Hotspots of mutation and breakage in dog and human chromosomes

Caleb Webber[1] and Chris P. Ponting

MRC Functional Genetics Unit, Department of Human Anatomy and Genetics, University of Oxford, Oxford OX1 3QX, United Kingdom

Sequencing of the dog genome allows an investigation of the location-dependent evolutionary processes that occurred since the common ancestor of primates and carnivores, ~95 million years ago. We investigated variations in G+C nucleotide fraction and synonymous nucleotide substitution rates (K_s) across dog and human genomes. Our results show that dog genes located either in subtelomeric and pericentromeric regions, or in short synteny blocks, possess significantly elevated G+C fraction and K_s values. Human subtelomeric, but not pericentromeric, genes also exhibit these elevations. We then examined 1.048 Gb of human sequence that is likely not to have been located near a primate telomere at any time since the common ancestor of dog and human. We observed that regions of highest G+C or K_s ("hotspots"; median sizes of 0.5 or 1.3 Mb, respectively) within this sequence were preferentially segregated to dog subtelomeres and pericentromeres during the rearrangements that eventually gave rise to the extant canine karyotype. Our data cannot be accounted for solely on the basis of gradually elevating G+C fractions in subtelomeric regions as a consequence of biased gene conversion. Rather, we propose that high G+C sequences are found preferentially within dog subtelomeres as a direct consequence of chromosomal fission occurring more frequently within regions elevated in G+C.

[Supplemental material is available online at www.genome.org.]

Over their evolution, genome sequences accumulate small-scale nucleotide substitutions, insertions, and deletions, and larger-scale rearrangements and translocations. Each effect has acted differentially among genomes from diverse species, and between and within chromosomes of the same species (Wolfe et al. 1989; Matassi et al. 1999; Mouse Genome Sequencing Consortium 2002). Our understanding of the rates of, and correlations among, these evolutionary processes has greatly benefited from comparative analyses of human, mouse, and rat genomes (Mouse Genome Sequencing Consortium 2002; Rat Genome Sequencing Project Consortium 2004). However, mouse and rat genomes are highly derived with respect to that of the common ancestor of eutherian mammals (CAE), both in terms of their highly rearranged genomes, and in the relatively large number of nucleotide substitutions in selectively neutral sites they have accumulated. In contrast, the karyotype of the CAE ($2n = 46$) has since been altered by only 12 rearrangements and one reciprocal translocation along the primate lineage to humans (Wienberg 2004). Thus, compared with the derived karyotypes of murid rodents, the human karyotype is relatively ancestral.

Chromosomal rearrangements, such as inversions and translocations, require double-stranded breaks. These were first supposed to occur randomly in chromosomes (Nadeau and Taylor 1984), although examinations of conserved synteny blocks in human and mouse chromosomes now indicate the presence of fragile regions that possess high propensities for breakage (Pevzner and Tesler 2003). Although others have challenged this conclusion (Trinh et al. 2004), the existence of fragile sites is consistent with observations that non-B DNA in human chromosomes is particularly susceptible to deletion events in human disease (Bacolla et al. 2004).

Nucleotide substitution rates at neutral sites are greatly affected by their sequences' CpG content, since methylated cytosine in a CpG dinucleotide is hypermutable (Cooper and Youssoufian 1988; Sved and Bird 1990). These rates are also strongly correlated with the G+C fraction (Mouse Genome Sequencing Consortium 2002; Hardison et al. 2003; Fryxell and Moon 2004). The G+C fraction of isochores, long (>300 kb) regions of relatively homogeneous base composition (Filipski et al. 1973; Bernardi et al. 1985), appears to have declined significantly prior to the appearance of the first boreoeutherian animal (Belle et al. 2004), the earliest common ancestor of all primates, rodents, and carnivores (Springer et al. 2003). However, isochore G+C values in primate and carnivore lineages appear to have declined only slightly thereafter (Belle et al. 2004).

In primate and rodent genomes, local rates of recombination are known to be positively correlated with both G+C fraction (Eyre-Walker 1993; Fullerton et al. 2001; International Human Genome Sequencing Consortium 2001) and neutral rates (Lercher and Hurst 2002; Mouse Genome Sequencing Consortium 2002; Hardison et al. 2003; Hellmann et al. 2003). These correlations have prompted speculation that recombination is mutagenic, and that its increase drives elevation of G+C fractions (Eyre-Walker 1993; Hardison et al. 2003; Meunier and Duret 2004), perhaps by "biased gene conversion" (BGC) (Lamb 1986; Brown and Jiricny 1987, 1988; Eyre-Walker 1993; Galtier et al. 2001; Marais 2003). In this BGC model, chromosomal locations where recombination is highest (such as human subtelomeric regions) (Kong et al. 2002) should, over time, increase their G+C contents while decreasing nucleotide substitution rates (Marais 2003).

We recently observed a hitherto unforeseen effect of subtelomeric location on nucleotide substitution rates. We com-

[1] **Corresponding author.**
E-mail caleb.webber@human-anatomy.oxford.ac.uk; fax 44-1865-282651.

Article and publication are at http://www.genome.org/cgi/doi/10.1101/gr.3896805.

pared chicken gene sequences with their human orthologs in terms of K_s, the number of silent (synonymous) nucleotide substitutions per synonymous site, an estimate of the neutral substitution rate (International Chicken Genome Sequencing Consortium 2004). We found that the average K_s value for genes on the small avian chromosomes ("microchromosomes") was significantly elevated compared with the average K_s value for genes on the larger chromosomes ("macrochromosomes"). We reasoned that this effect might have been driven by BGC, with a substitution rate increase for recombination-susceptible sequences close to telomeres. Such a rate increase has previously been observed for subtelomeric genes in *Saccharomyces cerevisiae* (Winzeler et al. 2003), although not for *Caenorhabditis* nematode genes (Stein et al. 2003). Indeed, we found elevated K_s values in regions <10 Mb from the ends of assembled macrochromosomes, to a level that was indistinguishable from K_s values obtained from genes in the chicken microchromosomes. We proposed that the microchromosomes' elevation in K_s values is a direct result of these chromosomes being deficient in genes that are located distant from telomeric ends.

The availability of the dog (*Canis familiaris*) genome sequence (Lindblad-Toh et al. 2005) now enables a fresh perspective to be gained on the correlations and causal relationships between evolutionary rates, recombination, G+C fractions, and physical location. The karyotype ($2n = 78$) of the dog, *C. familiaris*, is substantially rearranged with respect to the CAE. From chromosome painting experiments of carnivores, it appears that the high-numbered acrocentric karyotypes of extant canids ($2n = 74$–78) arose from a fragmentation of the ancestral carnivore karyotype ($2n = 42$), mostly involving pericentric inversions followed by the fission of chromosomes at their centromeres (Nash et al. 2001; Wienberg 2004).

The dog genome sequence presents an opportunity both to reconstruct evolutionary events on the lineage to dog and to infer with greater precision past variations in base composition and evolutionary rates. In particular, the derived dog karyotype allows us to test the hypothesis that sequences proximal to a telomere experience elevated nucleotide substitution rates because of higher recombination and BGC rates, thereby causing a concomitant increase in the fraction of G+C nucleotides. This hypothesis predicts that dog genes located near recently derived telomeres have accumulated considerably greater G+C content at fourfold degenerate sites than their human orthologs, which have been located far from a telomere since the last common ancestor of dog and human. Similarly, genes located in human subtelomeres would be expected to have increased their G+C fraction at fourfold degenerate sites relative to dog non-subtelomeric orthologs.

Our results show that dog genes that are located in either subtelomeric or pericentromeric regions possess elevated G+C nucleotide fractions and synonymous rates relative to dog interstitial genes. Moreover, the rank order correlation of G+C at fourfold degenerate sites of dog and human orthologs remains high despite these rearrangements. While we conclude that recombination-driven BGC has occurred in the vicinity of dog telomeres, no evidence for this was found near the more ancestral human telomeres.

We considered an alternative hypothesis, that chromosome fission during the fragmentation of the canid karyotype occurred preferentially within ancestrally high G+C regions. This model provides two testable predictions: (1) that G+C bias in chromosome breakage contributed to the observed elevations in G+C fractions within dog subtelomeres and pericentromeres; and (2) that high G+C regions suffered more numerous breakages and thus were preferentially segregated to shorter synteny blocks. Each of these predictions is supported by dog and human genome comparisons. We thus propose a high G+C "fragile breakage" model for the evolution of the canine karyotype.

Results

Data

We considered two sets of orthology relationships representing single, nonduplicated and omni-present orthologous genes in mammalian genomes. The first ortholog data set, which we refer to as D5, contains single orthologs in five species, namely, chicken, mouse, rat, human, and dog. D5 consists of data derived for the chicken genome sequencing project (International Chicken Genome Sequencing Consortium 2004) augmented by 7670 dog genes that have single orthologs in the other four species and that we predicted from the dog genome assembly (see Methods). A second ortholog data set, termed D2, consists of 13,738 single (1:1) orthologous genes in dog and human genomes, and was derived from phylogenetic analyses (L. Goodstadt and C.P. Ponting, in prep.). Data set D2 facilitated chromosomal mapping of quantities between dog and human genomes.

Fourfold degenerate sites (so-called 4D sites) at the third position of codons encoding eight amino acid types were identified as previously (Hardison et al. 2003). GC4D, the fraction of G or C bases at these sites, was calculated for D5 and D2 orthologs, as was GC53, the G+C fraction for 10 kb 5′-upstream and 3′-downstream of transcriptional start and stop sites. The physical distance of a gene to the nearest telomeric end of assembled chromosomal sequence (without spanning the centromere) was assumed to approximate well the true distance to the chromosome's telomere. Similarly, the distance to the centromere was assumed to be the minimum number of bases between the gene and centromere coordinates taken from the UCSC table browser (Karolchik et al. 2004). The neutral rate of nucleotide substitution was assumed to equal K_s, the number of synonymous substitutions per synonymous site, as estimated using codeml from Yang's PAML package (Yang 1997). Variations in GC4D, GC53, and K_s were noted for all dog (Supplemental Fig. 1) and human (Supplemental Fig. 2) autosomes. We show these variations for a representative chromosome (Chromosome 1) in each species in Figures 1 and 2.

G+C fraction and K_s

As seen for other mammalian genomes (Mouse Genome Sequencing Consortium 2002; Rat Genome Sequencing Project Consortium 2004), the G+C fraction varies greatly across the dog genome (Fig. 1). The G+C fraction averaged over all 4D sites in the dog assembly is 0.590, whereas it is slightly lower (0.570) in human. However, average G+C fractions 10 kb upstream and downstream of genes are equivalent for both species (0.450).

G+C fractions are known to correlate significantly between human and murid rodent orthologs (Mouchiroud et al. 1988; Mouse Genome Sequencing Consortium 2002). Among pairs of the five species (data set D5) investigated, we found that GC4D is most correlated between dog and human (Table 1). Surprisingly, it is marginally more correlated between these two species, which diverged ~95 million years ago (Mya) (Springer et al. 2003), than

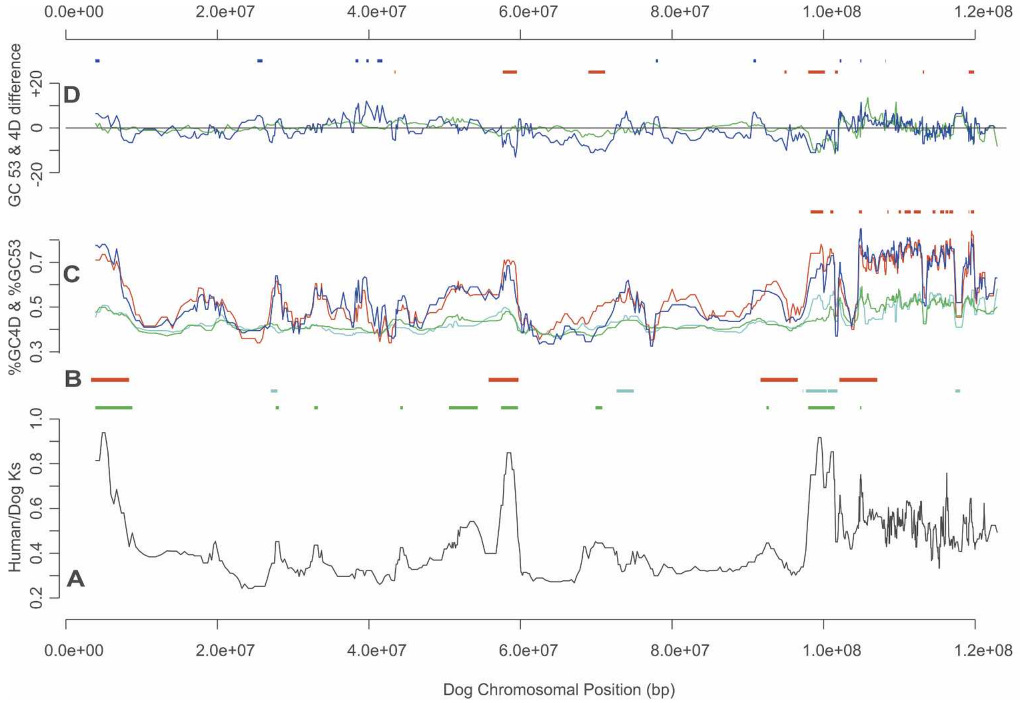

Figure 1. Variations in dog and human K_s, and different G+C fractions, as functions of distance (in base pairs) along dog Chromosome 1. These quantities are shown as median values for 10 gene overlapping windows (see Methods). (*A*) The variation in K_s values (in black) of dog and human orthologs along this chromosome. K_s value hotspots are indicated in green above *A*. (*B*) The syntenic locations (see Methods) of human telomeres (in red) on dog Chromosome 1. Short synteny blocks (<4 Mb) are indicated below in light blue. (*C*) Variations in GC53 or GC4D fractions (as percentages). %GC53 and %GC4D values of dog genes are shown in light blue and red, respectively, whereas %GC53 and %GC4D values of their human orthologs are shown in green and dark blue, respectively. %GC4D hotspots, exceeding the 80th centile for the whole chromosome (see Methods), are marked above *C* in red. (*D*) The differences in %GC53 (ΔGC53, dark blue), and in %GC4D (ΔGC4D, green), between dog and human orthologs. Above *D*, ΔGC4D hotspots (i.e., dog regions elevated in GC4D, with respect to human orthologs; see Methods) are indicated in red, whereas ΔGC4D cold spots (i.e., dog regions suppressed in GC4D, with respect to human orthologs; see Methods) are shown in blue.

between rat and mouse lineages that share a considerably more recent ancestor (~15 Mya) (Springer et al. 2003).

Median K_s values between these species' pairs reveal that, on average, fewer nucleotide substitutions have accumulated at silent sites in the lineages to dog and human (median $K_s = 0.36$), than they have to the lineages to mouse and human (median $K_s = 0.60$) (Mouse Genome Sequencing Consortium 2002), despite the carnivore lineage being an out-group to both primates and rodents. This arises because of the well-known higher substitution rates in murid rodents than in other mammals (Mouse Genome Sequencing Consortium 2002).

We further investigated the known positive correlation between neutral rates and G+C fraction (Mouse Genome Sequencing Consortium 2002; Hardison et al. 2003; Hellmann et al. 2003), among the five vertebrates. We find the ranked correlation to be greatest for dog and human orthologs (Table 2); among individual dog chromosomes, the correlation coefficient rises to a value of 0.80 (CFA6 and CFA31). Once more, the correlation is least for mouse and rat orthologs, despite these species sharing a more recent common ancestor.

Distance dependencies on G+C fraction and K_s

Previously we observed for chicken–human gene alignments (International Chicken Genome Sequencing Consortium 2004) that G+C fractions and K_s values are substantially elevated in regions proximal to the ends of assembled chromosomes ("subtelomeric regions"). In the present study, we find elevations of G+C fractions (Table 3) and of K_s values (Table 4) in the subtelomeric regions of both dog and human chromosomes. In contrast, such elevations are barely perceptible for either rat or mouse chromosomes. Elevation of G+C within human subtelomeres likely accounts for the negative correlation of G+C fraction with chromosomal size (Duret et al. 2002).

Elevation of K_s values in subtelomeric regions is most pronounced in human subtelomeres, and is least in murid rodent subtelomeres (Table 4). Because of these greatly reduced distance dependencies, rodent G+C fractions and K_s data will not be considered further.

These elevations are not maintained uniformly within subtelomeric regions. By plotting median values of GC4D and K_s for bins containing a minimum of 200 genes (data set D2), we observed that each of these quantities declines logarithmically from the ends of assembled dog or human chromosomes (Fig. 3); GC53 distributions mirrored those of GC4D (data not shown). These relationships are all statistically significant (Table 5).

Table 1. Correlation coefficients (Spearman's ρ) between 1:1 orthologs' GC4D fractions

	Human	Dog	Mouse	Chicken
Human	—			
Dog	0.945	—		
Mouse	0.836	0.818	—	
Chicken	0.539	0.595	0.539	—
Rat	0.825	0.810	0.937	0.524

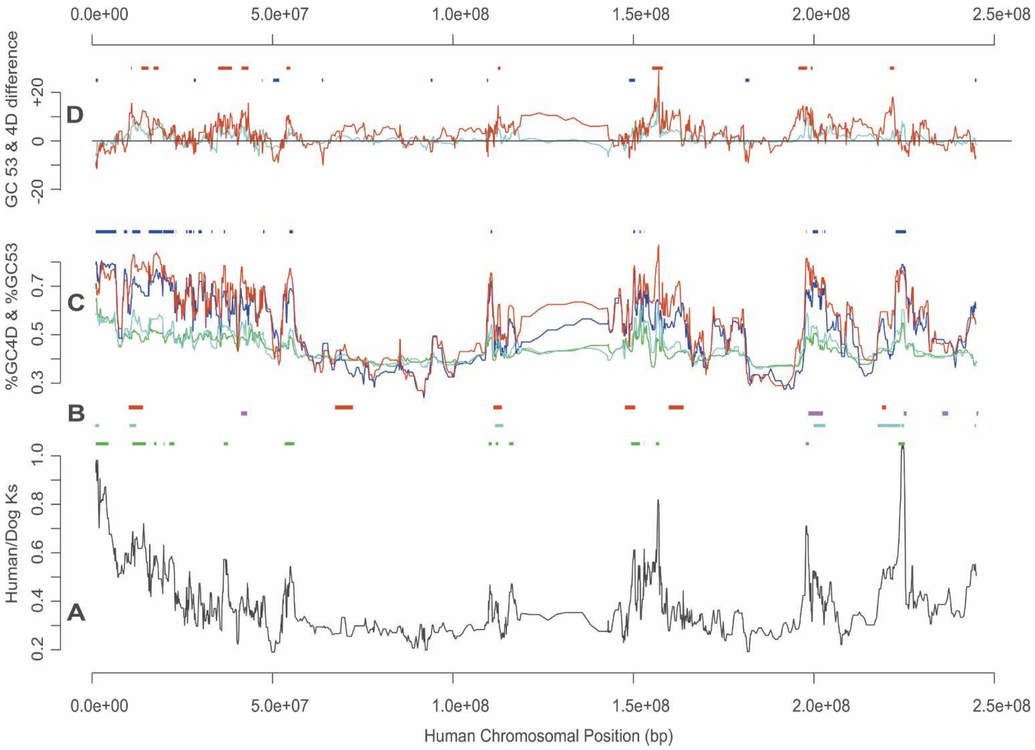

Figure 2. Variations in dog and human K_s, and different G+C fractions, as functions of distance (in base pairs) along human Chromosome 1. These quantities are shown as median values for 10 gene overlapping windows (see Methods). (*A*) The variation in K_s values (in black) of human and dog orthologs along this chromosome. K_s value hotspots are indicated in green above *A*. (*B*) The syntenic locations (see Methods) of dog telomeres (in red) and dog centromeres (in pink) on human Chromosome 1. Short synteny blocks (<4 Mb) are indicated below in light blue. (*C*) Variations in GC53 or GC4D fractions (as percentages). %GC53 and %GC4D values of human genes are shown in green and dark blue, respectively, whereas %GC53 and %GC4D' values of their dog orthologs are shown in light blue and red, respectively. %GC4D hotspots, exceeding the 80th centile for the whole chromosome (see Methods), are marked above *C* in dark blue. (*D*) The differences in %GC53 (ΔGC53, green), and in %GC4D (ΔGC4D, dark blue), between dog and human orthologs. Above *D*, ΔGC4D hotspots (i.e., dog regions elevated in GC4D, with respect to human orthologs; see Methods) are indicated in red, whereas ΔGC4D cold spots (i.e., human regions elevated in GC4D, with respect to dog orthologs; see Methods) are shown in blue.

Elevations in these quantities, and their declinations over physical distance, are substantially more pronounced in human subtelomeres than they are in dog subtelomeres (Fig. 3; Table 5). These thus reflect the findings shown in Table 1. Nevertheless, not all human chromosomes exhibit such a pronounced effect. Human Chromosome 1q subtelomere (Fig. 2), for example, as well as each of the individual subtelomeres of 3p, 3q, 4q, 9p, 11q, 13p, 14p, 15p, 15q, 17p, 18p 19q, 20p, and 21p, do not exhibit substantial K_s value elevations in subtelomeric regions (Supplemental Fig. 2). Nevertheless, taken together, these subtelomeric regions' genes still exhibit a perceptible and significant (Spearman's $\rho = 0.15$; $P = 9.0 \times 10^{-11}$) elevation in their K_s values. Consequently, G+C fraction and K_s value elevations occur, to varying extents, within most human subtelomeric regions.

Unexpectedly, we also observed strong and significant declines in genes' GC4D and K_s values from dog centromeres, but not from human centromeres (Fig. 3; Table 5). Consequently, we broadened our subsequent studies to consider GC4D and K_s value elevations in both subtelomeric and pericentromeric regions of the dog genome.

G+C fraction and K_s values are elevated in dog subtelomeric regions

From the perspective of the recombination-driven BGC model, the elevations of these quantities in dog subtelomeric and pericentromeric regions were surprising. Each of the dog chromosomes has been formed in the last 60 million years (Myr) from a mosaic of two to four segments from chromosomes of the common ancestor of the carnivores (Nash et al. 2001), and thus these chromosomal ends have only been derived recently by large-scale chromosomal rearrangements. If, owing to recombination-driven BGC, rises in G+C fraction and K_s value occur relatively slowly over time, then it appeared unlikely that sequence has dwelt for sufficient time at the ends of dog chromosomes for this effect to have become so pronounced.

If, in contrast, more rapid and substantial changes in base composition within high G+C regions occurred during the past ~100 Myr, then these are not consistent with either the high correlation of dog and human orthologs' GC4D values

Table 2. Correlation coefficients (Spearman's ρ) between K_s and GC4D for five species pairs

Species 1	Species 2	Correlation coefficient: GC4D (Species 1) and K_s	Correlation coefficient: GC4D (Species 2) and K_s
Chicken	Human	0.404	0.529
Dog	Human	0.681	0.640
Mouse	Human	0.507	0.468
Mouse	Rat	0.264	0.277
Chicken	Dog	0.393	0.536

Table 3. Median values of G+C fraction at 4D sites (GC4D) for 1:1 orthologs from five organisms, partitioned according to their locations in close proximity (<5 Mb) to telomeric ends, or in regions far (>9 Mb) from telomeric ends

Species	Subtelomeric median GC4D fraction	Interstitial median GC4D fraction
Chicken	0.54	0.44
Dog	0.68	0.61
Human	0.69	0.55
Mouse	0.58	0.55
Rat	0.57	0.55

($\rho = 0.945$) (Table 1), or with the findings of others that the human isochore structure is ancestral (Galtier and Mouchiroud 1998; Eyre-Walker and Hurst 2001; Belle et al. 2004). Thus, we do not expect rapid variation in base composition within high G+C regions at subtelomeres and pericentromeres.

Significant coincidence of G+C "hotspots" and chromosomal breakpoints

We considered whether these high G+C fraction regions were segregated to subtelomeres and pericentromeres during the fragmentation of the canid karyotype as a direct consequence of chromosomal fission occurring preferentially within such regions (the "fragile breakage" model) (Fig. 4). To investigate this model, we needed next to consider the locations of canine genes whose human orthologs are located in high G+C regions. Moreover, we needed to correlate G+C fraction, K_s value, chromosomal location, and chromosomal rearrangements at scales smaller than the ~10-Mb distances for which we observed G+C fraction and K_s value elevations within subtelomeric and pericentromeric regions (Fig. 3). GC4D and K_s values vary erratically as a function of chromosomal location, although their longer-range variations are observable using windowing methods (Figs. 1 and 2; Mouse Genome Sequencing Consortium 2002; Hardison et al. 2003).

Consequently, we defined long-range maxima (hotspots) in either G+C fraction or K_s value, using a sliding window of 10 genes (see Methods); these are two to three orders of magnitude longer (Supplemental Table 1) than previously described recombination hotspots (Jeffreys et al. 2001). As expected from our previous findings (Fig. 3), hotspots either in G+C fraction, or in K_s value (which are strongly correlated) (Table 2), were found to coincide with human subtelomeric, and dog subtelomeric and pericentromeric regions, more than expected by chance alone (data not shown).

However, we were most interested in the current chromosomal locations of dog genes whose single human orthologs have persisted in chromosomal interstitial regions throughout human history, since at least the CAE. In particular, we wished to track the location, in the dog genome, of human G+C fraction and K_s value local maxima within these interstitial regions. Thus, we first delineated 1.048 Gb of human sequence from the interstitial regions (>9 Mb from assembled chromosomal ends) of nine chromosomes, namely, HSA1, 5, 6, 9, 11, 13, 17, 18, and 20, which are known to have escaped major rearrangement (fusion or fission) since the CAE (Wienberg 2004). We refer to the human genes in such regions, and their canine single orthologs, as ancestral interstitial (AI) genes. In the following analysis we discarded all sequence except human and dog regions containing AI genes.

Next, we identified 113 human hotspots containing AI genes that exhibit the highest 20% of GC4D windowed values in each chromosome (see Methods). By mapping these hotspots to canine chromosomes, we identified 24 of these 113 high G+C hotspots that have been relocated, during the evolution of the canid karyotype, to among 17 subtelomeric regions of dog chromosomes. Using a randomization model (see Methods), we found that this is a higher number than expected by chance ($P = 0.037$). Thus, assuming that extant dog telomeres derived from fissuring events, we infer that ancestral regions high in G+C had a significantly greater propensity for fissuring during the evolution of the canid karyotype.

We also identified 15 human AI-gene-containing regions that are in conserved synteny with a dog pericentromeric region. Of these 15 regions, 12 (80%) are coincident with one or more of 113 human AI gene GC4D hotspots, which again is unexpected by chance alone ($P = 2.7 \times 10^{-6}$). Thus, our findings strongly suggest that ancestral interstitial hotspots have preferentially been rearranged to form extant subtelomeric and pericentromeric regions.

Significant coincidence of K_s hotspots and chromosomal breakpoints

We also performed a similar analysis applied to 66 human AI gene-containing regions characterized by significantly high K_s-value maxima (see Methods). As expected, given the strong correlations between dog and human G+C fractions, and G+C fraction and K_s value (Table 2), these K_s hotspots also are found preferentially both at dog subtelomeres and at pericentromeres: 14 of 66 high K_s value regions map to 17 dog subtelomeric regions ($P = 5.8 \times 10^{-6}$) and 9 of 66 high K_s value regions map to 15 dog pericentromeric regions ($P = 6.4 \times 10^{-5}$). The more significant coincidence of K_s peaks to fragile sites, over GC peaks, may in part be a consequence of the narrower biological variation in GC4D, compared to K_s, combined with fewer sampled sites, which acts to increase the sampling error and to limit G+C peak detection.

G+C fraction and K_s values are elevated in short synteny blocks

The fragile breakage model predicts that short synteny blocks in the dog genome exhibit higher GC4D and K_s values than longer blocks. This follows directly from our observations that chromosomal breakage in the dog lineage occurred preferentially in regions associated with high GC4D and K_s values. We thus investigated whether genes' GC4D and K_s values are correlated with the size of dog synteny blocks in which they are located. Indeed,

Table 4. Median K_s values for 1:1 ortholog pairs (D5 data set), partitioned according to their presence in subtelomeric regions (<5 Mb to telomeric ends of assembled chromosomes) or interstitial regions (regions >9 Mb from telomeres)

Species	Subtelomeric/ subtelomeric	Subtelomeric/ interstitial	Interstitial/ subtelomeric	Interstitial/ interstitial
Chicken/human	3.15	1.95	2.40	1.53
Dog/human	0.69	0.35	0.67	0.31
Dog/chicken	2.15	2.08	1.98	1.64
Mouse/human	0.96	0.62	0.78	0.54
Mouse/rat	0.19	0.17	0.19	0.19

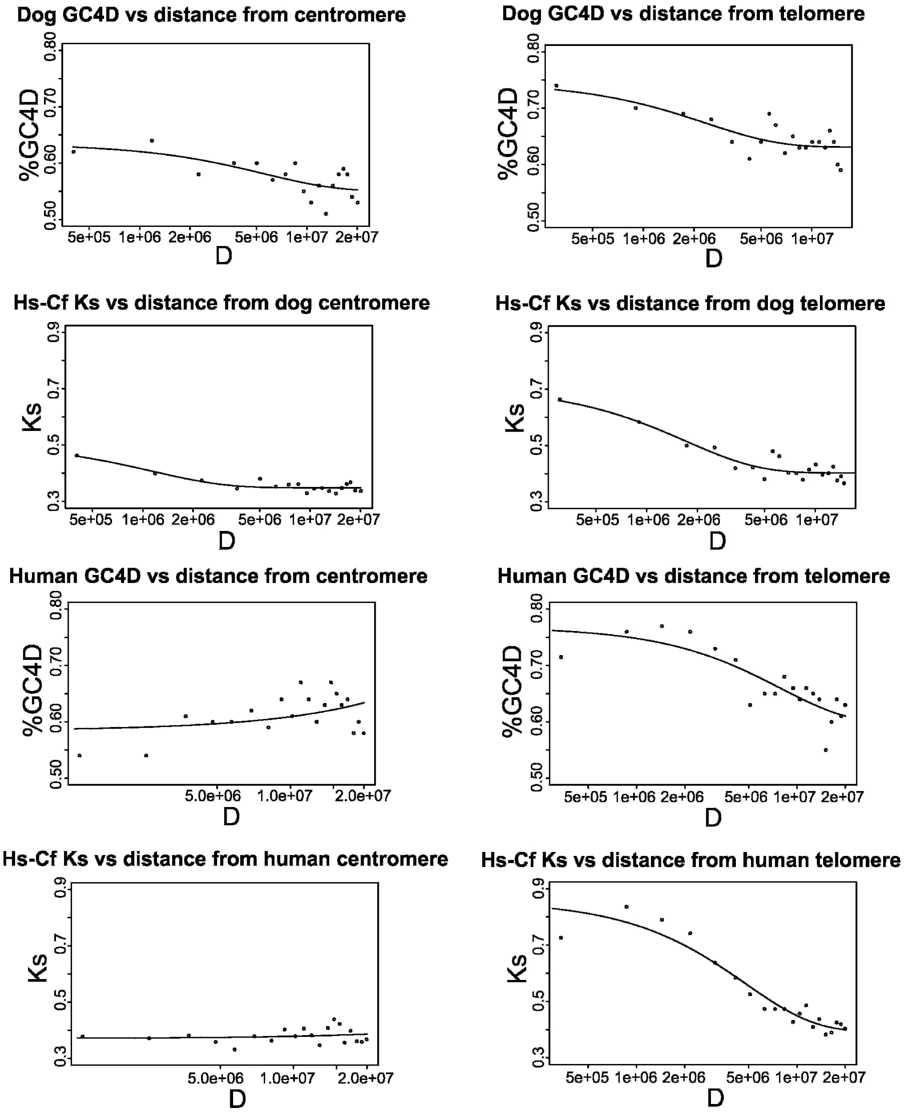

Figure 3. Variations in median K_s and GC4D percentage of dog and human single orthologs (set D2) with respect to log distance to a dog or human telomere or centromere. Values were obtained from nonoverlapping bins of 200–250 sequential genes. K_s was calculated between dog–human 1:1 orthologs (see Methods). *D* is the distance in bases. All plots have been fitted with a first-order exponential, $y = A + B\ \exp(X/C)$, except those plotted with respect to distance from the human centromere, which are fitted with straight lines.

we found significant negative correlations between synteny block size and either dog GC4D (Spearman's $\rho = -0.217$; *P*-value < 2.2×10^{-16}) or K_s value (Spearman's $\rho = -0.220$; *P*-value < 2.2×10^{-16}). Similar significant correlations (data not shown) were observed for correlations with dog GC53 or human GC4D or human GC53.

We then investigated whether these quantities increase with distance toward a synteny breakpoint, in a manner similar to that seen for subtelomeric sequence (Fig. 3). Indeed, sequence approaching synteny breaks does exhibit significantly higher G+C content ($\rho = 0.21$, $P < 2.2 \times 10^{-16}$) and K_s ($\rho = 0.20$, $P < 2.2 \times 10^{-16}$) values (Fig. 5). This indicates that increases in these quantities toward synteny breaks represent the more general case of a phenomenon that manifests also, as a special case, elevations within subtelomeric sequence.

Significant coincidence of ancestral high K_s value regions and high ΔGC4D regions

We next considered a prediction of the recombination-driven BGC model that G+C fraction has increased more within hotspots than elsewhere. For this analysis, we calculated values of ΔGC4D, the GC4D fraction of a dog gene over and above that of its human ortholog. We then identified 84 regions of human AI-gene-containing sequence that exhibited the highest ΔGC4D values (see Methods); human genes in ΔGC4D maxima thus possess significantly lower GC4D values, on average, than their dog orthologs. In agreement with the recombination-driven BGC model, we find that of 32 dog pericentromeric and subtelomeric regions that are in conserved synteny with human AI-gene-containing regions, 18 coincide with one or more of these 84 ΔGC4D maxima; this number is more than expected by chance ($P = 5.2 \times 10^{-4}$). In contrast, we found no significant coincidence between ΔGC4D minima (representing regions where the human GC4D values, on average, exceed those of their dog orthologs) and dog pericentromeric and subtelomeric regions ($P = 0.54$). Similarly, we find that dog GC4D hotspots tend to have increased their G+C content, relative to their human orthologs, when they are located close to chromosomal ends. ΔGC4D is significantly elevated in the 20 Mb approaching dog telomeres (Spearman's $\rho = 0.146$, *P*-value < 2.2×10^{-16}).

However, contrary to the recombination-driven BGC model, ΔGC4D is not correlated ($|\rho| < 0.05$) with distances to human telomeres and centromeres; it is also not correlated with distance to dog centromeres.

Thus, ancestral interstitial sequence that suffered a breakpoint and rearrangement during canid evolution is not only elevated in extant gene GC4D values, but this GC4D elevation in dog sequence significantly exceeds the elevation of its human orthologs. Obviously, these regions, which we propose have been substantially elevated in GC4D in the common ancestor of dog and human, either have reduced their GC4D values in the human lineage, or have increased their GC4D values in the canine lineage, or both.

Discussion

Our results demonstrate location-dependent effects on nucleotide composition and substitution rates in both human and dog genomes. We observed that GC4D and K_s values are significantly elevated within human subtelomeric regions, and that these elevations are greater than those seen for dog, mouse, and rat sub-

Table 5. Correlations (Spearman's ρ/*P*-value) of gene properties with increasing distance from centromeres or telomeres

Property	Human subtelomere	Dog subtelomere	Human pericentromere	Dog pericentromere
K_s	$-0.39/<2.2 \times 10^{-16}$	$-0.23/2.2 \times 10^{-5}$	$-0.02/0.34$	$-0.1/2.6 \times 10^{-11}$
GC4D	$-0.27/<2.2 \times 10^{-16}$	$-0.18/4.8 \times 10^{-3}$	$0.06/1.0 \times 10^{-4}$	$-0.09/5.7 \times 10^{-9}$
GC53	$-0.38/<2.2 \times 10^{-16}$	$-0.27/2.7 \times 10^{-5}$	$-0.08/2.0 \times 10^{-7}$	$-0.14/<2.2 \times 10^{-16}$

Correlations were calculated for D2 genes (unbinned) within 20 Mb of either a centromere or a telomere.

telomeric regions. These findings immediately suggest that human sequence, which has been relatively intransigent to chromosomal rearrangement, might have steadily increased its G+C fraction over time within subtelomeric regions, to a greater extent than seen for the subtelomeres of genomes, such as those of dog, mouse, and rat, which have suffered from substantial rearrangements. In this BGC scheme, the stability of the karyotype of human ancestors would have resulted in steadily increasing G+C fractions in subtelomeric regions that possess high rates of recombination (Duret et al. 2002). This is because recombination, when coupled to biased gene conversion, increases the incorporation of G or C bases at a mismatch site (Galtier et al. 2001). Higher G+C proportions have been linked to higher neutral rates mainly because of the hypermutability of the CpG dinucleotide (Cooper and Youssoufian 1988; Sved and Bird 1990), although this link has recently been questioned (Fryxell and Moon 2004). Such a scheme would be consistent with an origin of mammalian high G+C isochores from within stable subtelomeric regions in early chordate genomes.

Importantly, our findings cannot all be explained by this recombination-driven BGC scheme. In particular, the scheme predicts higher G+C fractions among human subtelomeric genes than their dog orthologs (i.e., a positive correlation between ΔGC4D and distance from telomeric end), because the human karyotype is relatively ancestral whereas that of the dog is derived. Although such a significant positive correlation occurs, it is relatively small (Spearman's $\rho = 0.045$) and at a similar level to genes approaching human centromeres ($\rho = 0.046$), where recombination typically is suppressed (Kong et al. 2002).

Moreover, this scheme predicts divergent GC4D values for human and dog orthologs that have been subjected to different amounts of recombination. Nevertheless, we observed an extremely high correlation (Spearman's $\rho = 0.945$) between human and dog orthologs' GC4D values (Table 1). This indicates that a gene's G+C fraction, for these two species, is an ancestral property, and thus has not altered substantially in rank order since their common ancestor ~95 Mya. An extant human subtelomeric gene that exhibits a high G+C fraction is likely to have possessed a relatively high G+C value in the genomes of the earliest boreoeutherian ancestor and its predecessors, regardless of its location in these ancestral chromosomes. This is supported by a previous observation that the average G+C fraction at the third codon positions (GC3) of 41 genes has reduced by only 2.3% for the primate lineage, and 2.1% for the carnivore lineage since the CAE (Belle et al. 2004), and by the similarity between average GC4D values for dog and human genes (0.590 and 0.570, respectively).

If G+C fractions, and thus K_s rates, have not been elevating substantially and progressively within the subtelomeres of chromosomes in the human lineage, then what other evolutionary process might have caused such effects? We suggest an alternative ("fragile breakage") scheme to account for our findings (Fig. 4). We propose that high G+C sequences are preferentially relocated to telomeres as a direct consequence of chromosomal fission occurring more frequently within regions that are elevated in G+C. In this model, new telomeres generated by a chromosomal fission within such regions would often be flanked by high G+C sequence. Centromeres would also be more likely to neighbor a high G+C region because of de novo formation of a centromere ("neocentromerization") (Amor and Choo 2002; Ventura et al. 2004; Warburton 2004) at the nascent chromosomal end, and because of telomere–centromere inversions, which are known features of canid evolution (Fig. 4; Nash et al. 2001).

In support of this scheme, we found that extant dog subtelomeric and pericentromeric sequences arose preferentially by fission of chromosomal sequence that was ancestrally enriched in G+C bases. We identified significant correlations between sub-

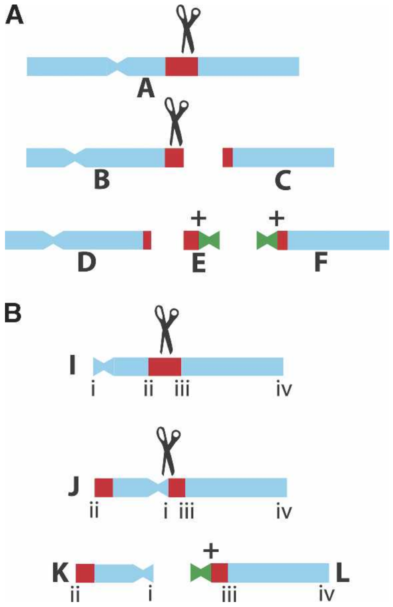

Figure 4. The "fragile breakage" model for the evolution of the canine karyotype. Schematic models of chromosome breakage within hotspots for metacentric (*A*) and acrocentric (*B*) chromosomes, showing how initially interstitial hotspots may be redistributed to subtelomeric and pericentromeric locations. (*A*) A metacentric Chromosome A breaks at the hotspot (shown in red), thereby forming Chromosome B and fragment C that lacks the original centromere. The hotspot fragment now resides within the subtelomeric regions of B and C. Further breakage within the subtelomeric hotspot of B may give rise to Chromosome D and, through neocentromerization, the microchromosome E. Fragment C may acquire a centromere adjacent to its hotspot. The hotspot, once interstitial in Chromosome A, is now distributed across two subtelomeric, and one pericentromeric, regions. (*B*) Acrocentric Chromosome I fissures at a hotspot (shown in red), enabling a pericentric inversion that relocates the centromere medially, giving rise to the metacentric Chromosome J. The hotspot fragments now reside in the proximal subtelomeric, and in the quartal pericentromeric, regions of J. A subsequent break at the centromere gives rise to acrocentric Chromosomes K and L. The interstitial hotspot is again redistributed to the subtelomeric and pericentromeric regions.

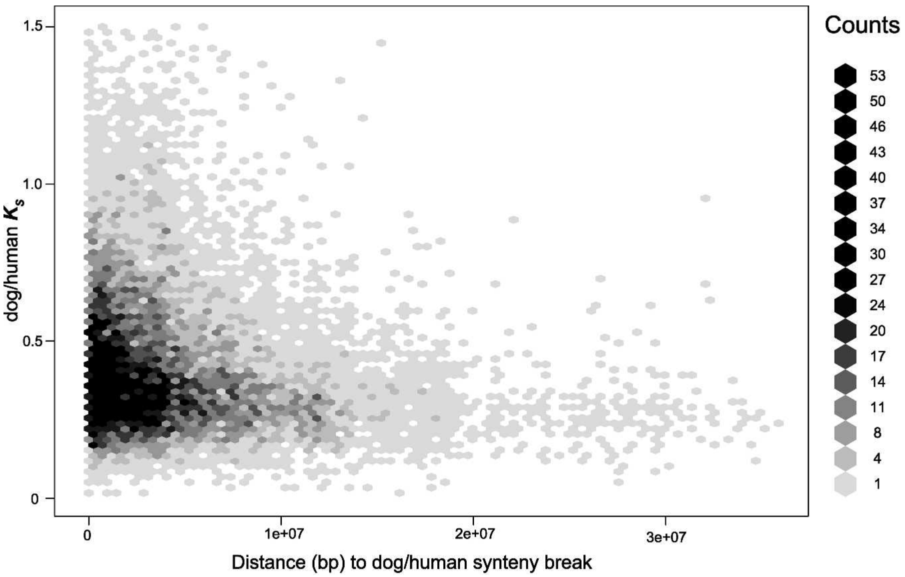

Figure 5. Density of dog/human orthologs' K_s values versus distance of the dog ortholog to a dog/human synteny breakpoint. The density color key is shown to the *right* of the plot. Increasing K_s and gene G+C content at fourfold degenerate positions (data not shown) are each significantly correlated with decreasing distance to a human/dog synteny breakpoint.

telomeric and pericentromeric regions in the dog genomes, and their orthologous regions in human that possess elevated G+C and K_s values, despite these human regions being unlikely to have been located close to telomeres since the CAE. We also note that our proposal is consistent with studies of human chromosomes that show that breakpoints in human chromosomes occur preferentially in telomeres (Yu et al. 1978; Stoll 1980) and in G+C-rich G-light regions (Aula and von Koskull 1976; Nakagome and Chiyo 1976; Stoll 1980; Abeysinghe et al. 2003). The proposal may be valid for other genomes besides those of human and dog because double-stranded breaks are also known to occur predominantly in high G+C regions in the yeast *Saccharomyces cerevisiae* (Baudat and Nicolas 1997; Gerton et al. 2000). While we cannot formally discount the possibility that such breaks preferentially occur in high G+C regions as a result of a dependency on a third quantity with which G+C and K_s both covary, we believe it possible that nonrandom breakage occurs directly because of nonuniform base composition often acting over megabase scales.

Finally, the fragile breakage model is entirely consistent with significant negative correlations between synteny block size and either GC4D or K_s values. Furthermore, we find significant negative correlations between either G+C content or K_s and distance to a synteny breakpoint (Fig. 5). It appears that ancestral sequence with high G+C fraction was particularly susceptible to breakage and subsequent rearrangement throughout the recent evolution of the canid karyotype. Although we recognize that our model for chromosome evolution does not account for the stability of closely related feline and human karyotypes (Wienberg et al. 1997), it does explain the preponderance of high G+C sequence in the smallest chromosomes of chicken (Auer et al. 1987; International Chicken Genome Sequencing Consortium 2004), as such sequence is more likely to suffer fissions.

Recent findings have overturned the random-breakage model of chromosomal evolution (Nadeau and Taylor 1984). Breakpoints often appear to be clustered, implying "reuse" of breakpoints, in a model termed fragile breakage (Pevzner and Tesler 2003; Bailey et al. 2004). One indicator of fragility is the occurrence of segmental duplication in orthologous sequence (Bailey et al. 2004). Our findings suggest that fragility is also associated with high G+C fraction. If so, high G+C regions appear not only to be highly susceptible to nucleotide substitution, insertions, deletions, and recombination (Hardison et al. 2003; Taylor et al. 2004), but also to chromosomal breakage.

Our results do not counter the hypothesis of G+C elevation at subtelomeric regions due to increased recombination and biased gene conversion, although we observed scant evidence of such elevation within human subtelomeres. Rather, they highlight correspondences between G+C and K_s hotspots and the evolution of canid chromosomes, and represent consequences of mutational processes that have shaped the canid, and perhaps other, karyotypes. We have shown that G+C hotspots in the common ancestor of dog and human either have reduced their GC4D values in the human lineage, or have increased their GC4D values in the canine lineage, or both. Further studies of the G+C changes in the mammalian lineage are likely to reveal the relative contributions of these two evolutionary processes to G+C fraction elevations within subtelomeres, pericentromeres, and fragile breakpoints.

Methods

Gene sets

We used two gene sets for our analyses. The first (denoted D5) was a set of 6800 1:1:1:1:1 chicken:human:mouse:rat:dog orthologs. This consisted of a set of 8164 1:1:1:1 chicken:human:mouse:rat orthologs, as described previously (International Chicken Genome Sequencing Consortium 2004), augmented with their predicted single dog orthologs. The four-way International Chicken Genome Sequencing Consortium set consists of Ensembl genes (Hubbard et al. 2005) based on the *Homo sapiens* NCBI34, *Mus musculus* NCBI30, *Rattus norvegicus* Baylor v2.1, *Gallus gallus* (WUSTL Feb. 2004 release). Dog orthologs were predicted by first aligning all transcripts between each human and mouse 1:1 ortholog pair using BLAST (Altschul et al. 1997). Next, the human transcript that aligned with the highest bit-score density (bit-score per aligned length) was used to query the dog genome (Broad v1), initially with Exonerate (Slater and Birney 2005), and refined subsequently with GeneWise (Birney et al. 2004). In all, 17,598 predictions representing 8066 queries were returned. Where predictions overlapped by >20% of their length, the prediction with the highest GeneWise score was retained; 12,167 predictions representing 7988 queries remained. The K_s value between each top hit and the initial human query was calculated with Codeml from the PAML package (Yang 1997; http://abacus.gene.ucl.ac.uk/software/paml.html), essentially as previously (Mouse Genome Sequencing Consortium 2002) (two genes were removed at this stage as a result of alignment problems). Likely dog processed pseudogenes were identified as intron-less predictions where the human template possesses at least one intron 10 or more codons from either translational end; subsequently, these were removed. The highest scoring transcript for each of the remaining 7751 genes was then added to the four-way International Chicken Genome Sequencing Consortium set to form the 1:1:1:1:1 D5 orthology set. In all but 15 of these dog predicted orthologs, these proved to represent reciprocal-best-BLASTp-hits to their predicted human orthologs' sequences. The median human–dog orthologs' K_s value for the set was 0.36, and the median amino acid percentage identity was 92.75%; these values are similar to those obtained by others (Lindblad-Toh et al. 2005). Where positional information was required for all genes within the set, 951 orthologous quintuplets containing one or more genes located on an unassembled chromosome were removed.

A second ortholog set (denoted D2) was obtained from phylogeny-based orthology predictions by L. Goodstadt and C.P. Ponting (in prep.) between Ensembl genes based on the *H. sapiens* NCBI35 and *C. familiaris* (Broad v1) genomes. The D2 set consisted of 13,747 orthologous gene pairs. This number is reduced to 11,713 pairs where placement on an assembled chromosome for both orthologs is required. As above, all K_s values were estimated using Codeml (Yang 1997). The median human–dog orthologs' K_s value for the D2 set was 0.372, the median K_a/K_s ratio was 0.107, and the median amino acid percentage identity was 89.4%.

Maxima (hotspots) determination

G+C nucleotide fractions were calculated both at fourfold degenerate (4D) sites (GC4D) and at sites 10 kb upstream and downstream of the transcriptional start site (GC53). In order to identify regional maxima in G+C fraction values, a sliding window containing 10 D2 genes was translated across each assembled dog and human autosome. Variations in quantities were examined for autosomes only, because mammalian X-chromosomes have persisted without fusion or fission at least since the CAE (Kohn et al. 2004). Within each window the median GC4D, the median K_s value between dog and human orthologs, the GC4D values of their orthologs, and the median difference between the genes' GC4Ds were recorded, along with the window's location, defined as the mean position of its genes' midpoints between transcriptional start and end bases.

K_s and ΔGC4D values' maxima (hotspots) were defined using similar procedures. First, for each chromosome the mean K_s or ΔGC4D value averaged over all gene windows was calculated. An initial set of hotspots was then defined as the locations of gene windows whose median K_s or ΔGC4D values were 2.0 standard deviations greater than the chromosomal mean; for normal distributions this threshold delineates the highest 2.3% of the data. Results (data not shown) obtained at higher thresholds, namely, 2.5 and 3.0 standard deviations, were similar to those described here. The resulting maxima were then aggregated by requiring at least three consecutive windows whose values all lay above threshold, and adjacent bins were amalgamated to form broader hotspots. An additional set of more localized peaks was defined by repeating this procedure using a sliding 20-Mb window across each chromosome. The resulting hotspots from both procedures were then combined. Summary statistics for the size distributions of these peaks are provided in Supplemental Table 1.

The larger localized variance (see Figs. 1 and 2; Supplemental Figs. 1 and 2) and limited range of GC4D values yielded fewer hotspots as compared to K_s and ΔGC4D hotspots. Consequently, we instead defined GC4D hotspots' sets as those windows whose median GC4D value exceeded the 60th, 70th, 80th, 90th, 95th, and 99th centiles of their corresponding whole chromosomal distribution. Adjacent maxima above each centile threshold were amalgamated to form broader peaks. Results comparable to those described here were obtained for the remaining G+C thresholds (data not shown). Hence, for simplicity, results from only the 80th centile threshold are presented here. Summary statistics for the size distributions of G+C peaks at the 80th centile are presented in Supplemental Table 1.

Conserved synteny

Conserved synteny was defined using dog versus human 500-kb synteny maps obtained from the Dog Genome Sequencing Consortium (Lindblad-Toh et al. 2005). Human subtelomeres were defined from the relevant genome assembly as regions within 5 Mb of the end of each autosome sequence. For each of the five human acrocentric chromosomes, only one subtelomeric region was defined. Regions within the dog genome exhibiting conserved synteny to human subtelomeres (<5 Mb from the assembled telomeric end) were also recorded.

For the dog genome, pericentromeric regions were defined as regions <5 Mb from the proximal end of the autosomes, while subtelomeric regions were defined as regions <5 Mb from the quartal ends of autosomes. The position of each human genomic region in conserved synteny to a dog pericentromeric or subtelomeric region was recorded. Interstitial regions were defined as sequence >9 Mb from both assembled telomeric and centromeric ends.

Significance of spatial coincidences

The likelihood of observing at least a given number of overlaps between a pair of data sets was evaluated using randomized simulations and *Z*-scores. For each of the two data sets, a set of identically sized nonoverlapping fragments was drawn from an identical sample space, and the number of overlaps between the two randomized sets counted. This procedure was repeated 10,000 times for each test, and the distributions of randomized overlap frequencies checked for signs of kurtosis. A *Z*-score was then derived for the observed number of genomic overlaps, from which a normalized probability was calculated.

Additional statistical analysis

Statistical analysis not described above was performed using R (http://cran.r-project.org/).

Acknowledgments

We are indebted to Ensembl, Kerstin Lindblad-Toh, Tarjei Mikkelsen, and others of the Dog Genome Sequencing Consortium, for assistance, and to Leo Goodstadt and Andreas Heger for helpful discussions. We thank all the reviewers, whose comments were invaluable in improving this manuscript.

References

Abeysinghe, S.S., Chuzhanova, N., Krawczak, M., Ball, E.V., and Cooper, D.N. 2003. Translocation and gross deletion breakpoints in human inherited disease and cancer I: Nucleotide composition and recombination-associated motifs. *Hum. Mutat.* **22:** 229–244.

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. 1997. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res.* **25:** 3389–3402.

Amor, D.J. and Choo, K.H.A. 2002. Neocentromeres: Role in human disease, evolution, and centromere study. *Am. J. Hum. Genet.* **71:** 695–714.

Auer, H., Mayr, B., Lambrou, M., and Schleger, W. 1987. An extended chicken karyotype, including the NOR chromosome. *Cytogenet. Cell Genet.* **45:** 218–221.

Aula, P. and von Koskull, H. 1976. Distribution of spontaneous chromosome breaks in human chromosomes. *Hum. Genet.* **32:** 143–148.

Bacolla, A., Jaworski, A., Larson, J.E., Jakupciak, J.P., Chuzhanova, N., Abeysinghe, S.S., O'Connell, C.D., Cooper, D.N., and Wells, R.D. 2004. Breakpoints of gross deletions coincide with non-B DNA conformations. *Proc. Natl. Acad. Sci.* **101:** 14162–14167.

Bailey, J.A., Baertsch, R., Kent, W.J., Haussler, D., and Eichler, E.E. 2004. Hotspots of mammalian chromosomal evolution. *Genome Biol.* **5:** R23.

Baudat, F. and Nicolas, A. 1997. Clustering of meiotic double-strand breaks on yeast chromosome III. *Proc. Natl. Acad. Sci.* **94:** 5213–5218.

Belle, E.M., Duret, L., Galtier, N., and Eyre-Walker, A. 2004. The decline of isochores in mammals: An assessment of the GC content variation along the mammalian phylogeny. *J. Mol. Evol.* **58:** 653–660.

Bernardi, G., Olofsson, B., Filipski, J., Zerial, M., Salinas, J., Cuny, G., Meunierrotival, M., and Rodier, F. 1985. The mosaic genome of warm-blooded vertebrates. *Science* **228:** 953–958.

Birney, E., Clamp, M., and Durbin, R. 2004. GeneWise and Genomewise. *Genome Res.* **14:** 988–995.

Brown, T.C. and Jiricny, J. 1987. A specific mismatch repair event protects mammalian cells from loss of 5-methylcytosine. *Cell* **50:** 945–950.

———. 1988. Different base/base mispairs are corrected with different efficiencies and specificities in monkey kidney cells. *Cell* **54:** 705–711.

Cooper, D.N. and Youssoufian, H. 1988. The CpG dinucleotide and human genetic disease. *Hum. Genet.* **78:** 151–155.

Duret, L., Semon, M., Piganeau, G., Mouchiroud, D., and Galtier, N. 2002. Vanishing GC-rich isochores in mammalian genomes. *Genetics* **162:** 1837–1847.

Eyre-Walker, A. 1993. Recombination and mammalian genome evolution. *Proc. R. Soc. London Ser. B* **252:** 237–243.

Eyre-Walker, A. and Hurst, L.D. 2001. The evolution of isochores. *Nat. Rev. Genet.* **2:** 549–555.

Filipski, J., Thiery, J.P., and Bernardi, G. 1973. An analysis of the bovine genome by Cs_2SO_4-Ag density gradient centrifugation. *J. Mol. Biol.* **80:** 177–197.

Fryxell, K.J. and Moon, W.J. 2004. CpG Mutation rates in the human genome are highly dependent on local GC content. *Mol. Biol. Evol.* **22:** 650–658.

Fullerton, S.M., Bernardo Carvalho, A., and Clark, A.G. 2001. Local rates of recombination are positively correlated with GC content in the human genome. *Mol. Biol. Evol.* **18:** 1139–1142.

Galtier, N. and Mouchiroud, D. 1998. Isochore evolution in mammals: A human-like ancestral structure. *Genetics* **150:** 1577–1584.

Galtier, N., Piganeau, G., Mouchiroud, D., and Duret, L. 2001. GC-content evolution in mammalian genomes: The biased gene conversion hypothesis. *Genetics* **159:** 907–911.

Gerton, J.L., DeRisi, J., Shroff, R., Lichten, M., Brown, P.O., and Petes, T.D. 2000. Inaugural article: Global mapping of meiotic recombination hotspots and coldspots in the yeast *Saccharomyces cerevisiae*. *Proc. Natl. Acad. Sci.* **97:** 11383–11390.

Hardison, R.C., Roskin, K.M., Yang, S., Diekhans, M., Kent, W.J., Weber, R., Elnitski, L., Li, J., O'Connor, M., Kolbe, D., et al. 2003. Covariation in frequencies of substitution, deletion, transposition, and recombination during eutherian evolution. *Genome Res.* **13:** 13–26.

Hellmann, I., Ebersberger, I., Ptak, S.E., Pääbo, S., and Przeworski, M. 2003. A neutral explanation for the correlation of diversity with recombination rates in humans. *Am. J. Hum. Genet.* **72:** 1527–1535.

Hubbard, T., Andrews, D., Caccamo, M., Cameron, G., Chen, Y., Clamp, M., Clarke, L., Coates, G., Cox, T., Cunningham, F., et al. 2005. Ensembl 2005. *Nucleic Acids Res.* **33:** D447–D453.

International Chicken Genome Sequencing Consortium. 2004. Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature* **432:** 695–716.

International Human Genome Sequencing Consortium. 2001. Initial sequencing and analysis of the human genome. *Nature* **409:** 860–921.

Jeffreys, A.J., Kauppi, L., and Neumann, R. 2001. Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex. *Nat. Genet.* **29:** 217–222.

Karolchik, D., Hinrichs, A.S., Furey, T.S., Roskin, K.M., Sugnet, C.W., Haussler, D., and Kent, W.J. 2004. The UCSC Table Browser data retrieval tool. *Nucleic Acids Res.* **32 (Suppl 1):** D493–D496.

Kohn, M., Kehrer-Sawatzki, H., Vogel, W., Graves, J.A., and Hameister, H. 2004. Wide genome comparisons reveal the origins of the human X chromosome. *Trends Genet.* **20:** 598–603.

Kong, A., Gudbjartsson, D.F., Sainz, J., Jonsdottir, G.M., Gudjonsson, S.A., Richardsson, B., Sigurdardottir, S., Barnard, J., Hallbeck, B., Masson, G., et al. 2002. A high-resolution recombination map of the human genome. *Nat. Genet.* **31:** 241–247.

Lamb, B.C. 1986. Gene conversion disparity: Factors influencing its direction and extent, with tests of assumptions and predictions in its evolutionary effects. *Genetics* **114:** 611–632.

Lercher, M.J. and Hurst, L.D. 2002. Human SNP variability and mutation rate are higher in regions of high recombination. *Trends Genet.* **18:** 337–340.

Lindblad-Toh, K., Wade, C.M., Mikkelsen, T.S., Karlsson, E.K., Jaffe, D.B., Kamal, M., Clamp, M., Chang, J.L., Kulbokas III, E.J., Zody, M.C., et al. 2005. Genome sequence, comparative analysis and haplotype structure of the domestic dog. *Nature* (in press).

Marais, G. 2003. Biased gene conversion: Implications for genome and sex evolution. *Trends Genet.* **19:** 330–338.

Matassi, G., Sharp, P.M., and Gautier, C. 1999. Chromosomal location effects on gene sequence evolution in mammals. *Curr. Biol.* **9:** 786–791.

Meunier, J. and Duret, L. 2004. Recombination drives the evolution of GC-content in the human genome. *Mol. Biol. Evol.* **21:** 984–990.

Mouchiroud, D., Gautier, C., and Bernardi, G. 1988. The compositional distribution of coding sequences and DNA molecules in humans and murids. *J. Mol. Evol.* **27:** 311–320.

Mouse Genome Sequencing Consortium. 2002. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420:** 520–562.

Nadeau, J.H. and Taylor, B.A. 1984. Lengths of chromosomal segments conserved since divergence of man and mouse. *Proc. Natl. Acad. Sci.* **81:** 814–818.

Nakagome, Y. and Chiyo, H. 1976. Nonrandom distribution of exchange points in patients with structural rearrangements. *Am. J. Hum. Genet.* **28:** 31–41.

Nash, W.G., Menninger, J.C., Wienberg, J., Padilla-Nash, H.M., and O'Brien, S.J. 2001. The pattern of phylogenomic evolution of the *Canidae*. *Cytogenet. Cell Genet.* **95:** 210–224.

Pevzner, P. and Tesler, G. 2003. Human and mouse genomic sequences reveal extensive breakpoint reuse in mammalian evolution. *Proc. Natl. Acad. Sci.* **100:** 7672–7677.

Rat Genome Sequencing Project Consortium. 2004. Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature* **428:** 493–521.

Slater, G.S. and Birney, E. 2005. Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* **6:** 31.

Springer, M.S., Murphy, W.J., Eizirik, E., and O'Brien, S.J. 2003. Placental mammal diversification and the Cretaceous–Tertiary boundary. *Proc. Natl. Acad. Sci.* **100:** 1056–1061.

Stein, L.D., Bao, Z., Blasiar, D., Blumenthal, T., Brent, M.R., Chen, N., Chinwalla, A., Clarke, L., Clee, C., Coghlan, A., et al. 2003. The genome sequence of *Caenorhabditis briggsae*: A platform for comparative genomics. *PloS Biol.* **1:** E45.

Stoll, C. 1980. Nonrandom distribution of exchange points in patients with reciprocal translocations. *Hum. Genet.* **56:** 89–93.

Sved, J. and Bird, A. 1990. The expected equilibrium of the CpG dinucleotide in vertebrate genomes under a mutation model. *Proc. Natl. Acad. Sci.* **87:** 4692–4696.

Taylor, M.S., Ponting, C.P., and Copley, R.R. 2004. Occurrence and consequences of coding sequence insertions and deletions in mammalian genomes. *Genome Res.* **14:** 555–566.

Trinh, P., McLysaght, A., and Sankoff, D. 2004. Genomic features in the breakpoint regions between syntenic blocks. *Bioinformatics* **20:** i318–i325.

Ventura, M., Weigl, S., Carbone, L., Cardone, M.F., Misceo, D., Teti, M., D'Addabbo, P., Wandall, A., Bjorck, E., de Jong, P.J., et al. 2004. Recurrent sites for new centromere seeding. *Genome Res.* **14:** 1696–1703.

Warburton, P.E. 2004. Chromosomal dynamics of human neocentromere formation. *Chromosome Res.* **12:** 617–626.

Wienberg, J. 2004. The evolution of eutherian chromosomes. *Curr. Opin. Genet. Dev.* **14:** 657–666.

Wienberg, J., Stanyon, R., Nash, W.G., O'Brien, P.C., Yang, F., O'Brien, S.J., and Ferguson-Smith, M.A. 1997. Conservation of human vs. feline genome organization revealed by reciprocal chromosome painting. *Cytogenet. Cell Genet.* **77:** 211–217.

Winzeler, E.A., Castillo-Davis, C.I., Oshiro, G., Liang, D., Richards, D.R., Zhou, Y., and Hartl, D.L. 2003. Genetic diversity in yeast assessed with whole-genome oligonucleotide arrays. *Genetics* **163:** 79–89.

Wolfe, K.H., Sharp, P.M., and Li, W.H. 1989. Mutation rates differ among regions of the mammalian genome. *Nature* **337:** 283–285.

Yang, Z. 1997. PAML: A program package for phylogenetic analysis by maximum likelihood. *Comput. Appl. Biosci.* **13:** 555–556.

Yu, C.W., Borgaonkar, D.S., and Bolling, D.R. 1978. Break points in human chromosomes. *Hum. Hered.* **28:** 210–225.

Web site references

http://cran.r-project.org/; R, Statistical Analysis Package.

http://abacus.gene.ucl.ac.uk/software/paml.html; Phylogenetic Analysis by Maximum Likelihood (PAML) package.

Received March 2, 2005; accepted in revised form June 16, 2005.

Hotspots of mutation and breakage in dog and human chromosomes

Caleb Webber and Chris P. Ponting

Genome Res. 2005 15: 1787-1797
Access the most recent version at doi:10.1101/gr.3896805