American Genetic Association

Journal of Heredity, 2017, 45–52
doi:10.1093/jhered/esw039
Symposium Article
Advance Access publication July 16, 2016

OXFORD

Symposium Article

Chromosome-Specific Centromere Sequences Provide an Estimate of the Ancestral Chromosome 2 Fusion Event in Hominin Genomes

Karen H. Miga

From the Center for Biomolecular Science and Engineering, University of California, 501 Engineering 2 Building, UC Santa Cruz, 1156 High Street, Santa Cruz, CA 95064.

Address correspondence to K. H. Miga at the address above, or e-mail: khmiga@soe.ucsc.edu.

Received March 2, 2016; First decision May 7, 2016; Accepted June 20, 2016.

Corresponding editor: Catherine L. Peichel

Abstract

Human chromosome 2 is a product of a telomere fusion of two ancestral chromosomes and loss/degeneration of one of the two original centromeres. Genomic signatures of this event are limited to inverted telomeric repeats at the precise site of chromosomal fusion and to the small amount of relic centromeric sequences that remain on 2q21.2. Unlike the site of fusion, which is enriched for sequences that are shared elsewhere in the human genome, the region of the nonfunctioning and degenerate ancestral centromere appears to share limited similarity with other sites in the human genome, thereby providing an opportunity to study this genomic arrangement in short, fragmented ancient DNA genomic datasets. Here, chromosome-assigned satellite DNAs are used to study shared centromere sequence organization in Denisovan and Neandertal genomes. By doing so, one is able to provide evidence for the presence of both active and degenerate centromeric satellite profiles on chromosome 2 in these archaic genomes, supporting the hypothesis that the chromosomal fusion event took place prior to our last common ancestor with Denisovan and Neandertal hominins and presenting a genomic reference for predicting karyotype in ancient genomic datasets.

Subject areas: Genomics and gene mapping
Key words: centromere, chromosome fusion, hominin, repeats, satellite DNA

Introduction

Human chromosome 2 results from an end-to-end fusion of two ancestral primate chromosomes, determined by comparative cytogenetic and genomic-based studies to have occurred within the human evolutionary lineage (Yunis and Prakash 1982; IJdo et al. 1991; Avarello et al. 1992). Genomic evidence of this chromosome fusion is supported by the presence of inverted telomeric repeats at the precise site of telomere–telomere fusion, as well as a block of degenerate centromeric satellite sequences on the long arm of chromosome 2 that marks the region of the relic, ancestral centromere (IJdo et al. 1991; Avarello et al. 1992). Both loci are defined by the prevalence of repeats, which confound standard mapping approaches aimed to investigate the genomic signature of these events in short-read, ancient genomes.

Previous efforts to explore the basis for the chromosome 2 fusion event have focused exclusively on a single sequence marker representing the head-to-head joining of the telomeric hexameric repeat (GGGGTT) at the site of telomere–telomere fusion (Meyer et al. 2012). However, with short-read sequence data, it is difficult to determine if this single marker is strictly found within the proposed fusion site, as it may be found in low abundance in an additional location(s) in hominin genomes. In contrast to the single marker available at the site of telomere–telomere fusion, the relic centromeric sequences that remain on 2q21.2 provide an opportunity to identify a larger set of informative sequence markers, useful in tracking the satellite DNA organization associated with the degenerate centromere as observed in modern human genomes.

Centromeric regions in primate genomes are defined at the sequence level by long, typically multimegabase sized arrays of alpha satellite DNA, an AT-rich tandem repeat with a canonical repeat unit length of ~171 bases (Manuelidis and Wu 1978; Rosenberg et al. 1978; Willard and Waye 1987). Genomic characterization of alpha satellite DNAs benefits from the substantial sequence divergence between individual repeat units, or "monomers," measured to be ~80% when sampled genome wide (Alexandrov et al. 2001; Rudd and Willard 2004; Hayden et al. 2013). These divergent alpha satellite DNAs are commonly organized into multimonomeric repeat units, or higher order repeats (HORs), which are largely specific to a single chromosome-assigned array (Willard 1985; Willard and Waye 1987). These chromosome-specific satellite DNAs offer a useful panel of experimental markers that are commonly used in human karyotype and other genetic studies (Willard 1985).

These chromosome-assigned HOR arrays are observed to turnover rapidly over short evolutionary time, resulting in lineage-specific satellite DNA profiles when compared to closely related primates (Archidiacono 1995). For example, in modern humans, chromosome 2 contains a single centromeric HOR array (D2Z1) that has been shown by sequence hybridization studies to share limited sequence similarity with the higher order array in the chimpanzee and other great ape genomes (Haaf and Willard 1992; Warburton et al. 1996). Therefore, careful study of alpha satellite DNAs provides an opportunity to confidently assign chromosome-specific and human-specific sequence patterns of both active and degraded centromeric sites.

Here I evaluate human-specific alpha satellite sequences from the relic centromere region to examine the presence or absence of the chromosome 2 fusion in available ancient genomes (Green et al. 2010; Reich et al. 2010; Meyer et al. 2012; Prufer et al. 2014). In doing so, low-copy, locus-specific satellite markers at the site of centromere degradation on 2q21.2 in modern humans were identified that are missing from the chimpanzee genome. Using this panel, one is able to survey available hominin genomes, Denisovan (*Homo sapiens ssp. Denisova*) and Neandertal (*Homo sapiens neanderthalensis*), to predict a molecular signature that is indicative of the centromere degeneration event. Additionally, chromosome-specific centromeric satellite sequence profiles were evaluated between hominin genomes and modern human, including the satellite array specific to the active centromere on modern human chromosome 2. Although this work focuses on the sequence of the high-coverage Denisovan and Neandertal genomes (Meyer et al. 2012; Prufer et al. 2014), these tools and satellite markers should be useful in studies of lower-coverage genomes (e.g., Green et al. 2010; Reich et al. 2010) enabling similar studies of genome and karyotype evolution to be performed in other ancient genomes of varying sequence coverage.

Materials and Methods

Genomic Annotation of Relic Centromere Site in Human and Chimpanzee Assemblies

Assembled repetitive sequences at the site of the relic centromeric region on human 2q21.2 (GRCh38), as well as the p-arm and q-arm adjacent to chimpanzee (CSAC 2.1.4/panTro4; The Chimpanzee Sequencing and Analysis Consortium, 2005) centromere 2B, were characterized using prior RepeatMasker annotation, available through the UCSC Genome Table Browser (Smit et al. 1996; Karolchik et al. 2004). Full-length monomers of alpha satellite were identified by HMMER (version 3; Eddy 2009), with training using a previously published 171-bp alpha satellite consensus sequence in both orientations (Waye and Willard 1987), as previously described (Hayden et al. 2013). This study is sensitive to incorrectly parsed monomers; therefore, special attention was given to the spacing between directly adjacent repeats in an effort to monitor and correct unsupported start and end assignments. Global Needleman–Wunsch alignments between human alpha satellite monomers at 2q21.2 and chimpanzee alpha satellite DNAs found on either p-arm or q-arm flanking centromeres were performed using EMBOSS (Olson 2002) Needle software (http://www.ebi.ac.uk/Tools/psa/emboss_needle/nucleotide.html), with a gap penalty of 1 and gap extension of 0.5). Visualization of all monomer homology relationships is provided as files interpreted by the UCSC genome browser (bed files; Speir 2016).

Alpha Satellite Sequence Databases From Available Ancient Hominin Genomes

Published genome sequences from two ancient hominin groups, Neandertals and Denisovans, were used that represent datasets of high coverage (estimated ~31-fold coverage based on confidently mapped data) from Denisovan (Meyer et al. 2012): NCBI Short Read Archive "SRX103808" and ~41x coverage Neandertal (Prufer et al. 2014): European Nucleotide Archive "ERP002097" and low-coverage genome sequencing (1.9-fold genomic coverage from the Denisovan phalanx; Reich et al. 2010; European Bioinformatics Institute under STUDY accession "ERP000119"); and a total of 1.3-fold derived from three Croatian Neandertals (Green et al. 2010; European Nucleotide Archive "ERP000318").

Alpha satellite sequence libraries were identified for each genomic dataset in two steps: first, using published alignments (*bam files) released with each sequencing project, sequences overlapping known alpha satellite annotation in the human assembly (GRCh37 or hg19) were collected (Karolchik et al. 2004; Speir et al. 2016). Secondly, remaining unmapped sequences were screened and identified as alpha satellite if they contained an exact match with a nonredundant 18-mer obtained from quality-masked (phred score 30) whole genome shotgun (WGS) reads from the CSAC Pan_troglodytes-2.1.4 (GCA_000001515.4) genome assembly (The Chimpanzee Sequencing and Analysis Consortium, 2005) and human genomes (HuRef; Levy et al. 2007; Hayden et al. 2013). The shortest exact seed match (18 bases) that could confidently predict alpha satellite DNAs from AT-rich matched control set (with an empirical P value of < 0.01) was identified. Chimpanzee and human alpha satellite sequence libraries were reformatted as a list of all possible, nonredundant 18-mers representing both forward and reverse orientations that were not observed in any sequence read outside of those containing alpha satellite. The resulting "seeding 18-mers" are expected to represent all known alpha satellite sequence variation in human and chimpanzee genomes. In total, 13 658 616 unique alpha 18-mers were used, of which 4 624 047 (33.85%) are specific to the human genome and 5 340 975 (39.10%) are specific to the chimpanzee genome, with 3 693 594 (27.05%) identified in both human and chimpanzee genomes. Unmapped sequences from the ancient hominin genomes were considered alpha satellite sequences if they contained an exact match to one or more in the human/chimpanzee 18-mer seeding library.

Resulting alpha satellite sequence libraries from ancient hominin genomes were reformatted into 24-mers with corresponding genomic frequency (the observed number of exact matches for any given 24-mer relative to the total number of possible 24-mers in the given genome) to be used in comparative analysis with degraded centromere 24-mer panels and alpha satellite HOR array–specific 24-mers orientation (JELLYFISH, v2.0; Marçais and Kingsford 2011).

Sequence Panel for Centromere Relic Region on 2q21.2 in Modern Humans

To generate a complete list of candidate sequences, human-specific alpha satellite DNAs in the relic centromere region on 2q21.2 were converted into a list of nonredundant *k*-mers, where *k* = 24 bases, in both forward and reverse orientations (JELLYFISH, v2.0; Marçais and Kingsford 2011). Exact matches in the assembled human genome, outside of the expected alignments from the degraded centromere region, were determined and filtered out using a Burrows–Wheeler Aligner (BWA) (Li and Durbin 2009). To test if paralogous sequences were present in the human genome, yet missing from the current human genome assembly, I screened through 10 human genomes (30× coverage) representing diverse populations (Meyer et al. 2012; NCBI Short Read Archive "SRX103808"). The corresponding read depth of this region was compared to four single-copy sites in the human genome (Supplementary Figure 1): 1) *CTCF* locus; chr16:67562407-67639183 (76,776 bp); 2) *BRCA1* locus; chr17:43044295-43125370 (81,705 bp); 3) *CENPA* locus; chr2:26747310-26833279 (85,969 bp); and 4) *XIST* locus; chrX:73820656-73852753 (32,097 bp). Control regions were converted to *k*-mer libraries, where *k* = 24 bases, in both forward and reverse orientations (JELLYFISH, v2.0; Marçais and Kingsford 2011). For each control *k*-mer, a genomic frequency estimate was determined, that is, the *k*-mer count divided by the total number of possible *k*-mers in a given genome. Frequencies were used to generate a distribution of *k*-mer frequencies to guide single-copy estimates and predictions. 24-mers from the centromere 2 relic region were included in the study if they were within 2 standard deviations (*SD*s) from the mean value for the single-copy frequency distribution. Finally, any 24-mer was removed that had an exact match within a panel of *Pan troglodytes troglodytes* genomes (Sequence Read Archive Project ID: "SRP018689", SRX243495, and SRX243496; Prado-Martinez et al. 2013).

Chromosome-Specific Markers for Centromeric Satellite Arrays in Modern Humans

A list of nonredundant *k*-mers, where *k* = 24 bases, in both forward and reverse orientations (JELLYFISH, v2.0; Marçais and Kingsford 2011), were generated using 29 published HOR consensus sequences with previous experimental evidence to support chromosome-specific assignment in modern human genomes (Alexandrov et al. 2001; Supplementary Table 1). Redundant 24-mers were eliminated if they were present in more than one HOR consensus list. Additionally, previously published HuRef WGS alpha satellite–containing read libraries were aligned to GRCh38 centromeric reference models to generate individual read libraries for each HOR array (Levy et al. 2007; Hayden et al. 2013; Miga et al. 2014; Rosenbloom et al. 2015). HOR array read libraries were reformatted into 24-mers in a similar manner to the initial consensus sequence. Consensus 24-mers were considered nonredundant (that is, specific to a given array) when they were only observed within the specified consensus sequence, and were only observed within 24-mer listing generated from the aligned HuRef WGS reads that map specifically to the corresponding reference model array in GRCh38. GenBank accessions for consensus sequences and corresponding reference models in GRCh38 are available in Supplementary Table 1. For each dataset—from the Neandertal, Denisovan, sample modern human (HuRef) genome, or chimpanzee genome—the number of reads assigned to each grouping based on at least one exact match to our HOR-specific 24-mer panel was counted. Frequencies for each HOR-assigned grouping were calculated by dividing the number of reads assigned to a single HOR array with the total number of HOR reads for each dataset. Pearson correlation values (r^2 values) were calculated for each pairwise listing of HOR read frequencies.

Results

Genomic Characterization of Ancestral Centromere on 2q21.2

The relic centromere site on human 2q21.2 (GRCh38 chr2:132208802-132250410) is a ~41-kb region enriched with degenerate alpha satellite DNAs. The placement of the ancestral centromere is parsimonious with syntenic gene order directly adjacent to the chimpanzee centromere location on chromosome 2B (CSAC 2.1.4/panTro4; The Chimpanzee Sequencing and Analysis Consortium 2005): where ANKRD30BL (NR_027019) is present on panTro4 chr2B p-arm (chr2B:132,773,774-132,787,642; chr2:132,147,907-132,161,955) and ZNF806 (NM_001304449.1) on the q-arm (chr2B:136,216,934-136,228,530; chr2:132,307,144-132,318,747). To perform a comparative genomic study of the alpha satellite sequences, pairwise alignments were generated among 184 full-length monomers present in human 2q21.2 to monomers in the flanking regions of the chimpanzee chromosome 2B p-arm (329 monomers) and q-arm (284 monomers). Ten blocks of sequence homology were observed, involving 89 monomers that shared high sequence identity (defined as greater than 95% identity), had consistent repeat orientation and relative spacing between adjacent monomers when compared with sequences on either the p-arm or q-arm adjacent to the centromere assigned gap in the chimpanzee chromosome 2B assembly (as illustrated in Figure 1). Monomers that are similar to alpha satellite on the chimpanzee p-arm are organized in a forward orientation, whereas monomers that share high identity with sequences on the q-arm are observed in the reverse orientation. The 224 bases in the human genome map at the site of the potential rearrangement, that is the sequence that separates monomers homologous to the p- and q-arm, provides a split, or partial sequence alignment to both sides of the centromere in the chimpanzee genome (p-arm chr2B:132879629-132879674, 45 bp, 91.6% and q-arm chr2B:136146263-136146412, 150 bp, 98.7%).

In addition to alpha satellite sequences, two interspersed repeats were characterized within the centromere relic region: a LINE element insertion (L1PA3, 5957 bp) and a SVA element insertion (SVA-E, 2571 bp). The L1PA3/LINE element is present in the chimpanzee genome, mapping to the q-arm directly adjacent to centromere 2B, (with 97.3% sequence identity when aligning the complete L1PA3 element and 1 kb of flanking DNA at 5′ and 3′ junction). The SVA-3 repeat is human specific, with date estimates of ~3.46 MYA (Wang et al. 2005), which contributes to the molecular signature of the degenerate centromere in modern humans. This repeat also adds to a panel of informative sequence markers within the centromeric relic region that are available to query the organization in ancient genomes.

Short Sequence Markers in Human 2q21.2 Predict the Presence of Degraded Ancestral Centromere in Ancient Hominin Genomes

To test if ancient hominin genomes have sequences consistent with the presence of the relic centromere, a panel of informative markers

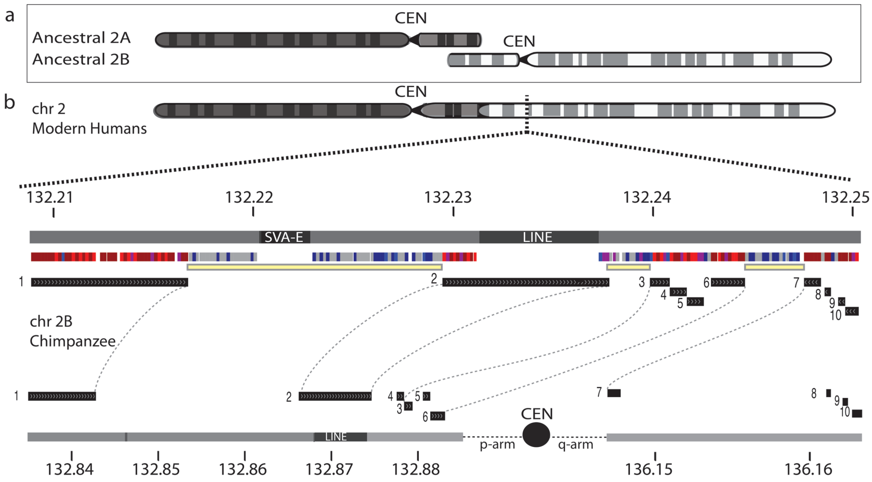

Figure 1. Genomic characterization of relic centromere region on chromosome 2q21.2 in modern humans. Human chromosome 2 is shown in (**a**) as a product of a telomere fusion of two ancestral chromosomes (Ancestral A and Ancestral B), each with their own functioning or active centromeres (labeled CEN). Following the chromosome fusion event, chromosome 2 in modern humans (**b**) has retained one active centromere region, with the other site containing ~41 kb (GRCh38 chr2:132208802-132250410) of degraded alpha satellite sequences due to the loss of one of the two original centromeres. The region is predominantly alpha satellite DNA (shown in gray track, with two interspersed transposable elements: SVA-E and LINE/L1PA3. Full-length individual alpha satellite monomers from the relic centromere region were aligned to alpha satellite in the chimpanzee genome assembly (CSAC Pan_troglodytes-2.1.4 (GCA_000001515.4); panTro4). Monomers are labeled based on global alignment sequence identity, where: ≥98% is dark red, ≥96% is light red, ≥94% is purple, ≥92% is blue; ≥90% is dark blue; and <90% is shown in grey. Ten blocks of homology are identified based on high sequence identity in monomer alignments (with a threshold of 95%) and with consecutive ordering and orientation relative to the chimpanzee genome assembly. Remaining human-specific regions are highlighted in yellow with dotted lines providing borders with regions of homology in the chimpanzee genome assembly (shown directly below). Individual homology blocks are mapped to the corresponding location in the chimpanzee chromosome 2B assembly, spanning either the p-arm or the q-arm, with the location of the centromere indicated by a black circle.

were generated to degenerate the alpha satellite sequences found on 2q21.2 that are specific to modern human and absent in chimpanzee. Due to the short read lengths in the ancient genome datasets, alpha satellite sequences were reformatted into nonredundant *k*-mers (where *k* = 24 base pairs with 1-bp overlap), representing both forward and reverse strands. Any 24-mer was removed that was observed to have an exact match outside of the relic centromere region in modern human genomes. Further, to address the concern of potential paralogous copies present on other locations in the human genome that could confound our study, 24-mers were selected that had low copy number estimates or were within the genomic frequency distribution for single-copy sites in the human genome (Supplementary Figure 1) and were determined to be consistently low across 10 individual 30× coverage genomes representing individuals from diverse human populations (Figure 2a). Remaining candidate 24-mers were designated to be “human specific” if they did not have an exact match when surveyed across two full-coverage *Pan troglodytes troglodytes* (chimpanzee) genomes (SRX243495 and SRX243496). In total, using this approach, 1005 sequence markers were identified capable of evaluating short-read ancient DNA databases for shared sequences at the site of the degenerate centromere on chromosome 2 in modern humans (Supplementary Data 1).

Using this panel of markers, ancient hominin genomes (Denisovan, high [~30x; Meyer et al. 2012] and low [1.9×; Reich et al. 2010] coverage, and Neandertal, high [~42×; Prufer et al. 2014] and low [1.3×; Green et al. 2010] coverage) were surveyed for the presence and relative abundance of each 24-mer identified as “human specific.” Exact matches with 99.2% of the 24-mer sequence markers (997/1005) were detected in the high-coverage Denisovan genome and 91.4% of sequence markers (922/1005) in the high-coverage Neandertal genome that span the regions previously determined to be human specific (as indicated in Figure 2b). Marker abundance (that is, the number of times a given 24-mer is observed genome wide) was highly similar between Denisovan and a panel of modern human genomes with the same sequence coverage. Coverage estimates of Neandertal were lower than that of Denisovan, but were within an expected distribution for single-copy sites in the Neandertal genome (Supplementary Figure 2). Assessment of low-coverage Denisovan and Neandertal genomes provided consistent results, but with lower support, as expected due to the reduced sequence alignments in the region (Supplementary Figure 3). In addition to alpha satellite, sequence support for junctions between alpha satellite and the two transposable elements in the region, SVA-E and LINE/LIPA, were documented (Supplementary Figure 4). In summary, the observation of the 2q21.2 24-mer sequence markers in all hominin genomic datasets is consistent with the hypothesis that the chromosomal fusion event occurred prior to the last common ancestor with Denisovan and Neandertal.

Hominin Centromeric Satellite Profiles Are Observed in Similar Proportions to Modern Humans But Distinct From Chimpanzee

To evaluate whether ancient hominin centromeric satellite patterns for many of the large, chromosome-assigned HOR arrays were similar to modern humans, the presence and proportion of chromosome-specific alpha satellite subsets were assessed genome wide in both Denisovan and Neandertal. Alpha satellite sequence databases were generated for available ancient hominin genomes by identifying reads that contain at least an 18-bp exact match with comprehensive alpha satellite sequence libraries obtained from both human and chimpanzee genomes (described in *Materials and Methods*). Based on this early assessment, all three ancient hominin genomes appeared to share more sequences with modern humans than with chimpanzee, with only ~4.5% and 8% of sequences determined to

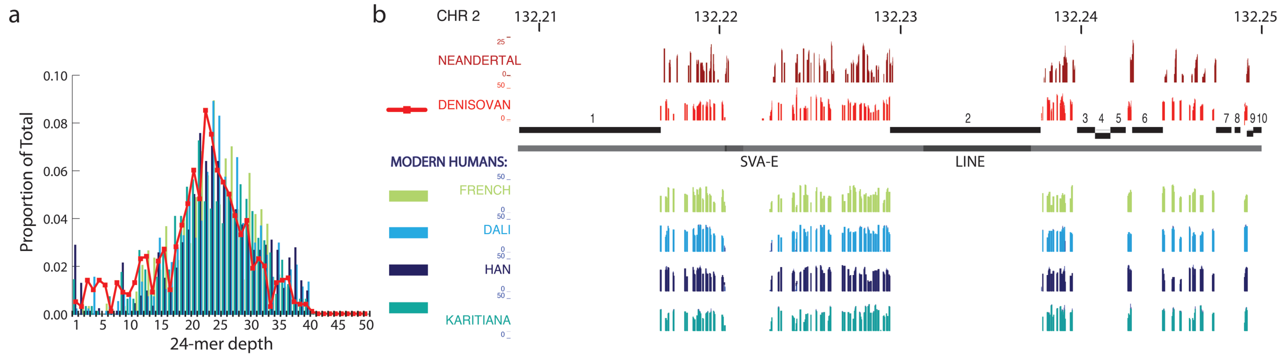

Figure 2. Evidence for the presence of the relic centromere region in Denisovan. Data analysis of 1005 alpha satellite markers determined to be human specific for centromere relic region for Denisovan and modern human genomes. Markers are expected to be effectively single, or low-copy in the genome based on the depth of coverage of each 24-mer in the genome, supported by read depth distribution (**a**) shown for modern humans: French (HGDP00521, light green), Dali (HGDP01307, light blue), Han (HGDP00778, dark blue), and Karitiana (HGDP00998, teal). The distribution for the Denisovan genome (shown in red) supports that the sequence markers provide low-copy data. The genome coordinates for each 24-mer marker are provided in (**b**) for the entire relic centromere region on human 2q21.2 (GRCh38 chr2:132208802-132250410). Regions of known homology with the chimpanzee genome assembly (panTro4; labeled 1–10) are indicated as regions omitted from our study. Repeat masker data for this region are used to display alpha satellite in gray, with location of SVA-E and LINE/LIPA3 locations indicated. Four modern human genomes are displayed as a histogram (min-max range: 0–40), to demonstrate the relative placement and observed abundance. The Denisovan data (shown in red) are found throughout the relic centromere region and are similar in sequence abundance to that of modern humans.

be chimpanzee specific in Denisovan and Neandertal, respectively; the remaining >90% of sequences support were human-specific sequences and/or sequences shared between human and chimpanzee genomes (Supplementary Figure 5).

HOR arrays in modern humans differ in sequence abundance and general chromosomal localization from those in the chimpanzee genome (Archidiacono et al. 1995). Therefore, to investigate if the presence and relative abundance of chromosome-assigned alpha satellite arrays are similar between human and ancient hominin genomes, a panel of 6293 informative 24-mers for 22 previously published centromeric arrays were designed in modern human genomes (Supplementary Table 1 and Supplementary Data 2; Alexandrov et al. 2001). Using these markers, nonoverlapping sets of sequence reads were identified that could be assigned to a given HOR array, defined as having at least one exact match to an array-specific marker from among the 22 published sequences.

Initial assessment of these markers using 10 full-coverage modern human genomes provided highly concordant read databases between individuals (with average r^2 value of 0.93, Supplementary Table 2, Figure 3a), with an average read database size of 214 Mb per genome. In contrast, using the same panel of markers on two ~20–30× chimpanzee genomes, ~100× fewer bases (with 2.38 Mb on average in chimpanzee) were identified. The sequence databases created in the chimpanzee genome, however, provided consistent results, with an r^2 value of 0.99 (Supplementary Table 3, Figure 3b). This supports previous observations that chimpanzee and human alpha satellite monomers share similar evolutionary histories and satellite DNA subfamily assignments (Alexandrov et al. 2001), but have very different abundance and chromosome assignments genome wide (Archidiacono et al. 1995). Correlation between human and chimpanzee sequence libraries provided a lower correlation average r^2 value of 0.70 (Supplementary Table 3).

Similar to modern human genomes, when characterizing Denisovan HOR alpha satellite read data, relatively the same number of bases (211 Mb) was assigned and provided an average r^2 value of 0.90 compared to 10 modern human genomes and an average r^2 value of 0.61 compared to the chimpanzee genomes (Supplementary Table 4, Figure 3c,d). The high-coverage Neandertal genome, however, provided only an average r^2 value of 81% and roughly 5× fewer bases (39.6 Mb; Supplementary Table 4). A similar trend was observed when surveying the low-coverage Neandertal genome. Notably, even this low-coverage database containing 0.28 Mb provided an average r^2 value of 0.91 when compared to 10 modern human genomes and an average r^2 value of 0.67 when compared to the chimpanzee genomes (Supplementary Table 4).

The active centromere present on modern human chromosome 2 is defined by a chromosome-specific alpha satellite HOR (8-mer, D2Z1), which is organized into a homogenized, multimegabase array that emerged after the last common ancestor with chimpanzee (Haaf and Willard 1992; Warburton et al. 1996). To study the evolution of this particular array in more detail, 2830 Denisovan read alignments were determined to contain at least one D2Z1-informative 24-mer to the HOR consensus sequence. Average percent identity of D2Z1 read alignments was observed to be 96.98%, and alignments spanned the entire HOR repeat unit.

Discussion

Efforts to study karyotype evolution in ancient hominin genomic datasets are challenged by the inability to generate end-to-end chromosomal assemblies. Of particular interest, it remains unknown when the end-to-end fusion of two ancestral primate chromosomes took place in the human lineage and how this chromosomal evolution influenced gene flow between ancient hominin populations. Previous genomic signatures of this event have been limited to inverted telomeric repeats at the precise site of chromosomal fusion and to the 41-kb relic centromeric region in modern human chromosomes 2q21.2 (IJdo et al. 1991; Avarello et al. 1992). Here I demonstrate the use of chromosome-assigned centromeric satellite markers to predict karyotype in ancient hominin genomic datasets. To do so, I identified a panel of low-copy, locus-specific satellite markers at the site of the relic centromere on human 2q21.2 that are missing from the chimpanzee genome. These data strengthen previous studies that focused on a single marker (telomeric hexameric repeat: "GGGTT") at the site of telomere–telomere fusion (Meyer et al. 2012). One key advantage of this work is that in contrast to a single sequence, the data presented here use a comprehensive panel of ~1000 informative markers to provide evidence for the presence of the relic centromere

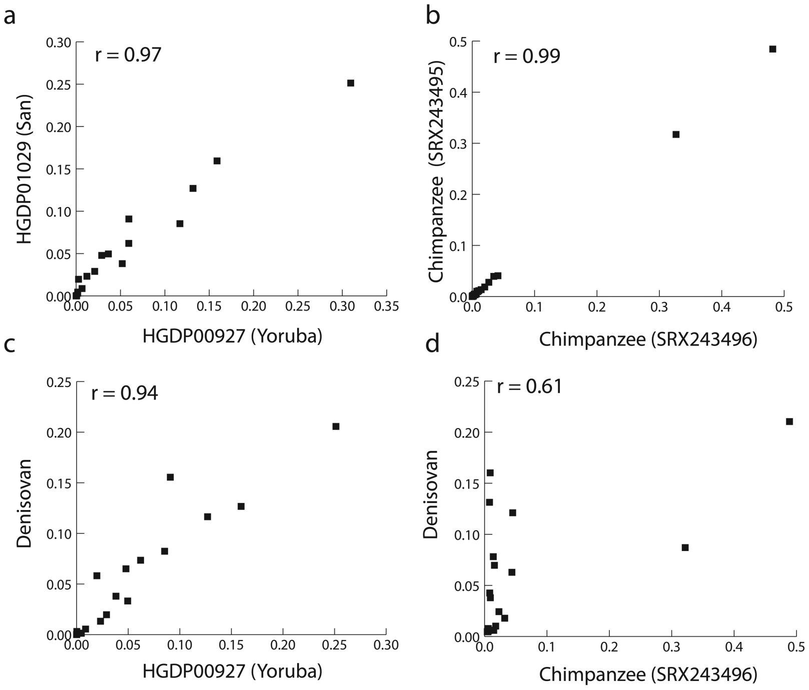

Figure 3. Ancient hominins share similar chromosome-assigned HOR satellite DNAs with modern humans. Previously characterized HOR panel was used to identify the presence and relative abundance chromosome-assigned satellite sequences in modern human, Denisovan, and chimpanzee genomes. For each genome, sequence-read libraries were identified for each of the 22 HOR repeats, and the relative proportion, or the number of reads assigned to a particular chromosome-specific HOR normalized by the total number of reads, was determined for each satellite family per genome. (**a**) Pearson correlations between the relative proportion of the HOR sequences in modern humans genomes are high, with an r^2 value of 0.97 observed between a San (HGDP01029) and Yoruba (HGDP00927) individual genome. Similarly, in (**b**), one observes a high Pearson correlation between two chimpanzee genomes (SRX243495 and SRX243496), with an r^2 value of 0.99. The Denisovan genome is more highly correlated with modern human HOR satellite DNA proportions (**c**) r^2 value of 0.94, shown here for Yoruba, HGDP00927 than chimpanzee (**d**) r^2 value of 0.61, shown here for SRX243496.

region in ancient hominin genomes. By increasing the number of markers, one is able to track sequences from the relic centromere even in low-coverage, short sequence-read genomic datasets.

However, it is important to note that this interpretation assumes a model in which the centromere sequence rearrangement—resulting in a human-specific genomic signature of satellite DNAs after the fusion—followed the loss of function at the ancestral centromere. In an alternative hypothesis, the human-specific relic centromeric sequences could have been shared with the last common ancestor with chimpanzee, yet lost in the extant chimpanzee genomes. Although it is not possible to rule out this alternative hypothesis, the prediction of satellite DNA instability and loss after a fusion event is consistent with studies of dicentric fusions in cancer (Berger and Busson-Le Coniat 1999; Stimpson et al. 2012), of karyotype stability in plant populations (Salse 2012; Lysak 2014), and of loss of centromere function associated with the depletion of heterochromatic chromatin, which results in aberrant recombination and tandem repeat instability (Peng and Karpen 2007).

In addition to the alpha satellite sequences at the relic centromere site, the predominant satellite array (D2Z1) associated with the active centromere on human chromosome 2 is distinct from the HOR array on the syntenic chimpanzee chromosome 2A centromere (Haaf and Willard 1992; Warburton et al. 1996). Using 167 D2Z1-specific markers, I provide evidence for the presence of a similar centromere 2 array in ancient hominin genomes, with high percent identity to the HOR repeat sequence and at the genome-wide frequency expected from a panel of modern humans. Additionally, using markers designed to identify individual HORs specific to individual human chromosomes, it appears that Denisovan have similar proportion of the chromosomal HOR satellite DNAs (Willard and Waye 1987; Alexandrov et al. 2001). Although it is possible that the satellite arrays are rearranged to different locations in the ancient hominin genomes, the more parsimonious interpretation would be that centromere sequences in ancient hominins are similar to those of modern humans genome wide. Notably, only a small fraction of ancient hominin alpha satellite sequences have not been observed in modern humans. Rather than presenting as novel centromeric multimegabase sized arrays, these chimpanzee-like sequences appeared to have a low abundance in the genome (data not shown) and are predicted to represent minor alpha satellite arrays (read depth estimates of ~5–50 kb). It is likely that the ancient sequence variants associated with these novel arrays were lost in the human lineage, or alternatively, these sequences may not yet be fixed in the human population, and thus may not be present in the small number of individuals included in our current study.

Here I provide evidence that Denisovan and Neandertal genomes share similar centromere sequence profiles with modern humans, thus demonstrating that the chromosome 2 fusion event took place before the last common ancestor with Denisovan and Neandertal. Previous studies have estimated that Denisovan individuals diverged from modern Africans ~0.8 million years ago, assuming 6.5 million years for the human–chimpanzee divergence (Reich et al. 2010). The timing of the fusion event is supported by estimates that the fusion event took place between ~0.74–3 MYA, using fixed substitutions that demonstrate extreme AT to GC bias in the human lineage since divergence from the common ancestor with the chimpanzee

(Dreszer et al. 2007). It remains unclear, however, if the fusion event was fixed within ancient hominin populations at that time, or if ancestral chromosomes 2A and 2B were still prevalent within ancient hominin populations. The insertion of a human lineage–specific SVA element (SVA-E subfamily) at the relic centromere region provides an estimate of the timing of the event to ~3.5 MYA (with a range of ~2.5–4.5 MYA; Wang et al. 2005).

This timing estimate is also consistent with the observation that both Denisovan and Neandertal share similar chromosome-specific HOR satellite profiles with modern humans. It is likely that the satellite DNA similarity between hominins and modern humans reflects the limited accumulation of nucleotide divergence and/or satellite array turnover that have occurred over a relatively short evolutionary timescale. However, satellite DNAs are expected to evolve concertedly through a process known as molecular drive, by which mutations are homogenized in a genome and ultimately fixed in an interbreeding population at an increased rate (Dover 1989), which leads to highly divergent satellite DNAs copy number and nucleotide sequence that are species specific (Ugarković and Plohl 2002). Although the timing of satellite homogenization through molecular drive remains unknown, the finding that Denisovan and Neandertal genomes share similar genome-wide satellite sequence profiles with modern humans, including satellites specific to the chromosome 2 centromere, could suggest the lack of interference in gene flow by differences in karyotype.

Efforts to further explore centromere satellite evolution and karyotype estimates in the human lineage may benefit from additional satellite-based genome-wide analyses of additional ancient hominin genomic datasets, using the data and approaches presented in this study.

Supplementary Data

Supplementary material can be found at http://www.jhered.oxfordjournals.org/.

Acknowledgments

I gratefully acknowledge Dr HF Willard and Dr RE Green for helpful evaluation of this manuscript. Early stages of the study were initiated with the support and mentorship of Dr HF Willard.

References

Alexandrov I, Kazakov A, Tumeneva I, Shepelev V, Yurov Y. 2001. Alpha-satellite DNA of primates: old and new families. *Chromosoma*. 110:253–266.

Archidiacono N, Antonacci R, Marzella R, Finelli P, Lonoce A, Rocchi M. 1995. Comparative mapping of human alphoid sequences in great apes using fluorescence in situ hybridization. *Genomics*. 25:477–484.

Avarello R, Pedicini A, Caiulo A, Zuffardi O, Fraccaro M. 1992. Evidence for an ancestral alphoid domain on the long arm of human chromosome 2. *Hum Genet*. 89:247–249.

Berger R, Busson-Le Coniat M. 1999. Centric and pericentric chromosome rearrangements in hematopoietic malignancies. *Leukemia*. 13:671–678.

The Chimpanzee Sequencing and Analysis Consortium. 2005. Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature*. 437:69–87.

Dover GA. 1989. Linkage disequilibrium and molecular drive in the rDNA gene family. *Genetics*. 122:249–252.

Dreszer TR, Wall GD, Haussler D, Pollard KS. 2007. Biased clustered substitutions in the human genome: the footprints of male-driven biased gene conversion. *Genome Res*. 17:1420–1430.

Eddy SR. 2009. A new generation of homology search tools based on probabilistic inference. *Genome Inform*. 23:205–211.

Green RE, Krause J, Briggs AW, Maricic T, Stenzel U, Kircher M, Patterson N, Li H, Zhai W, Fritz MH-Y, *et al*. 2010. A draft sequence of the Neandertal genome. *Science*. 328, 710–722.

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

Hayden KE, Strome ED, Merrett SL, Lee HR, Rudd MK, Willard HF. 2013. Sequences associated with centromere competency in the human genome. *Mol Cell Biol*. 33:763–772.

IJdo JW, Baldini A, Ward DC, Reeders ST, Wells RA. 1991. Origin of human chromosome 2: an ancestral telomere-telomere fusion. *Proc Natl Acad Sci USA*. 88:9051–9055.

Karolchik D, Hinrichs AS, Furey TS, Roskin KM, Sugnet CW, Haussler D, Kent WJ. 2004. The UCSC Table Browser data retrieval tool. *Nucleic Acids Res*. 32:D493–D496.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics*. 25:1754–1760.

Levy S, Sutton G, Ng PC, Feuk L, Halpern AL, Walenz BP, Axelrod N, Huang J, Kirkness EF, Denisov G, *et al*. 2007. The diploid genome sequence of an individual human. *PLoS Biol*. 5:e254.

Lysak MA. 2014. Live and let die: centromere loss during evolution of plant chromosomes. *New Phytologist*. 203:1082–1089.

Manuelidis L, Wu JC. 1978. Homology between human and simian repeated DNA. *Nature*. 276:92–94.

Marçais G, Kingsford C. 2011. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics*. 27:764–770.

Meyer M, Kircher M, Gansauge MT, Li H, Racimo F, Mallick S, Schraiber JG, Jay F, Prüfer K, de Filippo C, *et al*. 2012. A high-coverage genome sequence from an archaic Denisovan individual. *Science*. 338:222–226.

Miga KH, Newton Y, Jain M, Altemose N, Willard HF, Kent WJ. 2014. Centromere reference models for human chromosomes X and Y satellite arrays. *Genome Res*. 24:697–707.

Olson SA. 2002. EMBOSS opens up sequence analysis. European Molecular Biology Open Software Suite. *Brief Bioinform*. 3:87–91.

Peng JC, Karpen GH. 2007. H3K9 methylation and RNA interference regulate nucleolar organization and repeated DNA stability. *Nat Cell Biol*. 9:25–35.

Prado-Martinez J, Sudmant PH, Kidd JM, Li H, Kelley JL, Lorente-Galdos B, Veeramah KR, Woerner AE, O'Connor TD, Santpere G, *et al*. 2013. Great ape genetic diversity and population history. *Nature*. 499:471–475.

Prufer K, Racimo F, Patterson N, Jay F, Sankararaman S, Sawyer S, Heinze A, Renaud G, Sudmant PH, de Filippo C, *et al*. 2014. The complete genome sequence of a Neanderthal from the Altai Mountains. *Nature*. 505:43–49.

Reich D, Green RE, Kircher M, Krause J, Patterson N, Durand EY, Viola B, Briggs AW, Stenzel U, Johnson PLF, *et al*. 2010. Genetic history of an archaic hominin group from Denisova Cave in Siberia. *Nature*. 468:1053–1060.

Rosenberg H, Singer M, Rosenberg M. 1978. Highly reiterated sequences of SIMIANSIMIANSIMIANSIMIANSIMIAN. *Science*. 200:394–402.

Rosenbloom KR, Armstrong J, Barber GP, Casper J, Clawson H, Diekhans M, Dreszer TR, Fujita PA, Guruvadoo L, Haeussler M, *et al*. 2015. The UCSC Genome Browser database: 2015 update. *Nucleic Acids Res*. 43:D670–D681.

Rudd MK, Willard HF. 2004. Analysis of the centromeric regions of the human genome assembly. *Trends Genet*. 20:529–533.

Salse J. 2012. In silico archeogenomics unveils modern plant genome organisation, regulation and evolution. *Curr Opin Plant Biol*. 15:122–130.

Smit AFA, Hubley R, Green P. 1996–2010. RepeatMasker Open-3.0. Available from http://wwwrepeatmaskerorg

Speir ML, Zweig AS, Rosenbloom KR, Raney BJ, Paten B, Nejad P, Lee BT, Learned K, Karolchik D, Hinrichs AS, *et al*. 2016. The UCSC Genome Browser database: 2016 update. *Nucleic Acids Res*. 44:D717–D725.

Stimpson KM, Matheny JE, Sullivan BA. 2012. Dicentric chromosomes: unique models to study centromere function and inactivation. *Chromosome Res*. 20:595–605.

Ugarković D, Plohl M. 2002. Variation in satellite DNA profiles–causes and effects. *EMBO J*. 21:5955–5959.

Wang H, Xing J, Grover D, Hedges DJ, Han K, Walker JA, Batzer MA. 2005. SVA elements: a hominid-specific retroposon family. *J Mol Biol*. 354:994–1007.

Warburton PE, Haaf T, Gosden J, Lawson D, Willard HF. 1996. Characterization of a chromosome-specific chimpanzee alpha satellite subset: evolutionary relationship to subsets on human chromosomes. *Genomics*. 33:220–228.

Waye JS, Willard HF. 1987. Nucleotide sequence heterogeneity of alpha satellite repetitive DNA: a survey of alphoid sequences from different human chromosomes. *Nucleic Acids Res*. 15:7549–7569.

Willard HF. 1985. Chromosome-specific organization of human alpha satellite DNA. *Am J Hum Genet*. 37:524–532.

Willard HF, Waye JS. 1987. Chromosome-specific subsets of human alpha satellite DNA: analysis of sequence divergence within and between chromosomal subsets and evidence for an ancestral pentameric repeat. *J Mol Evol*. 25:207–214.

Yunis JJ, Prakash O. 1982. The origin of man: a chromosomal pictorial legacy. *Science*. 215:1525–1530.