An analysis of the gene complement of a marsupial, *Monodelphis domestica*: Evolution of lineage-specific genes and giant chromosomes

Leo Goodstadt,[1] Andreas Heger, Caleb Webber, and Chris P. Ponting

MRC Functional Genetics Unit, University of Oxford, Department of Physiology, Anatomy, and Genetics, South Parks Road, Oxford OX1 3QX, UK

The newly sequenced genome of *Monodelphis domestica* not only provides the out-group necessary to better understand our own eutherian lineage, but it enables insights into the innovative biology of metatherians. Here, we compare *Monodelphis* with *Homo* sequences from alignments of single nucleotides, genes, and whole chromosomes. Using PhyOP, we have established orthologs in *Homo* for 82% (15,250) of *Monodelphis* gene predictions. Those with single orthologs in each species exhibited a high median synonymous substitution rate ($d_S = 1.02$), thereby explaining the relative paucity of aligned regions outside of coding sequences. Orthology assignments were used to construct a synteny map that illustrates the considerable fragmentation of *Monodelphis* and *Homo* karyotypes since their therian last common ancestor. Fifteen percent of *Monodelphis* genes are predicted, from their low divergence at synonymous sites, to have been duplicated in the metatherian lineage. The majority of *Monodelphis*-specific genes possess predicted roles in chemosensation, reproduction, adaptation to specific diets, and immunity. Using alignments of *Monodelphis* genes to sequences from either *Homo* or *Trichosurus vulpecula* (an Australian marsupial), we show that metatherian X chromosomes have elevated silent substitution rates and high G+C contents in comparison with both metatherian autosomes and eutherian chromosomes. Each of these elevations is also a feature of subtelomeric chromosomal regions. We attribute these observations to high rates of female-specific recombination near the chromosomal ends and within the X chromosome, which act to sustain or increase G+C levels by biased gene conversion. In particular, we propose that the higher G+C content of the *Monodelphis* X chromosome is a direct consequence of its small size relative to the giant autosomes.

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

The newly sequenced genome ($2n = 18$; 3.6 Gb) of the South American gray short-tailed opossum (*Monodelphis domestica*) (Mikkelsen et al. 2007) allows initial comparisons of its predicted gene set, and its chromosomes, with those of humans. *Monodelphis* is a metatherian mammal (marsupial) whose lineage split from that of eutherians (placental mammals) ~170–190 million years ago (Mya) (Kumar and Hedges 1998; Woodburne et al. 2003). Since then, metatherians and eutherians have acquired distinct physiological and behavioral features. However, they still share many ancestral therian characters, most notably lactation using mammary papilla, and the bearing of live young without using a shelled egg.

Monodelphis is a small (80–155 g) and nocturnal marsupial. In the wild, it is terrestrial, present in low population densities, and feeds mainly on invertebrates and small vertebrates (Streilein 1982b). In common with murid rodents, reproduction occurs throughout the year, females enter oestrus following exposure to male odors (Fadem and Rayve 1985), and both sexes rely heavily on pheromonal communication (Streilein 1982a). Unlike murid rodents, however, male animals use skin and glandular secretions rather than urine odors for marking, possibly in order to conserve water, since some populations of *Monodelphis* are found in semiarid environments (Streilein 1982b; Zuri et al. 2005).

Much of the anatomical, physiological, and behavioral differences between metherian and eutherian mammals may be due to protein coding genes present in lineage-specific duplicates. These genes may either share together the functions of the progenitor (“subfunctionalization”) or have each acquired innovative roles (“neofunctionalization”) (Ohno 1970; Hughes 1994; Lynch and Conery 2000; Lynch and Force 2000). In the genomes of sequenced eutheria, the majority of the protein coding genes that are specific to the human (*Homo sapiens*), mouse (*Mus musculus*), or rat (*Rattus norvegicus*) fall into a few well-defined functional classes (Lander et al. 2001; Mouse Genome Sequencing Consortium 2002; Rat Genome Sequencing Project Consortium 2004). These include chemosensation (in particular, olfaction and pheromone detection), reproduction (including placental growth factors and pheromones), toxin degradation (by enzymes such as cytochrome P450s), and immunity and host defense (such as T-cell receptors, immunoglobulins, and alpha-/beta-defensins) (Emes et al. 2003; Castillo-Davis et al. 2004). These functions are critical to the survival and reproduction of adults. By way of contrast, transcription factors or genes that are involved in embryonic development are rarely lineage specific, as these are usually retained without duplication or loss, and without extensive sequence divergence, in each of these mammalian lineages (Mouse Genome Sequencing Consortium 2002; Rat Genome Sequencing Project Consortium 2004; Lindblad-Toh et al. 2005).

We sought to identify lineage-specific gene duplicates or “inparalogs” (Sonnhammer and Koonin 2002) without recourse to an out-group species by reconstructing the phylogenetic rela-

[1] **Corresponding author.**
E-mail leo.goodstadt@dpag.ox.ac.uk; fax 44-1865-272420.

Article published online before print. Article and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.6093907.

tionships for all *Homo* and of *Monodelphis* genes. Our PhyOP pipeline (Goodstadt and Ponting 2006) infers orthology and paralogy relationships among all predicted transcripts of all *Monodelphis* and *Homo* genes using synonymous substitution rates (d_S) as a distance metric. Within coding sequence, synonymous sites are the least subject to selection and are a better proxy for evolutionary distances than the protein similarity scores used in many other approaches to orthology prediction. This is especially important for lineage-specific paralogs, many of which have been subject to repeated bouts of adaptive evolution, leading to divergent sequence. PhyOP does not rely on conserved synteny, so that the degree of chromosomal rearrangement across different lineages or in different parts of the genome does not influence orthology prediction. Instead, disruptions in gene order conservation can be used as one of the metrics to identify retrotransposed pseudogenes that are, in general, randomly integrated into the genome. The degree of past selection among lineage-specific genes can then be deduced from estimates of their d_N/d_S values, defined as the number of nonsynonymous substitutions per nonsynonymous site (d_N) relative to the number of synonymous substitutions per synonymous site (d_S). The codeml program from the PAML package (Yang 1997) reliably estimates d_S values up to ~2.5 (Goodstadt and Ponting 2006), and thus, is well suited for investigating mammalian orthologs or mammal-specific paralogs.

Monodelphis autosomes are huge. The smallest, chromosome 6 (MDO6), is roughly the same size as the largest previously sequenced eutherian chromosome, human chromosome 1 (HSA1). The *Monodelphis* chromosome 1 is three times larger. By way of contrast, the *Monodelphis* chromosome X (MDOX), at 60.7 Mb, is less than half the size of any eutherian X chromosome that has yet been sequenced. During recombination, there is an obligatory minimum of one chiasma per chromosomal arm (Pardo-Manuel de Villena and Sapienza 2001). Therefore, all else being equal, recombination rates are expected to be greater in chromosomal arms that are shorter (especially *Monodelphis* X chromosomal arms) than in those that are longer (the large *Monodelphis* autosomal arms). Higher recombination rates are proposed to drive increases in G+C content due to biased gene conversion (BGC) (Duret et al. 2006). Regions of higher G+C content in eutheria and in chicken also often exhibit higher nucleotide substitution rates (d_S) (Buchmann et al. 1983; Mouse Genome Sequencing Consortium 2002; Hillier et al. 2004; Chimpanzee Sequencing and Analysis Consortium 2005), consistent with one theoretical model of BGC (Piganeau et al. 2002). d_S rates and G+C content among *Monodelphis–Homo* ortholog pairs can thus further illuminate the complex inter-relationships between recombination, substitution rates, and nucleotide composition.

Our results highlight *Monodelphis* inparalogs that are likely to contribute to the distinctive biology of metatherians. We also take advantage of our large predicted set of 12,817 one-to-one orthologs between *Monodelphis* and *Homo* to compare silent substitution (d_S) rates for the large autosomes of *Monodelphis* with those for its much smaller X chromosome. G+C content and d_S are found to be elevated not only in the X chromosome, but also in the 10-Mb subtelomeric regions of all chromosomes. Finally, using *Trichosurus vulpecula* sequences, we show that the disparity of silent substitution rates between the subtelomeric regions and chromosome interiors has been most acute in the metatherian lineage. We propose a model linking nucleotide content, substitution, and recombination rates with the propensity to evolve large chromosomes.

Results

An improved set of *Monodelphis* gene predictions

We augmented a set of 19,888 *Monodelphis* genes from Ensembl with 657 additional gene predictions in the MonDom3 genome assembly (see Methods). These additional predictions were enriched in *Monodelphis* inparalogs. Our assignments of orthology and paralogy revealed a further 130 genes representing adjacent paralogs that had been merged erroneously. We were also able to identify and discard 1402 putative pseudogenes. These included retrotransposed copies of multi-exon genes; genes with multiple disruptions to their coding sequence; those showing sequence similarity to retroviral and transposable elements such as LINE1; and also some noncoding sequence erroneously predicted from the reverse strand of presumably functional protein coding sequence. This resulted in a final protein coding gene count of 18,639. By comparison, our similarly estimated minimum gene count in *Homo sapiens* is 20,806, which is comparable to what we previously obtained comparing the dog and human gene sets (Goodstadt and Ponting 2006).

Homo–Monodelphis orthologs

Orthology between *Monodelphis* and *Homo* genes, together with *Monodelphis*- or *Homo*-lineage-specific paralogy, were assigned using PhyOP (Goodstadt and Ponting 2006). A total of 82% (15,250) of *Monodelphis* genes were predicted to have *Homo* orthologs; 12,817 of these were orthologous to only a single *Homo* gene ("one-to-one orthologs") (Fig. 1). The full set of genes together with orthology and paralogy relationships are available

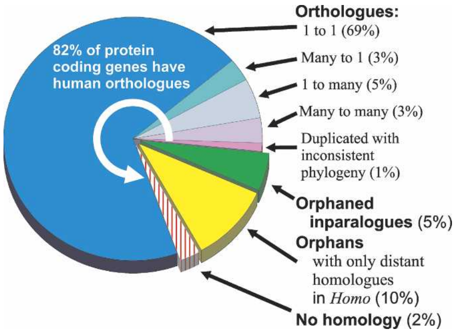

Figure 1. Proportion of *Monodelphis domestica* genes with *Homo* orthologs. The majority (82%) of protein coding genes have PhyOP-predicted orthologs among human genes. The classes of orthology relationships are indicated by counts of *Homo* orthologs followed by *Monodelphis*. Thus, a one-to-many relationship refers to a single *Homo* gene that is orthologous to several *Monodelphis*-specific duplications. Many of the genes with no predicted orthology with a *Homo* gene are nevertheless closely related to one or more *Monodelphis* genes with a low d_S, indicating that these are *Monodelphis*-specific duplications. These are labeled as "Orphaned inparalogues." This class includes genuine biological losses in the eutherian or human lineage, as well as heuristic failures in the PhyOP-prediction pipeline. The latter are often related to errors in gene predictions in difficult cases where there are many adjacent tandem duplications, such as among KRAB-ZnF genes. Similarly, many of the *Monodelphis* genes which have no plausible candidate ortholog in the human genome (labeled "Orphans with only distant homologues in *Homo*") are likely to involve problematic gene predictions. The *Monodelphis* genes in the "No homology" class do not have significant BLAST alignments ($E < 10^{-5}$) covering the majority of the sequence (75% of the shorter of the pair of sequences, and involving more than 50 residues). Most of these appear to be fragmentary, chimeric, or erroneous gene predictions, or noncoding sequence.

Table 1. Characteristics of one-to-one orthologs predicted for *Homo sapiens* and *Monodelphis domestica*

d_N / d_S	0.086 (0.044–0.152)
d_N	0.095 (0.044–0.174)
d_S	1.02 (0.76–1.44)
Amino acid sequence identity	81.0% (69.8%–90.0%)
Pairwise alignment coverage of the longer sequence	94.2% (80.4%–98.7%)

	Homo sapiens	*Monodelphis domestica*
Number of exons	9 (5–15)	9 (5–14)
Sequence length (codons)	471 (302–745)	445 (283–701)
Unspliced transcript length (bp)	27,241 (9888–66,806)	25,365 (8162–66,808)
G+C content at 4D sites	56.9% (41.3%–70.7%)	48.7% (37.0%–60.6%)

Shown are median values and, in parentheses, lower and upper quartiles.

from http://wwwfgu.anat.ox.ac.uk/download/monodelphis. The median d_S value for one-to-one orthologs was 1.02 (Table 1). As expected from the earlier divergence of birds, and the later branching of other eutherian mammals from the human lineage, this value is intermediate between that for human and chicken (1.66; Hillier et al. 2004), and the median d_S for human and dog (0.36; Goodstadt and Ponting 2006) or human and mouse (0.60; Mouse Genome Sequencing Consortium 2002).

The median value of d_N/d_S for *Homo* and *Monodelphis* one-to-one orthologs was 0.086. This is lower than the median d_N/d_S estimated for the primate lineage (0.112), but comparable to the median d_N/d_S rate for the mouse lineage (0.088; Mouse Genome Sequencing Consortium 2002). As deleterious mutations are more effectively purged among species with larger effective population sizes (Ohta 1973), we can infer from these d_N/d_S rates that the effective population sizes in the marsupial lineage, since the last common ancestor with eutherians, have been similar to those within the murid rodent lineage, and have been larger than those within the human lineage since the primate-rodent last common ancestor.

Conserved synteny

Within the nine *Monodelphis* chromosomes there are 415 "macro-synteny" blocks (see Methods), within which fine-scale gene order and transcriptional orientation have been largely preserved in human chromosomes. There are, however, a number of large-scale rearrangements, such as inversions and translocations (Fig. 2). This number of synteny blocks is similar to that found by whole-genome alignment methods (Mikkelsen et al. 2007). We find that half of all *Monodelphis* one-to-one orthologs reside in macro-synteny blocks containing 82, or fewer, *Homo* one-to-one orthologs, which is considerably smaller than the equivalent numbers, 151 and 167, in the dog and mouse, respectively (L. Goodstadt, unpubl.). Since eutherian karyotypes have been relatively stable along the human lineage (Wienberg 2004), it thus appears likely that considerable chromosomal rearrangements have occurred either in the lineage from the therian last common ancestor to the earliest eutherian and/or in the metatherian lineage to *Monodelphis*.

Substitution rates and G+C content are elevated in the metatherian X chromosome

We investigated G+C content since this strongly covaries with d_S among eutheria (Matassi et al. 1999; Lander et al. 2001; Hardison et al. 2003; Webber and Ponting 2005). Specifically, we considered GC4D, the G or C content of the third position of fourfold degenerate codons in genes, which correlates strongly with the G+C content of their surrounding regions (Eyre-Walker and Hurst 2001; Duret et al. 2002). We observed that the d_S values of *Monodelphis* and *Homo* one-to-one orthologs correlate well with their GC4D values (Spearman's $\rho = 0.51$ and 0.59, respectively). Similarly, *Monodelphis* and *Homo* orthologs' GC4D contents were also highly correlated (Spearman's $\rho = 0.73$). These rank correlations are intermediate between those for pairs of eutherians, and those for eutherians and chicken (Webber and Ponting 2005), confirming that a large part of mammalian G+C composition is ancestral.

We found, however, that the GC4D and d_S values of orthologs are not distributed uniformly among the nine *Monodelphis* haploid chromosomes. GC4D and d_S values are significantly elevated ($P < 10^{-14}$ and $P < 10^{-3}$, respectively) for the *Monodelphis* X chromosome relative to the autosomes (Fig. 3A,B; Table 2). These increases in GC4D and d_S values appear to be characteristic of metatherian rather than eutherian X chromosomes. No elevations in d_S were apparent for the portions of the human X chro-

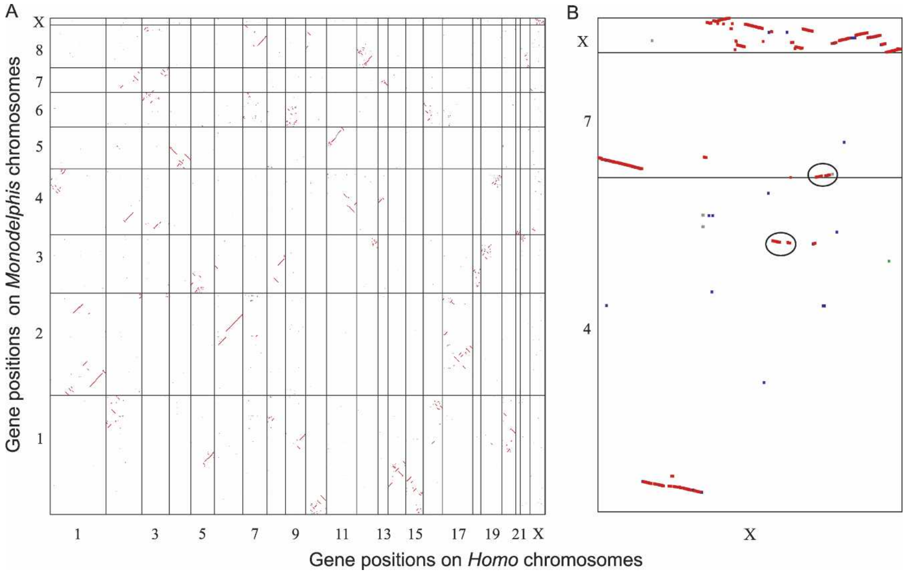

Figure 2. Oxford Grid (Edwards 1991) of PhyOP orthologs showing *Homo–Monodelphis* synteny blocks. (*A*) Genomic Synteny for all *Monodelphis* and *Homo* chromosomes. Genes are plotted in consecutive gene order along the *Monodelphis* chromosomes MDO1-8 and MDOX, and along the *Homo* chromosomes HSA1-22 and HSAX. One-to-one, one-to-many, many-to-one, and many-to-many *Homo*-to-*Monodelphis* orthologs are displayed as red, green, blue, and black dots, respectively. Diagonal lines represent genomic segments with conserved synteny. (*B*) The syntenic relationships for the ancient region of the *Homo* X chromosome that is syntenic to the *Monodelphis* X chromosome, and the more recently derived regions on the short arm of HSAX that are syntenic to MDO4 and MOD7. The circled regions have been placed on MDO4 and MDO7 in the current *Monodelphis* assembly 3, but FISH analyses map these regions to the Xq arm of MDOX (M. Breen and P. Water, pers. comm.).

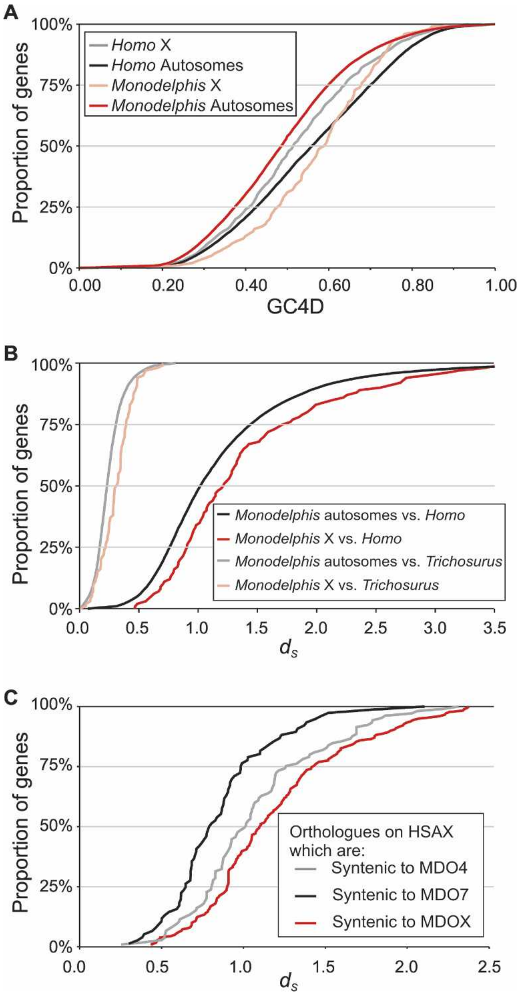

Figure 3 G+C Content at 4D sites (GC4D) and d_S in the *Monodelphis* and *Homo* X chromosomes. (*A*) GC4D values of *Monodelphis* genes in 1:1 orthology relationships with *Homo* genes. The cumulative distribution of GC4D for orthologs on the *Monodelphis* X chromosome is shifted to higher values compared with that for orthologs on *Monodelphis* autosomes, and genes on the *Monodelphis* X chromosome have a much increased median GC4D. This is exactly the opposite of the situation among *Homo* orthologs, where X chromosome genes have reduced GC4D. (*B*) The increased GC4D on the *Monodelphis* X chromosome is accompanied by an increase in median d_S and a rightward shift of the d_S cumulative distribution. This effect is even more pronounced when comparing orthologs between *Monodelphis* and *Trichosurus*, suggesting that much of the increase in d_S has taken place along the marsupial lineage. (*C*) The d_S for *Homo–Monodelphis* orthologs that lie on the *Homo* X chromosome and that are syntenic to the *Monodelphis* chromosome X or to chromosomes 4 and 7. Among orthologs on the *Homo* X chromosome, only those that are also found on the *Monodelphis* X chromosome have elevated d_S, confirming that this is largely a marsupial-specific phenomenon.

mosome that are syntenic to the *Monodelphis* chromosomes 4 and 7 and that appear to have conjoined the rest of the X chromosome early in eutherian evolution (Glas et al. 1999) (Figs. 2B, 3C; Supplemental Table 1). Moreover, nucleotide substitution rates and overall G+C contents are known to be suppressed, not elevated, in eutherian X chromosomes relative to their autosomes (Mouse Genome Sequencing Consortium 2002).

To further investigate whether the increase in X chromosome d_S is specific to metatheria, we compared *Monodelphis* genes with candidate orthologs from another marsupial, the Australian silver-gray brushtail possum (*Trichosurus vulpecula*). We aligned 111,634 *Trichosurus* expressed sequence tags (ESTs) to *Monodelphis* predicted genes and derived orthology relationships using a heuristic based on least divergence (smallest d_S value; see Methods). The median d_S value between the marsupial orthologs that are autosomal in *Monodelphis* is 0.28 ($n = 7804$), whereas that for orthologs, which are X chromosomal in *Monodelphis*, is 0.33 ($n = 93$) (Fig. 3C). (X chromosomal content is largely conserved between these two species [Rens et al. 2001].) These differences are highly significant ($P < 10^{-5}$). These results would be consistent with X chromosome elevation of d_S being largely a characteristic of the metatherian, rather than the eutherian, lineage.

Substitution rate and G+C content elevations in subtelomeric regions

We had previously noted a similar elevation of GC4D and d_S values for the smaller microchromosomes of chicken, relative to their larger macrochromosomes (Hillier et al. 2004). Increased GC4D and d_S values for chicken microchromosomes appear to be related to a more general phenomenon: Sequence from chromosome interiors located away from telomeres is associated with reduced G+C and lowered d_S. It seemed possible that the smaller proportion of interstitial sequence in the interior of the *Monodelphis* X chromosome could best explain the increased GC4D and d_S values. If so, we expect decreased GC4D and d_S in the autosomes simply because of their greater fractions of interstitial sequence.

Indeed, we find that in the *Monodelphis,* the median GC4D and d_S values of genes within 10 Mb of all assembled chromosomal telomere ends (0.67 and 1.37, respectively) are 43% and 38% higher than they are within interstitial regions (0.47 and 0.99, respectively) (Fig. 4; Table 3; Supplemental Fig. 1). These differences are highly significant ($P < 10^{-14}$ and $P < 10^{-3}$). We then investigated whether similar elevations were apparent for *Trichosurus* genes whose *Monodelphis* orthologs are within 10 Mb from a chromosomal end. The median d_S value (0.42; $n = 252$) of such *Trichosurus* genes is higher still than both the median d_S value (0.33; $n = 93$) between X chromosomal genes for these species and the median d_S value (0.25; $n = 7900$) for autosomal genes. The 65% increase in subtelomeric silent substitutions in *Monodelphis–Trichosurus* comparisons was substantially higher than that in *Monodelphis–Homo* orthologs (39%), indicating that elevation of d_S in subtelomeric regions is a characteristic of the metatherian, rather than the eutherian, lineage.

Increased efficacy of selection within high G+C regions

We had reason to believe that the same positional biases would be found for evolutionary rates (d_N/d_S), and that all of these observations arise from correlations with high-recombination rates (see Discussion). For the set of 12,898 1:1 *Monodelphis–Homo* orthologs and for the set of 6713 *Monodelphis–Trichosurus* ortholog alignments that had at least 100 aligning codons, we observed significant negative rank correlations between G+C and d_N/d_S ($P < 1 \times 10^{-6}$) (Fig. 5).

We then considered whether *Monodelphis* genes contained within high G+C regions have unusually short introns, as might be expected from previous observations linking high recombination rate and decreased intron length (Duret et al. 1995; Montoya-Burgos et al. 2003). Indeed, median intron lengths fell by fourfold for increasing G+C ($P < 1 \times 10^{-6}$) (Fig. 5). Significant

Table 2. Characteristics of genes from either the *Monodelphis* X chromosome or the autosomes MDO1-8 and their *Homo* orthologs

	Homo–Monodelphis 1:1 orthologs			
	X chromosome	Autosomes	Change on the X chromosome	*P*-value[a]
d_N / d_S	0.095 (0.047–0.150)	0.087 (0.044–0.154)	+8.7%	0.67
d_N	0.119 (0.057–0.203)	0.094 (0.043–0.173)	+26.3%	1.2×10^{-2}
d_S	1.213 (0.907–1.714)	1.005 (0.752–1.411)	+20.4%	$<10^{-3}$
Amino acid sequence identity	76.1% (65.7%–81.0%)	81.1% (70.0%–90.1%)	−6.3%	1.6×10^{-3}
G+C content at 4D sites	59.5% (49.7%–67.7%)	48.0% (36.5%–59.6%)	+24.0%	$<10^{-14}$
Intron length (bp)	14,909 (3215–42,971)	22,914 (6815–59,634)	−34.9%	$<10^{-3}$

Shown are median values and, in parentheses, lower and upper quartiles.

[a] *P*-values for the likelihoods that X chromosomal and autosomal distributions were equivalent were calculated using the Kolmogorov-Smirnov test.

reductions in both d_N/d_S and intron lengths are also seen for *Monodelphis–Homo* orthologs from the subtelomeres of *Monodelphis* chromosomes and from chromosome X (Tables 2, 3).

Lineage-specific biology and *Monodelphis* paralogs

Using PhyOP, we identified 2733 *Monodelphis* and 4105 *Homo* genes that have each arisen from duplications in their respective lineages since their last common ancestor. In the primary publication presenting the *Monodelphis* genome (Mikkelsen et al. 2007), we describe how these inparalogs are likely to participate in immunity or host defense, chemosensation, toxin degradation, and reproduction (see Supplemental Tables 2–6). The median d_N/d_S value for all *Monodelphis* inparalogs (0.51) (Table 4) is sixfold greater than that for *Monodelphis–Homo* orthologs (0.086), indicating that evolution of these genes has occurred under greatly relaxed constraints or widespread and recurrent episodes of adaptation.

KRAB zinc fingers have a different duplication time profile

As in other mammals, the KRAB zinc finger gene family has expanded rapidly in *Monodelphis*, with at least 350 members (Supplemental Table 5). Although we now know such genes have an ancient origin prior to the emergence of vertebrates (Birtle and Ponting 2006), the opossum is the most distantly related organism to humans known to have such a greatly expanded repertoire. Most *Monodelphis* inparalogs have low divergences (d_S values), suggesting that they are the result of recent duplications relative to the origin of the metatherian lineage (Fig. 6). KRAB zinc finger genes are the exception to this general rule. Among all functional classes, these genes appear to have experienced a burst of duplication at $d_S \sim 0.14$. Following the proposal of others (Vogel et al. 2006), the preferential localization of KRAB zinc finger genes in heterochromatic sequence may have led to a recent reduction in the duplication or loss of these genes during recombination.

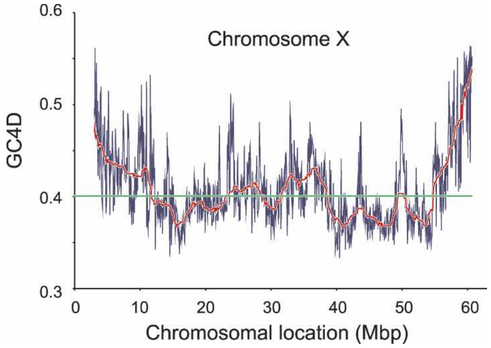

Figure 4. G+C content along the *Monodelphis* X chromosome. This plot shows the increase in G+C content at the telomeric end of the long arm of the *Monodelphis* X chromosome. The thin blue lines show the G+C content for adjacent 50-kbp windows along MDOX, while the thick red line is a running average of 50 such windows. The average G+C content for the entire chromosome is shown by the horizontal green line.

Chemosensation

The large expansions in gene families involved in chemosensation in the *Monodelphis* lineage relative to human provide ample evidence of their importance in nocturnal foraging and pheromonal communication (Supplemental Tables 2, 3). *Monodelphis* olfactory receptors (ORs), and V1R or V2R vomeronasal receptors have experienced numerous episodes of gene duplications, presumably as adaptive responses to changes in their environments and to conspecific competition. Those clades that experienced two or more gene duplications contain 468, 50, and 110 duplicated lineage-specific OR, V1R, and V2R genes, respectively. The large expansion of vomeronasal receptors may have been concomitant with the acquisition of unique structural adaptations in the *Monodelphis* vomeronasal organ, including a nuzzling "pad" thought to facilitate uptake of odorants (Poran 1998).

Three clusters of lipocalins have also been expanded, including one whose orthologs encode the major urinary protein pheromone in mice. However, *Monodelphis* uses skin and glandular secretions rather than urine for scent marking (Zuri et al. 2005). If these genes represent *Monodelphis* pheromones, they are thus likely to exhibit very different tissue-expression profiles. Another lipocalin cluster is orthologous to developmentally regulated milk protein genes in the tammar wallaby (Trott et al. 2002). Beta-microseminoprotein, an abundant constituent of seminal plasma, has been duplicated extensively in the *Monodelphis* lineage, resulting in 12 copies, whereas all other mammals (except New World monkeys that have three) have one (Makinen et al. 1999). There is evidence for positive selection at six sites among these *Monodelphis* inparalogs (data not shown), suggesting a role in conspecific competition during fertilization (Swanson and Vacquier 2002).

Immunity-related genes are evolving the fastest

The fastest evolving *Monodelphis* lineage-specific genes have roles in immunity and host defense. Their median d_N/d_S value of 0.80

Table 3. Characteristics of one-to-one orthologs that are found on the subtelomeric regions of *Monodelphis domestica* chromosomes

	Homo-Monodelphis 1:1 orthologs				*Trichosurus–Monodelphis* 1:1 orthologues			
	Subtelomere	Interstitial	Change at subtelomeres	*P*-value[a]	Subtelomere	Interstitial	Change at subtelomeres	*P*-value[a]
Number of 1:1 orthologs	629	11,604			252	7,900		
d_N/d_S	0.071 (0.039–0.125)	0.088 (0.045–0.155)	−18.7%	$<10^{-6}$	0.114[b] (0.041–0.230)	0.134[b] (0.050–0.277)	−14.9%[b]	0.095[b]
d_N	0.108 (0.053–0.186)	0.093 (0.043–0.173)	+16.1%	$<10^{-3}$	0.042 (0.016–0.098)	0.033 (0.011–0.073)	+37.3%	$<10^{-2}$
d_S	1.374 (1.047–1.957)	0.990 (0.746–1.386)	+38.7%	$<10^{-15}$	0.420 (0.300–0.527)	0.254 (0.188–0.336)	+65.4%	$<10^{-15}$
Amino acid sequence identity	78.4% (66.9%–87.4%)	81.2% (70.1%–90.1%)	−3.5%	$<10^{-5}$	91.0% (81.9%–95.8%)	92.6% (85.6%–96.7%)	−1.7%	<0.02
Monodelphis G+C content at 4D sites	67.2% (58.6%–76.9%)	47.1% (36.1%–58.5%)	+42.8%	$<10^{-15}$	59.4% (50.8%–68.5%)	43.0% (33.3%–53.5%)	+38.2%	$<10^{-15}$
Intron length (bp)	11,199 (3739–29,699)	23,563 (7099–60,932)	−52.5%	$<10^{-15}$	17,921 (5431–39,136)	27,058 (9047–66,219)	−33.8%	$<10^{-5}$

Shown are median values and, in parentheses, lower and upper quartiles. *P*-values are calculated using the Kolmogorov-Smirnov test.

[a] *P*-values were calculated using the Kolmogorov-Smirnov test.

[b] The elevated d_N/d_S for *Trichosurus–Monodelphis* orthologs is likely to be due to EST sequencing errors appearing as nonsynonymous changes. This would also explain the lower significance (higher *P*-value) for the decrease in subtelomeric d_N/d_S for these two species.

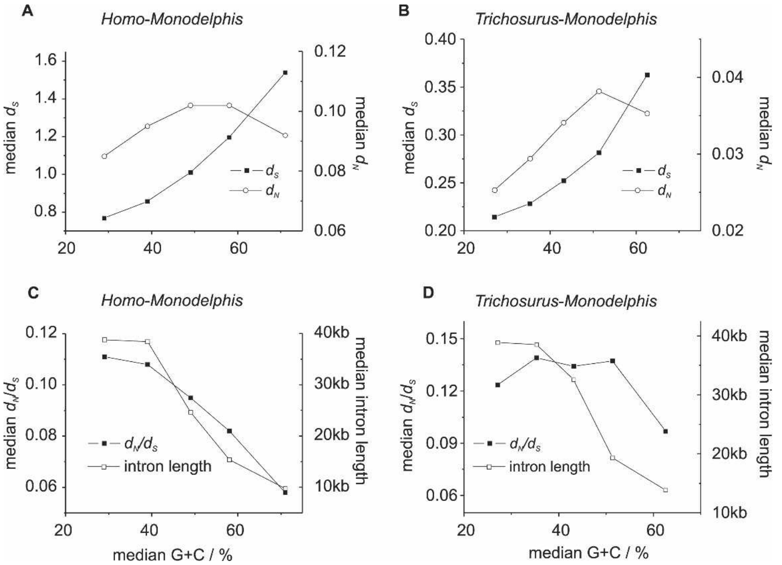

Figure 5. d_N, d_S, d_N/d_S, and intron length vary with G+C content at 4D sites among *Monodelphis–Homo* and *Monodelphis–Trichosurus* 1:1 orthologs. (*A,C*) The variation of the median values for d_N, d_S, d_N/d_S, and *Monodelphis* intron lengths in *Monodelphis–Homo* 1:1 orthologs. (*B,D*) The same relationships for 1:1 orthologs between the two marsupials *Monodelphis* and *Trichosurus*. The orthologs were divided into five equally populated classes according to *Monodelphis* G+C content at 4D sites (GC4D). We found that median d_N/d_S dropped from 0.12 for *Trichosurus* orthologs and 0.11 for *Homo* orthologs in the lowest G+C class (G+C <31.5% and <34.5%, respectively) to a median d_N/d_S of 0.10 and 0.058 in the highest G+C class (G+C >56.1% and >63.5%, respectively). The higher median d_N/d_S values for orthologs from the two marsupial species is likely to stem, in part, from *Trichosurus* EST sequencing errors that disproportionately affect d_N over d_S. Genes with high GC4D also exhibit higher median d_S and d_N values and have reduced intron lengths.

(Mikkelsen et al. 2007) is exceptionally high and perhaps indicates that the usual mammalian "genetic arms race" (Dawkins and Krebs 1979) with pathogens and parasites has been particularly severe in this marsupial lineage. Many immunoglobulin (IG) domain–containing proteins, such as IG chains, butyrophilins, leukocyte IG-like receptors, T-cell receptor chains, and carcinoembryonic antigen-related cell-adhesion molecules were found to be greatly expanded in the *Monodelphis* lineage (Supplemental Table 4). The chemokine *CCL4* is also greatly expanded in *Monodelphis*, with a total of five copies. *CCL4* inhibits infection by retroviruses such as HIV-1 in humans and may play a similar role in *Monodelphis* (Menten et al. 2002). Although lineage-specific duplication and adaptation of pancreatic RNases have previously been associated with dietary adaptations in foregut-fermenting herbivorous mammals (Zhang et al. 2002), the modest expansion to three homologs in *Monodelphis* may serve an immunological rather than a dietary role, since this opportunistic omnivore possesses only a relatively simple alimentary canal (see also Yu and Zhang 2006).

Dietary adaptation

In other ways, the *Monodelphis* genome does exhibit evidence for past adaptation to dietary changes. The six copies of the single exon hypoxanthine phosphoribosyltransferase homologs on chromosome 4 (~346 Mb) may reduce the loss of nitrogen via urinary excretion of allantoin, as suggested previously (Noyce et al. 1997), contributing to the marsupial tolerance of nitrogen-poor diets (Hume 1982). Other genes that have been duplicated in the genome are *SLC39A4*, encoding a zinc transporter whose expression is up-regulated in mouse under conditions of dietary zinc deficiency (Dufner-Beattie et al. 2003) and thiamine pyrophosphokinase 1 homologs, which are likely to be involved in the salvage of thiamine (vitamin B1). The duplication of various genes encoding gastric enzymes in the *Monodelphis* lineage have been discussed elsewhere (Mikkelsen et al. 2007).

Many other *Monodelphis*-specific genes have functions that fall outside of the typical mammalian themes of chemosensation, reproduction, immunity, and detoxification (Table 5). Notable among these genes are those encoding proteins involved in mucus production (*CLCA1* and *MUC16*), splicing factors (*SMG5*, *YTHDC1*, *SMG6*, and *CWC22* [*KIAA1604*]), lysosomal enzymes (cathepsin L and *GALC*), renins, and multiple keratins. The identification of these genes should now allow greater scrutiny of their contributions to *Monodelphis*- and marsupial-specific biology.

Finally, we were interested in whether orthologs of *Didelphis marsupialis* DM43 and DM40, which confer natural resistance to snake venoms in this related marsupial (Neves-Ferreira et al. 2000), could be found in *Monodelphis*. However, despite exhaustive searches of the current genome assembly, no substantially similar sequences to DM43 and DM40 were identified, perhaps indicating that these genes have evolved particularly rapidly.

Discussion

With the newly sequenced genome of *Monodelphis domestica* comes tremendous potential for attributing genetic and genomic variation to metatherian- or eutherian-specific traits. We have

Table 4. **Characteristics of genes that have undergone lineage-specific duplications either in *Homo sapiens* and/or *Monodelphis domestica***

	Homo sapiens	*Monodelphis domestica*
d_N/ d_S among inparalogs	0.617 (0.345–0.972)	0.506 (0.345–0.685)
d_N/ d_S with ortholog(s) in other species	0.202 (0.119–0.399)	0.208 (0.128–0.405)
d_N with ortholog(s) in other species	0.257 (0.162–0.469)	0.272 (0.176–0.475)
d_S with ortholog(s) in other species	1.236 (0.943–1.640)	1.247 (0.960–1.650)
Amino acid sequence identity with ortholog(s) in other species	61.5% (45.8%–74.0%)	60.4% (45.7%–72.8%)
Pairwise alignment coverage of the longer sequence with ortholog(s) in other species	76.9% (48.4%–93.9%)	79.5% (53.2%–95.0%)
Number of exons	3 (1–6)	3 (1–5)
Sequence length (codons)	226 (122–410)	318 (246–493)
Unspliced transcript length (bp)	3359 (933–11,569)	3288 (951–14,725)
Monodelphis G+C content at 4D sites	52.2% (43.2%–64.8%)	46.3% (37.0%–54.5%)

Shown are median values and, in parentheses, lower and upper quartiles.

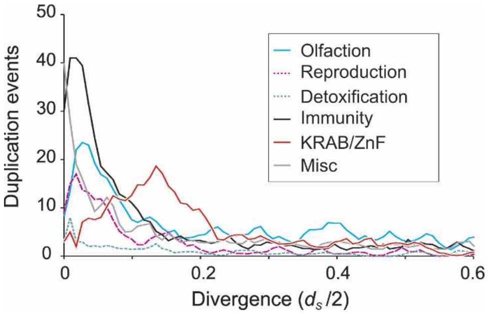

Figure 6. Inferred rates of duplication for *Monodelphis* inparalogs among different functional categories. The timing of duplication events for *Monodelphis*-specific inparalogs were inferred from the d_S-based rooted phylogenetic tree, and thus appears on a scale of $d_S/2$ units. Although d_S values vary substantially among genes (Fig. 3B), on average, a $d_S/2$ value of 0.25 represents ~90 million years. Genes in most functional classes (including olfaction, reproduction, detoxification, and immunity) follow a profile typical of other mammals with duplications peaking very recently. This would be in accordance with a rapid gene birth and death model. It is likely that reduction in the number of most recent duplications ($d_S/2 < 0.02$) is due partly to the collapse of nearly identical sequence in the assembly, and partly to underprediction of sequence-similar adjacent paralogs and overestimation of very small d_S values. The duplications of KRAB-ZnF genes are prominent in peaking at $d_S/2$ ~ 0.14.

shown that 82% of *Monodelphis* genes have demonstrable orthologs in a representative eutherian, *Homo sapiens*. Fifteen percent of *Monodelphis* genes appear to have arisen through gene duplications in the metatherian lineage. Many of these exhibit relatively little divergence (Fig. 6), indicating that they have arisen recently and may succumb eventually to inactivation and loss over longer time periods. This would be in agreement with what is seen in other mammals (International Human Genome Sequencing Consortium 2004) and metazoans (Lynch and Conery 2000), and would be consistent with a model of rapid birth and death of duplicate genes. Most *Monodelphis* inparalogs fall into the same broad functional classes seen in eutherian gene duplications. However, there are at least 283 duplicated genes (Table 5) whose functions do not belong to any of these popular categories. Each metatherian-specific gene is now available for further investigation of its contribution to the unique biology of this lineage (Samollow 2006).

Predicted orthologs from *Trichosurus vulpecula*

We have also predicted 8237 genes from the Australian marsupial *Trichosurus vulpecula* (Kerle et al. 1991) as orthologs of *Monodelphis* genes. These enable comparisons between a South American and an Australian marsupial whose lineages diverged, we estimate, 46–53 Mya, around the date (~45 Mya) at which land migration via Antarctica ceased between Australia and South America (Li and Powell 2001). These estimates were obtained by dividing these marsupials' median autosomal or X chromosomal d_S values by the equivalent values between *Homo* and *Monodelphis* and scaling by the estimated time (170–190 Mya) separating them from their last common ancestor (Kumar and Hedges 1998; Woodburne et al. 2003). This divergence time is considerably more recent than the previously estimated 60–70 Mya (Nilsson et al. 2004).

Protein coding gene counts

Our studies also contribute to an understanding of the differences in protein coding gene numbers encoded within different mammalian genomes. Previous estimates of the human gene count, for example, appear to have been inflated because of contributions from retrotransposed pseudogenes and transcribed, but noncoding, sequences (Goodstadt and Ponting 2006). Our previous lower-bound estimate of the human gene count, from a comparison with the predicted set of dog protein coding genes, was 19,700 (Goodstadt and Ponting 2006), and a comparable estimate (20,806) arises from this comparison with the predicted *Monodelphis* gene set. These estimates serve to re-emphasize that mammalian gene counts, in general, are little different from those of nonvertebrate metazoa such as nematode worms, and thus should not be considered as an appropriate measure of organismal complexity. The lower estimated gene count in *Monodelphis* (18,639) is likely to be the result of under-prediction rather than lineage-specific gene gains or losses. The missing genes may be partly attributed to the draft quality of the genome sequence. However, they probably mostly reflect the challenges of using distant-related eutherian evidence to find genes in a marsupial without the benefit of a large set of *Monodelphis* transcripts.

Decline in mammalian G+C content

It has been suggested that G+C content has declined substantially during metatherian evolution (Belle et al. 2004). This would certainly be consistent with the high G+C content of the platypus (*Ornithorhynchus anatinus*) (Margulies et al. 2005). Our results show that any such decline would have been most precipitous for the *Monodelphis* autosomes, particularly in their interstitial regions located well away from their telomeres. Decreases in G+C content are often assumed to be a consequence of the high mutation rate of cytosines in methylated CpG dinucleotides (Brown and Jiricny 1987). We note that *Monodelphis* possesses two inparalogs of the DNA (cytosine-5)-methyltransferase 1, whereas eutherians (and most other vertebrates) have only one. If these two copies together possess higher aggregate DNA methylation rates than their single eutherian counterpart, this could explain, at least in part, the stronger decline in G+C among metatheria than in eutheria.

High metatherian recombination rates in the short X chromosome and subtelomeric regions

The decline in G+C appears to have least affected the *Monodelphis* X chromosome and the subtelomeric regions of autosomes. Several lines of evidence suggest that this may be a consequence of biased gene conversion driven by female-specific recombination. Allelic gene conversion during meiotic recombination is proposed to be biased toward insertion of G or C over A or T, leading to an increase in G+C content within highly recombining regions (Duret et al. 2006).

The *Monodelphis* X chromosome can be expected to have a higher recombination rate because of its considerably (four- to 12-fold) smaller size and the obligatory minimum of one chiasma per chromosomal arm. This would be despite the lower recombination rate in *Monodelphis* females (Samollow et al. 2004). (Recombination in the X chromosome is, by definition, female specific.) Although we as yet lack comprehensive data, it appears that chiasmata in female meiotic cells in *Monodelphis* and other marsupials are concentrated close to telomeres (Bennett et al. 1986), contributing to a bias toward recombination at chromosome ends, and of course, within the X chromosome. This is exactly where the highest G+C values can be seen (Fig. 4). Chi-

Table 5. **Paralogous gene clusters in *Monodelphis* that have experienced at least two lineage-specific duplications and are not associated with functions relating to immunity and host defense, detoxification, reproduction, chemosensation, and KRAB–zinc finger-related transcription regulation**

Gene count			Chromosomes		
Monodelphis	*Homo*	d_S	*Monodelphis*	*Homo*	Description
27	1	1.2	8,Un,3	19	Mucin 16, serum marker for women with ovarian cancer
24			Un,3,5		Mas-related G-protein coupled receptors, implicated in nociception
10	1	1.1	Un,2	1	CLCA1, involved in mucus production
8			3		Protein SMG5 (EST1-like protein B), regulates telomerase and nonsense-mediated mRNA decay
7	1	1.4	Un,2	1	Renin (angiotensinogenase).
7			8,4,Un		COX6B2, cytochrome *c* oxidase subunit Vib
6	15	1.6	1	5	Protocadherin beta family, involved in brain development
6	4	1.6	7	2	Gamma crystallins
6	1	1.2	7	3	Liver arylacetamide deacetylase (AADAC)
6	1	1.4	2	6	Histones H3.1
5	1	0.8	8,1,5	4	Putative splicing factors YTHDC1
5	1	1.1	1	14	Galactocerebrosidase (lysosomal enzyme); mutated in Krabbe disease
5			1		Protocadherin gamma family, involved in brain development
5			3		LAG1 longevity assurance homolog 5; regulate synthesis of ceramides
5			Un,5		Telomerase-binding protein SMG6, regulates telomerase and nonsense-mediated mRNA decay
4	32	2.6	3	8,22,1	PRAMEs, Cancer-testis antigens
4	2	1.3	6	9	Cathepsin L-like lysosomal enzymes
4	1	1.2	2	17	Arachidonate 12-lipoxygenase, 12S-type, oxygenates carbon atoms of the fatty acid arachidonic acid
4	1	2.0	3	8	Melanoma-derived leucine zipper-containing extranuclear factor
4	1	1.6	7	3	Arylacetamide deacetylase-like 2 (AADACL2)
4	1	2.3	5	4	Sodium-dependent phosphate transport protein 2B
4	1	1.4	2	17	Keratin, type I cytoskeletal 10
4			Un,7		Intraflagellar transport 80 homolog (WD-repeat protein 56)
4			1,4		Hypoxanthine-guanine phosphoribosyltransferase (HGPRT) marsupial-specific retrogenes
4			Un,X		Mortality factor 4-like protein; histone acetylase complex subunit MRG15-2
3	22	1.3	8,Un	21,9,17,2, 22,1,18,10	Ankyrin repeat domain-containing proteins
3	5	1.2	2	17	Hair keratins, type I
3	5	1.3	2	17	Keratins, type I cytoskeletal
3	5	1.6	Un	12	Keratins, type II cytoskeletal
3	4	0.9	2	17	ATP-binding cassette, subfamily A transporters
3	4	1.3	Un	12	Keratins, type II cuticular
3	3	1.1	2	17	Type I inner root sheath-specific keratins 25
3	2	1.1	1	5	Solute carrier family 36 (proton/amino acid symporter) members
3	2	3.2	5	20	Oxytocin-neurophysin 1/vasopressin-neurophysin 2-copeptin
3	2	1.9	4	1	F-box only protein 6 and 44
3	1	0.9	4,2	2	Homolog of yeast pre-mRNA-splicing factor CWC22
3	1	1.9	1,3	20	Transcription factor SOX17, involved in endoderm formation
3	1	1.8	1	10	Pulmonary surfactant-associated protein D
3	1	0.9	5	4	Gastric alcohol dehydrogenase (retinol dehydrogenase)
3	1	1.0	8	10	Pre-mRNA-splicing factor 18
3	1	1.8	3	19	DNA (cytosine-5)-methyltransferase 1 (DNMT1) (unusually duplicated)
3	1	0.9	4	2	Ortholog of yeast vacuolar amino acid transporter 2 (unusually duplicated)
3	1	2.2	1,5	9	VAV2, a guanine nucleotide exchange factor for Rac
3	1	1.6	2	3	NmrA-like family domain containing 1 proteins
3	1	1.0	4	13	Potassium-transporting ATPase alpha chain 2 in skin and kidney
3	1	1.5	4	21	Cystatin B (Liver thiol proteinase inhibitor); mutated in progressive myoclonus epilepsy
3	1	1.0	4,3	3	LRRIQ2, leucine-rich repeats and IQ motif containing 2
3	1	2.0	1	2	Sepiapterin reductase, possesses role in the biosynthesis of tetrahydrobiopterin.
3	1	2.3	1,Un	2	C2orf44 ortholog, of unknown function
3	1	0.9	6,Un	16	Zymogen granule protein 16, secretory lectin ZG16
3	1	1.3	2	6	Histone H2B.f
3			4		Folate receptor 4 (delta) homologs
3			2		Gastricsins (pepsinogens c), gastric digestive proteinases
3			Un		ATP-binding cassette, subfamily A transporters
3			8,Un		Ankyrin repeat-containing proteins of unknown function
3			3		Secretor blood group alpha-2-fucosyltransferases
3			3		SLC39A4, solute carrier family 39, member 4; zinc transporter

Tables related to these latter functional categories are provided as Supplemental Tables 1–5. Where orthology relationships to human genes have been predicted, the gene count and chromosomal location of the corresponding human orthologs and the median d_S between orthologs are included.

asmata are more evenly distributed in male cells (Hayman et al. 1988).

Increased G+C content and recombination have previously been associated with increased numbers of synonymous substitutions (Wolfe et al. 1989; Matassi et al. 1999; Hardison et al. 2003; Webber and Ponting 2005). We see an increase in d_S not only in regions with high G+C content, but also specifically in the X chromosome and subtelomeric regions where we expect increased recombination. The comparisons of *Monodelphis*–*Trichosurus* orthologs suggest that these changes have occurred largely on the marsupial lineage.

Recombination promotes greater efficiency of selection

High recombination is believed to increase the efficiency of selection by disrupting interference between neighboring mutations, the "Hill-Robertson" effect (Hill and Robertson 1966). Because most nonsynonymous mutations are deleterious, this would tend to increase purifying selection. A higher recombination rate in the marsupial X chromosome and subtelomeric regions might then explain the reduced d_N/d_S among *Monodelphis* orthologs from these regions, as well as among genes with high G+C. The same evolutionary forces may explain the decrease in intron lengths in such *Monodelphis* regions, a phenomenon that has also previously been associated with high recombination (Duret et al. 1995; Montoya-Burgos et al. 2003).

Chromosomal rearrangements

The decline of G+C content in metatherian autosomes may also be associated with reduced rates of intra- or interchromosomal rearrangements. This is because synteny break regions, at least in eutheria, are enriched within regions exhibiting high G+C levels and d_S rates (Marques-Bonet and Navarro 2005; Webber and Ponting 2005). Thus, we might expect rearrangements to have occurred preferentially in the *Monodelphis* X chromosome and near the autosomal telomeres. Indeed, we note that one type of rearrangement, namely segmental duplication, is overrepresented in the X chromosome relative to the remainder of the genome (Mikkelsen et al. 2007).

This theory, building upon our own work (Webber and Ponting 2005) and that of others (Duret et al. 2002, 2006; Marques-Bonet and Navarro 2005), assumes that high recombination rates maintain high G+C levels, and that chromosomal regions of high G+C are unusually susceptible to breakage and consequent rearrangement. It provides five testable predictions. (1) G+C-poor chromosomes tend to be larger, which is the case for human (Duret et al. 2002), chicken (Hillier et al. 2004), and *Monodelphis* (this study) chromosomes. (2) G+C-poor chromosomes have experienced less recombination. Although more data are needed, there is evidence that this is indeed the case (Samollow et al. 2004). (3) G+C-rich regions preferentially segregate to chromosomal ends as a direct result of their susceptibility to breakage (Webber and Ponting 2005). As discussed above, regions enriched in G+C content, and with high (female-specific) recombination rates, exhibit a strong tendency to be located near telomeres. (4) Conversely, low G+C regions preferentially segregate to within chromosomal interiors and would be relatively refractive to breakage. The very limited number of chromosomal rearrangements observed among diverse marsupials would appear to support this (Rens et al. 2001). (5) Susceptibility to recombination is preserved, in part, across the mammalia; it is an ancestral, rather than a derived trait. This would explain the high rank order correlation between the GC4D values in *Monodelphis* and *Homo* (as they are between chicken and eutheria (Webber and Ponting 2005). Although genetic maps are only available for a few mammals, it is known that recombination rates in human, rat, and mouse syntenic sequence are moderately correlated (Jensen-Seaman et al. 2004). These five predictions will be available for testing upon the sequencing of additional genomes, such as those of the platypus, cattle, and songbird.

These issues of variable rates of recombination, mutation, and selection, together with the identification of genes that distinguish, say, Australian from American marsupials, will necessitate the sequencing of a second marsupial's genome. The *Monodelphis* genome sequence has provided a broad perspective of the features that distinguish metatherian from eutherian genomes. However, until a second genome of this distinctive order of mammals is sequenced, its idiosyncrasies will, by necessity, not be separable from general metatherian characteristics.

Methods

A more comprehensive set of *Monodelphis* gene predictions

We augmented a preliminary set of *Monodelphis* gene predictions from Ensembl with additional gene predictions using the Exonerate program (version 0.9) (Slater and Birney 2005) on the same MonDom3 genome assembly. Briefly, we used *Homo* protein coding transcripts (Ensembl release version 36 based on NCBI assembly 35) as templates for predicting *Monodelphis* transcripts. Exonerate predicted transcripts that overlapped an existing Ensembl prediction were discarded.

Lineage-specific paralogs present greater difficulties for gene prediction than other genes, due to their more rapid sequence divergence (Mouse Genome Sequencing Consortium 2002), and their frequent location in tandem clusters. We, therefore, initiated a second round of gene prediction using both *Homo* and *Monodelphis* sequences from gene families with *Monodelphis* lineage-specific duplications (see below) as templates.

Altogether, we were able to predict 657 additional *Monodelphis* genes to supplement Ensembl data. A more detailed description of the gene prediction pipeline is contained in the Supplemental information.

Inferring orthology and paralogy relationships

Orthology and paralogy relationships between *Monodelphis* and *Homo* genes were predicted using PhyOP (Goodstadt and Ponting 2006). This reconstructed the phylogeny for all *Monodelphis* and *Homo* transcripts using d_S as a proxy for their evolutionary distances. We collated all peptide sequences from *Monodelphis* (assembly 3) and *Homo* (Ensembl release 38 based on NCBI assembly 36) and identified homologs using BLASTP and an *E*-value upper threshold of 1×10^{-5}. We only discarded spurious and fragmentary alignments that were shorter than 50 residues or where <75% of the shorter sequence was included in the alignment. Homologs were clustered together and the number of synonymous substitutions per synonymous site (d_S) was calculated using the codeml program from the PAML package (Yang 1997) with default settings for pairwise analyses (F3X4). We took sets of sequences related by d_S values previously shown in simulation to have acceptable reliability ($d_S < 2.5$) and constructed rooted phylogenies using a modified version of the kitsch algorithm (applying the Fitch-Margoliash criterion) from the PHYLIP (Felsenstein 1981) suite of programs. Orthology relationships among the transcripts were inferred automatically by minimizing the number of duplications that must be invoked to reconcile the transcript phylogeny with the species tree.

The human gene set included a number of allelic variants on chromosomes 5, 6, and 22. We discarded such sequence unless the corresponding allele was missing from the same loci in the reference genome or unless the two alleles showed substantial divergence ($d_S > 0.5$).

We observed 33,446 *Homo* transcripts from 16,471 genes and 26,360 *Monodelphis* transcripts from 16,261 genes in orthology relationships. A single representative transcript for each gene was then chosen in order to map transcript phylogeny to orthology relationships between genes. We iteratively selected transcript pairs from both species with the lowest d_S value, while eliminating overlapping alternative transcripts. This procedure also allowed the identification of erroneously merged adjacent paralogs whose representative transcripts after separation did not overlap. The heuristic for the selection of representative transcripts necessarily left 347 *Homo* and 268 *Monodelphis* orthologous genes whose transcripts were inconsistent with the final representative phylogenies (referred to as "orthologs with inconsistent phylogeny").

Lineage-specific paralogs

We also sought to identify those "orphaned" genes that have been duplicated in the *Homo* or *Monodelphis* lineages (inparalogs), but whose ortholog in the other species is either not present or has not been predicted correctly. We selected clusters of transcripts without predicted orthology whose divergences and phylogeny indicate species-specific duplications, together with transcripts from orthologs with inconsistent phylogeny. We filtered out all pairwise relationships that are likely to predate the divergence between the *Monodelphis* and *Homo* lineages by using a d_S cut-off equivalent to the median d_S value (1.02) between predicted one-to-one orthologs. As in the case of the prediction of orthologs, sets of inparalogs were created by selecting transcript pairs from both species with the lowest d_S value, while eliminating overlapping alternative transcripts. This procedure similarly allowed the identification of erroneously merged adjacent paralogs.

As described above, we took *Homo* and *Monodelphis* sequences from duplicated gene families as templates for additional gene prediction. All analyses in this study used orthologs and paralogs inferred from this final gene set.

Identifying pseudogenes

Putative pseudogenes, mostly representing retro-transpositions, were identified by the presence of disruptions (defined by short introns [<10 bp] among Ensembl genes) and the loss of introns along with the absence of synteny (see below). We labeled as pseudogenes any nonsyntenic ("dispersed") gene with one or more disruptions, syntenic genes with multiple disruptions, and dispersed single exonic genes. Any ortholog families with Interpro matches for L1 transposable elements (IPR004244) were also identified as pseudogenes.

Pseudogenes are retrotransposed in random locations and tend, therefore, to be found on multiple chromosomes. For widely dispersed families of orthologs (with members on four or more chromosomes), we first attempted to reliably identify their original "parent" genes (genes from which retrocopies were derived). We selected orthologs that had three or more exons with matching exon boundaries across both species. In such cases, we could then go on to identify retro-transposed pseudogene family members containing two or fewer exons with nonmatching boundaries.

Manual curation of families with three or more inparalogs in the *Monodelphis* lineage identified another 432 candidate pseudogenes, including retrotransposed genes, retroviral sequences, and genes predicted on the wrong strand. We also labeled as noncoding all *Homo* genes that do not have an identifiable homolog among *Mus musculus* (Ensembl version 40.36a), *Canis familiaris* (Ensembl version 40.1i), or *Monodelphis* sequences. Many of these are likely to reflect spurious open reading frames called within the untranslated regions of real transcripts (E. Birney, pers. comm.)

Conserved synteny

The orthology relationships allowed us to identify areas of conserved synteny in the *Monodelphis* and *Homo* genomes. We constructed "micro-syntenic" blocks by grouping together successive genes with conserved gene order and orientation among predicted 1:1 orthologs in the other species. "Macro-syntenic" blocks could then be identified by concatenating contiguous micro-syntenic blocks that, after rearrangements and inversions, would have conserved gene order in the other species. Loss of synteny, especially in the identification of retrotransposed pseudogenes, was defined as a disruption of the gene order between both upstream and downstream neighbors of its orthologs in the other species by >50 genes.

Trichosurus orthologs of *Monodelphis* genes

We calculated d_S values between *Monodelphis* predicted transcripts and 111,634 ESTs from an Australian marsupial, the silver-gray brushtail possum (*Trichosurus vulpecula*). Alignments to the longest predicted *Monodelphis* transcript of each gene used tfasty and default values (Pearson 2000). Frame-shift positions in alignments were checked for indication of intron run-off and unusually low-sequence identity (<25%); such stretches of alignments were subsequently masked. d_S values were calculated using codeml (Yang 1997) and matches with a d_S value exceeding three times the lowest d_S match for that query were removed. Likely, paralog matches were eliminated by removing all alignments exceeding three times the overall median d_S value of 0.26. Each matching EST was then assigned to a particular query sequence by virtue of its lowest d_S value. All hits for a query were combined into a consensus sequence with conflicting positions masked. Subsequently, d_S values were estimated anew between this consensus sequence and the *Monodelphis* query sequence. We recovered alignments to ESTs for 8237 predicted *Monodelphis* transcripts. These exhibited an average 58% of nucleotides covered per transcript. A total of 343 predicted transcripts with EST alignments were located on unplaced contigs in the *Monodelphis* assembly. On average, 33% of transcript nucleotides were aligned to multiple ESTs.

Genes in subtelomeric regions

Subtelomeric regions were defined as the 10 Mb of sequence at the end of each assembled chromosome sequence. For the *Monodelphis* metacentric chromosomes (MDO1 and MDO2) (Svartman and Vianna-Morgante 1999), the tail ends of both arms were included. For the other acrocentric/subtelocentric chromosomes (MDO3-8, MDOX), only the tail ends of the long arms were used. Only genes that fell entirely within these defined regions were included in the analyses.

Statistical tests

We used the nonparametric Kolmogorov-Smirnov Test implemented in the R package (R Development Core Team 2006) to evaluate the statistical significance in comparing distinct distributions.

Acknowledgments

These studies were funded by the UK Medical Research Council. We thank all members of the *Monodelphis domestica* genome sequencing consortium, in particular, Kerstin Lindblad-Toh, Tarjei Mikkelsen, and Paul Samollow, for their assistance and helpful comments.

References

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

Bennett, J.H., Hayman, D.L., and Hope, R.M. 1986. Novel sex differences in linkage values and meiotic chromosome behaviour in a marsupial. *Nature* **323:** 59–60.

Birtle, Z. and Ponting, C.P. 2006. Meisetz and the birth of the KRAB motif. *Bioinformatics* **22:** 2841–2845.

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

Buchmann, P., Schneider, K., and Gebbers, J.O. 1983. Fibrosis of experimental colonic anastomosis in dogs after EEA stapling or suturing. *Dis. Colon Rectum* **26:** 217–220.

Castillo-Davis, C.I., Kondrashov, F.A., Hartl, D.L., and Kulathinal, R.J. 2004. The functional genomic distribution of protein divergence in two animal phyla: Coevolution, genomic conflict, and constraint. *Genome Res.* **14:** 802–811.

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

Dawkins, R. and Krebs, J.R. 1979. Arms races between and within species. *Proc. R. Soc. Lond. B. Biol. Sci.* **205:** 489–511.

Dufner-Beattie, J., Wang, F., Kuo, Y.M., Gitschier, J., Eide, D., and Andrews, G.K. 2003. The acrodermatitis enteropathica gene ZIP4 encodes a tissue-specific, zinc-regulated zinc transporter in mice. *J. Biol. Chem.* **278:** 33474–33481.

Duret, L., Mouchiroud, D., and Gautier, C. 1995. Statistical analysis of vertebrate sequences reveals that long genes are scarce in GC-rich isochores. *J. Mol. Evol.* **40:** 308–317.

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

Duret, L., Eyre-Walker, A., and Galtier, N. 2006. A new perspective on isochore evolution. *Gene* **385:** 71–74.

Edwards, J.H. 1991. The Oxford Grid. *Ann. Hum. Genet.* **55:** 17–31.

Emes, R.D., Goodstadt, L., Winter, E.E., and Ponting, C.P. 2003. Comparison of the genomes of human and mouse lays the foundation of genome zoology. *Hum. Mol. Genet.* **12:** 701–709.

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

Fadem, B.H. and Rayve, R.S. 1985. Characteristics of the oestrous cycle and influence of social factors in grey short-tailed opossums (*Monodelphis domestica*). *J. Reprod. Fertil.* **73:** 337–342.

Felsenstein, J. 1981. Evolutionary trees from DNA sequences: A maximum likelihood approach. *J. Mol. Evol.* **17:** 368–376.

Glas, R., Marshall Graves, J.A., Toder, R., Ferguson-Smith, M., and O'Brien, P.C. 1999. Cross-species chromosome painting between human and marsupial directly demonstrates the ancient region of the mammalian X. *Mamm. Genome* **10:** 1115–1116.

Goodstadt, L. and Ponting, C.P. 2006. Phylogenetic reconstruction of orthology, paralogy, and conserved synteny for dog and human. *PLoS Comput. Biol.* **2**: e133.

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

Hayman, D., Moore, H., and Evans, E. 1988. Further evidence of novel sex differences in chiasma distribution in marsupials. *Heredity* **61:** 455–458.

Hill, W.G. and Robertson, A. 1966. The effect of linkage on limits to artificial selection. *Genet. Res.* **8:** 269–294.

Hillier, L.W., Miller, W., Birney, E., Warren, W., Hardison, R.C., Ponting, C.P., Bork, P., Burt, D.W., Groenen, M.A., Delany, M.E., et al. 2004. Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature* **432:** 695–716.

Hughes, A.L. 1994. The evolution of functionally novel proteins after gene duplication. *Proc. Biol. Sci.* **256:** 119–124.

Hume, I.D. 1982. *Digestive physiology and nutrition of marsupials*. Cambridge University Press, Cambridge, UK.

International Human Genome Sequencing Consortium. 2004. Finishing the euchromatic sequence of the human genome. *Nature* **431:** 931–945.

Jensen-Seaman, M.I., Furey, T.S., Payseur, B.A., Lu, Y., Roskin, K.M., Chen, C.F., Thomas, M.A., Haussler, D., and Jacob, H.J. 2004. Comparative recombination rates in the rat, mouse, and human genomes. *Genome Res.* **14:** 528–538.

Kerle, J.A., McKay, G.M., and Sharman, G.B. 1991. A systematic analysis of the brushtail possum, *Trichosurus-vulpecula* (Kerr, 1792) (Marsupialia, Phalangeridae). *Aust. J. Zool.* **39:** 313–331.

Kumar, S. and Hedges, S.B. 1998. A molecular timescale for vertebrate evolution. *Nature* **392:** 917–920.

Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., et al. 2001. Initial sequencing and analysis of the human genome. *Nature* **409:** 860–921.

Li, Z.X. and Powell, C.M. 2001. An outline of the palaeogeographic evolution of the Australasian region since the beginning of the Neoproterozoic. *Earth-Science Reviews* **53:** 237–277.

Lindblad-Toh, K., Wade, C.M., Mikkelsen, T.S., Karlsson, E.K., Jaffe, D.B., Kamal, M., Clamp, M., Chang, J.L., Kulbokas III, E.J., Zody, M.C., et al. 2005. Genome sequence, comparative analysis and haplotype structure of the domestic dog. *Nature* **438:** 803–819.

Lynch, M. and Conery, J.S. 2000. The evolutionary fate and consequences of duplicate genes. *Science* **290:** 1151–1155.

Lynch, M. and Force, A. 2000. The probability of duplicate gene preservation by subfunctionalization. *Genetics* **154:** 459–473.

Makinen, M., Valtonen-Andre, C., and Lundwall, A. 1999. New World, but not Old World, monkeys carry several genes encoding beta-microseminoprotein. *Eur. J. Biochem.* **264:** 407–414.

Margulies, E.H., Maduro, V.V., Thomas, P.J., Tomkins, J.P., Amemiya, C.T., Luo, M., and Green, E.D. 2005. Comparative sequencing provides insights about the structure and conservation of marsupial and monotreme genomes. *Proc. Natl. Acad. Sci.* **102:** 3354–3359.

Marques-Bonet, T. and Navarro, A. 2005. Chromosomal rearrangements are associated with higher rates of molecular evolution in mammals. *Gene* **353:** 147–154.

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

Menten, P., Wuyts, A., and Van Damme, J. 2002. Macrophage inflammatory protein-1. *Cytokine Growth Factor Rev.* **13:** 455–481.

Mikkelsen, T.S., Wakefield, M.J., Aken, B., Amemiya, C.T., Chang, J.L., Duke, S., Garber, M., Gentles, A.J., Goodstadt, L., Heger, A., et al. 2007. Genome of the marsupial *Monodelphis domestica* reveals innovation in non-coding sequences. *Nature* **447:** 167–177.

Montoya-Burgos, J.I., Boursot, P., and Galtier, N. 2003. Recombination explains isochores in mammalian genomes. *Trends Genet.* **19:** 128–130.

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

Neves-Ferreira, A.G., Cardinale, N., Rocha, S.L., Perales, J., and Domont, G.B. 2000. Isolation and characterization of DM40 and DM43, two snake venom metalloproteinase inhibitors from *Didelphis marsupialis* serum. *Biochim. Biophys. Acta* **1474:** 309–320.

Nilsson, M.A., Arnason, U., Spencer, P.B., and Janke, A. 2004. Marsupial relationships and a timeline for marsupial radiation in South Gondwana. *Gene* **340:** 189–196.

Noyce, L., Conaty, J., and Piper, A.A. 1997. Identification of a novel tissue-specific processed HPRT gene and comparison with X-linked gene transcription in the Australian marsupial *Macropus robustus*. *Gene* **186:** 87–95.

Ohno, S. 1970. *Evolution by gene duplication*. Springer-Verlag, Heidelberg, Germany.

Ohta, T. 1973. Slightly deleterious mutant substitutions in evolution. *Nature* **246:** 96–98.

Pardo-Manuel de Villena, F. and Sapienza, C. 2001. Female meiosis drives karyotypic evolution in mammals. *Genetics* **159:** 1179–1189.

Pearson, W.R. 2000. Flexible sequence similarity searching with the FASTA3 program package. *Methods Mol. Biol.* **132:** 185–219.

Piganeau, G., Mouchiroud, D., Duret, L., and Gautier, C. 2002. Expected relationship between the silent substitution rate and the GC content: Implications for the evolution of isochores. *J. Mol. Evol.* **54:** 129–133.

Poran, N.S. 1998. Vomeronasal organ and its associated structures in the opossum *Monodelphis domestica*. *Microsc. Res. Tech.* **43:** 500–510.

R Development Core Team. 2006. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org.

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

Rens, W., O'Brien, P.C., Yang, F., Solanky, N., Perelman, P., Graphodatsky, A.S., Ferguson, M.W., Svartman, M., De Leo, A.A., Graves, J.A., et al. 2001. Karyotype relationships between distantly related marsupials from South America and Australia. *Chromosome Res.* **9:** 301–308.

Samollow, P.B. 2006. Status and applications of genomic resources for the gray, short-tailed opossum, *Monodelphis domestica*, an American marsupial model for comparative biology. *Aust. J. Zool.* **54:** 173–196.

Samollow, P.B., Kammerer, C.M., Mahaney, S.M., Schneider, J.L., Westenberger, S.J., VandeBerg, J.L., and Robinson, E.S. 2004. First-generation linkage map of the gray, short-tailed opossum, *Monodelphis domestica*, reveals genome-wide reduction in female recombination rates. *Genetics* **166:** 307–329.

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

Sonnhammer, E.L. and Koonin, E.V. 2002. Orthology, paralogy and proposed classification for paralog subtypes. *Trends Genet.* **18:** 619–620.

Streilein, K.E. 1982a. Behavior, ecology, and distribution of South American marsupials. In *Mammalian biology in South America* (eds. M.A. Mares and H.H. Genoways), pp. 231–250. University of Pittsburgh, Philadelphia, PA.

Streilein, K.E. 1982b. Ecology of small mammals in the semiarid Brazilian Caatinga. I. Climate and faunal composition. *Annals of Carnegie Museum* **51:** 79–107.

Svartman, M. and Vianna-Morgante, A.M. 1999. Comparative genome analysis in American marsupials: Chromosome banding and in-situ hybridization. *Chromosome Res.* **7:** 267–275.

Swanson, W.J. and Vacquier, V.D. 2002. The rapid evolution of reproductive proteins. *Nat. Rev. Genet.* **3:** 137–144.

Trott, J.F., Wilson, M.J., Hovey, R.C., Shaw, D.C., and Nicholas, K.R. 2002. Expression of novel lipocalin-like milk protein gene is developmentally-regulated during lactation in the tammar wallaby, *Macropus eugenii*. *Gene* **283:** 287–297.

Vogel, M.J., Guelen, L., de Wit, E., Peric-Hupkes, D., Loden, M., Talhout, W., Feenstra, M., Abbas, B., Classen, A.K., and van Steensel, B. 2006. Human heterochromatin proteins form large domains containing KRAB-ZNF genes. *Genome Res.* **16:** 1493–1504.

Webber, C. and Ponting, C.P. 2005. Hotspots of mutation and breakage in dog and human chromosomes. *Genome Res.* **15:** 1787–1797.

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

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

Woodburne, M.O., Rich, T.H., and Springer, M.S. 2003. The evolution of tribospheny and the antiquity of mammalian clades. *Mol. Phylogenet. Evol.* **28:** 360–385.

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

Yu, L. and Zhang, Y.P. 2006. The unusual adaptive expansion of pancreatic ribonuclease gene in carnivora. *Mol. Biol. Evol.* **23:** 2326–2335.

Zhang, J., Zhang, Y.P., and Rosenberg, H.F. 2002. Adaptive evolution of a duplicated pancreatic ribonuclease gene in a leaf-eating monkey. *Nat. Genet.* **30:** 411–415.

Zuri, I., Dombrowski, K., and Halpern, M. 2005. Skin and gland but not urine odours elicit conspicuous investigation by female grey short-tailed opossums, *Monodelphis domestica*. *Anim. Behav.* **69:** 635–642.

Received November 1, 2006; accepted in revised form February 21, 2007.

CSH PRESS GENOME RESEARCH

An analysis of the gene complement of a marsupial, *Monodelphis domestica* : Evolution of lineage-specific genes and giant chromosomes

Leo Goodstadt, Andreas Heger, Caleb Webber, et al.

Genome Res. 2007 17: 969-981 originally published online May 10, 2007
Access the most recent version at doi:10.1101/gr.6093907