Chromosoma (2006) 115:427–436
DOI 10.1007/s00412-006-0066-4

RESEARCH ARTICLE

Comparative genomic analysis links karyotypic evolution with genomic evolution in the Indian Muntjac (*Muntiacus muntjak vaginalis*)

Qi Zhou · Ling Huang · Jianguo Zhang · Xiangyi Zhao · Qingpeng Zhang · Fei Song · Jianxiang Chi · Fengtang Yang · Wen Wang

Received: 5 March 2006 / Revised: 30 March 2006 / Accepted: 31 March 2006 / Published online: 22 June 2006

Abstract The karyotype of Indian muntjacs (*Muntiacus muntjak vaginalis*) has been greatly shaped by chromosomal fusion, which leads to its lowest diploid number among the extant known mammals. We present, here, comparative results based on draft sequences of 37 bacterial artificial clones (BAC) clones selected by chromosome painting for this special muntjac species. Sequence comparison on these BAC clones uncovered sequence syntenic relationships between the muntjac genome and those of other mammals. We found that the muntjac genome has peculiar features with respect to intron size and evolutionary rates of genes. Inspection of more than 80 pairs of orthologous introns from 15 genes reveals a significant reduction in intron size in the Indian muntjac compared to that of human, mouse, and dog. Evolutionary analysis using 19 genes indicates that the muntjac genes have evolved rapidly compared to other mammals. In addition, we identified and characterized sequence composition of the first BAC clone containing a chromosomal fusion site. Our results shed new light on the genome architecture of the Indian muntjac and suggest that chromosomal rearrangements have been accompanied by other salient genomic changes.

Communicated by E.A. Nigg

Electronic Supplementary Material Supplementary material is available in the online version of this article at http://dx.doi.org/10.1007/s00412-006-0066-4 and is accessible for authorized users.

Qi Zhou, Ling Huang, Jianguo Zhang: these authors contributed equally to the paper. Sequence data from this article have been deposited in the GenBank Libraries under Accession No. DQ280153-DQ280188, DQ377335, DQ458964.

Q. Zhou · X. Zhao · W. Wang (✉)
CAS-Max Planck Junior Research Group,
Key Laboratory of Cellular and Molecular Evolution,
Kunming Institute of Zoology, Chinese Academy of Sciences,
#32 E. Jiao Chang Road,
Kunming, Yunnan 650223, People's Republic of China
e-mail: wwang@mail.kiz.ac.cn

L. Huang · J. Chi · F. Yang (✉)
Key Laboratory of Cellular and Molecular Evolution,
Kunming Institute of Zoology, Chinese Academy of Sciences,
Kunming, Yunnan 650223, People's Republic of China
e-mail: fy1@sanger.ac.uk

Q. Zhou · L. Huang · J. Chi
Graduate School of Chinese Academy Sciences,
Beijing 100039, People's Republic of China

J. Zhang · Q. Zhang · F. Song
Beijing Institute of Genomics, Chinese Academy of Sciences,
Beijing 100039, People's Republic of China

F. Yang
Wellcome Trust Sanger Institute,
Wellcome Trust Genome Campus, Hinxton,
Cambridge, CB10 1SA, UK

Introduction

Comparative genomics is now considered as a powerful tool to investigate species and individual variations in a functional and evolutionary context. The expansion of our knowledge on the organization of human genome and how it functions in health and disease benefits a lot from comparative studies with both distantly and closely related species (O'Brien et al. 1999). For example, comparison between genome sequences of human and other mammals has provided valuable insights into details of genomic rearrangement events (Pevzner and Tesler 2003) and identified primate-specific functional elements (Boffelli et al. 2003). Large-scale sequencing of vertebrate genomes has been initiated and some of them have or nearly have been accomplished over the past 15 years. To date, there are, in total, five mammals including

two primates, two rodents, and one carnivore species with relatively high-quality genome data (sequence coverage >threefold) released and detailed comparative analysis published (International Human Genome Sequencing Consortium 2001; Mouse Genome Sequencing Consortium 2002; Kirkness et al. 2003; Rat Genome Sequencing Project Consortium 2004; The Chimpanzee Sequencing and Analysis Consortium 2005). As for artiodactyls, the 0.66-fold draft sequence and corresponding analysis of pig have been recently published (Wernersson et al. 2005) and the sequencing project of cow (see http://www.hgsc.bcm.tmc.edu/projects/bovine/) is on its way. In respect that whole genome shotgun sequencing is high-cost and time-consuming, intermediate grade of finished genome sequence (Blakesley et al. 2004; Wernersson et al. 2005) or sequencing targeted region of the investigated species (Thomas et al. 2003) has turned out to be a cost-effective strategy which well meets the requirement of comparison and decoding of the genome under investigation.

We applied such strategy to the comparative analyses of the Indian muntjac (*Muntiacus muntjak vaginalis*) genome. Because of the unique karyotype and the phylogenetic position as an artiodactyl, such study is destined to contribute to comparative genomics. The Asian muntjacs have attracted the attention of many biologists because they exhibit the greatest chromosomal diversity among mammals, and several new muntjac species have been discovered since the 1980s (Ma et al. 1990; Evans and Timmins 1994; Giao et al. 1998; Aamato et al. 1999). Despite the high similarity in morphology and ability to mate among different species (Shi and Pathak 1981), the karyotypes in *Muntiacus* range from $2n$=6/7 (Wurster and Benirschke 1970) of the Indian muntjac to $2n$=46 (Wurster and Benirschke 1967) of the Chinese muntjac (*Muntiacus reevesi*). The Indian muntjac possesses the lowest known diploid chromosome number in mammals. The earliest fossil record of the Chinese muntjac dates back to the early Pleistocene (about two million years ago), while that of Indian muntjacs appeared in the late Pleistocene (about half a million years ago) (Ma et al. 1986), which was also supported by our molecular phylogenetic study (Wang and Lan 2000). The great changes of karyotypes of the muntjac species in such a short time have made them a valuable model for studying karyotypic and genomic evolution and their relationships with speciation events.

Besides the chromosomal number reduction in *Muntiacus*, early estimation of genome size by flow cytometry in the muntjac species also revealed that the haploid C-value of the Indian muntjac is 2.22 (about 2.17 Gb of genome size), whereas, the Chinese muntjac is 2.85 (Johnston et al. 1982; Wurster and Atkin 1972). This value of the Indian muntjacs is relatively small compared to other mammals (the C-values of humans, mice, dogs and cows are 3.50, 3.25, 2.80, and 3.70, respectively), and indeed, it is the lowest among the cervid species measured so far (see http://www.genomesize.com/results.php?page=1). Comparative cytogenetic studies have suggested that the ancestor of *Muntiacus* may have had a karyotype similar to that of the Chinese muntjac or a putative karyotype of $2n$=70 karyotype similar to that of *Hydropotes inermis* (Elder and Hsu 1988; Fontana and Rubini 1990; Yang et al. 1997a). Extensive tandem fusions, together with a few Robertsonian fusions, have mainly accounted for the reduction in chromosome numbers from the ancestral karyotype to the $2n$=6/7 of the Indian muntjac (Elder and Hsu 1988; Yang et al. 1995, 1997b), during which heterochromatin tends to be lost (Elder and Hsu 1988; Johnston et al. 1982). It is conceivable that the dynamics of "junk" DNA should correlate with variation of genome size, as for a broad range of vertebrates, noncoding and repetitive DNAs occupy great proportions of genome (Britten and Kohne 1968). For instance, the whole genome analysis of *Fugu rubripes*, which has the smallest vertebrate genome size measured to date, has demonstrated that an extremely low repeat content (<15%) and reduction of intron size dominantly contribute to the genome compression (Aparicio et al. 2002). Similarly, after examining the *indel* spectrum of non-long terminal repeat (LTR) elements in insects, Petrov et al. (2000) attributed the genome size reduction in *Drosophila* to a relatively higher rate of DNA loss than in crickets.

In this study, we have explored the Indian muntjac genome by a combination of bacterial artificial clones (BAC) sequencing and comparative analyses with other species. Firstly, we investigated whether such "junk" DNAs as introns are associated with compact genome size in the Indian muntjac, like in *Drosophila* and *Fugu* fish, as they were shown to be directly involved in evolution of genome size and influence chromosome evolution by participating in chromosome breakage, deletion, and inversion (Dimitri et al. 1997; Lim and Simmons 1994). Secondly, this study provides the second large scale comparative analyses of assembled draft sequence of artiodactyls besides the recently published 0.66-fold draft sequence of pigs (Wernersson et al. 2005). The comparative genomic study of the Indian muntjac with other mammals will provide a unique opportunity to understand the dynamics of genome architecture in the *Muntiacus* species.

Materials and methods

BAC localization and size estimation

Fibroblast cell line from a male Indian muntjac (*M. m. vaginalis*, KCB8002) with normal karyotype was used for constructing the genomic BAC library (Chi et al. 2005). The location of each BAC on a chromosome was determined by fluorescent in situ hybridization (FISH) as

previously described (Yang et al. 1997a). After that, signal detection was performed as follows: biotin-labeled probes were visualized with Cy3-avidin (1:500) while fluorescein isothiocyanate (FITC)-labeled probes with rabbit anti-FITC (1:200) and goat-anti-rabbit-FITC antibodies (1:100) as described previously (Chi et al. 2005). At last, FISH images of mounted slides were captured using the Genus system (Applied Imaging) from a cooled charge-coupled device (CCD) camera mounted on a Zeiss microscope (Axioplan2). Hybridization signals were assigned to specific chromosome regions defined by enhanced 4′6-diamidino-2-phenylindole (DAPI)-banding patterns.

Meanwhile, to estimate the insert size of selected clones, we purified the plasmid DNA with QIAGEN Large-Construct Kit and then digested it with 5 U *Not* I (New England BioLabs) overnight at 37°C. The digested DNA was subjected to pulsed-field gel electrophoresis (PFGE) at 14°C with an electric intensity of 6 V/cm for 10 h. The insert size for each BAC was finally estimated based on the low range pulsed-field gel (PFG) marker (New England BioLabs).

BAC sequencing, assembling and mapping

Thirty-seven BACs distributed at different regions of the muntjac chromosomes were selected and sequenced using a standard shotgun strategy. Purified plasmid DNA was first sheared with ultrasonic cell cruiser (NingBo Scientz Biotechnology) or Hydroshear instrument (GeneMachines). Resulted fragments were size-selected and extracted by agarose gel electrophoresis (QIAquick Gel Extraction Kit) and then inserted into pUC18 vector with T4 ligase (Promega). Sequence reads of randomly selected subclones were generated from single end except for three BACs (*bsbp*, *bsbq*, *bsbr*) of which both ends were generated. After generating more than threefold sequence redundancy, 17,615 high-quality reads were subsequently edited and assembled using Phred/Phrap/Consed program package (Ewing and Green 1998; Ewing et al. 1998; Gordon et al. 1998). Resultant contigs were then joined into scaffolds based on read-pair association and order information from shotgun clones (see Electronic Supplementary Material S1).

Then interspersed repeats were masked for the scaffolds of each BAC by RepeatMasker (version open-3.0.5, see http://www.repeatmasker.org) using appropriate muntjac repeat library (Repbase Update 9.11 RM database version 20050112, see http://www.girinst.org/Repbase_Update.html) (Jurka 2000). Masked scaffolds were compared to genome sequences (http://www.hgdownload.cse.ucsc.edu/goldenPath/bosTau1/bigZips/) of human (hg17, May 2004), mouse (mm5, May 2004), dog (canFam1, July 2004), and available scaffolds of cow (bosTau1, May 2004, http://www.hgsc.bcm.tmc.edu/projects/bovine/) using National Center for Biotechnology Information (NCBI) basic local alignment search tool (BLAST) algorithms (version 2.2.4) with default parameters (Altschul et al. 1990). Blast results were first clustered based on colinearity for further screening. Manual inspections readily identified spurious hits mainly result from repetitive sequences not identified by RepeatMasker and poor hits (identical base pairs < 100) due to lineage divergence. After removing these hits mentioned above, we ordered and oriented as many scaffolds as we could on comparison with genomes of other four mammals. Scaffolds with unique position on chromosome were then joined together and gaps between which were represented with 200 Ns.

Gene annotation

The process of muntjac gene predictions was mainly based on the Ensembl pipeline of gene annotation with some modifications (Curwen et al. 2004). In brief, three gene prediction algorisms were performed independently: FgenesH (Solovyev et al. 1995), TBLASTN (Altschul et al. 1990), and GeneWise (see http://www.sanger.ac.uk/Software/Wise2/) (Birney et al. 2004). The results were then integrated together to yield a summary of annotated genes.

FgenesH requires no input other than the assembled sequence of the muntjac as a procedure of ab initio prediction. TBLASTN allows similarity search against peptides database, and it was used with an E value of 1×10^{-6} as the cutoff for identifying potential human orthologous genes in the Ensembl protein dataset (ftp://ftp.ensembl.org/pub/current_homo_sapiens/data/fasta/pep/). A total of 690 human candidate peptides were produced, of which more than 25% of peptide length were maintained in those hits aligned with the muntjac genome. This part of peptides was subsequently subject to GeneWise against the muntjac genome sequences which additionally takes splice sites and frameshifts into account. In model organisms, the annotation results are validated and refined based on transcription data including full-length cDNA or expressed sequence tag (EST) sequences (Clark et al. 2003; Misra et al. 2002). As such a comprehensive resource is not available for the Indian muntjacs, we aligned ESTs and mRNAs of cow (http://hgdownload.cse.ucsc.edu/goldenPath/bosTau1/bigZips/) which is closely related to the muntjacs with BAC sequences as our final step of annotation. The best predictions are considered as overlapping results of above annotation steps. This set of predictions was then manually filtered for redundancy and resulted from alternative splicing or paralogous genes (see Electronic Supplementary Material S2).

Substitution rates calculation

Due to high probability of lack of selective constraint on pseudogenes, involvement of such genes will elevate overall substitution rates during the calculation. Nineteen putative genes were, thus, selected from refined annotation results upon stringent criteria to prevent potential influences of pseudogenes. First, according to the GeneWise results, those genes which were characterized with a score higher than 100 and without any stop codon in coding regions were retained. Second, conservatively, a predicted gene without any intron was excluded because it may be a candidate of processed pseudogene. Finally, 16 of these 19 genes have blast hits with transcription data of the cow, which suggests their probable roles of real genes. Sequences of orthologous genes in other mammals were directly retrieved from Ensembl ortholog prediction (see http://www.ensembl.org). The coding regions of the four mammals were clustered and aligned by ClustalW (Thompson et al. 1994). After that, the alignments were modified manually to ensure gap-free alignments. We used codeml in PAML software package v3.14 (Yang 1997) under free-ratio model (model=1) assuming no molecular clock (clock=0) to analyse the sequence and estimate the substitution rates for each gene. A generally accepted tree topology (Murphy et al. 2004; Thomas et al. 2003) for the investigated species (runmode=0) was adopted for calculating branch lengths with baseml under general-time-reversible model (model=7).

Results and discussion

Sequence assembly

A high-redundancy BAC library of the Indian muntjac comprising 124,800 BAC clones (Chi et al. 2005) was used in this study. Of the 403 randomly selected BAC clones that have been mapped to the chromosomes of the Indian muntjac by FISH, 37 BACs were subject to subsequent sequencing based on their positions to generate a representative dataset for the muntjac genome.

The shotgun sequencing effort on the 37 BACs at Beijing Genomics Institute (BGI) resulted in 17,615 distinct high-quality reads whose total length is 8.43 Mb. Combined with the estimated length for each BAC, we determined that the average coverage for draft sequences of these BACs is approximately 3.05-fold. Further assembly generated 1,275 contigs with an average N50 size of 3.08 kb, representing a total of 2.70 Mb nonredundant sequences of the Indian muntjac genome (see Electronic Supplementary Material S1). We ordered and oriented scaffolds derived from contigs into superscaffolds for each BAC based on alignments with genomes of other mammals including human, mouse, and dog (http://www.genome.ucsc.edu/). Particularly, the availability of draft sequences of the cow (*Bos taurus*) from the ongoing sequencing project (bosTau1, see http://www.hgsc.bcm.tmc.edu/projects/bovine/) which is the most closely related to the Indian muntjac among mammals with genome sequences released to date, greatly helped us to refine and check the order of scaffolds for each BAC. Correspondingly, we didn't integrate the sequence of another artiodactyl species (pig) due to its low sequencing coverage. We could eventually get scaffolds joint with gaps for 35 BACs constituting nonredundant sequence dataset with a total length of 2.20 Mb (Electronic Supplementary Material S1). We were not able to localize and orient scaffolds for two BACs, *bsat* and *bsaq* (Accession Number: DQ280186 and DQ280187, respectively), due to the poor alignments with genome sequences of other species. This is probably due to either fast evolutionary rate of sequences of these two BACs or the relatively low sequencing coverage. Gene annotation process with these scaffolds resulted in 55 putative genes (see Electronic supplementary material S2). It is noteworthy that no inconsistency for the arrangement of scaffolds was detected during comparison with different genomes, which suggested the order of assembled scaffolds are reliable and implies that genomic rearrangement barely occurs at the level of BACs (~100 kb) in the Indian muntjacs.

Comparative mapping results

To provide unambiguous anchoring information for each BAC and uncover sequence similarity pattern between the muntjac and other species, we masked all the repetitive elements from scaffolds of 37 BAC clones with RepeatMasker (version open-3.0.5, see http://www.repeatmasker.org) before performing BLAST search against human, mouse, and dog genomes. We found that the dog genome has the highest sequence similarity with the masked superscaffolds of the muntjac, representing 12.2% of the masked muntjac genome sequences that formed alignments with the dog. This proportion decreased to 9.8% in the human and 3.6% in the mouse, which probably correlates with the shorter divergence time between carnivore and artiodactyls (Murphy et al. 2004) and relatively higher mutation rate of the mouse (Mouse Genome Sequencing Consortium 2002).

Previous comparative genomic studies mostly focused on unveiling chromosomal mechanisms underlying the karyotypic evolution of the Indian muntjacs by comparative chromosome painting between the muntjacs and its closely (Yang et al. 1995, 1997a; Fronicke et al. 1997) or distantly (Yang et al. 1997b; Fronicke and Scherthan 1997) related species, which identified homologous regions and putative ancestral fusion sites of the muntjacs' chromosomes.

Table 1 Locations of each BAC on chromosomes of different species

GeneBank accession no.	BAC ID[a]	Indian muntjac[b]	Human[c]	Mouse[d]	Dog
DQ280153	bsaa	chr1-1b	chr1q23.3	chr1	chr38
DQ280154	bsab	chr1-4c	chr3p25.2	chr6	chr20
DQ280155	bsac	chr2-15	chr15q26.3	chr7	chr3
DQ280156	bsae	chr2-2c	chr16p13.13	**chr1**	chr6
DQ280157	*bsaf*	chr2-10	chr10q22.2	chr14	chr4
DQ280158	*bsag*	chr1-22	chr10p15.1	**chr13**	chr2
DQ280159	bsah	chr1-12	chr8q24.3	chr15	chr13
DQ280160	bsaj	chr(3+X)-X	chrXp22.2	chrX	chrX
DQ280161	bsak	chr1-3a	chr3p24.3	chr17	chr23
DQ280162	bsam	chr1-4a	chr18q12.3	chr18	chr7
DQ280163	bsan	chr1-4a	chr18q12.1	chr18	chr7
DQ280164	bsao	chr1-4a	chr18p11.22	chr18	chr7
DQ280165	bsap	chr1-4a	chr18p11.21	chr18	chr7
DQ280166	*bsar*	chr1-3a	chr1q31.2	**chr15**	chr38
DQ280167	*bsas*	chr1-3a	chr9q33.2	chr2	chr9
DQ280168	bsaw	chr1-3a	chr2q24.3	chr2	chr36
DQ280169	bsax	chr2-1a	chr12q23.3	chr10	chr10
DQ280170	*bsay*	chr2-1a	chr22q13.31	chr15	chr10
DQ280171	bsaz	chr2-1a	chr12p12.1	chr6	chr27
DQ280172	bsba	chr(3+X)-8	chr14q22.2	chr14	chr8
DQ280173	bsbb	chr(3+X)-8	chr15q22.31	chr9	chr30
DQ280174	bsbc	chr(3+X)-18	chr3p12.1	chr16	chr31
DQ280175	bsbe	chr(3+X)-8	chr15q23	chr9	chr30
DQ280176	bsbf	chr1-3b	chr6q25.1-25.2	chr10	chr1
DQ280177	*bsbg*	chr1-3b	chr8q24.22	chr15	chr13
DQ280178	bsbh	chr1-3b	chr6q23.2	chr10	chr1
DQ280179	*bsbj*	chr1-3b	chr3q22.1	chr9	chr23
DQ280180	*bsbk*	chr1-17	chr2p12	chr6	chr17
DQ280181	bsbl	chr1-17	chr17q23.2	chr11	chr9
DQ280182	bsbm	chr1-17	chr17q12	chr11	chr9
DQ280183	bsbn	chr(3+X)-8	chr14q24.3	chr12	chr8
DQ280184	bsbo	chr1-17	chr17q11.2	chr11	chr9
DQ280185	bsbp	chr(3+X)-8	chr14q24.3-31.1	chr12	chr8
DQ280188	bsbq	Fusion site	chr19q13.42	**chr11**	chr1
DQ458964	bsbr	chr(3+X)-9	chr13q22.2	chr14	chr22

[a] Eight BACs which shows syntenic relationship not reported before are shown in italics
[b] Location for each BAC is designated as chromosome number followed by a specific region ID shown in Chi et al. 2005
[c] Locations in other species are just shown as chromosome numbers or specific band information retrieved from NCBI mapview (http://www.ncbi.nlm.nih.gov/mapview/)
[d] Four BACs with inconsistent mouse–human synteny relationship are shown in bold

However, due to the limitation in resolution, chromosome painting is not adequate for assessing less conservative regions or syntenic relationship at a fine scale. Our data provide a unique opportunity for a detailed investigation of comparative genomic map between the Indian muntjac and other mammals.

We localized each BAC on the corresponding chromosomal regions of the human, the mouse, and the dog genomes (Table 1). The anchoring information provided by BLAST against genomes of these mammals revealed that a great part of the results are in accordance with previous comparative maps obtained from chromosome painting (Yang et al. 1997b) or in silico exploration (Mouse Genome Sequencing Consortium 2002; Kirkness et al. 2003). However, eight BACs exhibited homologous relationships with specific regions of the human chromosomes that have not been reported before (Table 1). Furthermore, as we aligned the sequence of each BAC separately with the human, the mouse, and the dog and obtained their locations independently in each species, we also identified four BACs with inconsistent human–mouse syntenic relationship as compared with a previous work (Table 1, Mouse Genome Sequencing Consortium 2002). The former result is not surprising in respect that cross-species chromosome painting lacks resolution for blocks less than 4 Mb in size (Murphy et al. 2004), while the latter inconsistencies

Table 2 Comparison of average intron lengths of muntjac and other mammals

	Muntjac	Human	Muntjac	Dog	Muntjac	Mouse
Investigated intron	87	95	80			
Average intron length (bp)	691	1,755	600	1,536	554	1,311

probably result from higher substitution rates of the mouse (Mouse Genome Sequencing Consortium 2002) or distant divergence between rodents and artiodactyls (Murphy et al. 2004). Three of these four BACs formed poor alignments (total alignable length <1 kb) with the mouse genome, which at last, led to the ambiguity of mapping results.

Significant reduction of intron size

The non-coding DNAs, such as introns, have been documented to sculpt the genome size. The correlation between intron size and genome size has been widely studied and verified across broad phylogenetic ranges including plants (Wendel et al. 2002), *Drosophila* (Petrov et al. 2000), and pufferfish (Aparicio et al. 2002; McLysaght et al. 2000). As previously described, the estimated genome size of the Indian muntjac is relatively small and revealed a compression of 36.6% in the human, 20.0% in the dog, and 40% in the cow. We annotated 54 genes in combination with EST data of the cow (see Materials and methods) based on the assembled sequences of these 37 BAC clones. This part of the muntjac genes enabled us to explore the impact of intron size on genome size.

The intron positions of the muntjac genes were determined by GeneWise (Birney et al. 2004) (see http://www.sanger.ac.uk/Software/Wise2/), which is widely imple-

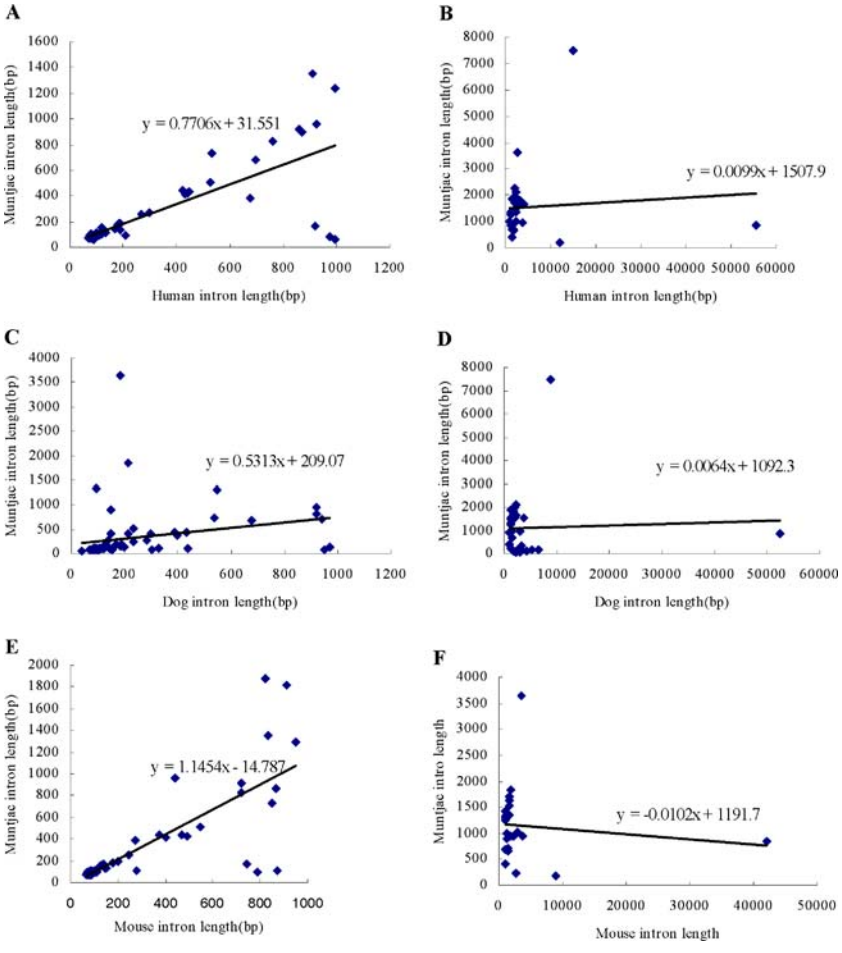

Fig. 1 Intron length comparison of Indian muntjac with other mammals. The regression line for all data is shown. **a**, **c** and **e** depict muntjac introns compared with human, dog, and mouse of which intron sizes are smaller than 1 kb. **b**, **d** and **f** involve those introns of human, dog, and mouse which are longer than 1 kb. It is evident that the slope of regression line in **a**, **c** and **f** is significantly larger than those in **b**, **d** and **f**, respectively

Table 3 Pairwise calculation of mean nonsynonymous substitutions (d*N*) in Muntjac and other mammals

	Human	Mouse	Dog	Muntjac
Human				
Mouse	0.1697 ± 0.2690			
Dog	0.1484 ± 0.2924	0.1222 ± 0.1329		
Muntjac	0.1333 ± 0.1351	0.2303 ± 0.2582	0.2020 ± 0.2826	

mented as a useful tool to exquisitely predict gene structure by similarity search against proteins. Fifteen genes with high reliability (GeneWise score >100) were selected from nonredundant predicted gene dataset. Orthologous genes in human, mouse, and dog were directly retrieved from Ensembl ortholog prediction dataset (see http://www.ensembl.org/). Pairwise comparison of intron lengths of the putative muntjac genes and their orthologs in other species were then performed. Excluding those with sequencing gaps between contigs or scaffolds, a total of 87 gap-free introns were used for comparison with their human orthologs. A more than twofold reduction of average intron size in the Indian muntjac was observed compared to the human, the dog, and the mouse (Table 2). Due to the non-normal distribution of intron size (Yu et al. 2002), we adopted the Wilcoxon Signed Ranks test and detected a significant difference with the human (87 introns, p=0.001) and the dog (95 introns, p=0.001). Although the test is not significant with the mouse (80 introns, p=0.369), which may be due to great variance of the mouse intron size, the mean intron sizes of the mouse genes is still more than two times larger than those of the muntjac genes.

We additionally looked into intron pairs with lengths higher and lower than 1 kb separately. The slope of regression curve for introns longer than 1 kb is smaller than that of introns shorter than 1 kb (Fig. 1). Seventy-nine percent of introns which are longer than 1 kb in the human become short in the muntjac. But for those introns shorter than 1 kb, only 54% become short in the muntjac. Similar results were also detected in the dog and the mouse. These results indicated that introns with large sizes might be more susceptible to deletion (Fig. 1). Altogether, our results demonstrated that the muntjac genes are characterized by a significant reduction in intron sizes and that genome size evolution also takes place within genes, which is consistent with the conclusion drawn in other species (Aparicio et al. 2002; Petrov et al. 2000; Wendel et al. 2002).

Rapid evolution of genes in the Indian muntjac

To better understand gene evolution patterns in the Indian muntjac, we selected reliably annotated muntjac genes and compared them with their orthologs in other mammals. Substitution rates at both synonymous and nonsynonymous sites were computed separately for each gene using their coding region sequences by PAML (Yang 1997, Table 3). The branch lengths of gene trees were estimated (Fig. 2) based on concatenated sequences of these genes.

Interestingly, we observed an elevated substitution rate in the muntjac across the 19 genes under inspection. Evolutionary distance estimated from synonymous (d_S) and nonsynonymous substitutions (d_N) indicates that the muntjac lineage possesses the longest branch length (Fig. 2). When we adopted the model that allows d_N/d_S ratio to vary along phylogenetic branches (model=1, free d_N/d_S ratio model) to look into substitution rates separately, we didn't detect significant sign of positive selection along the branch leading to the muntjac from the common ancestor of the dog and the muntjac (d_N/d_S=0.5585, p<0.01). But we do detect that the d_N value (0.1290) on this branch is the highest among all the branches and d_S value (0.2310) is approximate to the highest one (0.2790) on the mouse lineage. The χ^2 test with df=5 also indicates that this model fits the data significantly better than the one-ratio model (model=0, p=0.000989) which assumes constant rate along branches, suggesting that the muntjac genes evolved rapidly as reported in the mouse (Mouse Genome Sequencing Consortium 2002).

Such a pattern probably correlates with the significant short intron size in the Indian muntjac as has been shown in

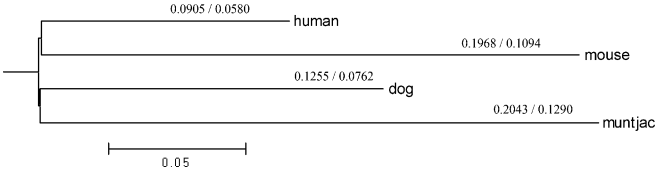

Fig. 2 Gene trees based on 19 concatenated genes from human, mouse, dog, and muntjac. Gene tree was constructed by *PAML* (Yang 1997) based on gap-free alignments of 19 genes. The length of each branch was calculated by baseml and d_N calculated by codeml. These two values were depicted as branch-length/d_N on the corresponding branches

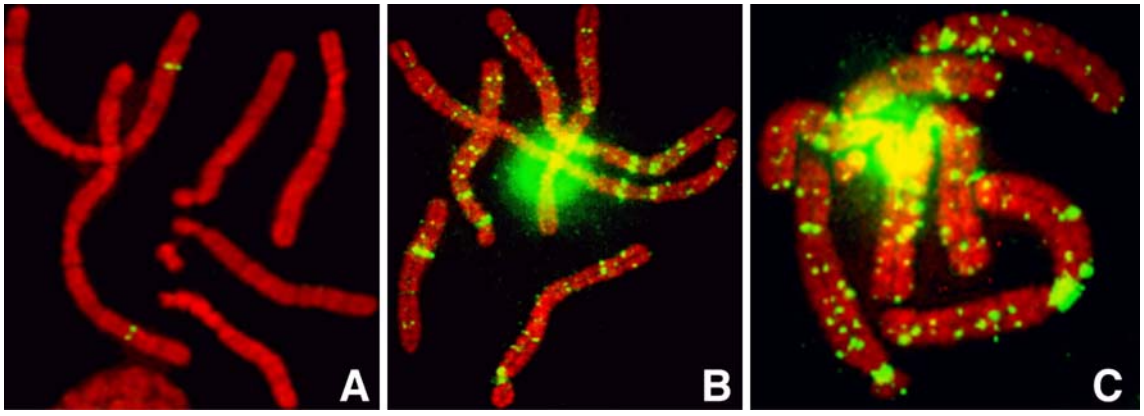

Fig. 3 Examples of fluorescent in situ hybridization (*FISH*). **a** *FISH* result using *BAC* clone *bsah* as the probe disclosing a specific signal at an Indian muntjac chromosome region. **b** *FISH* result for *BAC bsbq*, showing multiple signals at the putative chromosome fusion sites. Enhanced signal was specifically detected on *X*+3 chromosome and shown as green facular in the center of the figure. **c** *FISH* result using probes generated from the *PCR* products amplified by telomeric and centromeric specific primers

Drosophila (Marais et al. 2005). Similar to our results in the Indian muntjac, a weak negative correlation of intron size and evolutionary rate (measured by d_N) was found by analyses of 570 gene pairs in *Drosophila melanogaster* and *Drosophila yakuba*. Further investigation of expression profiles of these genes revealed positive correlation between the size of the first intron and expression levels. The author, thus, attributes such correlation to the selection on potential regulatory elements existing within an intron (Marais et al. 2005). Further validation of this hypothesis may be reached in the Indian muntjac by obtaining more knowledge about the expression of these muntjac genes.

Characterization of a BAC containing a chromosome fusion point

As previously described, *Cervidae* species may share a common ancestor with the putative karyotype of 2*n*=70 (Fontana and Rubini 1990), and tandem fusion (chromosomes fuse head to tail) process was considered as the dominant chromosomal rearrangement event in the *Muntiacus* genus (Lin et al. 1991; Lee et al. 1993; Yang et al. 1997b; Hartmann and Scherthan 2004). Interestingly, we found that a BAC clone (*bsbq*, see Table 1) probably comprises remnants of an ancient fusion event of the muntjac chromosomes, which is supported by two levels of evidence. First, this BAC revealed multiple FISH signals (Fig. 3b) at several putative chromosome fusion sites defined previously (Yang et al. 1997b). The second line of evidence is from the colocalization of centromeric and telomeric repeats in this BAC clone. Interstitial $(T_2AG_3)_n$ telomeric repeat (Lee et al. 1993) and centromeric satellite such as C5 (Lin et al. 1991) have previously been detected at the boundaries of these putative fusion sites. To detect such repeats in this BAC, we first extensively sequenced this BAC clone (4.76-fold sequencing coverage). Although all other gaps have been filled by polymerase chain reaction (PCR) amplification and subsequent sequencing, there still is a large one left for which we failed to close, indicating a presence of a highly repetitive sequences in this region. To

Table 4 Similar sequences in the public database with the TeloC-SatIA PCR product

Aligned length of PCR product (bp)	Identity	Accession no.	Description	Reference
574	91%	AY380827 (1,391 bp)	*Muntiacus reevesi micrurus* clone FM-satI 1.5 kb (2-2) satellite I DNA sequence	Lin et al. 2004
305	90%	AY322158 (400 bp)	*Muntiacus muntjak vaginalis* clone TGS400 telomere and satellite repeat sequence	Hartmann and Scherthan 2004
195	91%	AY322159 (225 bp)	*Muntiacus muntjak vaginalis* clone TGM225 telomere and satellite repeat sequence	Hartmann and Scherthan 2004
227	86%	CEU53516 (680 bp)	*Cervus elaphus canadensis* centromeric satellite DNA	Lee et al. 1997

overcome this problem, we then exploited PCR strategy with combinations of different centromeric and telomeric primer pairs described before (Hartmann and Scherthan 2004). Using plasmid DNA of this BAC clone as template, we obtained products from two pairs of primer combination. One pair is two telomere-specific primers $(C_3TA_2)_3$ (TeloC) and $(T_2AG_3)_3$ (TeloG), the other is $(C_3TA_2)_3$ (TeloC) against the Indian muntjac DNA satellite IA (SatIA, Bogenberger et al. 1982). The former G- and C-rich telomere primers resulted in a smear, which is consistent with the previous conclusion that the template contains telomeric repeats (Hartmann and Scherthan 2004). We obtained a 708-bp product (accession number: DQ377335) from the latter pair of primers, which was subsequently sequenced and subjected to BLAST search against the nucleotide database. This product shares high similarity with *M. reevesi* satellite I DNA sequence (accession number: AY380827, Lin et al. 2004) and previously identified telomeric and centromeric satellites sequence (Accession number: AY322158, Hartmann and Scherthan 2004) (Table 4). When hybridized to the Indian muntjac chromosomes, the probes generated from these PCR products also revealed both telomeric and centromeric signals (Fig. 3c). This provides substantial evidence for the presence of both centromeric and telomeric sequence in this BAC, which most likely represents the trace of a chromosome fusion event in the Indian muntjacs. Further in-depth investigation on this BAC and identification of the ligation point between the centromere and telomere will definitely help us to understand the molecular mechanism of how and why the muntjac chromosomes keep fusing.

Acknowledgements We thank Hongkun Zheng, Jun Li, Peixiang Ni, and Tao Feng in the Beijing Genomic Institute for assistance in preparing the perl scripts during the analysis. We also thank Huifeng Jiang, Yan Li, Yun Ding, Xin Li, and Haijing Yu in the Max-Planck Junior Research Group for incisive comments and discussions for preparing the manuscript and Xuebing Qi in the Comparative Genomics Group in Kunming Institute of Zoology for suggestions on performing PAML. This research was supported by key project grants of the Key Laboratory of Cellular and Molecular Evolution in Kunming Institute of Zoology and the National Natural Science Foundation of China (No. 30170506).

References

Aamato G, Egan, MG, Rabinowitz A (1999) A new species of muntjac, *Muntiacus putaoensis* (Artiodactyla: Cervidae) from northern Myanmar. Anim Conserv 2:1–7

Altschul SF, Gish W, Miller W, Meyers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215:403

Aparicio S, Chapman J, Stupka E, Putnam N, Chia J-m, Dehal P, Christoffels A, Rash S, Hoon S, Smit A, Gelpke MDS, Roach J, Oh T, Ho IY, Wong M, Detter C, Verhoef F, Predki P, Tay A, Lucas S, Richardson P, Smith SF, Clark MS, Edwards YJK, Doggett N, Zharkikh A, Tavtigian SV, Pruss D, Barnstead M, Evans C, Baden H, Powell J, Glusman G, Rowen L, Hood L, Tan YH, Elgar G, Hawkins T, Venkatesh B, Rokhsar D, Brenner S (2002) Whole-genome shotgun assembly and analysis of the genome of *Fugu rubripes*. Science 297:1301–1310

Birney E, Clamp M, Durbin R (2004) Genewise and GenomeWise. Genome Res 14:988–995

Blakesley RW, Hansen NF, Mullikin JC, Thomas PJ, McDowell JC, Maskeri B, Young AC, Benjamin B, Brooks SY, Coleman BI, Gupta J, Ho S-L, Karlins EM, Maduro QL, Stantripop S, Tsurgeon C, Vogt JL, Walker MA, Masiello CA, Guan X, Program NCS, Bouffard GG, Green ED (2004) An intermediate grade of finished genomic sequence suitable for comparative analyses. Genome Res 14:2235–2244

Boffelli D, McAuliffe J, Ovcharenko D, Lewis KD, Ovcharenko I, Pachter L, Rubin EM (2003) Phylogenetic shadowing of primate sequences to find functional regions of the human genome. Science 299:1391–1394

Bogenberger J, Schnell H, Fittler F (1982) Characterization of X-chromosome specific satellite DNA of *Muntiacus muntjak vaginalis*. Chromosoma 87:9–20

Britten RJ, Kohne DE (1968) Repeated sequences in DNA. Hundreds of thousands of copies of DNA sequences have been incorporated into the genomes of higher organisms. Science 161:529–540

Chi JX, Huang L, Nie W, Wang J, Su B, Yang F (2005) Defining the orientation of the tandem fusions that occurred during the evolution of Indian muntjac chromosomes by BAC mapping. Chromosoma 114:167–172

Clark MS, Edwards YJK, Peterson D, Clifton SW, Thompson AJ, Sasaki M, Suzuki Y, Kikuchi K, Watabe S, Kawakami K, Sugano S, Elgar G, Johnson SL (2003) *Fugu* ESTs: New resources for transcription analysis and genome annotation. Genome Res 13:2747–2753

Curwen V, Eyras E, Andrews TD, Clarke L, Mongin E, Searle SMJ, Clamp M (2004) The Ensembl automatic gene annotation system. Genome Res 14:942–950

Dimitri P, Arca B, Berghella L, Mei E (1997) High genetic instability of heterochromatin after transposition of the LINE-like I factor in *Drosophila melanogaster*. PNAS 94:8052–8057

Elder FFB, Hsu TC (1988) Tandem fusions in the evolution of mammalian chromosomes. In: Sandberg AA (ed) The cytogenetics of mammalian autochromosomal rearrangements. Alan R Liss, New York, pp 481–506

Evans TD, Timmins RJ (1994) News from Laos. Oryx 29:3–4

Ewing B, Green P (1998) Base-calling of automated sequencer traces using phred.II. error probabilities. Genome Res 8:186–194

Ewing B, Hillier L, Wendl MC, Green P (1998) Base-calling of automated sequencer traces using phred.I. accuracy assessment. Genome Res 8:175–185

Fontana F, Rubini M (1990) Chromosomal evolution in Cervidae. Biosystems 24:157–174

Fronicke L, Scherthan H (1997) Zoo-fluorescence in situ hybridization analysis of human and Indian muntjac karyotypes (*Muntiacus muntjak vaginalis*) reveals satellite DNA clusters at the margins of conserved syntenic segments. Chromosom Res 5:254–261

Fronicke L, Chowdhary BP, Scherthan H (1997) Segmental homology among cattle (*Bos taurus*), Indian muntjac (*Muntiacus muntjak vaginalis*), and Chinese muntjac (*M. reevesi*) karyotypes. Cytogenet Cell Genet 77:223–227

Giao PM, Tuoc D, Dung VV, Wikramanayake ED, Amato G, Arctander P, MacKinnon JR (1998) Description of *Muntiacus truongsonensis*, a new species of muntjac (Artiodactyla: Muntiacidae) from central Vietnam, and implications for conservation. Anim Conserv 1:61–68

Gordon D, Abajian C, Green P (1998) Consed: a graphical tool for sequence finishing. Genome Res 8:195–202

Hartmann N, Scherthan H (2004) Characterization of ancestral chromosome fusion points in the Indian muntjac deer. Chromosoma 112:213–220

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

Johnston FP, Church RB, Lin CC (1982) Chromosome rearrangement between the Indian muntjac and Chinese muntjac is accompanied by a deletion of middle repetitive DNA. Can J Biochem 60:497–506

Jurka J (2000) Repbase update: a database and an electronic journal of repetitive elements. Trends Genet 16:418–420

Kirkness EF, Bafna V, Halpern AL, Levy S, Remington K, Rusch DB, Delcher AL, Pop M, Wang W, Fraser CM, Venter JC (2003) The dog genome: survey sequencing and comparative analysis. Science 301:1898–1903

Lee C, Sasi R, Lin CC (1993) Interstitial localization of telomeric DNA sequences in the Indian muntjac chromosomes: further evidence for tandem chromosome fusions in the karyotypic evolution of the Asian muntjacs. Cytogenet Cell Genet 63:156–159

Lee C, Court DR, Cho C, Haslett JL, Lin CC (1997) Higher-order organization of subrepeats and the evolution of cervid satellite I DNA. J Mol Evol 44:327–335

Lim JK, Simmons MJ (1994) Gross chromosome rearrangements mediated by transposable elements in *Drosophila melanogaster*. Bioessays 16:269–275

Lin CC, Sasi R, Fan YS, Chen ZQ (1991) New evidence for tandem chromosome fusions in the karyotypic evolution of Asian muntjacs. Chromosoma 101:19–24

Lin CC, Chiang PY, Hsieh LJ, Liao SJ, Chao MC, Li YC (2004) Cloning, characterization and physical mapping of three cervid satellite DNA families in the genome of the Formosan muntjac (*Muntiacus reevesi micrurus*). Cytogenet Genome Res 105:100–106

Ma SL, Wang YX, Xu L (1986) Taxonomic and phylogenetic studies on the genus *Muntiacus*. Acta Theriol Sinica 6:191–208

Ma SL, Wang YX, Shi LM (1990) A new species of the genus *Muntiacus* from Yunnan, China. Zool Res 11:47–52

Marais G, Nouvellet P, Keightley PD, Charlesworth B (2005) Intron size and exon evolution in *Drosophila*. Genetics 170:481–485

McLysaght A, Enright AJ, Skrabanek L, Wolfe KH (2000) Estimation of synteny conservation and genome compaction between pufferfish (*Fugu*) and human. Yeast 17:22–36

Misra S, Crosby MA, Mungall CJ, Matthews BB, Campbell KS, Hradecky P, Huang Y, Kaminker JS, Millburn GH, Prochnik SE, Smith CD, Tupy JL, Whitfied EJ, Bayraktaroglu L, Berman BP, Bettencourt BR, Celniker SE, de Grey AD, Drysdale RA, Harris NL, Richter J, Russo S, Schroeder AJ, Shu SQ, Stapleton M, Yamada C, Ashburner M, Gelbart WM, Rubin GM, Lewis SE (2002) Annotation of the *Drosophila melanogaster* euchromatic genome: a systematic review. Genome Biol 3:RESEARCH0083

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

Murphy WJ, Pevzner PA, O'Brien SJ (2004) Mammalian phylogenomics comes of age. Trends Genet 20:631–639

O'Brien SJ, Menotti-Raymond M, Murphy WJ, Nash WG, Wienberg J, Stanyon R, Copeland NG, Jenkins NA, Womack JE, Marshall Graves JA (1999) The promise of comparative genomics in mammals. Science 286:458–481

Petrov DA, Sangster TA, Johnston JS, Hartl DL, Shaw KL (2000) Evidence for DNA loss as a determinant of genome size. Science 287:1060–1062

Pevzner P, Tesler G (2003) Genome rearrangements in mammalian evolution: lessons from human and mouse genomes. Genome Res 13:37–45

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

Shi LM, Pathak S (1981) Gametogenesis in a male Indian muntjac x Chinese muntjac hybrid. Cytogenet Cell Genet 30:152–156

Solovyev VV, Salamov AA, Lawrence CB (1995) Identification of human gene structure using linear discriminate functions and dynamic programming. In: Proceedings of the Third International Conference on Intelligent Systems for Molecular Biology 3:367–375

Thomas JW, Touchman JW, Blakesley RW, Bouffard GG, Beckstrom-Sternbe SM, Margulies EH, Blanchette M, Siepel AC, Thomas PJ, McDowell JC, Maskeri B, Hansen NF, Schwartz MS, Weber RJ, Kent WJ, Karolchik D, Bruen TC, Bevan R, Cutler DJ, Schwartz S, Elnitski L, Idol JR, Prasad AB, Lee-Lin SQ, Maduro VV, Summers TJ, Portnoy ME, Dietrich NL, Akhter N, Ayele K, Benjamin B, Cariaga K, Brinkley CP, Brooks SY, Granite S, Guan X, Gupta J, Haghighi P, Ho SL, Huang MC, Karlins E, Laric PL, Legaspi R, Lim MJ, Maduro QL, Masiello CA, Mastrian SD, McCloskey JC, Pearson R, Stantripop S, Tiongson EE, Tran JT, Tsurgeon C, Vogt JL, Walker MA, Wetherby KD, Wiggins LS, Young AC, Zhang LH, Osoegawa K, Zhu B, Zhao B, Shu CL, De Jong PJ, Lawrence CE, Smit AF, Chakravarti A, Haussler D, Green P, Miller W, Green ED (2003) Comparative analyses of multi-species sequences from targeted genomic regions. Nature 424:788–793

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

Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 22: 4673–4680

Wang W, Lan H (2000) Rapid and parallel chromosomal number reductions in muntjac deer inferred from mitochondrial DNA phylogeny. Mol Biol Evol 17:1326–1333

Wendel JF, Cronn RC, Alvarez I, Liu B, Small RL, Senchina DS (2002) Intron size and genome size in plants. Mol Biol Evol 19:2346–2352

Wernersson R, Schierup M, Jorgensen F, Gorodkin J, Panitz F, Staerfeldt H-H, Christensen O, Mailund T, Hornshoj H, Klein A, Wang J, Liu B, Hu S, Dong W, Li W, Wong G, Yu J, Wang J, Bendixen C, Fredholm M, Brunak S, Yang H, Bolund L (2005) Pigs in sequence space: a 0.66X coverage pig genome survey based on shotgun sequencing. BMC Genomics 6:70

Wurster DH, Atkin NB (1972) Muntjac chromosomes: a new karyotype for *Muntiacus muntjak*. Experientia 28:972–973

Wurster DH, Benirschke K (1967) Chromosome studies in some deer, the springbok, and the pronghorn, with notes on placentation in deer. Cytologia 32:273–285

Wurster DH, Benirschke K (1970) Indian muntjac, Muntiacus muntjak: a deer with a low diploid chromosome number. Science 168:1364–1366

Yang Z (1997) PAML: a program package for phylogenetic analysis by maximum likelihood. Comput Appl Biosci 13:555–556

Yang F, Carter NP, Shi L, Ferguson-Smith MA (1995) A comparative study of karyotypes of muntjacs by chromosome painting. Chromosoma 103:642–652

Yang F, O'Brien PC, Wienberg J, Ferguson-Smith MA (1997a) A reappraisal of the tandem fusion theory of karyotype evolution in Indian muntjac using chromosome painting. Chromosom Res 5:109–117

Yang F, Muller S, Just R, Ferguson-Smith MA, Wienberg J (1997b) Comparative chromosome painting in mammals: human and the Indian muntjac (*Muntiacus muntjak vaginalis*). Genomics 39:396–401

Yu J, Yang Z, Kibukawa M, Paddock M, Passey DA, Wong GK-S (2002) Minimal introns are not "junk". Genome Res 12:1185–1189