SUPPLEMENTARY INFORMATION

Mikkelsen *et al*. "Genome of the marsupial *Monodelphis domestica* reveals innovation in non-coding sequences"

Contents

Supplementary Notes

S1 – BAC library construction

A BAC library was constructed for the purpose of end-sequencing and subsequent anchoring of the WGS shotgun assembly. From the same partially inbred female that was selected for sequencing, a 15X coverage BAC library (384,000 total clones, 175 kb average insert size) was constructed using *Eco*RI partial digests of agarose-embedded DNA. This library was designated VMRC18.

For BAC library construction, nuclei were isolated from brain tissue using a Dounce homogenizer and sequential centrifugation steps. Nuclei were then embedded in Incert agarose and high molecular weight DNA was prepared *in situ* in the agarose. Generation of the library closely followed the cloning approach developed in P. de Jong's laboratory [1]. DNA fragments from the appropriate size fraction were cloned into the pCC1BAC vector (Epicentre Technologies). The ligation products were then transformed into DH10B (T1 resistant) electro-competent cells (Invitrogen). The library was arrayed into 1000 384-well microtiter dishes. Analysis of 120 randomly selected BACs via pulsed field gel electrophoresis of *Not*I-digested DNA indicated that the average insert size was around 175 kb.

This library and a previously constructed 10X coverage BAC library (VMRC6) from a male *Monodelphis domestica* specimen are available through the BACPAC Resources Center at the Children's Hospital Oakland Research Institute (http://bacpac.chori.org/).

S2 – Assembly versions

Starting in October 2004, a series of draft assemblies were made publicly available by the Broad Institute. The monDom4 assembly (January 2006), provided the basis for many of the analyses described in the main manuscript, while an updated version was subsequently released as monDom5 (October 2006). All versions can be accessed from GenBank using NCBI whole-genome shotgun project accession ID AAFR00000000, and from http://www.broad.mit.edu/ftp/pub/assemblies/mammals/monodelphis/. For read and coverage statistics see Supplementary Table S1.

Both monDom4 and monDom5 have the same sequence content and contig/scaffold structure, but monDom5 incorporated additional FISH data (see S3), which increased the percentage of genome in anchored scaffolds from 94.6% to 96.5%. Three scaffolds were reassigned to different chromosomes: Scaffolds 141 (2.0 Mb) and 151 (1.8 Mb), which were erroneously mapped to chromosomes 4 and 7, respectively, in monDom4, were reassigned to chromosome X in monDom5. Scaffold 108 (4.1 Mb) was reassigned from chromosome 5 to chromosome 8. Nine additional scaffolds that were unanchored in monDom4 could be anchored in monDom5: 1 to chromosome 3, 1 to chromosome 4, 1 to chromosome 6, and 6 to chromosome X. In addition there was minor shuffling of scaffold order and orientation within all chromosomes. The sequence content for chromosomes is described in Supplementary Tables S2 (monDom5) and S3 (monDom4).

S3 – Anchoring scaffolds

The assembly was anchored using primarily FISH mapping data. The linkage map data (200 SSLPs) was integrated with the FISH data and used to further order and orient supercontigs and to confirm the FISH data. For FISH mapping, two BAC clones were selected from every scaffold ≥ 1 Mb in size, and one clone each from scaffolds ≥ 500 kb. Additional clones were selected if the initial selections did not give unambiguous order and orientation.

In total, 402 BAC probes were prepared, hybridized and visualized using previously described protocols [2]. Metaphase chromosomes were prepared by mitotic stimulation of peripheral lymphocytes from each of two male and two female *Monodelphis domestica*. The precise cytogenetic band location of each probe was determined from analysis of no fewer than 10 metaphase spreads (20 chromosomes). Cytogenetic assignments were made according to the chromosome nomenclature of Pathak *et al*. [3].

Of the 395 clones mapped by FISH, 384 (97.2%) had a unique cytogenetic location and were assigned to cytogenetic band. The remaining 11 (2.8%) clones mapped to more than one cytogenetic location. To order clones that mapped within each band, the FISH analysis was repeated using groups of five clones. Clones that demonstrated overlapping signal in metaphase FISH were also analyzed using interphase FISH to resolve the precise clone order where possible. The full details of the cytogenetic anchoring of the

monDom5 genome assembly and detailed cytogenetic BAC map is published elsewhere [4].

In addition, the assembly was compared to the fingerprint map generated in the laboratory of M. A. Marra. The order of BAC end read placement in the assembly was compared to the ordering of clone in the fingerprint map (no analysis of individual fingerprint bands was performed). The main purpose was to validate the order of BACs within scaffolds and to look for regions where a single fingerprint contig would contain multiple assembly scaffolds thereby enabling further ordering and orientation of the assembly scaffolds. No additional sequence could be ordered and oriented in this way.

S4 – Assembly quality control

Several quality control steps were incorporated into the assembly process to ensure optimal continuity, structural integrity and nucleotide accuracy.

The continuity of the assembly is reflected in the large uninterrupted sequence 'contigs' (N50 length = 108 kb), which are joined into much larger 'scaffolds' (N50 length = 59.8 Mb). The high degree of continuity implies that the majority of genes should be present without sequence gaps. The coverage of the euchromatic genome is estimated to be ~98%, based on the small proportion of the assembly residing within spanned gaps.

The structural integrity of the assembly was assessed with a substantially revised version of the certification process first developed for the dog genome [5]. As before, internal inconsistencies in the assembly were used as evidence, but unlike the previous process which classified regions as either 'certified' (essentially equivalent to finished sequence) or 'uncertified' (containing assembly errors), the new process allows regions that are highly likely to be misassembled to be distinguished from those which are merely possibly incorrect.

The new certification model converts multiple forms of evidence into probabilities for the presence of different types of assembly errors across the genome: *structural problems*, where different regions of the genome have been joined (Supplementary Table S4) (A simplified measure of this is the fraction of reads that are placed consistently with their

partner. In the monDom5 genome assembly, 94.9% of reads have a consistent placement, which is similar to the 96.5% or reads with consistent placement in the dog genome.); *local deletions and insertions* (Supplementary Table S5), where either portions of the genome are missing or sequence from a different region has been inserted, but the overall structure is correct; and *high quality base errors* (Supplementary Table S6), where isolated bases are wrong, but reported as having a quality score of 50 or higher. Full details of the model will be published elsewhere (M. Grabherr, *in preparation*), but source code is available upon request, and a complete assembly markup is available from: http://www.broad.mit.edu/ftp/pub/assemblies/mammals/monodelphis/monDom5/ http://www.broad.mit.edu/ftp /pub/wga/misc/docs/AssemblyMarkup2.0.pdf

Approximately 87% of the opossum sequence resides in regions that are certified to be free of structural assembly errors in the monDom5 assembly. The remaining 13% are declared as 'uncertified', but only 2% of the genome is labeled as having a significant probability of containing an error. Of these latter regions, half reside in unanchored sequences. For comparison, by the new methodology 89% of the dog genome resides in certified regions in the most recent canFam2 assembly, while only 3% have a significant probability of error.

To assess nucleotide level accuracy, 11 finished BACs (1.66 Mb) from the sequenced opossum were obtained from Genbank. As each finished BAC contains only one haplotype, while the assembled genome contains a mixture of two haplotypes, we restricted the comparison to homozygous regions in the assembly. Such regions were located by grouping the SNPs identified from the two assembled haplotypes (see S5) into consecutive blocks of either ≥ 0.3 SNP/kb, or ≤ 0.01 SNP/kb. Blocks of less than 100 kb were ignored, as were blocks with over 50% of the bases excluded due to gaps or uncertified sequences. Those blocks with rates ≤ 0.01 SNP/kb were considered homozygous and span 900-kb. The finished BACs were aligned to the assembly using an alignment tool built into ARACHNE. Those alignments falling within regions declared homozygous (0.9 Mb) were accepted for the base accuracy estimate (Supplementary Table S7). A somewhat higher error rate can be expected in polymorphic regions of the genome.

S5 – Single Nucleotide Polymorphism discovery

SNPs were discovered in the *Monodelphis domestica* genome sequence as a standard part of the ARACHNE assembly process (Supplementary Table S8). Such SNPs represent cases where the single individual (I-1438) sequenced was heterozygous. Assembly SNPs were only called within certified regions (see S4), and if each allele were observed at least twice, implying that ≥ 4 separate shotgun reads aligned to the base containing the putative SNP.

Additional SNPs were discovered by low-coverage shotgun sequencing of three stocks frequently used in laboratory research (Supplementary Table S8): B5539 (SFBR Population 1; 104,042 passing reads), D3508 (SFBR Population 2; 100,000 reads), and C5181 (SFBR Population 5; 99,212 reads). Sequencing reads were trimmed to include only intervals of 250 bp or longer with a running 20 bp average quality score of at least 20. These intervals were aligned to the monDom4 assembly and SNPs were discovered using SSAHA-SNP [6] with a window size of 500bp and a ceiling of 0.02 SNP/bp in any single comparison. SNPs were required to have a Phred quality score of 23 or higher with a minimum quality score of 15 for each of five bases either side [7]. SNPs are available at http://www.broad.mit.edu/mammals/opossum/ and have been submitted to dbSNP.

To estimate the validation rate of the discovered SNPs, ten 200 kb regions (two regions each on chromosome 1 and 2 and one each on chromosome 3-8 and X) were chosen randomly from regions that included at least 40 SNPs and were more than 1 Mb away from a telomere. Within each region, 20 SNPs were selected for genotyping, with individual SNPs given preference if their source population was under-represented in the region. On the X chromosome, most available SNPs originated from the assembly. Validation genotyping was performed on a Sequenom Mass Spectrometry system. After removal of SNPs with low call rate (<75%), 126 SNPs remained. The average validation rate for the autosomes was 94.9% (n = 118). On chromosome X the validation rate was only 62% (n = 8), but this was likely due to an unfortunate choice of region, which turned out to be highly repetitive. Validation rates by source are shown in Supplementary Table S9.

The applicability of the SNP set for different laboratory stocks was assessed using representatives from SFBR Populations 1, 2, 3 and 5 (Supplementary Table S10). The polymorphism rates within Populations 1 and 2 were 75% and 60% respectively, suggesting that SNPs can be randomly selected for mapping purposes within these populations. It should be noted, however, that of 18 representatives of Population 2, twelve had some Population 1 admixture (<25%) and in the six pure Population 2 individuals, the polymorphism rate was 42%. The two parentals from each of Population 3 and 5 were genotyped, capturing the entire polymorphism rate within those lines (Population 3 at 10% and Population 5 at 4%). Note, however, that SNPs identified from Population 5 are likely to be useful for crosses generated between it and either Population 1 or 2.

We assessed linkage disequilibrium by calculating r^2 across the 200kb regions both within and across populations, using Haploview [8] (Supplementary Figure S3). SNPs with a minor allele frequency of less than 0.04 were excluded and SNPs at all distances were subjected to association analysis. The markers analyzed vary among populations, as Haploview automatically removes markers that are monomorphic within the group of individuals. Markers with a call rate of less than 75% were excluded. Linkage disequilibrium ($r^2 > 0.5$) extended less than 15 kb across populations.

S6 – Representation of segmental duplications in the assembly

To elucidate the effect of any bias in the analysis due to the draft nature of the opossum assembly, we first examined its overall read coverage landscape to try to locate trace "pile-ups". Such pile-ups might suggest an over-collapsed region in the assembly that concealed segmental duplication. The largest "pile-ups" are due to the incorporation of mitochondrial sequence into the assembly. However, excluding these, even if one assumed that every over-sampled base was actually part of an over-collapsed duplication (surely a conservative upper bound), the overall duplication rate would be boosted by less than half (to < 2.6%)- leaving it well under the duplication rate for human.

We next analyzed all traces that were omitted from the assembly. ~10% of the 38.8 million reads left after low quality and contaminating reads were not placed in the assembly. This is a similar fraction to that seen in other draft assemblies. The "Improved"

ARACHNE WGS assembly of the mouse genome (see below) left ~13% of reads unplaced, and the most recent dog assembly (canFam2) left ~6% unplaced. For comparison, assembly of the highly repetitive and polymorphic fungus *Puccinia graminis* left 30-40% of reads unplaced (E. Mauceli, unpublished data). The opossum genome is therefore unlikely to contain large, euchromatic segmentally duplicated regions that completely failed to assemble.

Finally, we examined a historical series of mouse assemblies to chart how our ability to discover segmental duplications changes with the quality of an assembly. The MGSCv3 assembly of the mouse genome [9] was primarily derived from whole genome shotgun traces, though it had a tiny amount of directed sequencing incorporated as well. This assembly was done with an older version of the ARACHNE software. An "Improved" assembly was done with the newer ARACHNE version that was used to assemble opossum, using the same whole genome shotgun traces as the MGSCv3 assembly, but omitting all directed sequencing data. Finally, we analyzed the most recent finished assembly available for mouse, mm8 (NCBI Build 36; February 2006), which benefits from several rounds of directed sequencing and finishing work. Supplementary Table S13 shows that the improvements made to the ARACHNE assembler allow for better recognition of segmental duplications, even without any finishing. 5.3% of the finished mm8 mouse assembly is covered by segmental duplication, as compared to only 3.5% in the "Improved" assembly most comparable to the current opossum assembly. Even if there were a roughly similar inflation of segmental duplication coverage in a finished opossum assembly (from 1.7% in the current assembly to 2.6% in a speculative finished assembly), opossum would remain far less duplicated than either mouse or human.

S7 – Synteny maps and assignment of breakpoints to lineages

We performed full genomic alignments of repeat masked sequence from the opossum genome (monDom5) against the human (hg18), dog (canFam2), mouse (mm8), rat (rn4) and chicken (galGal3) genomes using `PatternHunter` [10].

We sought to align the opossum genome sequence with the human genome. We found that ~10% (338 Mb) of the nucleotides can be directly aligned to the human sequence. This is much lower than for eutherians such as mouse (40%) or dog (~57%), likely

reflecting both actual turnover of genomic sequence and some difficulty in detecting neutrally evolving segments due to nucleotide divergence [11,12].

Following established methods [5,9], we identified collinear clusters of the identified synteny anchors, which were used to form larger syntenic segments in a hierarchical fashion. Segments that are larger than a given size in both genomes, and are comprised of at least 4 anchors at a given stage in the merging process, define a resolution-dependent, pair-wise synteny map between the two genomes. Some comparative statistics for the pair-wise maps are given in Supplementary Table S14.

We next merged and reconciled the pair-wise synteny maps to create a 5-way multi-species map. We identified the locations along the opossum genome (the reference genome) where any of the four pair-wise maps indicated a synteny breakpoint, and then, from the opossum locations, identified the corresponding locations in each of the four non-reference genomes. This allowed each genomic interval in opossum bracketed by synteny breakpoints to be associated with, and oriented with respect to, a genomic interval in each of the four non-reference genomes. The resulting multi-species synteny map represent blocks of homologous sequence that are present and contiguous in all five species.

The multi-species synteny map can be represented as a graph (Supplementary Figure S6): each block is represented by two nodes (- and +). Each pair of nodes have exactly one incoming and one outgoing edge for each species, which connect to the previous and next neighboring block edge in that species, respectively, as defined by their species-specific genomic ordering. This representation greatly simplifies the bookkeeping needed to relate breakpoints to evolutionary events.

At a resolution of 500 kb, our 5-way maps consist of 616 blocks, representing 1,232 synteny breakpoints. Most (75%) of these breakpoints are simple cases where the outgoing (or incoming) edges connect to only one or two neighboring blocks (chromosomal ends are not counted, so if the end of a block is the most distal segment of a chromosome in one or more species, there may be only one neighboring block). In these cases, synteny breakpoints can be unambiguously associated with a lineage, using

the principle of parsimony. The remaining breakpoints connected to either three (19%) or more than three (6%) neighboring blocks. At higher resolutions, we observe more breakpoints (2,126 at 100 kb), but also a higher percentage of simple breakpoints (80%).

We next focused specifically on the subset of simple cases involving purely internal breakpoints, *i.e.* those that do not involve a chromosome end in any of the five species (and therefore exactly two neighboring blocks). These represent 65% of all breakpoints in the 500 kb maps and 74% in the 100 kb maps. This subset is clearly associated with translocations and inversions (but not chromosome fusions or fissions), and can be unambiguously assigned to branches in the phylogenetic tree. This subset provides perhaps the most reliable estimate of the relative rates of such rearrangements between the eutherian lineages at any given resolution (Supplementary Figure S7a). Our 5-way synteny map largely confirms the trends reported previously [5] and, in particular, the observation that although the human and dog lineages have generally experienced a similar number of rearrangements, the human lineage appears to have experienced a greater number of smaller rearrangements. We also confirm that most of the larger rearrangements (≥ 500 kb) that have occurred in the mouse and rat lineages appear to have occurred in their last common ancestor. Finally, we verified that these trends are qualitatively unchanged when internal breakpoints with 3 alternative neighboring blocks are included (as multiple events; Supplementary Figure S7b).

S8 – Reconstruction of the ancestral boreoeutherian karyotype

Much of the structure of the ancestral eutherian karyotype can be inferred from syntenic relationships among extant mammals. Given that most of the synteny breakpoints in our 5-way map were clearly lineage-specific, we reasoned that it would be possible to reconstruct genomic regions that were once contiguous in the boroeutherian ancestor (ancestral to dog, human, mouse and rat) by joining syntenic blocks into larger groups by following their links in a hierarchical fashion that is consistent with the phylogenetic tree.

Our approach is summarized in the following table:

Reconstructed groups	Clear order and orientation?	500 kb map	100 kb Map	Connectivity evidence considered
Level 0: Blocks	Y	616	1063	n/a
Level 1	Y	467	682	O=D=H or D=H=M=R
Level 2	Y	101	118	O=D, D=H, or (O=H and not D=M=R)
Level 3	Y	74	77	D=M=R
Level 4	Y	67	66	Supported by M, R, C
Level 5	N	43	44	Links between level 4 groups: Terminal nodes that map interior of other groups

O = opossum; D = dog; H = human; M = mouse; R = rat; C = chicken.

Higher level information was used only when all lower level information was exhausted. Pairs of blocks that are adjacent in both the dog and human genomes are assumed to have been connected in the boroeutherian ancestor. Blocks that are adjacent in opossum and either dog and human, but not both, are considered evidence of a lineage-specific breakpoints, which can be ignored. After opossum, dog and human associations are exhausted, information from the murids are also considered. Finally, to better handle cases where the outgroup (the metatherian lineage to opossum) itself has experienced a rearrangement, we augmented the five-way maps to include synteny information from the chicken genome. In brief, the resulting 6-way maps consist of the same blocks as the five-way maps, but with extra links (edges in the graph) defined between blocks that align to the chicken genome. (Chicken is treated differently due to the relative sparseness of synteny anchors it provides).

Application of our reconstruction process to the initial set of 616 synteny blocks (500 kb resolution) led to the identification of 67 ancestral boreoeutherian groups in which all constituent blocks had a well-defined order and orientation. Allowing for ordering ambiguities, this could be further refined to 43 groups with strong evidence of ancestral connectivity (Supplementary Figure S8). 95% of the covered human sequence is found in the 30 largest groups. Notably, reconstructions from 500 kb and 100 kb maps lead to almost identical ancestral groups, despite the much larger number of initial blocks in the

100 kb maps (1063 vs. 616). In both cases, the reduction in fragmentation is dramatic and demonstrates that it is possible to reconstruct large (almost chromosome-sized) regions of the ancestral boroeutherian genome by a straightforward computational analysis of extant genomic sequence.

S9 – Gene prediction

A set of predicted genes in the opossum genome was built using a modified version of the standard Ensembl genebuild pipeline [13]. Previous species-specific sequence resources (opossum cDNA and proteins) were very limited, so the vast majority of gene models were based on `Genewise` [14] alignments of proteins from genetically distant species. Opossum and human cDNAs were aligned and used to add UTRs to the `genewise` predictions where possible. The gene models were assessed by generating sets of potential orthologs in other mammalian species, using the Ensembl homology pipeline. Potentially missing or partial gene predictions were identified by comparative analysis of the ortholog sets, and `Exonerate` [15] used to build new gene models based on the human ortholog peptide sequence, if needed.

The final Ensembl set contained 19,584 protein-coding gene models (plus 13 mitochondrial genes), as well as 148 miRNAs and 798 other ncRNAs (plus 23 mitochondrial ncRNAs) predicted based on homology to known eutherian instances. This set, including detailed descriptions of its derivation, and future updates are available from http://www.ensembl.org/Monodelphis_domestica/.

The set of protein-coding gene predictions analyzed in the main manuscript were generated by independent refinement of an intermediate Ensembl build with 19,888 predictions. First, additional gene predictions were generated by aligning human transcripts (Ensembl v40) to the opossum genome assembly using `Exonerate`. This yielded 367 additional genes that did not overlap with Ensembl predictions, which all possessed single human orthologs (as predicted by `PhyOP` [16]). Next, the opossum genome assembly was re-scanned for previously unnoticed lineage-specific genes. This scan identified 290 additional genes, where each was paralogous to a previously predicted opossum gene. A further 130 genes were also identified after the separation of erroneously merged adjacent paralogs. On the other hand, 979 genes with the

characteristics of pseudogenes were identified and removed during the process of predicting orthology and paralogy with PhyOP. These included apparently retrotransposed pseudogenes, coding sequence with multiple disruptions, as well as genes with sequence similarity to LINE-1. Manual curation of gene duplications revealed a further 423 predictions that appeared to be non-coding.

Accounting for additional, as well as likely non-coding or pseudogenic predictions resulted in an estimate of 18,648 genes. The lower gene count compared to the final Ensembl set is primarily due to a more stringent filtering of likely pseudogenes. The analyzed get set, and associated orthology and paralogy predictions, are available from http://wwwfgu.anat.ox.ac.uk:8080/download/monodelphis/1v2/orthologs/

A total of 3,111 human gene predictions from the analyzed Ensembl build had no clear homologs in the initial set analyzed here. Ongoing refinement of the human gene catalog has led to the identification of ~2,000 of these as poorly supported predictions based on pseudogene-like features and lack of experimental evidence (M. Clamp, *submitted*, see also [16,17]). These predictions were therefore dropped from the analysis. 623 of the remaining 1,083 human gene predictions were subsequently found have at least 25% of their coding regions covered by BLASTZ genome-wide alignments between human and opossum, indicating possible false negatives. A list of the relevant human gene predictions is available from ftp://ftp.broad.mit.edu/pub/papers/opossum_genome/

S10 – Families of paralogous CNEs

To look for evidence of generation of CNEs, we grouped highly conserved eutherian elements ($\log_2$-odds $\geq$ 60) into paralogous families. Each element was realigned to the human, mouse and dog genome sequences using BLASTN [18], and all significant hits ($E < 10^{-10}$) were recorded. CNE that had overlapping alignments in any of the three genomes were subsequently grouped into paralogous families using single-linkage clustering.

We identified 312 families of 4 or more paralogous eutherian CNEs (3,032 in total; or ~2.5% of all eutherian CNEs with $\log_2$-odds $\geq$ 60). Of these families, 81 (26%) overlapped segmental duplications with sequence similarity $\geq$ 90% in the human genome, and are therefore duplicated only in the human lineage.

Most of the remaining 231 families (2,542 in total) are distributed across multiple chromosomes and show no clear evidence of being part of more ancient segmental duplications (for example, CNEs from two different families are rarely found as nearby pairs in more than one genomic region). Strikingly, the families display a strong bias towards containing either exclusively eutherian-specific CNEs, or none at all (Supplementary Figure S10a). This suggests that the paralogous CNE families emerged in 'bursts' either before or after the Eutheria-Metatheria divergence, reminiscent of the spread of TEs. Indeed, 20 out of 35 completely eutherian-specific families include annotated TEs. Moreover, the 196 CNE families with opossum orthologs (which contains all but two of the ancient CNE families previously identified by us [19]) show a strong bias towards containing members that either all have orthologs in chicken, or none at all (Supplementary Figure S10b). However, the vast majority of these more ancient CNEs have no recognizable TE-like features. This again suggests that the fraction of eutherian CNEs overlapping annotated TEs is only a lower bound on the true fraction derived from such elements.

Supplementary Tables

Supplementary Table S1 – Read statistics (monDom5)

Insert size (kb)	Vector	Passing reads (M)	Paired reads (M)	Assembled reads (M)	Sequence coverage (X)	Sequence coverage > Phred20 (X)	Physical coverage (X)
4	Plasmid	29.1	28.5	26.6	6.06	5.11	17.6
10	Plasmid	7.4	7.3	6.6	1.62	1.35	7.4
40	Fosmid	1.8	1.7	1.6	0.328	0.252	7.6
200	BAC	0.5	0.4	0.4	0.0724	0.0454	5.2
TOTAL		38.8	38.0	35.2	8.08	6.76	37.8

Supplementary Table S2 – Distribution across chromosomes (monDom5)

Chromosome	Total bases (Mb)	Number of supercontigs	N50 supercontig size (Mb)	Size of gaps in supercontigs (Mb)	Total est. size (Mb)*
1	733	27	80.1	11.6	748
2	528	32	60.5	10.5	542
3	515	32	126.0	9.8	528
4	423	38	43.8	9.4	435
5	298	11	80.4	3.8	305
6	285	13	57.1	4.6	292
7	255	14	34.9	3.1	261
8	303	22	32.5	6.4	313
X	73	27	3.6	3.3	79
All	3413	216	60.5	62.5	3502

* Includes anchored sequence, spanned gaps, and centromeric sequence (3 Mb each).

Supplementary Table S3. Distribution across chromosomes (monDom4)

Chromosome	Total bases (Mb)	Number of supercontigs	N50 supercontig size (Mb)	Size of gaps in supercontigs (Mb)	Total est. size (Mb)*
1	733	27	80.1	11.6	748
2	528	32	60.5	10.5	541
3	513	31	126.0	9.8	526
4	418	37	43.8	9.3	430
5	302	12	80.4	4.1	309
6	238	13	59.1	4.0	245
7	256	15	34.9	3.2	263
8	299	21	32.5	6.0	308
X	55	19	4.7	2.6	61
All	3343	207	61.2	61.1	3431

* Includes anchored sequence, spanned gaps, and centromeric sequence (3 Mb each).

Supplementary Table S4. Distribution of possible misjoins

Possibility of error (%)	# of intervals	% genome	Cumulative % of genome (this level and higher)
0.0–0.1	187,540	2.74	13.30
0.1–1.0	76,825	4.74	10.56
1.0–5.0	34,935	4.12	5.77
5.0–10.0	2,666	0.65	1.65
10.0–20.0	975	0.56	1.01
20.0–30.0	158	0.18	0.45
30.0–40.0	51	0.11	0.27
40.0–50.0	12	0.04	0.16
50.0–60.0	23	0.10	0.12
60.0–100.0	4	0.03	0.03

CLEAN stretch N50 of probability 5.00% and up: 4,975,588.

DIRTY stretch N50 of probability 5.00% and up: 18,402.

Supplementary Table S5. Distribution of possible indels

Possibility of error (%)	# of intervals	% genome	Cumulative % of genome (this level and higher)
0.0–0.1	42,140	0.43	13.30
0.1–1.0	117,157	2.07	12.86
1.0–5.0	101,662	5.90	10.79
5.0–10.0	28,025	2.98	4.89
10.0–20.0	11,401	1.64	1.90
20.0–30.0	2,440	0.25	0.27
30.0–40.0	328	0.02	0.02
40.0–50.0	36	0.00	0.00
50.0–60.0	0	0.00	0.00
60.0–100.0	0	0.00	0.00

CLEAN stretch N50 of probability 5.00% and up: 197,946.

DIRTY stretch N50 of probability 5.00% and up: 6,800.

Supplementary Table S6. Distribution of possible base errors (Q50)

Possibility of error (%)	# of intervals	% genome	Cumulative % of genome (this level and higher)
0.0–0.1	190	0.00	13.30
0.1–1.0	3,863	0.05	13.29
1.0–5.0	77,597	1.08	13.24
5.0–10.0	50,274	0.80	12.16
10.0–20.0	52,307	1.09	11.37
20.0–30.0	32,130	0.87	10.27
30.0–40.0	20,408	0.79	9.40
40.0–50.0	15,412	0.86	8.61
50.0–60.0	12,729	1.09	7.75
60.0–100.0	38,279	6.66	6.66

CLEAN stretch N50 of probability 5.00% and up: 37,129.

DIRTY stretch N50 of probability 5.00% and up: 4,627.

Supplementary Table S7 – Accuracy of quality scores by comparison to ~0.90 Mb finished sequence

Reported quality (Monodelphis5.0)*	Fraction of assembly (%)	Actual quality**
0–9	0.16	4.4
10–19	0.37	11.9
20–29	0.60	22.2
30–39	1.28	28.9
40–49	4.01	37.6
50–	93.60	45.3
>40	97.61	44.9

* The distribution of quality scores for the entire assembly.

** The distribution of quality scores for 0.90 Mb of finished sequence. Both mismatches and indels were scored as errors.

Supplementary Table S8 - SNP discovery statistics

SFBR Population (library name)	Population Code	WGS Processed Reads	Passing Reads	Examined Bases[a]	SNPs discovered[a]	SNP rate (1/x bp) [b]
Pop 1 (S251)	B5539	105,592	104,042	41,511,966	111,322	358
Pop 2 (S252)	B3508	100,985	100,000	40,084,481	127,712	312
Pop 5 (S253)	C5181	100,220	99,212	38,868,170	274,562	140
monDom4	I-1438	N/A	N/A	N/A	775,686	N/A
Totals		306,797	303,254		1,285,404[b]	

[a] in anchored portion of assembly only.

[b] total unique SNPs

Supplementary Table S9 - SNP validation rates by discovery source

SFBR Population	Type	Monomorphic	N	Validation (%)
monDom4	All	5	20	75
Pop1	All	1	51	98
Pop2	All	0	24	100
Pop5	All	2	32	94
monDom4	Autosomal	3	13	77
Pop1	Autosomal	1	51	98
Pop2	Autosomal	0	24	100
Pop5	Autosomal	2	30	93
monDom4	X	3	7	57
Pop1	X	0	0	n/a
Pop2	X	0	0	n/a
Pop5	X	0	2	100

Supplementary Table S10 - Polymorphism rates of discovered SNPs across stocks

SFBR Population Genotyped	SNP Source	Total	Polymorphic	Monomorphic	Polymorphism rate (%)
Pop1	monDom4	18	9	9	50
Pop1	Pop1	51	48	3	94
Pop1	Pop2	24	21	3	87
Pop1	Pop5	32	10	22	31
Pop2	monDom4	20	6	14	30
Pop2	Pop1	51	32	19	63
Pop2	Pop2	24	14	10	58
Pop2	Pop5	32	10	22	31
Pop3	monDom4	20		20	0
Pop3	Pop1	47	3	44	6
Pop3	Pop2	22	1	21	4
Pop3	Pop5	31		31	0
Pop5	monDom4	19		19	0
Pop5	Pop1	47	1	46	2
Pop5	Pop2	22		22	0
Pop5	Pop5	32		32	0

Supplementary Table S11 – Sequence features by chromosome.

				Covered by transposable elements (%)					Covered by segmental duplications (%)[a]			
	Size (Mb)	G+C (%)	CpG (%)	Total	LINE	SINE	ERV	DNA	Total	Intra	Inter	Involving "chrUn"
chr1	745	37.8	0.9	52.4	29.0	11.1	10.4	1.8	0.8	0.6	0.1	0.2
chr2	538	38.0	1.0	52.5	28.8	10.9	10.8	1.8	1.1	0.8	0.2	0.4
chr3	524	37.3	0.9	53.7	29.7	9.9	12.1	1.7	1.5	1.2	0.2	0.6
chr4	432	37.7	0.9	52.6	28.9	10.4	11.4	1.8	1.7	1.3	0.2	0.8
chr5	301	37.2	0.9	52.6	30.3	9.5	11.0	1.6	1.1	0.8	0.2	0.3
chr6	289	38.1	1.0	51.3	28.5	11.2	9.6	1.9	0.7	0.4	0.2	0.2
chr7	257	36.7	0.9	51.1	29.5	9.2	10.6	1.6	0.7	0.4	0.3	0.3
chr8	309	37.8	1.0	50.4	28.4	10.4	9.7	1.7	0.9	0.5	0.2	0.3
chrX	76	40.9	1.4	50.10	31.6	9.6	7.2	1.6	3.3	2.8	0.2	1.9
All[b]	3475	37.7	0.9	52.2	20.2	10.4	10.7	1.7	1.7	0.8	0.2	1.0

[a] The sums of the coverage across categories do not equal the genome total because some bases are involved in multiple duplication types.
[b] Excludes unplaced sequences ("chrUn")

Supplementary Table S12 – Distribution of distances between segmental duplicons.

Distance	Opossum	Human
< 10 kb	382	391
10 kb - 100 kb	2510	2282
100 kb - 1 Mb	3180	2699
1 Mb - 10 Mb	609	3960
10 Mb - 100 Mb	929	3586
> 100 Mb	232	404

Intrachromosomal segmental duplications in opossum are more local than in human: the median distance between related duplicons is 175 kb in opossum, compared to 2.2 Mb in human. This is true whether the duplication is direct or inverted, and despite the far larger chromosome sizes in opossum.

Supplementary Table S13 – Influence of assembly quality on segmental duplications.

	Genome coverage (%)			
Mouse Assembly	All	Intra-chromosomal	Inter-chromosomal	Involving unplaced sequences
MGSCV3	2.5	0.8	0.3	1.7
"Improved"	3.5	2.3	2.4	0.0
mm8	5.3	4.9	0.7	0.1

The sum of the coverage across categories does not equal the genome total because some bases are involved in multiple duplication types.

Supplementary Table S14 – Synteny map statistics (500 kb resolution).

	Opossum vs. human	Opossum vs. dog	Opossum vs. mouse	Opossum vs. chicken	Human vs. dog	Human vs. mouse
Number of syntenic anchors	328,000	324,000	239,000	145,000	1,629,000	696,000
N50 anchor size (bp)	329	315	259	243	779	548
Mean anchor distance (kb)	9.8	9.9	13.4	20.8	1.3	1.3
Number of segments	500	510	647	539	229	316
N50 size of segments (Mb)	18.2	14.7	17.8	12.9	26.6	18.8
Genomic coverage [a] (%)	93.5	93.3	90.1	84.9	97.8	96.3

[a] Refers to coverage of opossum (excluding "chrUn", but including total estimated chromosome sizes, with heterochromatin, gaps, etc.), or human in the human vs. dog and mouse maps.

Supplementary Table S15 – X chromosome breakpoint assignments

Synteny Breakpoint Region (hg18)	Lineage assignment	Comments
chrX:17664033-17728023	Eutherian	
chrX:18140710-18167559	Eutherian	
chrX:19113330-19277336	Monodelphis	chr7-chr4 break
chrX:34585077-35879166	Eutherian	
chrX:36580943-37094099	Eutherian	
chrX:45964689-46318311	Undetermined	
chrX:46837291-48636117	Undetermined	Rodent rearrangement is not coincident.
chrX:49030496-49643173	Undetermined	
chrX:50573770-53469980	Undetermined	
chrX:54840248-64054182	Undetermined	Rodents have 3 rearrangements in region.
chrX:64728118-64814614	Monodelphis	
chrX:67862120-67971773	Eutherian	
chrX:68529469-68640747	Eutherian	
chrX:70712454-73551496	Undetermined	
chrX:74251225-74410807	Undetermined	
chrX:80506327-80548039	Monodelphis	
chrX:86441814-87111773	Monodelphis	
chrX:93641634-93972400	Monodelphis	
chrX:97795447-97816592	Monodelphis	
chrX:99750538-99785955	Monodelphis	
chrX:102970042-103558443	Eutherian	
chrX:105084179-105823390	Monodelphis	
chrX:106735159-106955260	Eutherian	
chrX:108123852-108771183	Monodelphis	
chrX:109447994-110271968	Monodelphis	
chrX:115978562-116017332	Monodelphis	
chrX:117794515-117992509	Eutherian	
chrX:119645095-120108904	Monodelphis	
chrX:125675125-125686887	Monodelphis	
chrX:134543187-135232062	Monodelphis	
chrX:145098764-146268646	Monodelphis	
chrX:151889748-152400162	Undetermined	
chrX:152791350-152826560	Undetermined	
chrX:153211237-153280871	Undetermined	

Based on 300-kb resolution synteny map. Assumes human/dog X chromosome represents the boreoeutherian ancestor. Chicken used as outgroup for lineage assignments. "Undetermined" assignments are in most cases due to lack of synteny anchors to chicken, or jumps to unanchored scaffolds in the opossum assembly.

Supplementary Table S16 – Ancient genes found in opossum, but lost in human or eutherians.

Opossum gene prediction	Comments
ENSMODG00000000115	Malate synthase
ENSMODG00000001080	Inosine/uridine-preferring nucleoside hydrolase
ENSMODG00000018409	Photolyase
ENSMODG00000017963	Fucolectin precursor (weak homology)
ENSMODG00000024867	Fucolectin precursor (weak homology)
ENSMODG00000008308	Fatty acid synthase
ENSMODG00000002691	Unknown
ENSMODG00000021523	Peroxidase-related

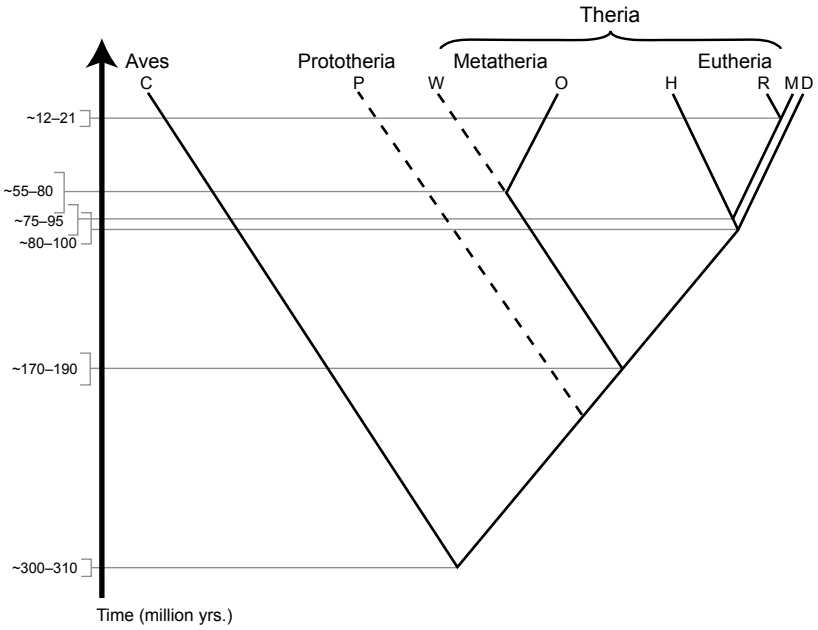
Gene predictions are identified by their Ensembl ID (version 42)

Supplementary Table S17 – Lineage-specific CNEs near key developmental genes.

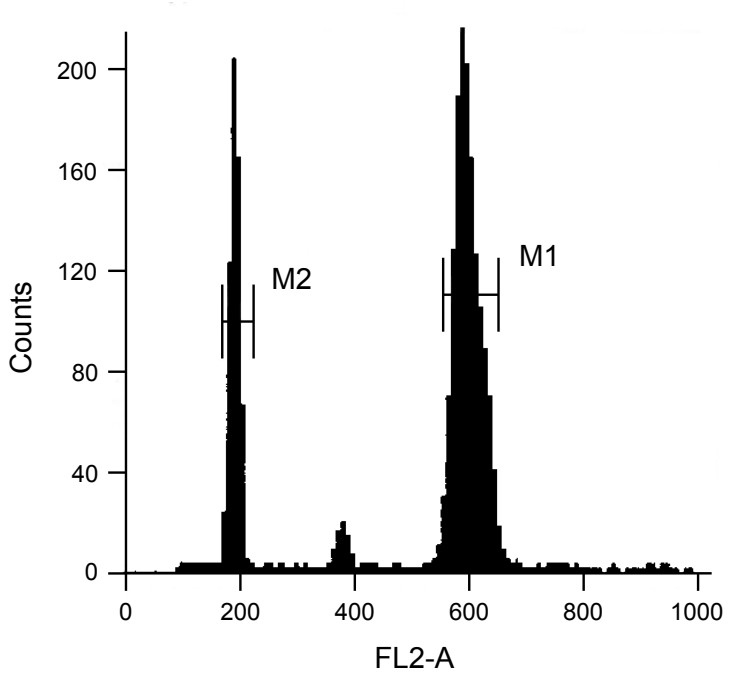
CNEs (ungapped syntenic regions)	Category	Within whole clusters	Within 100kb of a developmental gene
Amniote (human-chicken)	**All (%)**	54.3	15.0
	Absent in opossum (%)	38.7	8.5
	Bias	0.73	0.57
Eutherian	**All (%)**	35.2	9.6
	Eutherian-specific (%)	23.7	6.0
	Bias	0.67	0.62

The table shows the total proportions of amniote or eutherian CNEs found within the ≥ 1 Mb conserved gene deserts surrounding key developmental genes in the human genome, as defined in [5]. Amniote elements are represented by those identified by chicken-human alignments, and present or absent in opossum. Eutherian elements include all $\log_2$-odds scores. The bias towards loss and gain increase somewhat in regions close to the developmental genes.

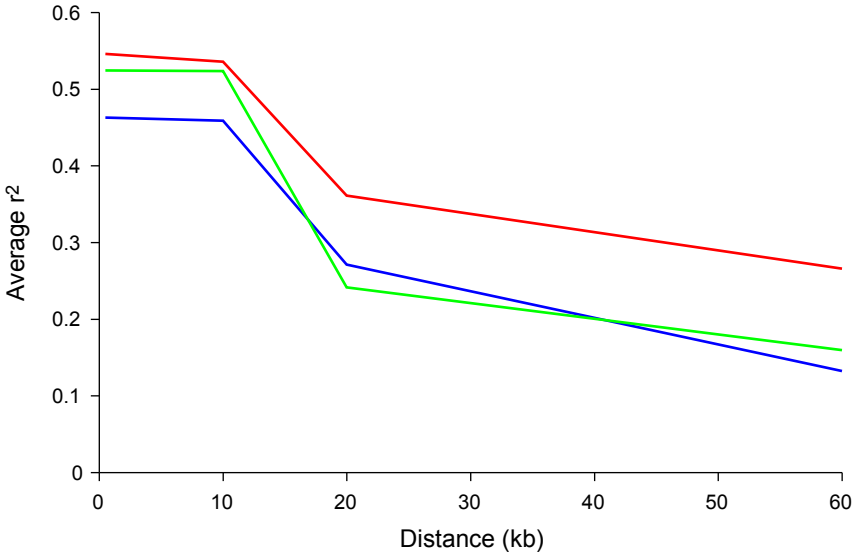
Mikkelsen et al.
Supp. Figure 1



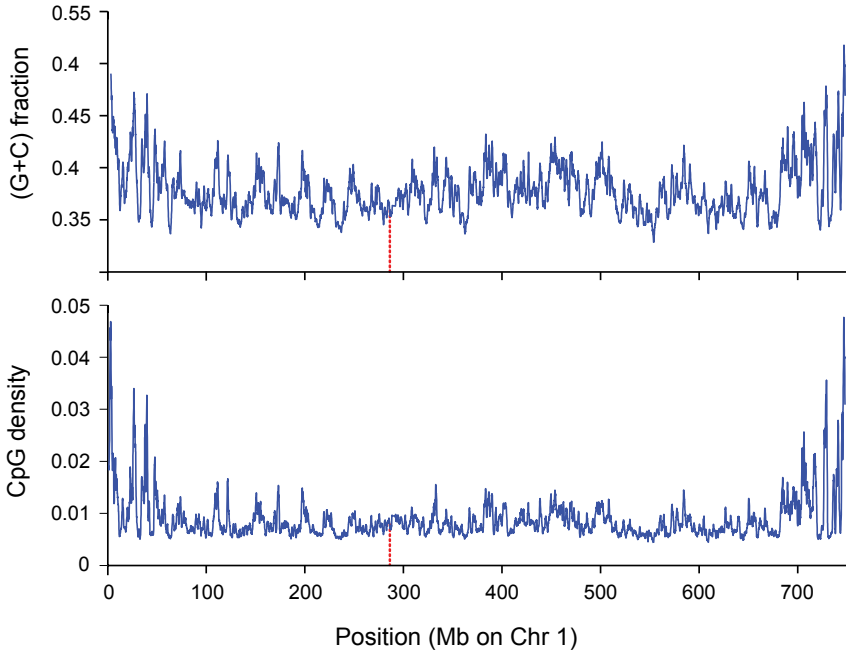
Mikkelsen et al.
Supp. Figure 2



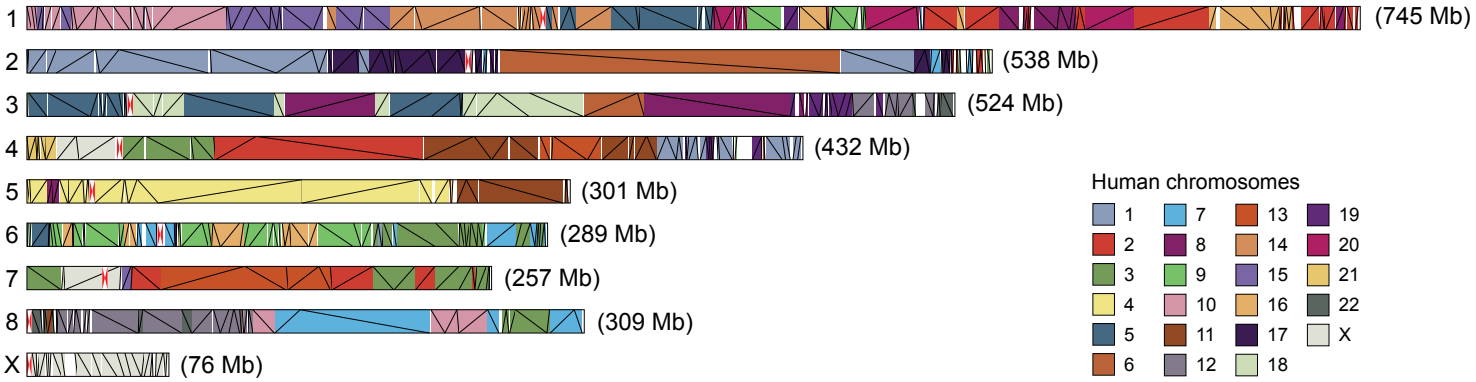
Mikkelsen et al.
Supp. Figure 3



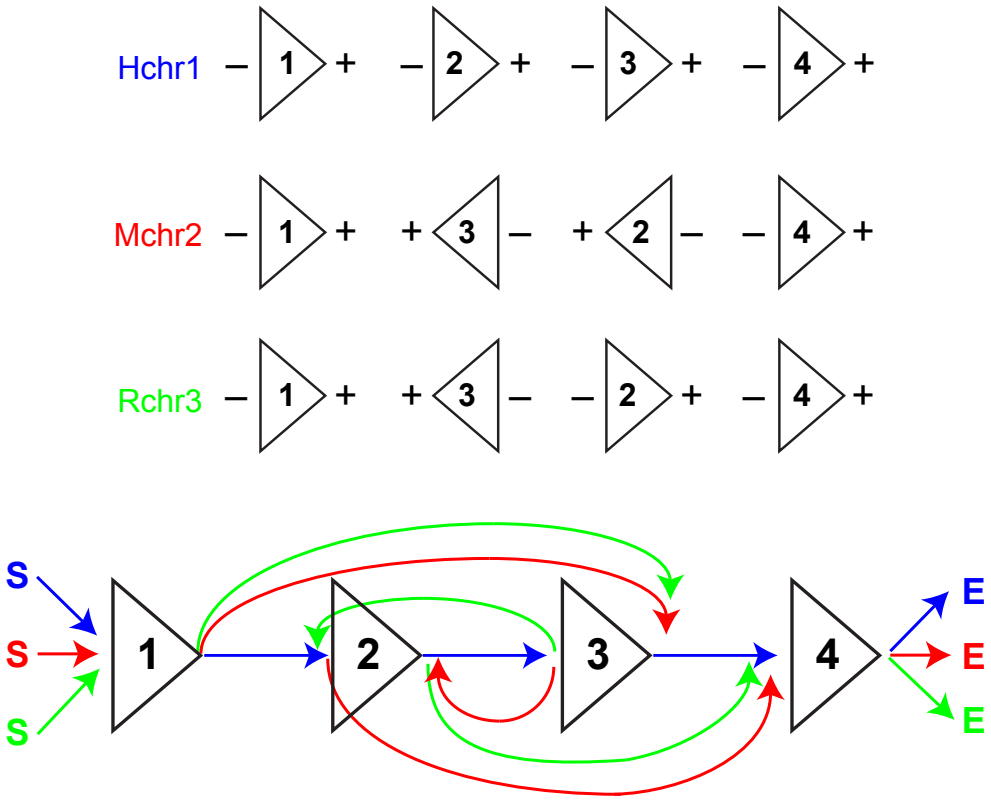
Mikkelsen et al.
Supp. Figure 4



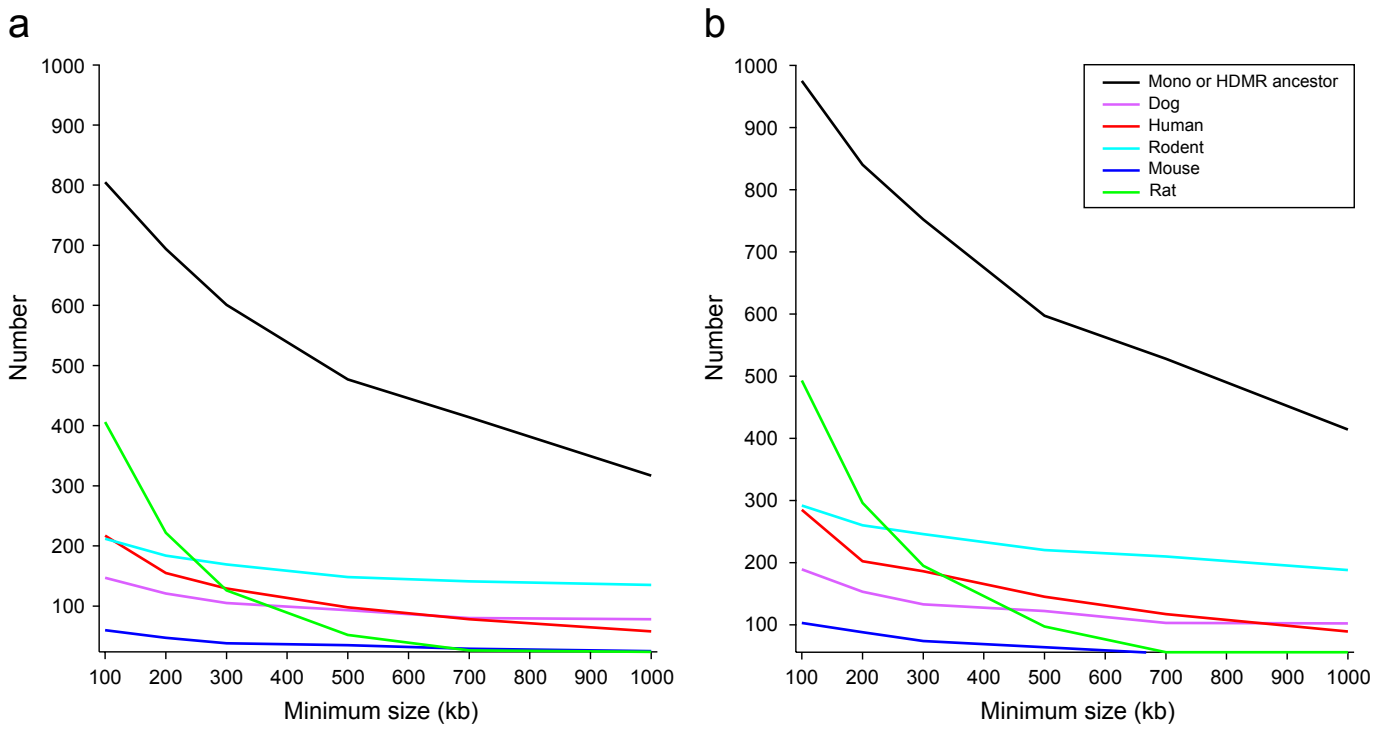
Mikkelsen et al.
Supp. Figure 5



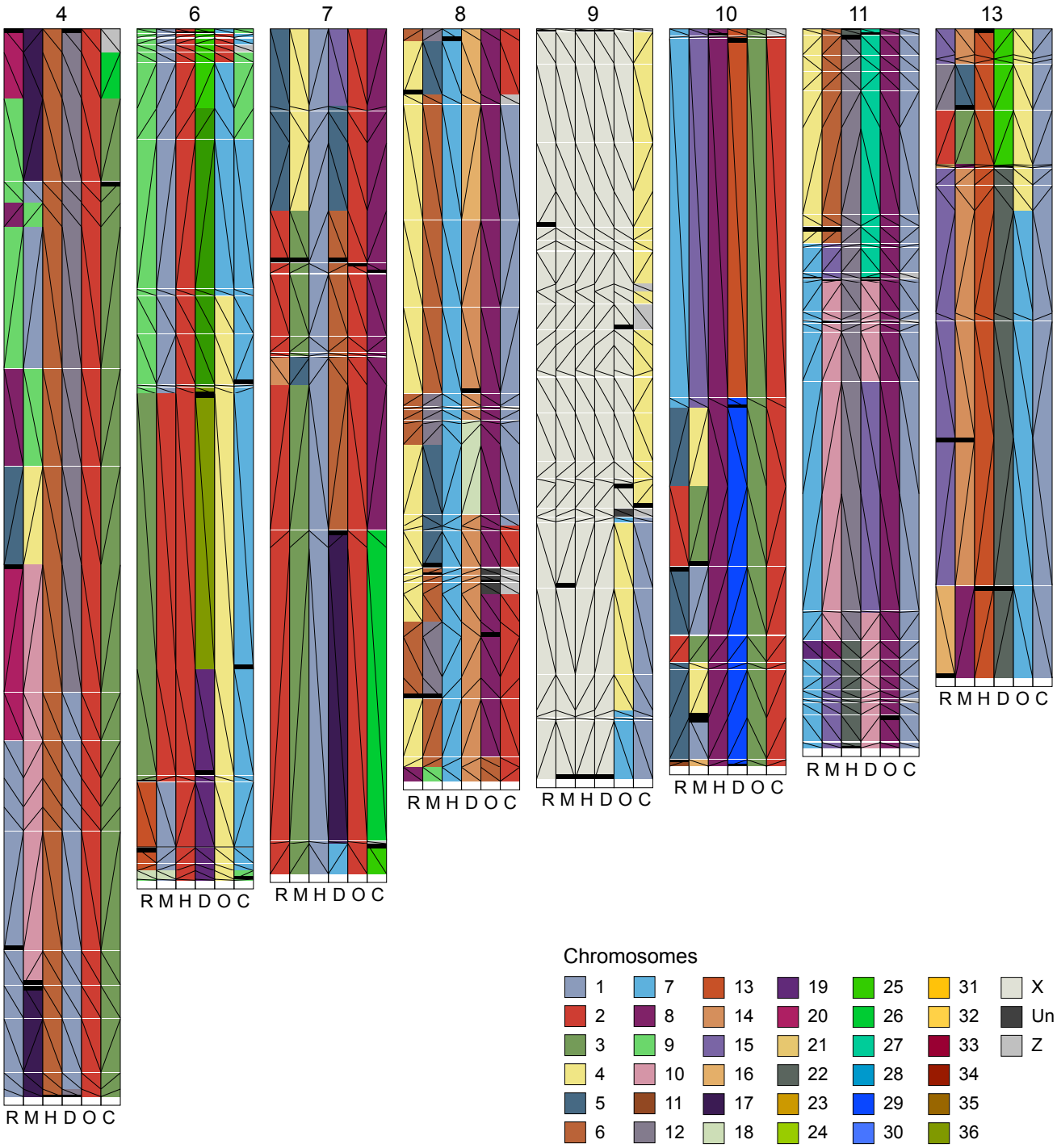
Mikkelsen et al.
Supp. Figure 6

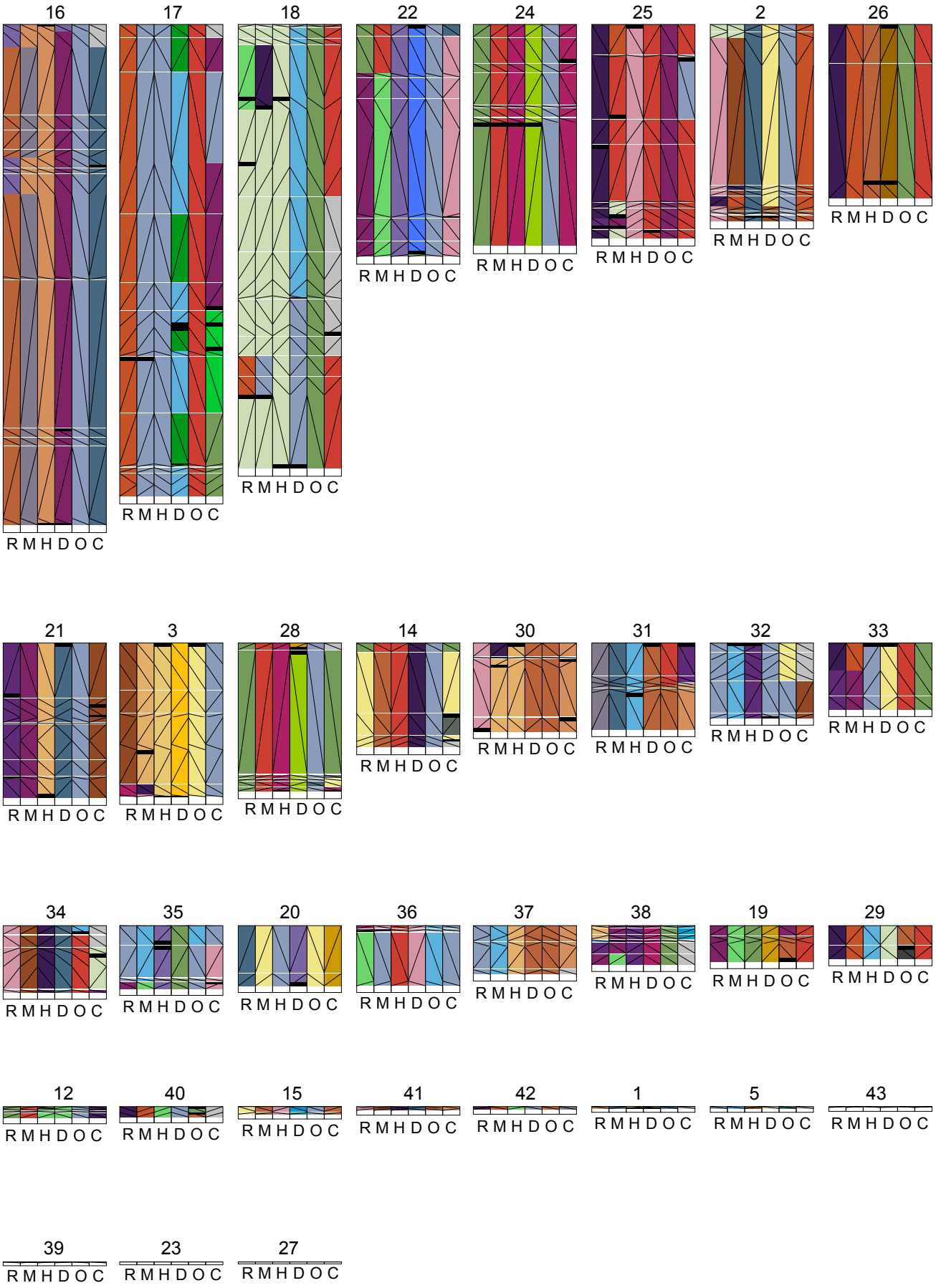


Mikkelsen et al.
Supp. Figure 7

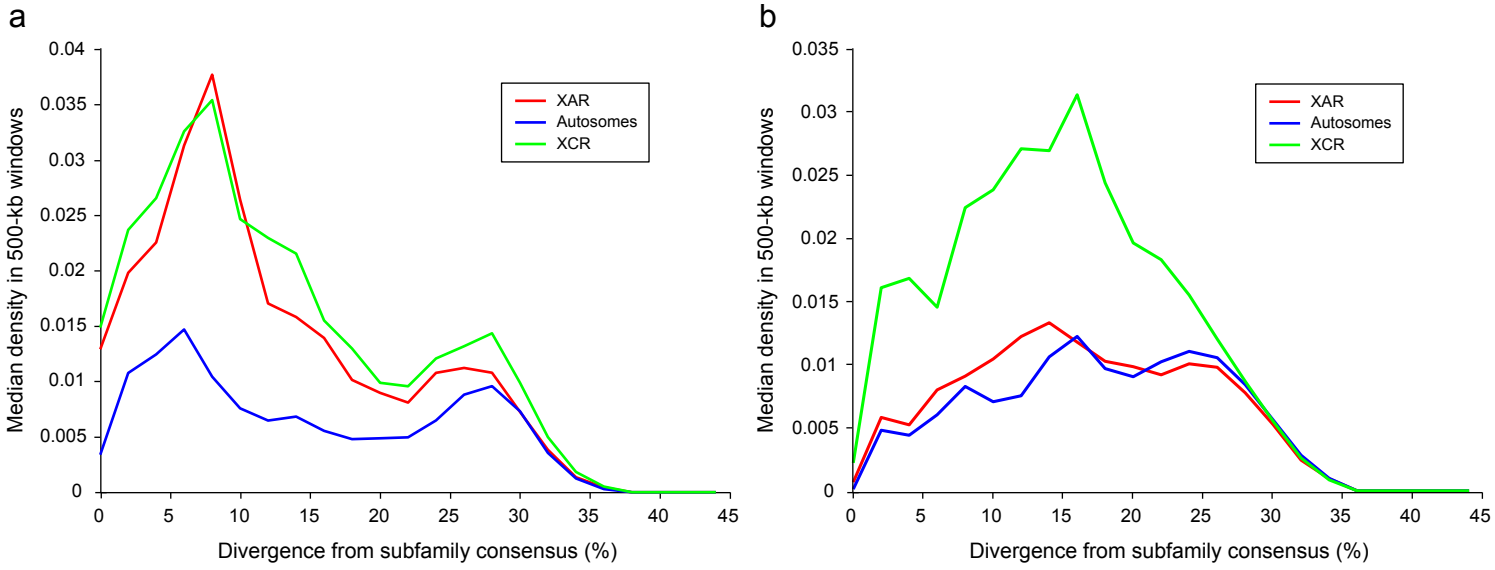


Mikkelsen et al.
Supp. Figure 8

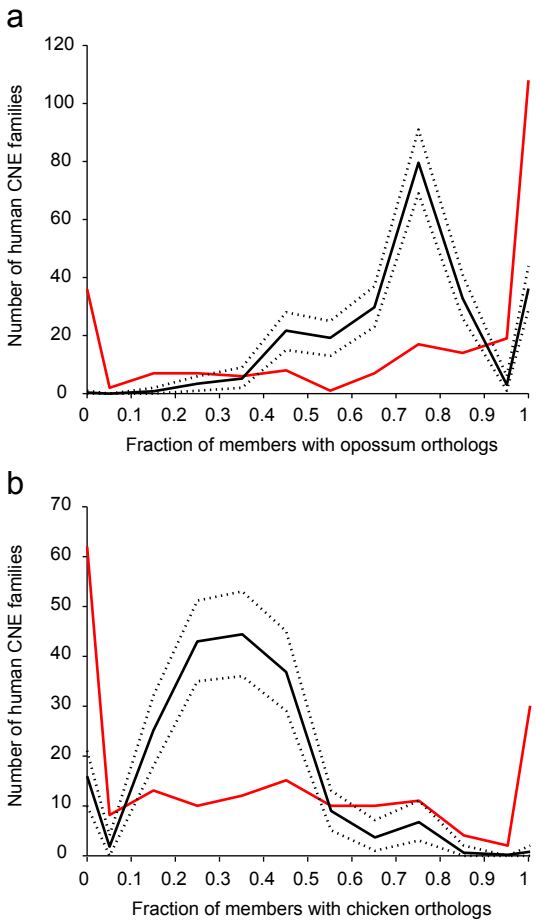



16
17
18
22
24
25
2
26
21
3
28
14
30
31
32
33
34
35
20
36
37
38
19
29
12
40
15
41
42
1
5
43
39
23
27
R M H D O C

Mikkelsen et al.
Supp. Figure 9



Mikkelsen et al.
Supp. Figure 10



Supplementary Figure Legends

Supplementary Figure S1 – Simplified phylogeny of amniotes. Infraclass Eutheria is represented by human (H), mouse, (M), rat (R) and dog (D). Infraclass Metatheria is divided into two extant lineages: the Australasian and the American, presented by wallaby (W) and opossum (O), respectively. Infraclass Prototheria is represented by platypus (P). Aves is represented by chicken (C). The two dashed lines indicate major lineages that are not yet represented by complete genome sequences. Divergence time estimates are based on [20-24].

Supplementary Figure S2 – Flow cytometric estimation of genome size. Whole blood from three female animals was obtained from P. Samollow (Southwest Biomedical Research Foundation). Buffy coats (white blood cells) were separated via centrifugation. The buffy coats were washed three times with PBS and used for flow cytometric determination of genome size. Cells were combined with chicken red blood cell nuclei, stained in propidium iodide, and analyzed by flow cytometry. Several thousand cells were analyzed in four separate experimental runs per animal using a FACSCalibur (BD Biosciences, San Jose, CA). The left peak (M2) represents the chicken red blood cells (2.33 pg per 2C nucleus) and the right peak (M1) represents the *M. domestica* sample. The number of events counted is given on the y-axis (x 1000) and the relative fluorescence intensity is given on the x-axis. Estimates were consistent both between runs and between individuals. From these data we concluded that the genome size of *M. domestica* is 3.5-3.6 pg.

Supplementary Figure S3 – Linkage disequilibrium (LD) in *Monodelphis domestica* laboratory stocks. LD (represented by r^2) was computed from ten genomic regions for all individuals genotyped (blue), individuals from SFBR Population 1 (red) and from SFBR Population 2 (green). LD falls rapidly at ~15 kb in all populations. In population 1 a plateau extending past 15kb can also be seen reflecting some long range LD within this laboratory population.

Supplementary Figure S4 – Sequence composition in the opossum genome. Distribution of (G+C)-content (upper) and CpG density (lower) in 1-Mb sliding windows

across opossum chromosome 1. The centromere is indicated by the dotted red line. Increased (G+C)-content and CpG densities are evident in the subtelomeric regions.

Supplementary Figure S5 – Opossum-human synteny map at 500 kb resolution. Segments on opossum chromosomes 1-8 and X are colored by their orthologous chromosomes in human. Diagonal lines show the extent of collinear syntenic segments. Large gaps in the map (white) typically correspond to extensive gene clusters where synteny is difficult to ascertain. Centromeres are indicated by the opposing red triangles. The estimated size of each chromosome is given on its right. The chromosomal assignments and size estimates reflect all available FISH data (assembly version monDom5).

Supplementary Figure S6 – A simplified multi-species synteny map (Breakpoint graph). (A) There are four ancestral blocks (labeled 1-4) that make up chromosome 1 of species H, chromosome 2 of species M and chromosome 3 of species R. The blocks appear in different order and orientation in each chromosome. (B) Synteny is represent as a graph where each block has two nodes (- and +) with exactly three incident edges each, totaling three incoming and three outgoing. Blocks at the ends of chromosomes are handled by allowing edges to start or end on special terminal nodes (S and E).

Supplementary Figure S7 – Synteny breakpoints assigned to lineages. (a) Number of simple breakpoints associated with given lineage and map resolution. (b) Number of simple or three-way breakpoints associated with a given lineage and map resolution.

Supplementary Figure S8 – Reconstructed boreoeutherian synteny groups. All 42 ancestral synteny groups are shown sorted by size. For each group, the corresponding chromosomal assignments for mouse (M), rat (R), human (H), dog (D), opossum (O) and chicken (C) are shown.

Supplementary Figure S9 – LINE/L1 density in human and mouse by approximate age of insertion. Curves show LINE/L1 density for autosomes, the XAR and the XCR in mouse (left) and human (right) as a function of divergence from the subfamily consensus, which approximates the age of insertion. In both species, the X chromosome enrichment

is particularly strong for relatively recent insertions. The XAR enrichment unique to mouse extends almost as far back as the XCR enrichment.

Supplementary Figure S10 – Phylogenetic distribution of paralogous CNE families with 4 or more members in human. a, The observed distribution of the fraction of members of eutherian CNE families with identifiable orthologs in opossum (red) is significantly biased towards 0 and 1, compared to the distribution expected if the presence and absence of opossum orthologs was uniformly distributed between the families (black). **b**, Discounting completely eutherian-specific CNE families, the observed distribution of the fraction of members of eutherian CNE families with identifiable orthologs in chicken is also significantly biased towards 0 and 1.

Supplementary Data

All assembly versions, with quality scores and markups are available from:
http://www.broad.mit.edu/ftp/pub/assemblies/mammals/monodelphis/

SNP maps are available from:
http://www.broad.mit.edu/mammals/opossum/snps.html

Gene predictions and orthology assignments are available from:
http://wwwfgu.anat.ox.ac.uk:8080/download/monodelphis/1v2/orthologs/

Segmental duplication annotations, synteny maps, reconstruction of the boreoeutherian karyotype and conserved element predictions are available from:
ftp://ftp.broad.mit.edu/pub/papers/opossum_genome/

Acknowledgements

Generation of the *Monodelphis domestica* sequence at the Broad Institute of MIT and Harvard was supported by grants from the NIH/NHGRI (Grants U54HG003067-03 and U54HG003067-04). BAC library construction was supported in part by NIH grant U01 HG02526 (C. T. Amemiya).

Members of the Opossum Genome Sequencing Consortium were supported by NIH grants R01 HD044078 (J. Greally), R24 RR014214 (P. B. Samollow), R01 DK065058 (J. L. VandeBerg), RR014214, 1R01LM007994-01 (P. V. Benos), IP20RR18754 (M. Baker), R33GM065612 (D. D. Pollock), R01GM59290 (M. A. Batzer), ES015165, ES13053, and T32-ES07031 (R. L. Jirtle) and 5 P41 LM006252-09 (J. Jurka); NSF grants MCB-0316255 (P. V. Benos), BCS- 0218338 (M. A. Batzer), EPS-0346411 (M. A. Batzer and D. D. Pollock) and MCB-0234930 (R. Miller); and additional awards from the Robert J. Kleberg, Jr. and Helen C. Kleberg Foundation, the State of Louisiana Board of Regents Support Fund, State of Colorado support funds, the Pittsburgh Foundation, TATRC/DoD, the UK Medical Research Council, and the Australian Research Council.

References cited in supplementary information

1 Osoegawa, K. et al., *Genomics* **52** (1), 1-8 (1998).
2 Breen, M. et al., *BMC Genomics* **5** (1), 65 (2004).
3 Pathak, S. et al., *Cytogenet Cell Genet* **63** (3), 181-184 (1993).
4 Duke, S. E. et al., *Chromosome Res* **in press** (2007).
5 Lindblad-Toh, K. et al., *Nature* **438** (7069), 803-819 (2005).
6 Ning, Z., Cox, A. J., and Mullikin, J. C., *Genome Res* **11** (10), 1725-1729 (2001).
7 Sachidanandam, R. et al., *Nature* **409** (6822), 928-933 (2001).
8 Barrett, J. C., Fry, B., Maller, J., and Daly, M. J., *Bioinformatics* **21** (2), 263-265 (2005).
9 Waterston, R. H. et al., *Nature* **420** (6915), 520-562 (2002).
10 Ma, B., Tromp, J., and Li, M., *Bioinformatics* **18** (3), 440-445 (2002).
11 Margulies, E. H., Chen, C. W., and Green, E. D., *Trends Genet* **22** (4), 187-193 (2006).
12 Margulies, E. H. et al., *Proc Natl Acad Sci U S A* **102** (13), 4795-4800 (2005).
13 Birney, E. et al., *Nucleic Acids Res* **34** (Database issue), D556-561 (2006).
14 Birney, E., Clamp, M., and Durbin, R., *Genome Res* **14** (5), 988-995 (2004).
15 Slater, G. S. and Birney, E., *BMC Bioinformatics* **6**, 31 (2005).
16 Goodstadt, L. and Ponting, C. P., *PLoS Comput Biol* **2** (9) (2006).
17 Goodstadt, L., Heger, A., Webber, C., and Ponting, C. P., *Genome Res*.
18 Altschul, S. F. et al., *J Mol Biol* **215** (3), 403-410 (1990).
19 Xie, X., Kamal, M., and Lander, E. S., *Proc Natl Acad Sci U S A* **103** (31), 11659-11664 (2006).
20 Kumar, S. and Hedges, S. B., *Nature* **392** (6679), 917-920 (1998).
21 Lee, M. S., *J Mol Evol* **49** (3), 385-391 (1999).
22 Phillips, M. J. and Penny, D., *Mol Phylogenet Evol* **28** (2), 171-185 (2003).
23 Springer, M. S., Murphy, W. J., Eizirik, E., and O'Brien, S. J., *Proc Natl Acad Sci U S A* **100** (3), 1056-1061 (2003).
24 Woodburne, M. O., Rich, T. H., and Springer, M. S., *Mol Phylogenet Evol* **28** (2), 360-385 (2003).