Genome Biology

Evolutionary-new centromeres preferentially emerge within gene deserts

Genome Biology 2008, **9**:R173 doi:10.1186/gb-2008-9-12-r173

Mariana Lomiento (mlomiento@biologia.uniba.it)
Zhaoshi Jiang (zhaoshi@u.washington.edu)
Pietro D'Addabbo (p.daddabbo@biologia.uniba.it)
Evan E Eichler (eee@gs.washington.edu)
Mariano Rocchi (rocchi@biologia.uniba.it)

ISSN 1465-6906

Article type Research

Submission date 15 July 2008

Acceptance date 16 December 2008

Publication date 16 December 2008

Article URL http://genomebiology.com/2008/9/12/R173

Evolutionary-new centromeres preferentially emerge within gene deserts

Mariana Lomiento*, Zhaoshi Jiang†, Pietro D'Addabbo*, Evan E. Eichler†‡, Mariano Rocchi*

Addresses: *Department of Genetics and Microbiology, University of Bari, Via Amendola 165/A, Bari 70126, Italy. † Department of Genome Sciences, University of Washington School of Medicine and ‡Howard Hughes Medical Institute, 1705 NE Pacific St., Seattle WA 98195, USA

Correspondence: Mariano Rocchi. Email: rocchi@biologia.uniba.it

Abstract

Background: Evolutionary-New Centromeres (ENCs) result from the seeding of a centromere at an ectopic location along the chromosome during evolution. The novel centromere rapidly acquires the complex structure typical of eukaryote centromeres. This phenomenon has played an important role in shaping primate karyotypes. A recent study on the ENC of macaque chromosome 4 (human 6) showed that the ENC domain was deeply restructured, following the seeding, with respect to the corresponding human region assumed as ancestral. It was also demonstrated that the region was devoid of genes. We hypothesized that these two observations were not merely coincidental and that the absence of genes in the seeding area constituted a crucial condition for the ENC fixation in the population.

Results: To test our hypothesis we characterized 14 ENCs selected according to conservative criteria. Using different experimental approaches we assessed the extent of genomic restructuring. We then determined the gene density in the ancestral domain where each ENC was seeded.

Conclusion: Our study suggested that restructuring of the seeding regions is an intrinsic property of novel evolutionary centromeres that could be regarded as potentially detrimental to the normal functioning of genes embedded in the region. The absence of genes, which was found to be of high statistical significance, appeared as a unique favorable scenario permissive of ENC fixation in the population.

Background

The centromere is a complex structure ensuring the proper segregation of chromosomes in mitosis and meiosis. It usually harbors large blocks of satellite DNA (alpha satellite in primates). In spite of their complexity, centromeres have been shown to be able to relocate along the chromosome during evolution. These novel centromeres are referred to as Evolutionary-New Centromeres (ENCs). The first ENC examples supported by molecular cytogenetic techniques were described in non-human primates, in orthologs to human chromosome 9 [1]. Since then, several other examples were reported in primates and other taxa [2-10]. The phenomenon implies the seeding of the novel centromere and the inactivation of the old one.

The emergence of an ENC has been hypothesized to be epigenetic in nature, that is, not accompanied by any sequence transposition. This conjectural view is supported by indirect evidence, primarily by parallels with clinical cases of human neocentromeres. These are ectopic, analphoid centromeres usually originating in chromosomal acentric fragments allowing for their mitotic survival as supernumerary chromosomes (for a review see Marshall *et al.* [11]). They originate as opportunistic events, secondary to a chromosomal rearrangement. The latter circumstance has been regarded as a strong evidence of their epigenetic nature. The detrimental phenotypic consequences of the aneuploid status frequently incurred by neocentromeres is thought to limit germline transmission and is, therefore, analogous to ENCs. Recently, however, two familial transmissions of autosomal neocentromeres, occurring in apparently normal individuals with otherwise normal karyotypes, were described [5, 12]. They have been considered as ENCs at initial stages.

ENCs are relatively frequent. In macaque, for instance, nine out of 21 centromeres are evolutionary new [6]; in donkey at least five originated after a relatively short evolutionary timeframe since the donkey/zebra divergence (less than 1 million years) [8]. The relatively high number of ENCs could suggest a scenario where the absence of selective constraint allows ENC fixation. The finding, in humans, that neocentromeres do not affect gene expression [13-16] appears in line with this view.

The insight on the progression dynamics of the ENC of macaque chromosome 4 (MMU4, human 6), recently provided by Ventura *et al.* [6], has disclosed a potentially different evolutionary scenario in ENC formation. A DNA region of approximately 250 kb was pinpointed as the ENC seeding region and was shown to have been deeply affected by a variety of mutational processes including extensive duplication on both sides of the centromere, massive insertions of small

stretches of alpha-satellite DNA, and microdeletions inferred by absence of specific STS amplification. It could be supposed that this process would strongly antagonize ENC fixation because such structural variation would significantly affect the physical integrity of genes or regulatory elements located within the seeding region. Not surprisingly Ventura *et al.* [6] observed that this region was devoid of genes. We hypothesized that this observation was not coincidental but crucial in understanding the genomic context of ENC formation.

To test this hypothesis, 14 primate ENCs were analyzed in order to (i) ascertain the presence of novel segmental duplications (SD) around the ENC suggestive of a restructuring process of the kind reported by Ventura *et al.* [6]; (ii) survey the gene density in the seeding regions. Our analysis strongly suggested that the restructuring of the neocentromeric region is an intrinsic property of ENC progression and, consequently, the highly significant absence of genes we have observed may represent a critical pre-requisite for ENC progression and fixation in the population. The 14 seeding regions were also analyzed for AT content.

Results

Search for ENCs

Published studies and our unpublished data on chromosomal evolution in primates were surveyed in search for ENCs. Identification of 31 ENCs was made: 15 in Catarrhini (Old World monkeys [OWMs] and Hominoidea) and 16 in Platyrrhini (New World monkeys, [NWMs]). The vast majority of the NWM ENCs apparently emerged at the breakpoint of a chromosomal fission or repositioned from a telomere to the other telomere (see, for instance, the evolution of chromosome 3 [5]). Centromeres of human acrocentrics 15 and 14 are examples of ENCs that originated at a breakpoint and at a telomere, respectively, following a chromosomal fission [3]. Their short arms consist of several megabases of acquired sequences. These circumstances suggested that telomeric ENCs could represent a different ENC category, with different progression dynamics. We therefore excluded these ENCs from the analysis and focused our investigation on the ENCs that emerged inside a chromosome and were not concomitant to a disruption of the seeding region.

Fourteen ENCs met these conservative criteria: one in woolly monkey (*Lagothrix lagothricha*, LLA, Atelinae, NWM), eight in OWMs [6], one in white-cheeked gibbon (*Nomascus leucogenys*, NLE) [17], one in orangutan (*Pongo pygmaeus*, PPY) [18], and three in humans (*Homo sapiens*, HSA) [5, 18, 19]. The ENC that emerged on chromosome 7 (human 8) of woolly monkey (NWM) has not been previously published. The evolutionary history of chromosome 8, supporting the emergence of an ENC in this primate, is summarized in Supplemental Figure 1 [see Additional

data file #1]; FISH examples in Supplemental Figure 2a, b [see Additional data file #2]. BAC clones used in the analysis are reported in Supplemental Table 1 [see Additional data file #3]. The eight ENCs found in macaque (Cercopithecinae) are also present in the silvered leaf monkey (*Trachypithecus cristatus*, TCR, Colobinae), indicating that all ENCs originated in the Cercopithecinae/Colobinae common ancestor. The rhesus macaque was used as representative of OWMs because its genome has been fully sequenced [20].

Reiterative FISH experiments with corresponding human BAC clones were performed in non-human primate metaphases in order to precisely map these ENCs on the human sequence used as a reference (build35 assembly, March 2004) (Example in Supplemental Figure 2c [see Additional data file #2]). The macaque sequence was used as a reference for the three human ENCs (rheMac2 release, January 2006). The position of the human ENCs in macaque was defined using macaque BAC clones hybridized to human metaphases. The results are summarized in Table 1. In some cases a BAC generated split signals on both sides of the centromere (Table 1 in bold), while flanking BACs gave a single signal on the expected pericentromeric side. The sequence corresponding to the splitting BAC was flagged as the ENC seeding region. In other cases the position of the ENC was defined by two overlapping BACs mapping on opposite sides of the ENC.

Ancestral organization of regions where ENCs were seeded

The human regions orthologous to the sequence domains where the non-human ENCs were seeded were investigated for evolutionary conservation against mouse and dog genomes by visually inspecting the UCSC Comparative Genomics Net tracks [21]. The analysis was performed in order to validate the human sequence as *bona fide* reference sequence with respect to the changes the ENC regions underwent during evolution. We performed a similar comparative analysis for macaque regions corresponding to the three human ENCs for which the macaque was used as a reference. In both human and macaque sequences, the analysis encompassed approximately 2 Mb on each side of the seeding point. Substantial differences were found only in mouse [breaks or inversions at regions corresponding to human chromosome 2 (85.7-86.7 Mb and 137.6-137.7 Mb), chromosome 8 (61.9-62.8 Mb), and chromosome 11 (88.4-89.2 Mb)]. No rearrangements were found in the dog, with the exception of the cluster of Olfactory Receptor (OR) genes located at 121.5-122.3 Mb in human chromosome 9 and absent in dog. The human/dog concordance strongly suggest that these rearrangements are derivative in mouse.

Tempo of ENC seeding (essential primate phylogeny is reported in Figure 1)

As mentioned, all eight ENCs found in OWMs were present in both macaque (Cercopithecinae) and silvered leaf monkey (Colobinae) species. Therefore, all originated before Cercopithecinae/Colobinae divergence, estimated to have occurred 16 million years ago (mya) [22]. The position of the centromere on chromosomes orthologous to HSA2q (MMU12), HSA13 (MMU17), and HSA18 (MMU18) is shared by Hominoidea and NWMs [23] [personal unpublished data]. The ENC seeding on these chromosomes, therefore, occurred in OWM (Cercopithecoidea) after their divergence from Hominoidea, which was approximately 23 mya [22]. It was not possible to precisely define the upper temporal limit of the remaining OWM ENCs because the position of the centromere on orthologous NWM and Hominoidea chromosomes showed discrepancies [1, 5, 19].

The ENC on orangutan chromosome 11 is *Pongo*-specific [18] and is shared by both orangutan subspecies (*Pongo pygmaeus abelii* and *Pongo pygmaeus pygmaeus*). Consequently it was seeded within the interval 4-14 mya (between Pongidae/Hominidae and PPY *abelii*/PPY *pygmaeus* splits, respectively). The HSA11 ENC is, very likely, *Hominidae*-specific [18]. Thus it dates within the interval 8-14 mya (after Pongidae/Hominidae split and before gorilla-pan-homo divergence, respectively). HSA3 and HSA6 ENCs are shared by great apes, so they date prior to 8 mya. Uncertainty on the ancestral position of the centromere in these chromosomes impinges on the uncertainty of the upper temporal limit of their occurrence [5, 19]. For the ENC of the woolly monkey (LLA7, NWM, Atelidae), we could define only the upper temporal limit of 22-23 mya, which is the estimated divergence time of the Atelidae (LLA) and Cebidae (CJA) lineages [24].

Search for segmental duplications around ENCs

SD analysis was straightforward for the three human ENCs (chromosomes 3, 6 and 11) due to the high quality of the sequence assembly within these human pericentromeric regions [25]. Duplications were found in the pericentromeric regions of all three human chromosomes. On chromosome 6 and particularly on chromosome 3, intrachromosomal duplications predominate. The duplication status of the sequenced macaque and orangutan genomes is less accurate with respect to humans because of the severe limitations intrinsic to the whole-genome shotgun sequencing assembly (WGSA) approach [26] in resolving high-identity duplications (Note: whole genome sequence data are not currently available for the white-cheeked gibbon and woolly monkey).

To circumvent, at least in part, these problems, we exploited complementary bioinformatic and molecular cytogenetic techniques because they are partially “assembly independent”. First, we examined each of the ENC regions for the presence of recent duplications in various primates

using the whole genome shotgun sequence detection (WSSD) [27], where whole genome shotgun (WGS) reads from each primate are mapped against the human reference genome (hg17). Table 2 lists WSSD positive intervals detected for each primate species. Segmental duplications were detected, for example, on MMU4 (HSA6), MMU17 (HSA13), and PPY11 (HSA11). We then selected and tested various BAC clones by FISH. Some duplication data already resulted from experiments aimed at identifying the seeding region using human BAC clones (see above). However, split signals on both sides of the centromere could be alternatively interpreted as due to a disruption of distinct, non-duplicated sequences composing the human BAC, as a consequence of the colonization of alpha satellite DNA. Additionally, orthologous human clones may not be suitable for the analysis because of the restructuring process that could have substantially altered the pericentromeric sequences within each species. Final, new material, not represented in human BACs, may exist within these locations due to lineage-specific interchromosomal duplications.

Considering these potential biases, we also selected species-specific BAC clones identified with different approaches. For macaque we took advantage of the data on MMU BAC clones available at the Bioinformatics Research Laboratory of the Baylor College of Medicine, Houston, TX, USA [28]. For orangutan (PPY) and white-checked gibbon (NLE), we queried appropriate BAC-end sequences from CHORI-276 (PPY) and CHORI-271 (NLE) BAC libraries using the Trace Archive database of the NCBI [29]. A BAC library was not available for the woolly monkey (LLA). The phylogenetic distance of this NWM species coupled with the potential degenerative consequences of pericentromeric restructuring processes prompted us to discard the woolly monkey from the pericentromeric duplication analysis. Relevant FISH results of species-specific BAC clones yielding duplicated signals around the ENCs are reported in Table 3 (all tested clones in Supplemental Table 2 [see Additional data file #4]); examples are in Figure 2 and in Supplemental Figure 2e, f [see Additional data file #2]. We discovered pericentromeric duplications mapping near the centromeres for almost all ENCs. One BAC-end of asterisked BACs in Table 3 and in Suppelmental Table 2 was identified, by RepeatMasker, as entirely composed of 171 bp alpha satellite repeats. No internal repeat was found truncated, and the homology with the apha satellite consensus ranged from 75 to 90%.

Two findings were of particular interest. Four nearly overlapping human BACs (RP11-543A19, -1043D14, -539I23, and -527N12) covering a region of 1.3 Mb (chr13: 61,111,769-62,699,203) around the MMU17 ENC gave duplicated signals around the centromere. Additionally, the two human BACs defining the ENC of MMU2 (HSA3) are 319 kb apart (Table 1). Three BACs spanning this interval (RP11-1089F10, -1142P11, and -10O22) failed to give any FISH signals in

macaque, suggesting a deletion of the corresponding region within the macaque lineage. To exclude the possibility of technical artifact, we mixed on the same slide human and macaque metaphases, added an excess of probe, and extended the hybridization time for three days. Again in these conditions, no signal was detected in macaque metaphases, while strong signals were present in human metaphases. We performed a BLAST sequence similarity using the human 319 kb region as query against macaque sequences deposited in the Trace Archive database §§. Only very small stretches (less than 1 kb) of homologous DNA were found externally located with respect to a central chr3:164,271,000-164,461,000 region (190kb) in which no homology was detected (data not shown). Additionally, the macaque BAC clone CH250-91J4, identified at the Baylor College database (see above), mapping at HSA chr3:164,777,357-164,967,209, that is slightly external to the "deleted" region, failed to yield any signal in human metaphases (data not shown). Altogether these data strongly suggest that the region is highly rearranged in macaque.

Gene content at ENC regions

We carefully analyzed the human genome (used as reference for non-human ENCs) and the macaque genome (used as reference for the three human ENCs) for annotated genes mapping within or in proximity of ENC seeding regions. The analysis was performed by querying the human and macaque RefSeq-related tracks of UCSC genome browser [21] (hg17 assembly, RheMac2 assembly). No RefSeq genes were identified within the seeding regions as defined above (Table 1). In order to assess the statistical significance of gene depletion in the regions where ENCs were seeded, we performed a gene/exon density simulation (see Methods) for 14 ENC regions. We found that the gene/exon density of the 14 ENCs is significantly depleted ($p<0.0001$) when compared to random simulated data (see Figures 3 A-C). Table 4 reports the most proximal and distal RefSeq genes with respect to the ENC seeding point. The distance between the two genes is reported in the second column. Clusters of olfactory receptor genes flank the ENCs of chromosomes MMU14 (HSA11), MMU15 (HSA9), and HSA11 (MMU14). These OR clusters were not considered because OR genes are extremely redundant and a large number of these are pseudogenic within the primate lineage. The inactivation of a few of them would unlikely have strong phenotypical consequences. It is worth noting, in this context, that more than half of OR genes became inactive in recent human evolution [30].

AT content

The precise location of some human neocentromeres has been achieved through CENP-A mapping by ChIP-on-chip experiments (reviewed by Marshall *et al.* [11]). AT content has been shown to be

one of the few common features shared by these neocentromeres. We calculated the AT content for the human domains corresponding to the ENC seeding regions as defined in Table 1. The results are reported in the last column of Table 1.

Discussion

The organization, evolution and function of eukaryotic centromeres represent a deficiency in our understanding of genome biology. The discovery of human clinical neocentromeres and ENCs has further complicated, on one hand, our understanding of the centromere. On the other hand, neocentromeres and ENCs have allowed an initial dissection of the centromere complexity. They have made evident, for instance, its epigenetic nature. The ENC analysis we have accomplished in the present study has contributed to the identification of factors that, very likely, play a crucial role in ENC progression and fixation in the population. We have provided strong evidence that the pericentromeric duplication activity is an intrinsic property of ENCs. This conclusion was mainly supported by FISH experiments using species-specific BAC clones that detected SDs around the centromere in almost all studied ENCs. A deep restructuring was particularly evident in MMU17 (human 13) and MMU2 (human 3). The latter ENC showed a large deletion. This observation is not unexpected and could be generated by allelic non-homologous recombination occurring in one side of the centromere. Our overall results indicate that deep restructuring is a feature inherent to pericentromeric duplication activity triggered by the ENCs. Our analysis also indicated that species-specific probes are the most appropriate for detecting potential interchromosomal duplications (see ENCs of MMU12, 13, 14, 15 and 17).

Contrary to what we detected in the ENC of MMU4 (human 6), where SDs were strictly intrachromosomal [6], we found that SDs associated with other ENCs were both inter and intrachromosomal (see, for example Figure 2b). Pericentromeric analysis in humans has indicated that the majority of SDs are interchromosomal. It could be hypothesized that intrachromosomal duplications arose first, followed by interchromosomal ones. This interpretation, however, clashes with the finding, in humans, that the interchromosomal versus intrachromosomal SD ratio usually increases approaching the centromere, with the exception of few chromosomes [31]. Interestingly, three of these exceptions (chromosomes 3, 6 and, partially, chromosome 11) correspond to ENCs. It can be hypothesized that these differences could be a reflection of the age of the ENCs. Intrachromosomal occur first but then as centromeres become established they begin to exchange between non-homologous chromosomes, such that eventually interchromosomal duplications outnumber the intrachromosomal.

Studies on selected human neocentromeres have shown that the chromatin remodeling, accompanying the neocentromere seeding, does not alter gene expression [13-16]. By analogy with ENC, the presence of genes would not negatively affect, *per se*, the ENC functioning. Our studies suggest that the subsequent duplication activity, implying deep restructuring, would, on the contrary, antagonize the ENC fixation. In this scenario, the only condition compatible with ENC fixation in the population would be either the lack of genes in the ENC seeding region or the presence of multi-copy gene family where loss would be tolerated. The study provided strong support for this scenario: the ENC seeding regions we have examined are significantly depleted of genes. The MMU17 (HSA13) ENC is of relevance in this context. It exhibits the largest gene desert (4.9 Mb) and one of thee largest duplicated regions (1.3 Mb). The non-casual matching is further reinforced by the analysis of the pattern of SDs around this repositioned centromere in three distinct regions showing large-scale variation in OWM species as reported by Cardone *et al.* [23]. This extensive variation could be interpreted as further evidence of relaxed constraint on duplication activity due to the large size of the gene desert.

In an individual heterozygous for an ENC, a meiotic exchange within the region delimited by the old and the novel centromeres would produce dicentric and acentric chromosomes, mimicking the consequences of a pericentric inversion. These events are supposed to affect the fitness of heterozygous carriers negatively. Meiotic drive in females in favor of the repositioned chromosome is a possible explanation, as reported for Robertsonian fusion in humans [32]. Genetic drift and population structure can also be hypothesized to have played an important role in neocentromere fixation.

The AT content of all gene deserts flanking the ENCs was higher than 59%, that is the average of the entire human genome [33] (see last column of Table 1). These findings, however, could just reflect the high AT content of gene-poor regions.

Conclusion

Our study strongly supports the hypothesis that the evolutionary fate of a repositioned centromere is largely dependent upon a low gene density of the seeding region. This feature appears to be the consequence of the peculiar dynamics of ENCs progression associated with extensive restructuring of the region, including deletions, that can be assumed as potentially detrimental in genic regions of the genome.

Materials and methods

Cell lines

Metaphase preparations were obtained from cell lines (lymphoblasts or fibroblasts) from the following ape species: common chimpanzee (*Pan troglodytes*, PTR), gorilla (*Gorilla gorilla*, GGO), Bornean orangutan (*Pongo pygmaeus pygmaeus*, PPY), white-cheeked gibbon (*Nomascus leucogenys*, NLE); OWMs: rhesus monkey (*Macaca mulatta*, MMU,), vervet monkey (*Chlorocebus aethiops*, CAE, Cercopithecinae), silvered leaf monkey (*Trachypithecus cristatus*, TCR, Colobinae); NWMs: wooly monkey (*Lagothrix lagothricha*, LLA, Atelidae), common marmoset (*Callithrix jacchus*, CJA, Callitricidae).

FISH experiments

DNA extraction from BACs was reported previously [2].Co-hybridization FISH experiments were performed essentially as described by Lichter *et al.* [34]. To suppress cross-hybridization signals due to repeat sequences, COT1 DNA (5ug) was added to the hybridization mixture. Digital images were obtained using a Leica DMRXA2 epifluorescence microscope equipped with a cooled CCD camera (Princeton Instruments, Princeton, NJ, USA). Cy3-dUTP, Fluorescein-dCTP, Cy5-dCTP and DAPI fluorescence signals, detected with specific filters, were recorded separately as grayscale images. Pseudocoloring and merging of images were performed using Adobe Photoshop™ software.

BAC-end sequence analysis

BAC-end sequences were retrieved from the Trace Archive database [29]. They were then analyzed using the RepeatMasker software [35] in order to identify BAC-ends partially or entirely composed of repeat sequences. The software provides information on the extension and type of repeat.

Primate segmental duplication characterization in ENC regions

In order to identify segmental duplication content in various primates, we used the previously described assembly-independent approach (whole genome shotgun sequence detection, WSSD) where whole genome shotgun sequence (WGS) reads [27] from each query primate genome were mapped against regions from the human genome reference sequence (build35) corresponding to the evolutionary-new centromeres. We considered regions of excess WGS read depth (≥ mean+1.5*SD) to represent putative duplicated regions in each primate. Due to different genomic sequence divergences between each primate and the human reference sequence, we used sequence identity thresholds of ≥88% to map macaque reads while ≥94% was used for alignment of reads from chimpanzee and orangutan.

Gene/exon density simulation

In order to statistically assess the depletion of gene density/exon in the regions where ENCs were seeded, we performed the gene/exon density simulation as follows. First, we computed the average gene/exon density for the 14 ENC regions based on their annotation within the human genome. This became our observed value for gene/exon density within ENC regions (red line in Figure 3). Next, we randomly selected the same number of gap-free basepairs (23.2Mbp) from the human genome and computed the average gene/exon density for these randomly selected intervals. We generated 10,000 replicates and plotted the distribution of gene/exon density based on this simulation. We computed an empirical p-value as the number of replicates with gene/exon density equal or lower than the observed density in 10,000 replicates. We repeated the analysis excluding ENCs that had been identified within the human lineage of evolution (n=3) and obtained similar results (data not shown). For genes, we considered the position of all human non-redundant genes (RefSeq gene n = 22,589) and their corresponding exons as determined by the UCSC genome browser [21]. As a second analysis to assess transcript density, we independently mapped the location of all spliced human ESTs (n = 4,246,559) to the human genome (build35) and selected the location of the highest alignment score. If an EST/transcript mapped to two or more locations with an equivalent score, one was selected at random, such that each transcript was assigned once and only once to the human genome. As part of this analysis, we clustered exons that overlapped as a result of alternative splicing and counted each cluster as a single exon.

Abbreviations

BAC, bacterial artificial chromosome; ChIP, chromatin immuno precipitation; ENC, evolutionary new centromere; FISH, fluorescence *in situ* hybridization; NWM, New World monkey; OR, olfactory receptor; OWM, Old World monkey; SD, segmental duplication; UCSC, University California Santa Cruz.

Authors' contributions

ML planned and carried out the molecular cytogenetic experiments; PD analyzed the bioinformatic data; ZJ and EEE performed the statystical analysis; MR designed, coordinated the study and wrote the paper. All authors read and approved the final manuscript.

Additional data files

The following additional data are available. Additional Figure 1 illustrates the evolutionary history of chromosome 8 in primates. Additional Figure 2 reports examples of FISH experiments.

Additional Table 1 lists the human probes used to track the evolutionary history of chromosome 8. Additional Table 2 lists the species-specific BAC clones used in FISH experiments to detect pericentromeric segmental duplications.

Acknowledgements

MiUR (Ministero Italiano della Universita' e della Ricerca) is gratefully acknowledged for financial support. This work was also supported by NIH grant GM058815 to EEE and a Rosetta Inpharmatics Fellowship (Merck Laboratories) to ZJ. EEE is an investigator of the Howard Hughes Medical Institute.

References

1. Montefalcone G, Tempesta S, Rocchi M, Archidiacono N: **Centromere repositioning**. *Genome Res* 1999, **9**:1184-1188.
2. Ventura M, Archidiacono N, Rocchi M: **Centromere emergence in evolution**. *Genome Res* 2001, **11**:595-599.
3. Ventura M, Mudge JM, Palumbo V, Burn S, Blennow E, Pierluigi M, Giorda R, Zuffardi O, Archidiacono N, Jackson MS, Rocchi M: **Neocentromeres in 15q24-26 map to duplicons which flanked an ancestral centromere in 15q25**. *Genome Res* 2003, **13**:2059–2068.
4. Carbone L, Ventura M, Tempesta S, Rocchi M, Archidiacono N: **Evolutionary history of chromosome 10 in primates**. *Chromosoma* 2002, **111**:267-272.
5. Ventura M, Weigl S, Carbone L, Cardone MF, Misceo D, Teti M, D'Addabbo P, Wandall A, Björck E, de Jong P, She X, Eichler EE, Archidiacono N, Rocchi M: **Recurrent sites for new centromere seeding**. *Genome Res* 2004, **14**:1696-1703.
6. Ventura M, Antonacci F, Cardone MF, Stanyon R, D'Addabbo P, Cellamare A, Sprague LJ, Eichler EE, Archidiacono N, Rocchi M: **Evolutionary formation of new centromeres in macaque**. *Science* 2007, **316**:243-246.
7. Band MR, Larson JH, Rebeiz M, Green CA, Heyen DW, Donovan J, Windish R, Steining C, Mahyuddin P, Womack JE, Lewin HA: **An ordered comparative map of the cattle and human genomes**. *Genome Res* 2000, **10**:1359-1368.
8. Carbone L, Nergadze SG, Magnani E, Misceo D, Francesca Cardone M, Roberto R, Bertoni L, Attolini C, Francesca Piras M, de Jong P, Raudsepp T, Chowdhary BP, Guerin G, Archidiacono N, Rocchi M, Giulotto E: **Evolutionary movement of centromeres in horse, donkey, and zebra**. *Genomics* 2006, **87**:777-782.
9. Kasai F, Garcia C, Arruga MV, Ferguson-Smith MA: **Chromosome homology between chicken (Gallus gallus domesticus) and the red-legged partridge (Alectoris rufa); evidence of the occurrence of a neocentromere during evolution**. *Cytogenet Genome Res* 2003, **102**:326-330.
10. Kobayashi T, Yamada F, Hashimoto T, Abe S, Matsuda Y, Kuroiwa A: **Centromere repositioning in the X chromosome of XO/XO mammals, Ryukyu spiny rat**. *Chromosome Res* 2008, **16**:587-593.
11. Marshall OJ, Chueh AC, Wong LH, Choo KH: **Neocentromeres: new insights into centromere structure, disease development, and karyotype evolution**. *Am J Hum Genet* 2008, **82**:261-282.
12. Amor DJ, Bentley K, Ryan J, Perry J, Wong L, Slater H, Choo KH: **Human centromere repositioning "in progress"**. *Proc Natl Acad Sci USA* 2004, **101**:6542-6547.
13. Yan H, Ito H, Nobuta K, Ouyang S, Jin W, Tian S, Lu C, Venu RC, Wang GL, Green PJ, Wing RA, Buell CR, Meyers BC, Jiang J **Genomic and genetic characterization of rice Cen3**

reveals extensive transcription and evolutionary implications of a complex centromere. *Plant Cell* 2006, **18**:2123-2133.
14. Nagaki K, Cheng Z, Ouyang S, Talbert PB, Kim M, Jones KM, Henikoff S, Buell CR, Jiang J: **Sequencing of a rice centromere uncovers active genes**. *Nat Genet* 2004, **36**:138-145.
15. Lam AL, Boivin CD, Bonney CF, Rudd MK, Sullivan BA: **Human centromeric chromatin is a dynamic chromosomal domain that can spread over noncentromeric DNA**. *Proc Natl Acad Sci USA* 2006, **103**:4186-4191.
16. Saffery R, Sumer H, Hassan S, Wong LH, Craig JM, Todokoro K, Anderson M, Stafford A, Choo KH: **Transcription within a functional human centromere**. *Mol Cell* 2003, **12**:509-516.
17. Roberto R, Capozzi O, Wilson RK, Mardis ER, Lomiento M, Tuzun E, Cheng Z, Mootnick AR, Archidiacono N, Rocchi M, Eichler EE: **Molecular refinement of gibbon genome rearrangement**. *Genome Res* 2007, **17**:249-257.
18. Cardone MF, Lomiento M, Teti MG, Misceo D, Roberto R, Capozzi O, D'Addabbo P, Ventura M, Rocchi M, Archidiacono N: **Evolutionary history of chromosome 11 featuring four distinct centromere repositioning events in Catarrhini**. *Genomics* 2007, **90**:35-43.
19. Eder V, Ventura M, Ianigro M, Teti M, Rocchi M, Archidiacono N: **Chromosome 6 phylogeny in primates and centromere repositioning**. *Mol Biol Evol* 2003, **20**:1506-1512.
20. Gibbs RA, Rogers J, Katze MG, Bumgarner R, Weinstock GM, Mardis ER, Remington KA, Strausberg RL, Venter JC, Wilson RK, Batzer MA, Bustamante CD, Eichler EE, Hahn MW, Hardison RC, Makova KD, Miller W, Milosavljevic A, Palermo RE, Siepel A, Sikela JM, Attaway T, Bell S, Bernard KE, Buhay CJ, Chandrabose MN, Dao M, Davis C, Delehaunty KD, Ding Y *et al*: **Evolutionary and biomedical insights from the rhesus macaque genome**. *Science* 2007, **316**:222-234.
21. **University of California Santa Cruz Genome Bioinformatics** [http://www.genome.ucsc.edu]
22. Raaum RL, Sterner KN, Noviello CM, Stewart CB, Disotell TR: **Catarrhine primate divergence dates estimated from complete mitochondrial genomes: concordance with fossil and nuclear DNA evidence**. *J Hum Evol* 2005, **48**:237-257.
23. Cardone MF, Alonso A, Pazienza M, Ventura M, Montemurro G, Carbone L, de Jong PJ, Stanyon R, D'Addabbo P, Archidiacono N, She X, Eichler EE, Warburton PE, Rocchi M: **Independent centromere formation in a capricious, gene-free domain of chromosome 13q21 in Old World monkeys and pigs**. *Genome Biol (www)* 2006, **7**:R91.
24. Schneider H, Canavez FC, Sampaio I, Moreira MA, Tagliaro CH, Seuanez HN: **Can molecular data place each neotropical monkey in its own branch?** *Chromosoma* 2001 **109**:515-523.
25. She X, Horvath JE, Jiang Z, Liu G, Furey TS, Christ L, Clark R, Graves T, Gulden CL, Alkan C, Bailey JA, Sahinalp C, Rocchi M, Haussler D, Wilson RK, Miller W, Schwartz S, Eichler EE: **The structure and evolution of centromeric transition regions within the human genome**. *Nature* 2004, **430**:857-864.
26. Eichler EE: **Segmental duplications: what's missing, misassigned, and misassembled-and should we care?** *Genome Res* 2001 **11**:653-656.
27. Bailey JA, Gu Z, Clark RA, Reinert K, Samonte RV, Schwartz S, Adams MD, Myers EW, Li PW, Eichler EE: **Recent segmental duplications in the human genome**. *Science* 2002, **297**:1003-1007.
28. **Bioinformatics Research Laboratory of the Baylor College of Medicine, Houston, TX** [http://brl.bcm.tmc.edu/pgi/rhesus/dataAccess.rhtml].
29. **National Center for Biotechnology Information** [http://www.ncbi.nlm.nih.gov/Traces]
30. Gilad Y, Bustamante CD, Lancet D, Paabo S: **Natural selection on the olfactory receptor gene family in humans and chimpanzees**. *Am J Hum Genet* 2003, **73**:489-501.
31. Bailey JA, Yavor AM, Massa HF, Trask BJ, Eichler EE: **Segmental duplications: organization and impact within the current human genome project assembly.** *Genome Res* 2001, **11**:1005-1017.

32. Pardo-Manuel de Villena F, Sapienza C: **Transmission ratio distortion in offspring of heterozygous female carriers of Robertsonian translocations**. *Hum Genet* 2001, **108**:31-36.
33. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, Funke R, Gage D, Harris K, Heaford A, Howland J, Kann L, Lehoczky J, LeVine R, McEwan P, McKernan K, Meldrim J, Mesirov JP, Miranda C, Morris W, Naylor J, Raymond C, Rosetti M, Santos R, Sheridan A, Sougnez C *et al*: **Initial sequencing and analysis of the human genome**. *Nature* 2001, **409**:860-921.
34. Lichter P, Tang Chang C-J, Call K, Hermanson G, Evans GA, Housman D, Ward DC: **High resolution mapping of human chromosomes 11 by in situ hybridization with cosmid clones**. *Science* 1990, **247**:64-69.
35. **RepeatMasker** [http://www.repeatmasker.org/]

Legends to Figures

Figure 1

The Figure shows the phylogenetic relationships of the species under study. Data on OWMs and Hominoidea are from Raaum *et al.* [22], while those on NWMs are from Schneider *et al.* [24].

Figure 2

(2a) Examples of FISH experiments using species-specific BAC clones yielding duplicated signals around the centromere. The CH250 and CH271 are BAC libraries specific for macaque and gibbon, respectively. The DAPI-stained chromosome without the signal is reported on the left to better show the morphology of the chromosome. (2b) FISH experiment using the BAC clone CH250-417O7 (MMU2) on a macaque metaphase, showing pericentromeric signals on several chromosomes.

Figure 3

Gene density simulations. The observed density of (a) genes (Refseq), (b) Refseq exons and (c) EST exons within the corresponding region of the 14 ENCs were compared against a simulated set of 10,000 regions distributed randomly within the human genome (Methods). A significant depletion of exons and genes was observed.

Table 1

Definition of the ENC seeding region in the reference genome

Chromosome	ENC position	size (kb)	p arm BAC	
Platirrhini				Position in HS
LLA7 (HSA8)	chr8:63,002,317-63,047,396	45	RP11-953L16	chr8:62,816,38
Catarrhini				
MMU2 (HSA3)	chr3:164,221,008-164,539,729	319	RP11-449O23	chr3:164,054,8
MMU4 (HSA6)	chr6:145,651,644-145,845,896	194	**RP11-474A9**	chr6:145,651,6
MMU12 (HSA2q)	chr2:138,847,788-138,947,383	99	RP11-343I5	chr2:138,777,1
MMU13 (HSA2p)	chr2:86,680,785-86,885,407	204	**RP11-722G17**	chr2:86,680,78
MMU14 (HSA11)	chr11:5,856,181-5,864,725	8	RP11-625D10	chr11:5,667,33
MMU15 (HSA9)	chr9:122,486,836-122,532,865	46	RP11-64P14	chr9:122,344,5
MMU17 (HSA13)	chr13:61,178,154-62,520,878	1343	***RP11-543A19***	chr13:61,111,7
MMU18 (HSA18)	chr18:50,313,129-50,360,135	47	RP11-61D1	chr18:50,155,7
NLE15 (HSA11)	chr11:89,446,995-89,488,776	42	RP11-529A4	chr11:89,286,3
PPY11 (HSA11)	chr11:20,180,424-20,332,556	152	**RP11-56J22**	chr11:20,180,4
HSA				Position in MM
HSA3 (MMU2)	chr2:14,301,434-14,386,749	85	CH250-111O10	chr2:14,301,46
HSA6 (MMU4)	chr4: 57,710,481- 57,863,274	153	**CH250-20M17**	chr4: 57,710,48
HSA11 (MMU14)	chr14:17,109,970-17,281,610	171	CH250-111J7	chr14:17,015,7

Seeding regions of the studied ENCs, defined by a splitting BAC (in bold) or by overlapping BACs mapping in opposite side of the centromere (p arm/q arm). In the latter case the overlapping portion of the two BACs was assumed as the seeding point. In MMU17 (human 13), several contiguous human BACs gave split signals. The Table reports the most external ones (in italics). The human genome was used as a reference genome for non-human primate ENCs. The macaque genome was used as a reference for the three human ENCs (see text).

Table 2

Duplication analyses in ENC regions

ENC	Start (HSA hg17)	End (HSA hg17)	Size	Non-redundant WSSD base pair (bp)		
				HSA	PTR	PPY
MMU2 (HSA3)	164,221,008	164,539,729	318,722	0	0	0
MMU4 (HSA6)	145,651,644	145,845,896	194,253	0	0	0
MMU12 (HSA2)	138,847,788	138,947,383	99,596	0	0	0
MMU13 (HSA2)	86,680,785	86,885,407	204,623	24,002	0	0
MMU14 (HSA11)	5,856,181	5,864,725	8,545	0	0	0
MMU15 (HSA9)	122,486,836	122,532,865	46,030	0	0	0
MMU17 (HSA13)	61,178,154	62,520,878	1,342,725	24,879	15,879	103,912
MMU18 (HSA18)	50,313,129	50,360,135	47,007	0	0	0
PPY11 (HSA11)	20,180,424	20,332,556	152,133	0	0	126,135

ENC	Start+1M	End+1M	Size	Non-redundant WSSD base pair (bp		
				HSA	PTR	PPY
MMU2 (HSA3)	163,221,008	165,539,729	2,318,722	0	0	0
MMU4 (HSA6)	144,651,644	146,845,896	2,194,253	0	0	0
MMU12 (HSA2)	137,847,788	139,947,383	2,099,596	0	0	17,001
MMU13 (HSA2)	85,680,785	87,885,407	2,204,623	1,227,738	309,321	0
MMU14 (HSA11)	4,856,181	6,864,725	2,008,545	0	0	13,379
MMU15 (HSA9)	121,486,836	123,532,865	2,046,030	0	0	0
MMU17 (HSA13)	60,178,154	63,520,878	3,342,725	160,4637	98,004	144,056
MMU18 (HSA18)	49,313,129	51,360,135	2,047,007	0	0	0
PPY11 (HSA11)	19,180,424	21,332,556	2,152,133	0	0	784,808

We estimate the number of duplicated basepairs predicted in each of the ENC intervals using the WSSD method; duplications >10 kb and >94% were detected with the exception of the macaque where a threshold of >88% was used due to the greater sequence divergence of the human and macaque genome. The analysis was performed separately for each of the four primate species. Two different ENC intervals were considered: a narrow interval, as defined in Table 1 (upper dataset) and a larger interval adding 1 Mbp to each side of the region (lower dataset).

Table 3

Species-specific BACs yielding duplicated signals oround ENCs

ENC	BAC	Position in HSA (May2004)
MMU13 (HSA2p)	CH250-565F19*	chr2:86,755,212-alphoid
	CH250-417O7	chr2:86,785,727-repeat
	CH250-371E19*	chr2:86,870,586-alphoid
MMU12 (HSA2q)	CH250-359C1	chr2:138,344,201-138,510,183
	CH250-158G21	chr2:138,478,651-138,621,067
	CH250-18F12*	chr2:138,643,711-alphoid
MMU14 (HSA11)	CH250-444O7*	chr11:5,861,684-alphoid
	CH250-499K18*	chr11:6,038,164-alphoid
MMU15 (HSA9)	CH250-221O11*	chr9:122,220,400-alphoid
MMU17 (HSA13)	CH250-310C22	chr13:61,479,136-61,591,608
	CH250-299M13	chr13:61,503,914-61,617,441
	CH250-115C9	chr13:61,540,997-61,676,877
MMU18 (HSA18)	CH250-322J6	chr18:50,437,322-repeat
NLE15 (HSA11)	CH271-140J13	chr11:89,572,864-repeat

Species-specific BAC clones yielding duplicated signals around the ENC. Their specific pericentromeric location, confirmed by FISH, was derived by their BAC-end(s) mapping. One BAC-end of asterisked BACs is entirely composed of alphoid repeats. The FISH signal, however, was not centromeric, indicating that the alphoid content of the BAC was marginal. See Figure 1 for examples.

Table 4

RefSeq genes flanking the ENCs

ENC	inteval	left gene		right
Platirrhini			**Position in HSA (hg17)**	
LLA7 (HSA8)	0,534 Mb	ASPH	chr8:62,699,652-62,789,681	FAM7
Catarrhini				
MMU2 (HSA3)	3.607 Mb	C3orf57	chr3:162,545,283-162,572,573	SI
MMU4 (HSA6)	0.772 Mb	UTRN	chr6:144,654,566-145,215,861	EPM2
MMU13 (HSA2p)	0.097 Mb	RNF103	chr2:86,742,174-86,762,636	RMD5
MMU14 (HSA11)	1.213 Mb	MMP26	chr11:4,966,000-4,970,233	C11or
MMU15 (HSA9)	0.423 Mb	PTGS1	chr9:122,212,783-122,237,535	PDCL
MMU12 (HSA2q)	0.485 Mb	HNMT	chr2:138,555,540-138,607,665	LOC3
MMU17 (HSA13)	4.888 Mb	PCDH20	chr13:60,881,822-60,887,282	PCDH
MMU18 (HSA18)	0.247 Mb	C18orf54	chr18:50,139,169-50,162,379	C18or
NLE14 (HSA11)	2.746 Mb	CHORDC1	chr11:89,574,265-89,595,854	MTNR
PPY11 (HSA11)	0.203 Mb	DBX1	chr11:20,134,336-20,138,446	HTATI
HSA			**Position in MMU (rheMac2)**	
HSA3 (MMU2)	0.641 Mb	EPHA3 in HSA (not annotated in MMU)	chr3:89,239,364-89,613,972 (MMU2:13,335,593-13,694,578)	PROS (L313
HSA6 (MMU4)	0.897 Mb	PRIM2A in HSA (not annotated in MMU)	chr6:57,290,381-57,621,334 (2 dup in MMU: MMU4:56,935,673-57,245,600 MMU11:20,043,342-20,044,345)	KHDR (not a
HSA11 (MMU14)	1.280 Mb	LRRC55 in HSA (not annotated in MMU)	chr11:56,705,797-56,714,154 (MMU14:16,226,175-16,234,557)	PTPRJ (not a

Position of the most proximal and distal genes with respect to each ENC seeding region, calculated in the reference genome (see text). The interval size, in Mb, between the two genes is reported in column 2.

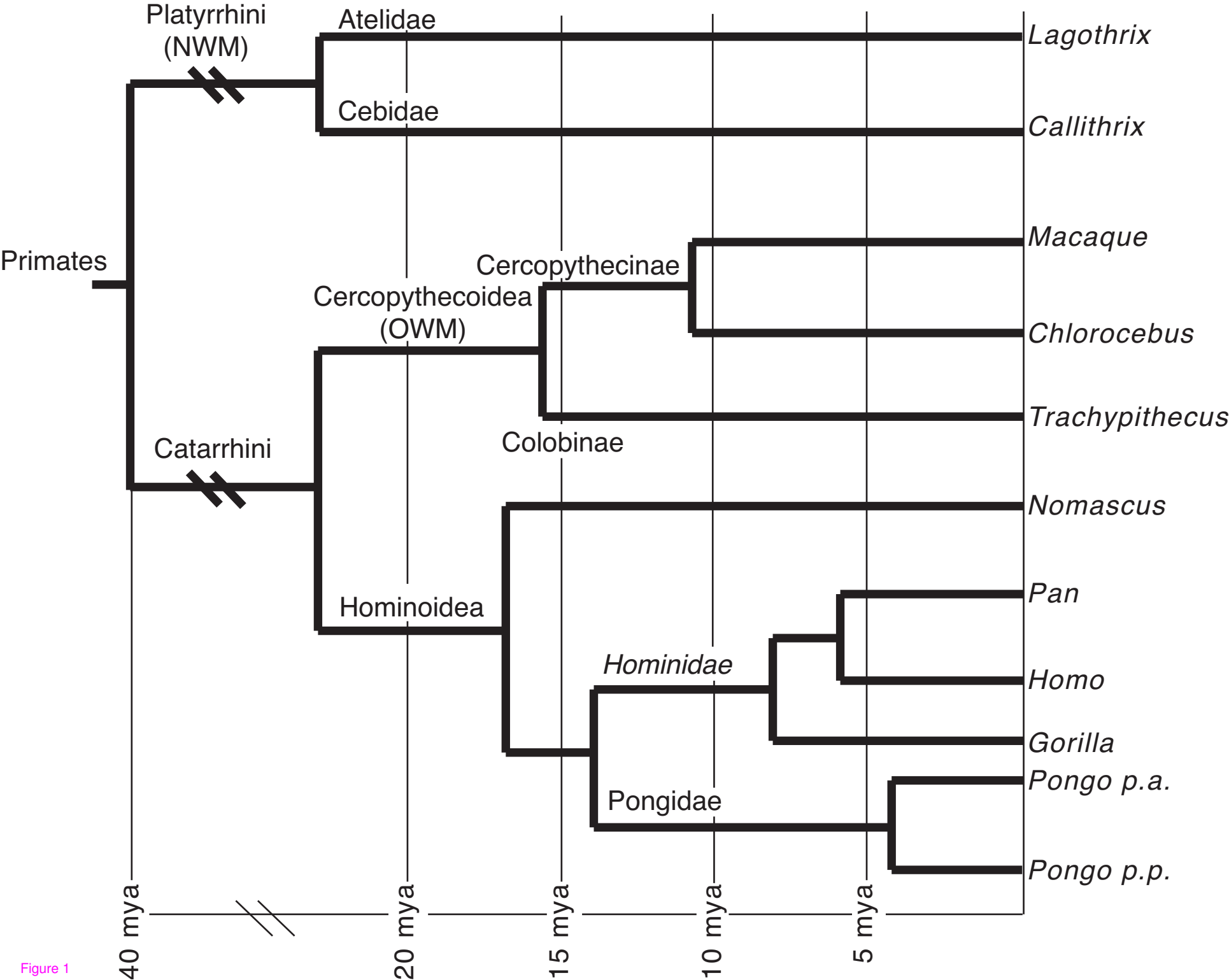

Platyrrhini
(NWM)
Atelidae
Cebidae
Lagothrix
Callithrix
Primates
Cercopythecinae
Cercopythecoidea
(OWM)
Macaque
Chlorocebus
Trachypithecus
Colobinae
Catarrhini
Nomascus
Hominoidea
Pan
Hominidae
Homo
Gorilla
Pongo p.a.
Pongidae
Pongo p.p.
40 mya
20 mya
15 mya
10 mya
5 mya

Figure 1

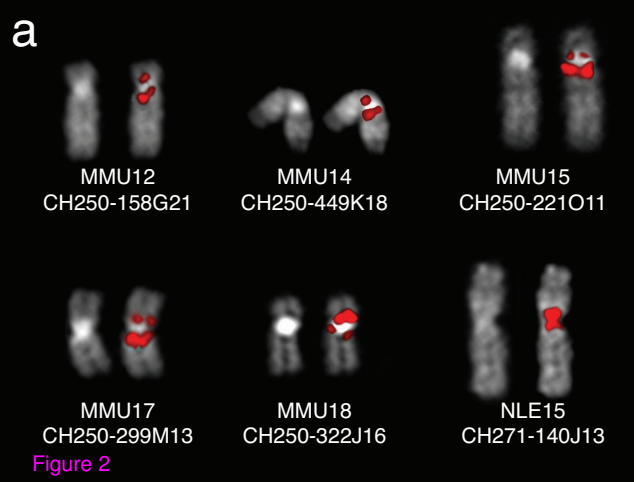
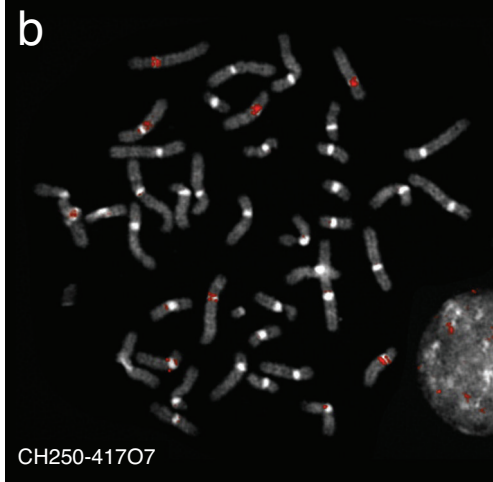
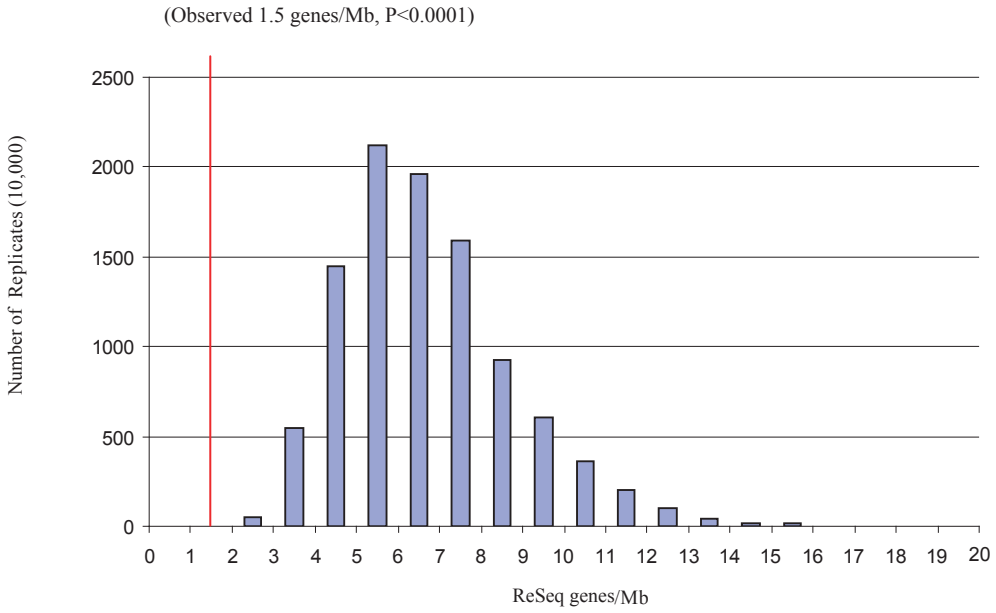

a
MMU12
CH250-158G21
MMU14
CH250-449K18
MMU15
CH250-221O11
MMU17
CH250-299M13
MMU18
CH250-322J16
NLE15
CH271-140J13

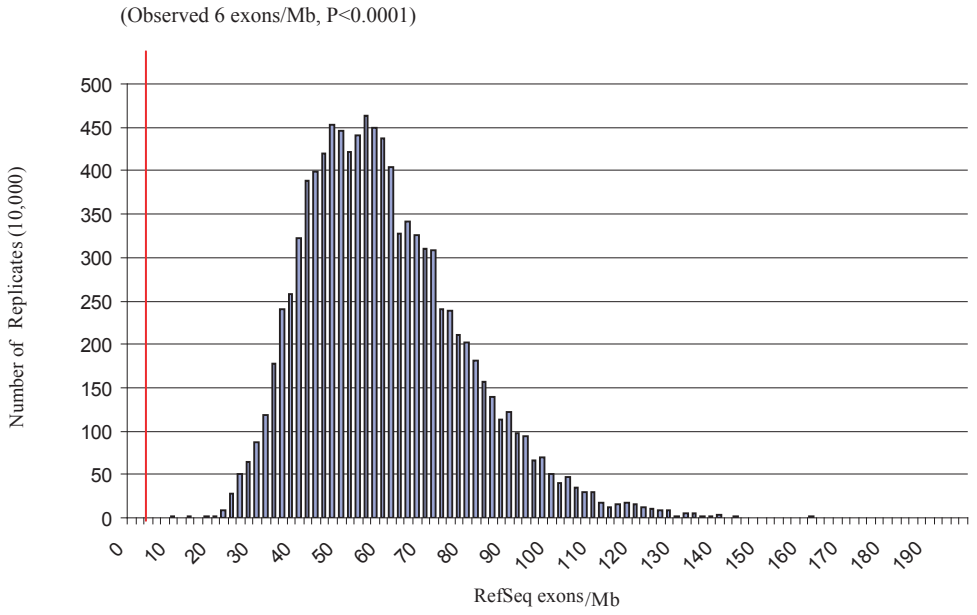
Figure 2


b
CH250-417O7

A


(Observed 1.5 genes/Mb, P<0.0001)
Number of Replicates (10,000)
2500
2000
1500
1000
500
0
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
ReSeq genes/Mb

B


(Observed 6 exons/Mb, P<0.0001)
Number of Replicates (10,000)
500
450
400
350
300
250
200
150
100
50
0
0 10 20 30 40 50 60 70 80 90 100 110 120 130 140 150 160 170 180 190
RefSeq exons/Mb

C

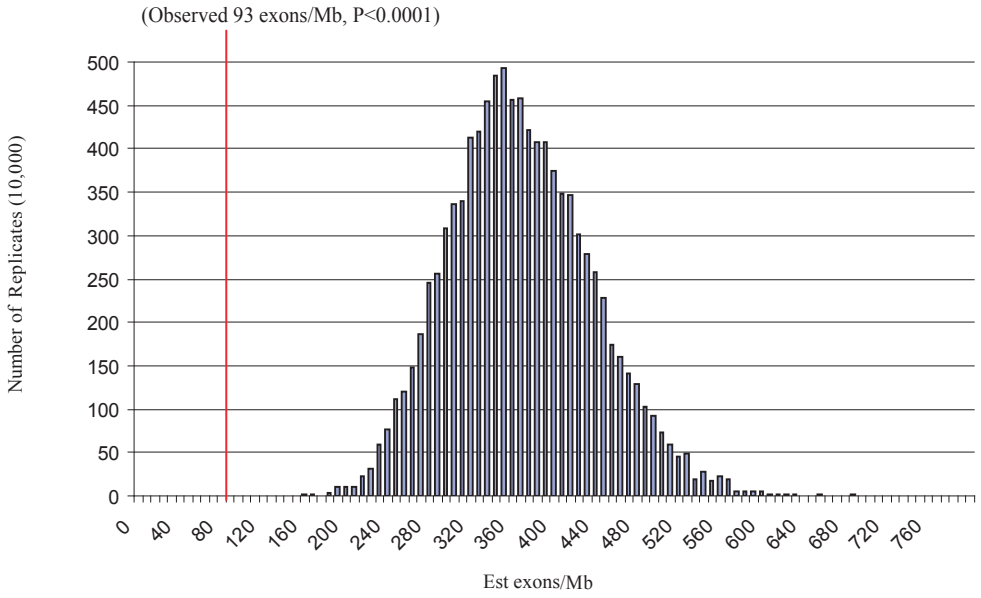

(Observed 93 exons/Mb, P<0.0001)
Number of Replicates (10,000)
500
450
400
350
300
250
200
150
100
50
0
0 40 80 120 160 200 240 280 320 360 400 440 480 520 560 600 640 680 720 760
Est exons/Mb

Figure 3

Additional files provided with this submission:

Additional file 1: additional_data_file_1.pdf, 387K
http://genomebiology.com/imedia/2104356565240553/supp1.pdf
Additional file 2: additional_data_file_2.pdf, 294K
http://genomebiology.com/imedia/1338822122405549/supp2.pdf
Additional file 3: additional_data_file_3.doc, 51K
http://genomebiology.com/imedia/1843421986240554/supp3.doc
Additional file 4: additional_data_file_4.doc, 89K
http://genomebiology.com/imedia/8566176872405545/supp4.doc