newly evolved and/or newly homogenized on the X, as they contain no L1 "fossil evidence" of a more ancient existence (Fig. 5). This contrasts with other sequences, such as regions of monomeric alpha satellite, that show clear evidence of being present for at least 7 to 8 million years (Fig. 5) and thus may have preceded higher-order arrays of alpha satellite as the functional primate centromere. The characteristics of the monomeric class of alpha satellite are consistent with those we would predict for an ancestral primate centromere sequence. The genomes of lower primates have been shown to contain monomeric alpha satellite that lacks both higher-order structure and certain sequence features such as binding sites for the centromere protein CENP-B (*28, 43*). Monomeric alpha satellite on the X chromosome spans the region joining the arm of the chromosome with the DXZ1 array, and monomeric alpha satellite has also been detected in the pericentromeric regions of at least six other human chromosomes (*7, 10, 27*). Highly homogeneous arrays of higher-order alpha satellite, such as DXZ1, are relatively recent additions to our genome, emerging near the Orangutan/Gorilla split (*5, 28*), at a time of apparent pericentromeric expansion. Predicted mechanisms of homogenization—in which amplification of a small group of monomers present within the monomeric stretch is followed by fixation through unequal crossover—would result in large higher-order repeat arrays (*5, 11, 44*). Homogenization of the sequence characteristics of the particular repeat during expansion would create a sequence class distinct from its progenitor. Our demonstration that higher-order alpha satellite and monomeric alpha satellite are distinct (Fig. 2) supports this mechanism. The age gradient revealed by L1 subfamilies (Fig. 5) supports the hypothesis that monomeric alpha satellite present within the pericentromeric regions of human chromosomes predates higher-order arrays of alpha satellite and thus may represent direct descendants of the ancestral primate centromere sequence. It seems likely that descendants of this ancestral

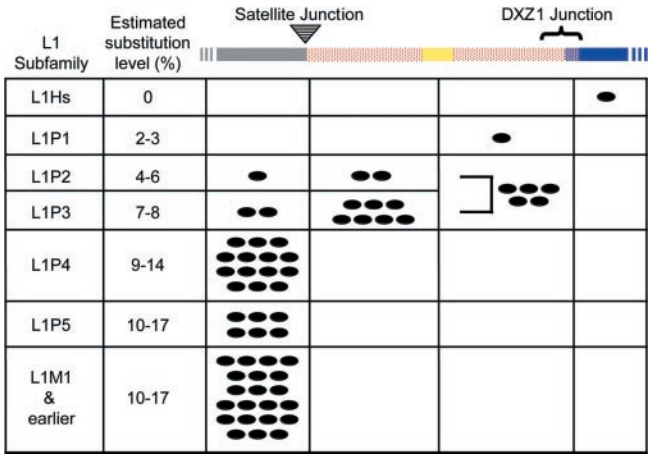

Fig. 5. Distribution of L1 elements detected within alpha satellite DNA in the Xp pericentromeric contig. L1 subfamily designations and estimated level of substitution relative to the current L1 element, L1Hs, are shown at the left, from (*41*). Each oval represents one L1 element observed within the section of the map region designated along the top, assigned to different L1 subfamilies based on sequence content (*41*). Genomic features of the map are shown in color, as in Fig. 1. A human-specific L1 element (L1Hs) is found only in DXZ1, while primate-specific elements (L1P subfamilies) are found in monomeric alpha satellite region. In contrast, the euchromatic region of Xp contains L1 elements from both ancient (L1M1 subfamily) and more recent evolutionary time.

sequence will be found at all human (and presumably other primate) centromeres (*28*). Whereas the analysis here is based on the use of L1 evidence to infer the age of different alpha satellite classes, it is also possible that L1 elements play a more active role in determining the relative competence for centromere function of different types of alpha satellite (*45*). Comparative genomic and functional analysis of the orthologous region of X chromosomes from higher and lower primate genomes (with and without higher-order alpha satellite arrays, respectively) may provide direct evidence on this point.

Centromere genomics, evolution and complexity. Conservation of X chromosome content throughout the mammalian radiation, as hypothesized by Ohno (*46*), has been upheld with few exceptions. Genetic maps of the X chromosomes of many species not only support conservation of content, but also reveal colinear gene order along the length of the chromosome (*47*). Notwithstanding this chromosomal isolation throughout evolution, the DNA responsible for the segregation of the X chromosome has changed dramatically since the mammalian radiation and even since the emergence of primates. Recent reports of centromere repositioning in prosimians (*48, 49*) may highlight early events in the establishment of the primate centromere. Our analysis of the genomic composition of the X pericentromeric region indicates that the functional centromere has continued to evolve with the expansion or addition of the higher-order alpha satellite array, DXZ1, during recent primate evolution. Analysis of the complete sequence of the pericentromeric region will provide unique insight into the evolution of alpha satellite sequences and, through comparative genomic analysis, into the evolution of other solutions to the centromere problem in mammals and other organisms.

Outside of yeasts, centromeres have been analyzed at the genomic or functional level in relatively few complex eukaryotic organisms to date (*1, 2*). While centromeric regions in many organisms appear to be dominated by satellite DNAs (*1, 2, 50, 51*), the competence of such sequences to form de novo centromeres has generally not been established. Given the wide diversity of satellite and other DNA sequences found in pericentromeric regions in different genomes, one may anticipate a similarly wide range of solutions to the need for functional centromeres.

Even in the human genome, the situation may be more complex than is apparent from our analyses of the X chromosome centromere. For example, the X pericentromeric region shows no evidence to date of the extensive duplications and paralogies that mark many centromere regions in the human genome (*7, 26, 52*). Further, it is apparent that centromere function is determined in part epigenetically (*1, 2, 12, 13*). Notwithstanding our identification of DXZ1 sequences as the functional centromere on the human X chromosome (and likely, therefore, on other normal human chromosomes), rare abnormal chromosomes that lack alpha satellite DNA altogether can segregate normally through the action of activated neocentromeres (*12, 53*).

As we strive to understand the organization and evolution of the human genome, any claims of a "complete" sequence will require full analysis of the pericentromeric and other heterochromatic regions of our chromosomes. The data shown here suggest that, notwithstanding their repetitive content and the need for a directed genomic approach, the satellite-containing centromeric regions can be mapped, sequenced and assembled. Complete assembly of centromere contigs should be feasible, and, as demonstrated here, will have both functional and evolutionary implications.

References and Notes

1. S. Henikoff, K. Ahmad, H. S. Malik, *Science* **293**, 1098 (2001).
2. B. A. Sullivan, M. D. Blower, G. H. Karpen, *Nature Genet. Rev.* **2**, 584 (2001).
3. International Human Genome Sequencing Consortium, *Nature* **409**, 860 (2001).
4. J. C. Venter *et al.*, *Science* **291**, 1304 (2001).
5. P. Warburton, H. Willard, in *Human Genome Evolution*, M. Jackson, T. Strachan, G. Dover, Eds. (BIOS Scientific, Oxford, 1996), pp. 121–145.
6. I. A. Alexandrov *et al.*, *Nucleic Acids Res.* **21**, 2209 (1993).
7. J. E. Horvath *et al.*, *Hum. Mol. Genet.* **9**, 113 (2000).
8. H. F. Willard, J. S. Waye, *Trends Genet.* **3**, 192 (1987).
9. M. M. Mahtani, H. F. Willard, *Genome Res.* **8**, 100 (1998).
10. J. Puechberty *et al.*, *Genomics* **56**, 274 (1999).
11. G. P. Smith, *Science* **191**, 528 (1976).
12. K. H. Choo, *Trends Cell Biol.* **10**, 182 (2000).
13. H. F. Willard, *Curr. Opin. Genet. Dev.* **8**, 219 (1998).
14. A. P. Miller *et al.*, *Hum. Mol. Genet.* **4**, 731 (1995).
15. J. S. Waye, H. F. Willard, *Nucleic Acids Res.* **13**, 2731 (1985).
16. Specific STSs were developed as described (*17*). STSs and clones were evaluated for Xp-specifity using PCR, Southern and fluorescence in situ hybridization (FISH) analyses. Primer sequences and amplification conditions, as well as BAC clone characteristics and library sources, can be found at http://genetics.gene.cwru.edu/willard/data.htm. X chromosome specificity was determined by PCR amplification from the Corriell Mapping Panel Version 2 (NIGMS Human Genetic Mutant Cell Repository, Camden NJ). Amplification from mouse/human somatic cell hybrid DNA containing either Xp or Xq was used to determine the Xp-specificity of STSs. Clones were also analyzed for Xp specificity by FISH to metaphase chromosomes from a patient cell line in which a translocation breakpoint within the DXZ1 array separates Xp and Xq in their entirety (GM03316, NIGMS Human Genetic Mutant Cell Repository, Camden NJ). High stringency Southern blots of BAC DNA and total human genomic DNA were performed essentially as described (*15*) and FISH was performed using standard protocols (*14*).
17. P. E. Warburton, G. M. Greig, T. Haaf, H. F. Willard, *Genomics* **11**, 324 (1991).
18. To walk toward the centromere from Xp11.21, STSs sWXD1280 (AL41952) and 662AR (G02071; Fig. 1), were used to query the high throughput genomic sequence (HTGS; www.ncbi.nlm.nih.gov/BLAST) and identified a fully sequenced PAC clone (344I7; AL024458). Sequence from this clone was used to query a BAC end sequence database (www.tigr.org/tdb), identifying an additional BAC clone (2303N20; CITB). To progress from DXZ1 out onto the short arm, the sequence of a single 2.0 kb higher-order repeat of DXZ1 (*15*) was used to query genome survey sequences (GSS) by BLAST. One clone (348G24; AL591484) shared significant identity (96%) to DXZ1 at one end (AQ528473) and had a sequence other than DXZ1 at its opposite end (AQ528470). STSs developed to this sequence mapped this clone to Xp. All remaining clones of the contig were isolated by standard wet bench library screening.
19. C. Lee, R. Critcher, J. G. Zhang, W. Mills, C. J. Farr, *Chromosoma* **109**, 381 (2000).
20. M. G. Schueler *et al.*, data not shown. High molecular weight human genomic DNA was digested with *Xho*I and separated by pulsed-field gel electrophoresis, and hybridized to alpha satellite probes at high stringency, essentially as described (*16*). *Xho*I fragment lengths predicted by the clone restriction map were detected in all human samples tested using probes from clones 432D18 and 169L1 (Fig. 1).
21. A. F. Smit, *Curr. Opin. Genet. Dev.* **6**, 743 (1996).
22. J. A. Bailey, L. Carrel, A. Chakravarti, E. E. Eichler, *Proc. Natl. Acad. Sci. U.S.A.* **97**, 6634 (2000).
23. G. M. Greig, C. B. Sharp, L. Carrel, H. F. Willard, *Hum. Mol. Genet.* **2**, 1611 (1993).
24. Celera assembly GA_x2HTBKQSN8H (*4*) spans 500 kb in proximal Xp and includes coverage of 265,268 bp of the satellite region described here. The dataset extends from the start of the satellite region through gamma satellite and terminates ~35 kb proximal to gamma satellite (Fig. 1). While the Celera contig assembly is completely in agreement with that described in Fig. 1, the DNA sequence is incomplete, as 17.2% of the sequence is N's. Nonetheless, >98% of the available sequence consists of repetitive DNA, including alpha satellite (30.4%), gamma satellite (14.7%), various interspersed repeats (4.1%) and a localized 35-bp repeat (33%) (*25*). Because the Celera sequence is limited to the distal portion of the satellite region described here, it overrepresents gamma satellite and the 35-bp satellite, while undersampling alpha satellite (*25*).
25. To derive an unbiased sample of the entire ~450 kb satellite region, sequence was determined from 33 independent BAC end sequences totaling 16.6 kb. Of these, 23 BAC end sequences were alpha satellite, 1 was gamma satellite, 6 were LINE elements, and 1 was a HERV repeat element. To further sample the transition region, shotgun libraries from four BAC clones (543P7, 432D18, 486J21, 348G24) were created by digestion with *Eco*RI or *Bam*HI. Whole digests were ligated in the presence of pUC and several subclones from each BAC clone were analyzed. A total of 38 independent sequences from 20 subclones sampled an additional 17 kb of sequence. Of these, 21 subclone sequences were alpha satellite, 10 were gamma satellite, and 3 were LINE sequences. Seven sequence samples detected no repeats by RepeatMasker analysis (http://ftp.genome.washington.edu/cgi-bin/RepeatMasker); however, upon analysis, six of these consisted of a novel GC-rich, 35-bp repeat by MEME analysis (http://meme.sdsc.edu/meme/website). Units of this repeat occur in a tandem, head-to-tail fashion and share ~85% pairwise sequence identities (M. G. Schueler, data not shown). Overall, analysis of the 33-kb of sequence sampled from across the region indicated a high percentage of alpha satellite sequences (62%), followed in frequency by gamma satellite (16%), interspersed elements (14%), and the 35 bp repeat (8%).
26. J. E. Horvath, S. Schwartz, E. E. Eichler, *Genome Res.* **10**, 839 (2000).
27. The tree depicted in Fig. 2A was generated by PAUP [D. Swofford D. Begle (Illinois Natural History Survey, Champaign, IL, 1993)]. Each taxa is a single 171-bp monomer of alpha satellite identified through our sample sequencing efforts (*25*), the available Celera sequence (*24*) and the 344I7 sequence (*18*). This tree results from an Heuristic search using maximum parsimony criteria and includes 155 taxa, with one monomer of African Green Monkey (AGM) alpha satellite [H. Rosenberg, M. Singer, M. Rosenberg, *Science* **200**, 394 (1978)] designated as the outgroup. Heuristic searches using maximum parsimony, likelihood, or distance criteria produce branches supporting two major clades. Of 1000 bootstrap replicates, 980 support the major

branch distinguishing the two clades. Analysis of the entire dataset (474 taxa) by heuristic search using maximum parsimony criteria yields the same major branches depicted in Fig. 2A. For the analysis of consensus sequences shown in Fig. 2B, 382 monomers from the monomeric clade (Fig. 2A) were aligned using ClustalW and a single X monomeric consensus sequence (Xmon) was derived. Bootstrap analysis was used to determine the relationship of Xmon to the consensus monomers established for the four human suprachromosomal families (*28*), again using AGM as an outgroup (Fig. 2B). Of 1000 bootstrap replicates, 92% support the major branch separating Xmon and AGM from the remaining consensus monomers. Our analysis indicates that Xmon is more closely related to AGM and to the M1 (suprachromosomal family 4) sequence than to any of the other human consensus monomers.

28. I. Alexandrov, A. Kazakov, I. Tumereva, V. Shepelev, Y. Yurov, *Chromosoma* **110**, 253 (2001).
29. S. J. Durfy, H. F. Willard, *Genomics* **5**, 810 (1989).
30. Specific positions within each DXZ1 monomer at which the sequence varies from a consensus DXZ1 monomer are designated signature sites. The analysis of signature sites was automated using a perl script, modifying the consensus identity index used previously (*15*). The script first identifies and then compares signature sites of test monomers with the specific signature sites of each DXZ1 monomer. Because any two higher-order repeats of typical DXZ1 (type 1) share 98 to 100% identity (*31*), the number of signature sites for each DXZ1 monomer varies slightly. Type 2, 3, and 4 DXZ1 repeats have numbers of signature sites that are outside the range of those detected in type 1 repeats.
31. Approximately 1000 to 2000 copies of the DXZ1 higher-order repeat exist on each X chromosome (*9*, *15*). Average sequence divergence was estimated from the sequence of seven complete (M. G. Schueler *et al.*, unpublished data) and over 40 partial repeat units (*29*), from at least nine different X chromosomes.
32. Type 2 and type 3 DXZ1 sequences show approximately the same degree of sequence divergence within each type as they do between types. This indicates that the gradient of divergence from type 1 DXZ1 is very steep and that the diverged types are no longer being homogenized to the same degree as type 1. This is consistent with a model in which the efficiency of homogenization decreases both as a function of physical distance from the type 1 array and as a direct consequence of the increased sequence dissimilarity itself. (Type 3 is represented by only a single repeat and thus cannot be evaluated from this perspective.)
33. D. J. Wolff *et al.*, *Am. J. Hum. Genet.* **58**, 154 (1996).
34. K. Gustashaw *et al.*, data not shown. The most proximal probe tested was 1050H14 (Fig. 1), containing both monomeric alpha satellite and types 3 and 4 DXZ1 sequences.
35. A. W. Higgins, M. G. Schueler, H. F. Willard, *Chromosoma* **108**, 256 (1999).
36. A. W. Higgins *et al.*, data not shown. The Xq contig (*54*) terminates an estimated few hundred kb from DXZ1. Pulsed-field gel analysis was carried out as in (*9*). Isochromosomes with breaks within the DXZ1 array showed both missing and altered restriction fragments relative to the parental X chromosome prior to isochromosome formation. Isochromosomes with breaks outside of DXZ1 on proximal Xq showed unaltered DXZ1 restriction patterns. Isochromosomes with breaks in DXZ1 confirm Darlington's original model of centromere misdivision [C. D. Darlington, *J. Genet.* **37**, 341 (1939)].
37. J. J. Harrington, G. Van Bokkelen, R. W. Mays, K. Gustashaw, H. F. Willard, *Nature Genet.* **15**, 345 (1997).
38. M. K. Rudd *et al.*, data not shown. A ~85 kb NotI fragment from BAC (242E23) containing DXZ1 higher-order repeats was isolated and cloned into the pPAC4 vector, containing a blasticidin-resistance gene [T. A. Ebersole *et al.*, *Hum. Mol. Genet.* **9**, 1623 (2000)]. The circular PAC was then transfected into HT1080 cells, as described (*37*). Blasticidin-resistant colonies were isolated and screened cytogenetically by FISH for DXZ1-containing artificial chromosomes. Tests of mitotic stability were carried out in the presence and absence of blasticidin. After 30 days, clones containing an artificial chromosome in >60% of cells were scored as stable (*37*). Indirect immunofluorescence for CENP-E staining was performed as described (*37, 39*).
39. B. A. Sullivan, S. Schwartz, *Hum. Mol. Genet.* **4**, 2189 (1995).
40. A. M. Laurent *et al.*, *Genomics* **46**, 127 (1997).
41. A. F. Smit, G. Toth, A. D. Riggs, J. Jurka, *J. Mol. Biol.* **246**, 401 (1995).
42. Observed LINE elements (Fig. 5) include those detected through our sequence sampling efforts and from sequence within 344I7 or other sequenced clones available from sequence databases (*4, 18, 25*). Only L1 elements that were imbedded within alpha satellite were considered. A directed effort was made to identify LINE elements within the DXZ1 array. DNA from 45 BAC stabs (from the RPCI-13 library) positive for DXZ1 by PCR and hybridizing to 348d (type 2 DXZ1; Fig. 1) at high stringency were digested with BamHI, followed by Southern analysis using probes to the 5′ and 3′ ends of L1 at reduced stringency. Two clones yielded both the diagnostic DXZ1 2.0-kb BamHI fragment and hybridization with L1 probes following colony purification. Fragments from both BAC clones that hybridized with L1 probes were subcloned and sequenced. Restriction enzyme digestion patterns and sequence were identical between the subclones originating from the two BAC clones, indicating that these two clones sample the same LINE element.
43. I. G. Goldberg, H. Sawhney, A. F. Pluta, P. E. Warburton, W. C. Earnshaw, *Mol. Cell. Biol.* **16**, 5156 (1996).
44. L. Donehower, D. Gillespie, *J. Mol. Biol.* **134**, 805 (1979).
45. A. M. Laurent, J. Puechberty, G. Roizes, *Chromosome Res.* **7**, 305 (1999).
46. S. Ohno, *Nature* **244**, 259 (1973).
47. S. J. O'Brien *et al.*, *Science* **286**, 458 (1999).
48. M. Ventura, N. Archidiacono, M. Rocchi, *Genome Res.* **11**, 595 (2001).
49. G. Montefalcone, S. Tempesta, M. Rocchi, N. Archidiacono, *Genome Res.* **9**, 1184 (1999).
50. T. D. Murphy, G. H. Karpen, *Cell* **82**, 599 (1995).
51. G. P. Copenhaver *et al.*, *Science* **286**, 2468 (1999).
52. J. A. Bailey, A. M. Yavor, H. F. Massa, B. J. Trask, E. E. Eichler, *Genome Res.* **11**, 1005 (2001).
53. P. E. Warburton, *Trends Genet.* **17**, 243 (2001).
54. M. G. Schueler *et al.*, *Genomics* **66**, 104 (2000).
55. Funded by a March of Dimes Franklin Delano Roosevelt Award and by NIH grants HD32111 and HG00107 (H.F.W.). M.K.R. is supported by NIH training grant HD07518. We thank E. Eichler, B. Grimes, J. Bailey, B. Sullivan, and J. Dunn for helpful discussions and assistance. H.F.W. particularly thanks J. Waye and G. Dover for stimulating thoughts regarding satellite DNA structure, function, and evolution.

3 August 2001; accepted 10 September 2001

Replication Dynamics of the Yeast Genome

M. K. Raghuraman,[1]*† Elizabeth A. Winzeler,[3]* David Collingwood,[2]* Sonia Hunt,[1] Lisa Wodicka,[4]‡ Andrew Conway,[5] David J. Lockhart,[4]§ Ronald W. Davis,[6] Bonita J. Brewer,[1] Walton L. Fangman[1]

Oligonucleotide microarrays were used to map the detailed topography of chromosome replication in the budding yeast *Saccharomyces cerevisiae*. The times of replication of thousands of sites across the genome were determined by hybridizing replicated and unreplicated DNAs, isolated at different times in S phase, to the microarrays. Origin activations take place continuously throughout S phase but with most firings near mid–S phase. Rates of replication fork movement vary greatly from region to region in the genome. The two ends of each of the 16 chromosomes are highly correlated in their times of replication. This microarray approach is readily applicable to other organisms, including humans.

The replication of eukaryotic chromosomes is highly regulated. Replication is limited to the S phase of the cell cycle; and within S phase, initiation of replication is controlled with respect to both location and time. The sites of initiation, called replication origins, have been best characterized in the budding yeast *Saccharomyces cerevisiae*, in which a functional assay based on plasmid maintenance has allowed identification of potential origins of replication [autonomous replication sequence elements (ARSs)]. There are estimated to be ~200 to 400 ARSs in the yeast genome (*1*, *2*), and most, but not all, function as chromosomal origins (*3*). The few origins investigated at the sequence level usually encompass ~200 base pairs (bp); most contain a perfect match or a one-base mismatch to an 11-bp ARS consensus sequence (ACS) (*4*, *5*). However, the presence of an ACS is not sufficient to predict an origin of replication: There are many more ARS consensus sequences in the genome than origins

[1]Department of Genetics, [2]Department of Mathematics, University of Washington, Seattle, WA 98195, USA. [3]Genomics Institute of the Novartis Research Foundation, 3115 Merryfield Row, San Diego, CA 92121, USA. [4]Affymetrix, 3380 Central Expressway, Santa Clara, CA, USA. [5]Silicon Genetics, 935 Washington Street, San Carlos, CA 94070, USA. [6]Department of Biochemistry, Stanford University School of Medicine, Stanford, CA 94305, USA.

*These authors contributed equally to this work.
†To whom correspondence should be addressed. E-mail: raghu@u.washington.edu
‡Present address: Aventa Biosciences, 4757 Nexus Centre Drive, Suite 200, San Diego, CA 92121, USA.
§Present address: The Salk Institute for Biological Studies, Laboratory of Genetics, 10010 North Torrey Pines Road, La Jolla, CA 92037, USA.