Table of Contents

1. Supplementary Notes S1-S24

Sequence Assembly S1…… 2

Integration of sequence assembly and physical map S2….. 3

Creation of chromosomal "AGP" files S3……. 4

Assembly quality assessment S4……. 5

Read depth analysis S5…… 6

Segmental duplications S6…….. 7

Repeat content analysis S7……. 8

Chimaeric read analysis S8 …….. 8

Cloning biases analysis S9 ……. 9

G+C content analysis S10…….. 9

Estimates of heterozygosity rates in platypus S11…… 10

ncRNA content analysis S12…….. 10

cDNA sequencing S13…….. 11

Gene predictions S14……... 11

Gene orthology S15…….. 12

Orthology assignment S16…….. 12

Nucleotide substitution rates between orthologs S17……... 14

Gene evolution S18…….. 15

Phylogenetic position of platypus S19…… 16

Interspersed repeats S20……. 18

Population structure analysis S21…….. 19

Microsatellites analysis S22…… 20

G+C fraction in various mammalian species S23……. 23

CpGs at promoters and other regulatory elements S24….. 23

2. Supplementary Tables 1-11

3. Supplementary Figures 1-14

4. Supplementary References

1. Supplementary Notes

Sequence Assembly S1. Early genome size estimates for the platypus suggested a genome size of 2.7Gb[1] which, provided a C-value estimate of 3.0. The C-value describes the DNA content of a genome in picograms per haploid genome. The tau value, the number of basepairs per genome, should be C*Na(Avogadro's num)*(mean molar mass of a basepair)^-1. Molar mass of a pair of nucleotides is 660 g/mol. Therefore, genome size for platypus is C-value (pg): 3.0 and genome size for platypus in base pair = 3.0* 10 ^-12 * 6.022*10 ^23/660 = 2.7Gb.

Based on comparisons of true genome size to C values (Accessory Table A1) of mouse, opossum, chimpanzee and chicken all were overestimates by on average ~20%. This suggests a more accurate genome size for platypus is ~2.4Gb.

A more recent genome size estimate (J.S. Johnston, unpublished data) is 2.35Gb. In this biochemical analysis, the platypus genome was compared to the chicken genome. The 1C genome is 206% of the chicken genome. If one assumes the chicken (hen) genome size of 1C (hen) = 1213mb (2C = 2.33 pg), this produces a genome size for the platypus of 2352 +/- 10.6mb. Additionally, a syntenic region was sequenced in both platypus and human[2]. The region in human was 1.65Mb and in platypus was 1.26Mb. This would suggest the platypus genome should be ~76% of the size of the human genome or ~2.3Gb.

In another recent genome size estimate (P. Obrien, W. Rens and M. Ferguson-Smith, unpublished data) using the method reported by Trask *et al.*[3], flow cytometry was used to determine the lengths in basepairs of individual platypus chromosomes. Flow cytometry measures DNA content and is not influenced by the degree of chromosome contraction potentially giving more accurate sizing. When adding the base pair lengths for each chromosome together, the total size was 1.92Gb.

The platypus (*Ornithorhynchus anatinus*) genome was sequenced to a total of 6.25X total coverage or ~6X phred Q20 redundancy (using 2.4Gb as the estimate of genome size) from plasmid, fosmid and BAC end sequences (Acessory Table A2).

The combined sequence reads were assembled using the PCAP software[4], 300 parallel PCAP jobs ran on a cluster of AMD® opteron™ computers (AMD, Sunnyvale, CA, USA) with 2~8Gb Random Access Memory (RAM) and dual processors. PCAP was then used to process the overlaps (bdocs, bclean), calculate the layout (bcontig) using HP itanium with 96 RAM and 4 processors, an to generate the conensus sequence using 300 parallel processes (bconsen) on 8G RAM clusters of AMD opterons. The stringent parameters used for each step are provided at the end of this document.

After the initial assembly with PCAP, we modified the read pair constraints iteratively after recalculating the statistics on read pair distances from the assembly. While the contiguity improved some, the platypus assembly was still fragmented as compared to other large vertebrates such as chimpanzee and chicken where we had used PCAP to assemble them, see Table A3. N50s calculated for the subset of largest contigs/supercontigs that total the genome size indicated by the number in parentheses in the 1st column (e.g. Platypus (1.7Gb) indicates that the largest supercontigs totaling up to 1.7Gb were used for the N50 calculations). We noted that the platypus has

relatively less coverage from fosmid ends reads (which provide scaffolding at the mid-range size of ~40kb). To understand whether this caused the increased fragmentation of this genome relative to other assemblies, we removed fosmid ends reads from the chimpanzee genome and reassembled the genome using PCAP. Even after removing the fosmid end data from the chimpanzee, however, the platypus assembly was still less contiguous suggesting that the fewer number of fosmid ends is not the major reason for increased fragmentation of the genome.

Thus the read quality was reasonable (84% good quality reads), and the average read length (703 bases), with an average of 591 Q20 bases, is similar to what we have found with other data sets (not shown). The chaff rate, or number of reads which could not be assembled, is relatively low at 8% (although in chicken the rate was ~4%). However, the assembly has a larger number of contigs (343,384 contigs > 1 kb) and supercontigs (143,543 supercontigs > 1 kb) than would be expected for this level of coverage based on our experience with other genomes (Accessory Table A3).

In this assembly we identified several sources of small amounts of vector and other contamination. For example, close examination of platypus gene alignments with other species, focusing on the degradome of platypus, revealed best statistical matches to a small number of sequences from the protozoa, *Theileria* (X. Puente, personal communication). Based on these initial observations a thorough evaluation of all contigs was carried out. In summary, ~1076 supercontigs totaling nearly 3M were flagged as potential *Theileri*a sequence. Of those, 230 supercontigs (half of the sequence) had more than one contig aligning to Theileria (either only Theileria or more strongly to *Theileria* than to vertebrate). The remaining 846 supercontigs (again about half of the sequence) are supercontigs (could be singletons) that have only one contig aligning more strongly to *Theileria* than vertebrate. Other contaminants that were removed included eleven small primate contigs were identified (Bob Harris, personal communication), one *C. elegans* and one *Z. mays* chloroplast contig (Paul Kitts, personal communication). Of the submitted contigs 1,122 AAPN01 contigs match sequences from the *Ornithorhynchus anatinus* mitochondrial genome (NC_000891.1). These mitochondrial sequences remain in the submitted data.

Integration of sequence assembly and physical map S2. The draft assembly was aligned to the fingerprint map using shared BAC-end sequences (BES) with additional linking information provided by *in silico* digestion of supercontig sequences as described[5]. There were a total of 379,614 BES that could be used for this purpose. The predicted *Hind*III fragments from each assembly supercontig were used to create overlapping artificial clones of approximately 200 Kb, with successive clone overlap of 60Kb. The artificial *in silico* clones were added to a copy of the FPC fingerprint database and compared to experimental clones at a coincidence score cutoff of 1e-06. The threshold for establishing collinearity of the assembly and fingerprint maps was 1) a minimum of six shared BES links or four BES links from a minimum of three different BACs or 2) a minimum of two supercontig *in silico* clones matching clones in the given fingerprint contig, provided there were supporting BES links. 70% of all links were required to be consistent and no topological constraints were violated. These heuristically-determined parameters were chosen to minimize the number of topologically impossible contig combinations and marker inconsistencies in this and other projects[6,7] 86% of the fingerprint map and 82% of the assembly were aligned. This alignment allowed us to construct 'ultracontigs'

containing ordered and oriented supercontigs. 49% of the assembly was assigned to ultracontigs, providing a 270% increase in scaffold length over the supercontigs alone. BAC clones were chosen from the larger ultracontigs to use in FISH, with 279 clones successfully assigned chromosome locations. These data suggested 6 ultracontig breaks. Assignment of sequence scaffolds to platypus chromosomes follows the agreed nomenclature based on size, cytological landmarks, hybridization of chromosome-specific DNA and definition by hybridization with anchor BACs[8].

Creation of chromosomal "AGP" files S3. As described above, a whole genome shotgun assembly was performed using stringent parameters to avoid global mis-assemblies. We then linked the assembly to the physical map (as detailed above) to, where possible, order and orient supercontigs into ultracontigs. Using the new ultracontigs, we used the underlying WGS assembly read pairing data to refine supercontig order and orientation (note that when supercontigs are small, the physical map does not always definitively provide orientation). We then filtered the following small contigs before creating the chromosomal files: (1) all singleton contigs that were <500 basepairs and did not have a hit to the EST set, (2) supercontigs <2kb that were not part of an ultracontig and did not have an EST hit at >95% identity and (a) were >97% identical over >97% of their length to other larger ultracontigs or (b) were >95% repetitive (based on RepeatMasker).

The remaining ultracontigs and supercontigs were initially ordered along the chromosomes using the FISH data generated by this project (Australian National University, The University of Adelaide, and Washington University Genome Sequencing Center). The FISH data placed 198 supercontigs on individual chromosomes and identified one chimeric ultracontig (we then broke that ultracontig into its two constituent pieces). The FISH data provided gross localization data (identifying either a single band or sometimes just the arm such as 2q), and, in general (except for two cases where two probes were designed from the same ultra/supercontig and happened to fall in different bands), were not helpful in providing orientation information. After this gross ordering stage using the FISH data, we returned to using the underlying read pairing data to attempt to order and orient ultracontigs within bands/arms (for example, if three supercontigs were assigned to 2q, we had to determine order and orientation of those supercontigs within 2q). When there were multiple "grouped" (by arm or band) but unordered/unoriented super/ultracontigs and no platypus-specific data (i.e. no read pair data or EST data), we looked for a consensus order/orientation from comparative alignments to human (build36)[9], chicken (galGal3)[7], opossum (monDom4)[10] and dog (canFam2)[11] to aid in assigning order and orientation. As a final step, comparisons to newly available EST data (30,644 EST assemblies) were used to extend and in some cases create additional small ultracontigs (by placing neighboring supercontigs next to each other when they were linked by unique alignments with EST data). This process extended/confirmed 60 ultracontigs and created an additional 215 small ultracontigs. All super/ultracontigs that were not localized to a chromosome by virtue of FISH were assigned to the unlocalized chromosome, "chrUn."

The final chromosomal assembly is composed of 205,534 supercontigs (of those, using the physical map 4,197 supercontigs were organized into 689 ultracontigs) covering 1.84Gb of actual sequence (without including estimated gap sizes) or almost 2.0Gb including gap sizes. Of the 1.84Gb, 437Mb (1507 supercontigs organized into

145 ultracontigs) have been anchored and ordered along platypus chromosomes using the physical map in combination with FISH data. The N50 statistic is defined as the length L such that 50% of all nucleotides are contained in contigs of size at least L. The N50 number is 298 and the N50 size is 967kb. The amount of sequence localized per chromosome is provided in Accessory Table A4.

Assembly quality assessment S4. To assess assembly quality, we analyzed genome coverage, assembly accuracy and rates of mis-assembly. However, these analyses were compromised by the outbred platypus populations given that polymorphisms could masquerade as errors. Estimates of platypus genome size center around 2.3Gb[1,2] (J.S. Johnston, unpublished data) including centromeric and telomeric sequence (Supplementary Notes S1). We estimated the genome coverage of the assembly using other sources of platypus sequence including finished BACs, additional WGS data, and a set of mRNAs. In a set of 97 finished BAC sequences from an unrelated male platypus (13.2Mb) and three from the female platypus used for whole genome assembly (WGA,428kb), 90% of finished bases could be aligned with the draft whole genome shotgun assembly indicating the missing sequence was not restricted to centromeres and telomeres. As a part of our efforts to characterize the platypus genome, we used 454 instrumentation[2] to sequence an additional 0.04X WGS coverage and to develop a cDNA resource (Supplementary Notes S3;S13). Of the 454-based WGS sequence, 93% aligned to the assembly, but this is likely an overestimate given it may share some of the biases of the WGA and repeated sequence that was collapsed in the assembly may still produce alignments. Of the 2,677 cDNA assemblies larger than 1kb (totaling ~4 Mb), 92% were identified in the current genome assembly. While this also likely overestimates coverage since it avoids repeated sequence, it nevertheless is an important indicator of the representation of the transcribed sequences in the assembly.

While accuracy estimates are difficult given the heterozygosity of the female, we estimated the nucleotide accuracy of the genome using the alignments to finished clones. Comparison with female sequence revealed overall substitution, deletion and insertion rates of 0.04%, 0.01%, and 0.01%, respectively, but these are probably primarily polymorphisms, not errors. Comparisons with the male showed higher discrepancy rates but these are also likely dominated by polymorphisms (sub=0.09%, del=0.01%, ins=0.01%).

To assess structural accuracy of the assembly, we analyzed the alignments of the 13.2Mb of finished platypus BAC clone sequence against 1,895 platypus contigs from 604 supercontigs. Six small contigs (~0.7 per Mb) have not been positioned within large supercontigs; these are not strictly errors but do affect the utility of the assembly. Three cases were found where a supercontig should have been inserted inside another supercontig (~0.3 per Mb). Only four order discordances (misordered sequence contigs within a supercontig) were discovered translating to ~200kb of these events in the genome, and no misoriented contigs were detected.

To attempt to understand the nature of the missing sequence, we analyzed in detail the content of the sequence that was covered by the whole genome sequence as well as the sequence that was "uncovered" or in "gaps" in the whole genome assembly and found it higher in both G+C and repetitive content. Overall G+C content of those

"covered" sequences was 44% while it was 50% for those regions not covered. The distribution is as shown Accessory Table A5.

RepeatMasker identified 43% of the covered sequence as repetitive and 65% of the uncovered sequence as repetitive. The repetitive content is shown in Accessory Table A6. As seen here, there are almost 2X as many bases in LINEs in the "uncovered" regions than in the "covered" regions and more than 2X as many bases in simple sequence repeats and low complexity sequences. We further analyzed the 50bp immediately adjacent to the regions not covered by alignments to the WGS sequence. We found those regions to be enriched in tandem repeats as well as in bases masked by RepeatMasker relative to the genome as a whole. Specifically, there were 7.1% of bases annotated as tandem repeats in these 50bp flanking regions as compared to 2.1% for the genome as a whole. And 67.3% of those bases overlap bases masked by RepeatMasker (32.2% LINEs and 30.7% SINEs as compared to 21.7% LINEs and 22.7% SINEs). Thus, while the content of the "missing bases" are enriched for LINEs, the content of the "edges" of the gaps while enriched for repetitive bases, are not biased towards LINEs or SINEs. It is not surprising that the sequence is stopping within LINEs and SINEs given that their average G+C content in the platypus genome is 55% (57% for LINE2 alone). In the human genome the average G+C content for LINEs is 44% and for SINEs is 56%. Finally, as would also be expected, the average G+C content at the "edges" of the gaps in coverage are higher than that of the genome as a whole. The average G+C content for the 100bp and 50bp flanking each gap in coverage is 49%; for the 25bp and 10bp flanking each gap it is 50% (as compared to 45% for the entire set of finished BACs).

Read depth analysis S5. We analyzed the following two types of read depth distribution (Accessory Fig. A1). In type 1, we "step" along the contig and count the read coverage for each contig base, from the underlying reads aligned at good quality base positions (>Q20). In type 2, we take the middle point of the read as representation of that read and see how many other reads cover that point in this read. The ratio of type 2 average coverage to type 1 average coverage provides an indication of the distribution of the reads for a region, i.e. whether they are more concentrated (possibly collapsed) or more spread out.

For platypus we find a ratio of 1.47 which is higher than those for other genomes such as chimp (1.27), chicken (1.27) and the worm, *C. remanei* (1.34). The read depth ratio indicates that there could be more regions collapsed in platypus in turn causing the total size of the assembly to be an underestimate of genome size (thus reducing the amount of sequence we are actually missing).

As shown in the type 1 read depth distribution (Accessory Fig. A2), the platypus does not have a peak on the mean value of read depth (chicken: 8.5, chimp: 7.0, platypus: 8.82). The shape of chicken and chimp, however, are more similar to a Poisson distribution with modes at the mean. While the platypus has many bases of low coverage, it also has many more bases at the tail. The higher percentage at a low depth reflect the genome assembly has many more short contigs with lower than average coverage, the tail with high coverage reflect possible short stretches of collapsed repeats.

As shown in the type 2 read depth distribution, the platypus data appear "flatter" than the other assemblies with a higher percentage centered around the mean depth than the other genomes represented here. It also has a tail with a higher percentage coverage than all the other genomes compared. Given that the distribution of platypus read depth in Accessory Fig. A3 is more like a random distribution than that in Accessory Fig. A2, this also indicates the genome assembly has more "piled up" or possibly collapsed regions. We therefore attempted to analyze these regions having higher than expected coverage to determine whether they did appear to be repetitive elements and to possibly obtain a better genome size estimate.

As seen here, the regions that appear to be collapsed are not rich in genome-wide repeats and have a lower than genome average G+C content level. The total size of these high depth regions provides a rough estimate that the current assembly underestimates the size of the genome by ~120Mb, bringing the total assembled sequence to ~1.96Gb (1.84Mb+120Mb=1.96Gb).

Segmental duplications S6. To further scrutinize the platypus genome assembly (ornAna1) we used two different approaches [12,13] designed to detect genomic duplicated sequences >1kb with >90% sequence identity. When requiring that the detection of duplicated regions align in apparent "unique" regions of the platypus genome, a total of 190.4Mb (10.4%; Accessory Table A7) of non-redundant sequence was detected as potentially duplicated by self-comparison of the platypus genome. The majority (67%) of alignment pairs were between 1-2 kb in length (Accessory Fig. A4) with a mean sequence identity of 94.4% (Accessory Fig. A5). This excess of short sequence alignments may reflect either uncharacterized transposon/retrotransposon repeat families or the presence of many short segmental duplications, similar to those in chicken. In order to validate larger, more identical duplications in the platypus genome, we examined all regions of the platypus genome assembly for excess sequence read depth. Using this method, we found that only a small fraction (1% or 16.6Mb) of the platypus genome corresponds to larger segmental duplications >10 kb in length and >94% sequence identity. Of 111 genomic regions containing tandemly duplicated genes (see below), 14 overlap these segmental duplications, far more than the ~1 expected by chance. The pattern of duplications is similar to what has been observed for many non-primate mammalian genomes. Most of the larger duplications that have been mapped to chromosomes are organized in a tandem configuration (Accessory Fig. A6). For example, if one considers all duplicated sequences >5 kb in length and >90% in identity, 88% (114/130) of the pairwise alignments map within 1Mb of each other. Of the two interchromosomal duplicates, one corresponded to a nearly complete copy of the mitochondrial genome (11kb) duplicated to the platypus X chromosome (X1) and likely represents a mitochondrial introgression into the nuclear genome. Due to the complexity of correctly assembling duplicated sequence[14], we should emphasize that our survey of the segmental duplication landscape is preliminary. Only ~10% of the detected duplicated sequence has been assigned to a particular chromosome (Supplementary Fig. A6). Nevertheless, the data provide the first genome-wide view of platypus segmental duplications and a useful guide for those interested in platypus-specific gene expansions and targeted finishing of the genome. All of these data are available at http://eichlerlab.gs.washington.edu/database.html and http://www.genome.ucsc.edu.

Repeat content analysis S7. We counted the bases in the original reads that PCAP identified as highly repetitive[4]; there are 2.825 billion bases in repeats that are represented more than 90 times, or about 15% of the total input read bases defined as highly repetitive by the assembler. While a low percentage (8%) of the total reads were unplaced, of the unplaced reads, the assembler annotated 56% as not placed due to repeats. The average G+C content of the unplaced reads is 47%. Average G+C content of the repeat library for platypus is 53%.

As described in the main text, the density of interspersed repeats (IRs) in the platypus genome is far higher than in any vertebrate genome characterized thus far. We aligned each of the 97 finished platypus BAC clones against itself (using PRINTREPEATS, score=50) to identify repeats within the clone and found 26% of the platypus BACs align against themselves whereas that number for a similarly sized set of chimp BACs is only 17% and for chicken is only 3%. When analyzing specifically tandem repeat content (minimum size of a single unit in the tandem repeat set to 10bp, and maximum size 100bp) there are approximately twice as many bases in tandem repeats in platypus as there are in chimp and more than five times as many in platypus as in chicken. The average size of those tandem repeats is almost twice as large in platypus as compared to chimp or chicken but the total number of tandem events per kb is slightly lower in platypus than chimp. Further, the distance between the copies of the tandem is on average larger than that in chimp or in chicken with a larger standard deviation as well. Tandems were measured using two methods for identifying tandem repeats, one was "tandem" and the other was "trf"[15]. Inverted repeats are known to cause structures which lead to cloning bias. And tandem repeats can also be unstable; they can be collapsed by *E. coli*, for example, leading to cloning bias. Repeats have also been known to lead to problems during sequencing.

Chimeric read analysis S8. Chimeras are sequences that contain segments from two different portions of the genome. In the assembly, a read "f" is identified as chimeric if it has an internal position "pos" satisfying: (1) a region immediately before (after) pos is similar to other reads; (2) all the similarity ends(starts) around pos and majority of those similarities show a "long overhang." There are two primary types of chimeric reads shown in Accessory Fig. A7: (1) local repeat structure causes a "loop" to form, and the resulting chimer contains two sequences that are near each other in the genome, and (2) two sequences from different regions of the genome are ligated during the cloning process.

Overall, there were 0.73% of reads identified as possible chimeras in platypus (196,963/26,954,275) as compared to 0.25% in chimpanzee (84,480/34,014,260) or three times more in platypus than in chimp. In platypus, these 196,953 chimeric reads are from 92,148 plates (total plates=171,917) or 53% of the plates. There are only 1,139 plates with more than 10 chimeric reads (totaling 16,255 of the chimeras or 8% of the total). Thus, the chimeric reads do not appear to be plate specific and thus not a problem during production of plates.

Finally, we checked whether the chimeric reads appear at contig ends and could possibly affect the contiguity. Given that the number of contig ends affected is similar in both chimp and platypus, the chimeric reads are not a likely major contributor to the fragmentation of the assembly (Accessory Table A8).

Cloning biases analysis S9. To estimate the possible effects of cloning bias, we evaluated the orientation of reads at the ends of contigs. If, in neighboring contigs, all reads point away from each other, for example, we would reason that there is a cloning bias in this region possibly caused by inverted repeats. For the reads at the contig end, we defined the following combinations of orientations as shown in Accessory Fig. A8. The first arrow represents orientation of the read at the "contig end" of the "first" contig. The second two represent the orientation of the first two reads of the neighboring "next" contig downstream. For the "contig start" read, if there is cloning bias resulting in no clone being available for that portion of the genome, then we would see a large number of "1_1_1 <--- ---> --->", or all pointing away from the gap. For the "contig end" read, if there is a cloning bias, we would see many orientation combinations of "0_1_1 <--- <--- ---->", that is, the previous reads point away from the gap as the contig ends, and reads in the next "contig start" read are also pointing away from the gap. The overall results for each orientation combination are tallied in Accessory Table A9.

There are only 8.08% of platypus "contig start" combinations of type "1_1_1" whereas in chimp there are 7.37%. There are 8.2% of the platypus "contig end" combinations of type "0_1_1" while chimpanzee has 7.31%. In fact, in the gaps between contigs, there is a slightly lower percentage of gaps with no spanning clone that there are in chimpanzee. Thus, the effect of cloning bias in platypus is similar to that in chimpanzee and, thus, should not be a factor contributing to the more fragmented assembly.

G+C content analysis S10. We measured GC content of each fragment in platypus assembly OrnAna1. For each fragment we divided the number of bases that were either C or G by the number of bases, not counting any ambiguous bases. Repeat-masked bases were included in the counts. Average G+C is flat at around 47% from 2,000bp to 30,000bp. Below 2,000 it generally rises, to about 53% at 500bp. Above 30,000 it generally drops, to about 43%.

From the fundamental Central Limit Theorem, we know a population of variables x with some distribution, mean, and standard deviation; random samples are taken from this population of an acceptable size n. The sample mean essentially is equivalent to the population mean, suppose the population has real GC content with a mean, then random samples of subset has a mean close to the real one provided small size is acceptable.

We used the 454 sequencing platform to generate 1 million 100bp reads (0.04X coverage) for the platypus genome. For each of the eight batches of reads, the G+C content of these reads was 47% which is higher than both the genome average (45%) and higher than the read average (43%). We had previously used both the ABI 3730 and the 454 platforms to generate data for several bacterial genomes. We calculated their GC content of both types as shown in Accessory Fig. A9.

We can see the 454 and 3730 data both share similar G+C content as opposed to the platypus where the 3730 reads show 43.8% and the 454 show 47.4%. Finally, we compared G+C content of the read set versus the G+C content of the assembly for several genomes (Accessory Fig. A10). The platypus data show a larger difference

between the %G+C of the reads and that of the genome as a whole as compared to the other genomes shown here.

We also measured the G+C content of the various lengths of the contig ends as well as for regions of high read depth and low read depth in the platypus assembly and found the following distributions (Accessory Fig. A11a). The G+C content is higher on contig ends in general than in the genome as a whole. As is evident here, the G+C content rises as you approach the contig ends confirming earlier observations during the finishing of BACs and in the alignments with finished BACs. This is also consistent with the notion that the high G+C content is causing premature termination of reads during the sequencing process. Further, we find that the regions of the assembly with the lowest read depth (Accessory Figure A11b) also are enriched for sequence even higher in G+C than the contig ends also indicative of biases against the G+C rich sequence.

Estimates of heterozygosity rates in platypus S11. From the platypus chosen for sequencing, we chose 100 primer pairs for heterozygosity testing. From this we sequenced 69,738 bases and found 52 points of heterozygosity including two indels which were counted as one difference per site. Thus, the rate of heterozygosity based on these data is 1 per 1341 bases.

To enable future genetic mapping studies of platypus populations we created a large database of single nucleotide polymorphisms (SNPs). One area of immediate application of this SNP data would be to investigate any genetic correlates or even disease resistance loci of infection with the fungus *Mucor amphibiorum*, which poses a significant potential threat to platypuses in Tasmania. *M. amphibiorum* occurs throughout the Australian range of the platypus, but only platypuses in Tasmania are affected, with fungal infections producing large ulcerative lesions, leading to significant morbidity and mortality, probably from consequent bacterial infection.

SSAHA2 identified 1,120,308 SNPs (1 in 1644 bases) and PCAP identified 675,550 SNPs (1 in 2726 bases). Significant overlap was observed between these independent analyses. We experimentally tested a total of 141 SNPs from the set of 550,203 SNPs that overlapped from the two analysis techniques. Of these 138 were confirmed (97.9%) demonstrating the high accuracy of those SNPs identified by both methods. Discovered SNPs have been submitted to dbSNP.

Based on the differences in the assembly itself, PCAP (our assembly tool) identified 675,550 SNPs in the 1.84Gb of sequence or a heterozygosity rate of 1 in 2726 bases and SSAHA2 identified 1,120,308 SNPs or 1 in 1644 bases.

ncRNA data analysis S12. From 21,890 sequenced cDNA clones with exclusion of unreadable or very short sequences, empty vectors, *E. coli* contaminations and other ambiguities yielded 18.676 clones; among these 30 known U snRNAs (1390 sequences), 75 tRNAs (2124 sequences), 2 isoforms of 7SL SRP (signal recognition particle) RNA (5 sequences), 59 parts of rRNAs (11059 sequences), 98 mRNAs (128 sequences), 5 Y RNA sequences (106 sequences), one RNAseP RNA (2 sequences), one 7 SK nuclear RNA (14 sequences), one spacer (6 sequences) and 31 microRNA candidates (80 sequences) were identified and excluded from a more detailed analysis.

Further 3140 sequences contained 166 snoRNAs (109 C/D-box (2985 sequences) and 57 H/ACA-box (155 sequences) snoRNAs).

cDNA sequencing S13. In order to add experimental support to the *in silico* gene predictions for the platypus genome we sought to produce cDNA sequence data from a platypus fibroblast cell line. RNA extraction was performed on a cell pellet from one T75 cell culture flask using the RNeasy Mini Kit (Qiagen). The RNA was treated with Turbo DNase (Ambion) and measured for quantity and quality via spectrophotometry and gel electrophoresis.

Library construction was performed using a variation of the Clontech SMART system in which the 5' and 3' PCR adapters (5' sequence; 3' sequence) contain type IIs restriction enzyme sites (*Mme*I). The RT conditions were modified to include high-temperature cycling for greater efficiency. Post-RT amplification of the library was performed with a single PCR primer (seq) that maintained the *Mme*I site at the 5' and 3' termini. This product was titrated for the optimum number of PCR cycles, to avoid over-cycling of the product.

The optimally-cycled product was then normalized using a duplex-specific nuclease (DSN) that preferentially digests double-stranded DNA in the presence of single-stranded DNA (Trimmer; Evrogen). In short, the cDNA library DNA is boiled and allowed to re-anneal for approximately 5 hours in a buffered salt solution. During this time the high-copy molecules re-anneal while the low-copy molecules maintain the single-stranded state. At the end of the incubation period, the nuclease is added and the dsDNA molecules are digested, leaving the single-stranded material behind as a template for re-amplification using the single primer discussed above. Another PCR cycle titration was performed at this stage to prevent over-cycling.

As mentioned above, the 5' and 3' terminii of the cDNA molecules contained *Mme*I restriction enzyme sites. The inclusion of such sites allowed for cleavage of the poly-A tail from the 3' end of cDNAs as well as removal of both 5' and 3' adapter sequences in the final purification stage. The use of biotin-tagged adapters permitted purification of *Mme*I-digested library products with M280 streptavidin beads (Invitrogen). This normalized and purified product was then deemed ready for 454 sequencing library construction, which was performed according to the standard 454-FLX library protocol.

Gene predictions S14. Predicting gene models on *Ornithorhynchus anatinus* presents a unique challenge due to its large evolutionary distance from most of the mammalian sequence evidence available. A variety of approaches have been used in order to produce the highest quality gene models from the evidence available.

Initially the standard Ensembl annotation pipeline was used[16]. The process is started by aligning both the limited platypus protein set and mammalian proteins from Uniprot to the genome and making CDS predictions based on these seeds using genewise[16]. The small number of platypus cDNAs and human cDNAs were also aligned to the genome using exonerate and these alignments were used to add UTR information to the CDS structures. A comparative analysis with human, mouse, dog and chicken Ensembl gene sets helped find missing one to one orthologs and partial transcript structures. For the problem cases identified, transcript structures generated

by alignment of human and chicken Ensembl peptides were then used to both fill gaps in the gene set and extend partial structures.

After the initial annotation cDNA data from 454 sequencers became available they were also aligned to the genome using exonerate and used to add UTR information to the CDS predictions from genewise reference here. cDNA alignments in locations where there were no protein based predictions were examined for long open reading frames and used to fill potential gaps in the gene set.

After merging all the transcript predictions into a non-redundant set of loci it became clear there was an over prediction of alternative splice forms: the merge resulted in 38,155 transcript models. This is thought to be due to the distance of the evidence from the genome meaning two almost identical sequences produce distinct models rather than merging to produce the same model. This problem was solved by filtering the alternative isoforms on the basis of several criteria, including the quantity of evidence supporting each exon, presence of methionine, stop signals and the number of non consensus splice sites.

Finally, the gene set was assessed to remove processed pseudogenes. This process looks at genes which contain many frameshifts, introns less than 10bp long, and also introns which have high repeat content. This process removed 201 genes from the protein coding set.

Once the Ensembl annotation process was complete other annotations were added to the gene set, such as the manual curation of olfactory receptors (personal communication, Tsviya Olender and Doron Lancet). These manually annotated gene models were used in locations where the Ensembl pipeline had failed to produce a prediction. The final set contains 18,597 genes, 27,557 transcripts and 186,394 unique exons. The majority of these models, 16,091, were based at least in part on protein evidence from Uniprot. 2,507 genes have no support from the Uniprot proteins. Most of these genes are based on either the 454 cDNAs which added 676 genes to the set and the PhyOp process[17] which added 718 genes to the set. There were also models supported by platypus proteins and orthology evidence.

Gene orthology S15. Assignments were made for protein-coding genes predicted by Ensembl for six amniotes (Accessory Table A10). At least one orthology assignment can be made for the majority of genes in each genome (Supplementary Table 5). The number of assigned orthologs ranges from 82% for platypus to 94% for dog. This is consistent with previous findings, based on orthology analysis, that the true gene count is lower than the current ENSEMBL sets[17,18]. The number of simple ortholog sets, which contain exactly one gene per species, decreases with increasing divergence (Accessory Table A11). Orthology and paralogy assignments across these six species, and phylogenetic trees, are available to download from http://genserv.anat.ox.ac.uk/clades/amniota.

Orthology assignment S16. Orthology assignment followed a procedure implemented previously[19]. Orthologs were predicted in three steps: (1) orthologs were predicted between pairs of genomes using PhyOP[17], (2) pairwise orthologs were combined into clusters, and (3) clusters were split into orthologous groups using chicken as the outgroup.

For each pair of genomes, each translated transcript was aligned against every other translated transcript using BLASTP[20]. Alignments with *E*-values exceeding 10^{-5} or covering less than 75% of the smaller sequence were removed. The remaining alignments were weighted according to a normalized bit score

$s_{ij} = 1 - \frac{max(s'_{ij}, s'_{ji})}{min(s'_{ii}, s'_{jj})}$, where s'_{ij} is the bit score for a BLASTP alignment between sequence *i* and sequence *j*. Orthologs between transcripts were then assigned using PhyOP, a tree-based orthology assignment procedure. Orthology relationships between transcripts were then translated into orthology relationships between genes.

Pairs of orthologous genes were then grouped into clusters in a graph clustering procedure. We constructed a graph with genes as vertices and vertices are connected if the adjacent genes have been predicted to be orthologous or in-paralogous. Clusters were given by connected components in this graph.

Genes in each cluster were multiply aligned using MUSCLE[21]. Genes with multiple transcripts were collated into a string of non-redundant exons from all transcripts concatenated in sequence. Genes were translated and aligned in amino acid space and afterwards back-translated into nucleotide sequences.

Phylogenetic trees were built with NJTree providing the established phylogeny for the amniota[22] and using the "-best" option. Incomplete and inconsistent genes confuse the tree building procedure and were eliminated using a heuristic procedure: If two genes from the same species do not overlap in their multiple alignment, then the shorter gene was removed; if it was adjacent to the longer gene and in consistent orientation on the genome. Each tree was then split into orthologous groups using chicken sequences as outgroups.

We obtained 16,807 orthologous groups from the clustering procedure. These groups include both the 7,587 $(1:1)^n$ simple ortholog sets and orthologous groups with duplications and deletions. As a consequence of using NJTree, $(1:1)^n$ simple orthology gene trees almost always correspond to the species phylogeny.

Rate estimation was performed with PAML[23] using the Goldman & Yang (1994) model[24]. In all cases, codon frequencies were estimated from the nucleotide composition at each codon position (F3X4 model). The parameters estimated were the transition/transversion ratio (κ), synonymous substitution rate (d_S), non-synonymous substitution rate (d_N) and the ratio of synonymous to non-synonymous substitutions (ω). Rates were not allowed to vary across sites. For each tree we computed the minimum, maximum, median and average distance to root. Trees vary in their height. The variance can be explained by longer branches leading to chicken and mouse (Accessory Fig. A12). The increased synonymous substitution rate in mouse has been observed previously, while the elongated branches in chicken reflect ambiguity in placing the root.

In all analyses, the multiple alignments were masked and filtered. The translated sequences were masked using seg. All codons with a single masked residue were discarded. Then, poorly aligned columns were removed by filtering using GBlocks with default options[25], but permitting half of all character in a column to be gaps.

Sequences and accession codes for platypus gene predictions discussed in the manuscript are found in the accessory file named Supporting_Sequences_and _codes.doc. Ensembl codes correspond to the January 2007 gene build.

Nucleotide substitution rates between orthologs S17. Previously, when comparing genome pairs, we have reported genome-wide d_S and d_N/d_S values as medians over all orthologs. When considering multiple species rate estimation is more complicated, as estimates can be made from multiple branches and tree topologies can vary (Accessory Fig. A12).

We computed genome-wide *dS, dN* and *dN/dS* using four methods (Accessory Table A12). Method 1 provides the median values from pairwise comparisons of all 1:1 orthologs. Method 2 provides the median values from pairwise comparisons of 1:1 orthologues drawn only from the simple $(1:1)^n$ ortholog set. Method 3 provides the median values from rate estimates inferred across the species phylogeny for each simple $(1:1)^n$ orthologue set. Method 4 provides values inferred from 20 samples of 200 concatenated multiple alignments from the simple $(1:1)^n$ ortholog set.

Rates calculated using the four methods are fairly consistent in their rank (Accessory Fig. A13), although they do vary considerably between methods (Accessory Table A13). We note that estimated d_S rates from methods that fit data to the species phylogeny are consistently smaller than those from pairwise estimates, and estimated d_N rates are consistently higher. As a consequence, when d_N/d_S values are inferred across the species phylogeny these are typically lower than median values inferred between species pairs. There is no clear theoretical foundation for estimating a single genome-wide aggregate substitution rate or selection strength. Substitution rates vary with nucleotide composition, which itself changes between lineages. Choosing an aggregate value to represent the genome-wide substitution rate is thus an arbitrary choice. One possibility is to display the variation across all orthologous groups and report a median value (this is a consequence of the non-normal distribution of substitution rates across orthologs). Here, each gene's contribution is equal. This has been the traditional way of reporting genome wide estimates.

An alternative is to estimate rates for all orthologues simultaneously in a single model producing a maximum likelihood estimate. A full model would allow rate variation across sites, across branches in the tree, across genes and genomic background. Such models, though easy to formulate, are difficult to fit because of their parameter richness and thus are not in use. We therefore have applied a reduced model (Method 4) that models variation across branches, to a large data set, namely a concatenated multiple sequence alignment of 200 strict $(1:1)^n$ ortholog set alignments. The model averages over all coding positions in the genome such that each gene contributes to the estimate in a manner which is proportional to its length. By sampling repeatedly

(20 times) from the set of orthologous groups we obtain an estimate of variation in the data. In the manuscript, we report values for both Method 3 and Method 4 (Accessory Table A14; Supplementary Fig. 1). Both methods do not examine biases due to the model. A full Bayesian analysis to compute posterior probabilities is not feasible due to its excessive computational cost. Terminal lineage d_N/d_S values (Method 4) are least for *Mus musculus*, and most for *Homo sapiens*, which accurately reflect the known differences in their effective population sizes. In particular, differences in terminal lineage d_N/d_S values (Method 4) between the three groups *H sapiens* + *O. anatinus*, *M. musculus,* and *C. familiaris*+*M. domestica* + *G. gallus* are statistically significant ($P < 0.01$, $n = 20$, ANOVA).

Gene evolution S18. The platypus genome is unique among sequenced mammalian genomes in containing approximately 11 paralogs of xanthine dehydrogenase/oxidase (*XDH* gene). Of these paralogues, 4 have been placed on platypus chromosome X1, in conserved synteny with human *XDH* (*HSA2*). In addition to its housekeeping role in purine metabolism, eutherian *XDH* appears to have roles in the secretion of milk lipids[26]. As its protein level correlates with the maturation of mouse mammary tissue in pregnancy[27] the platypus *XDH* paralogs may assist in regulating milk lipid content during lactation.

The platypus uses electro- and mechano-reception, and not vision, for underwater foraging[28,29]. The molecular identity of these receptors remains unknown, and there is no evidence for a large expansion of, for example, a voltage-sensitive calcium channel gene family in the platypus genome. Two novel *TRPV6*-like transient receptor potential cation channel genes were observed in the platypus genome. However, these are not monotreme innovations, despite being absent from eutherian and marsupial genomes, as they are also present in the *Xenopus tropicalis* genome.

The monotreme eye is relatively ancestral in form: it is the eutherians whose eyes have exhibited significant changes[30]. Consequently, the platypus cones still contain oil-droplets, as do sauropsids. Consistent with this ancestral state, the platypus genome encodes several ancient genes, present in fish, which have been lost from eutherians. These include the shortwave-sensitive-2 (*SWS2*) opsin gene[30,31] and two genes of no known function that, on the basis of cDNA information, are expressed predominantly in the fish retina (Table 1). A fourth such gene (Table 1), a paralogue of *ATP6AP1*, is a likely vacuolar ATP synthase subunit which, when mutated in zebra fish, results in a "*bleached blond*" mutant, exhibiting pigmentation defects, particularly of the retinal epithelium[32,33].

To evaluate the calthelicidin antimicrobial peptide gene family relationships a neighbour-joining tree, with chicken sequences as an outgroup was constructed using MEGA3.1. The following amino acid sequences were used: Chicken1NP_001001605.1, chicken2 NP_001020001.1, human NP_004336.2, mouse NP_034051.1, monkey NP_001028681.1, chimp NP_001065283.1, dog NP_001003359.1, cattle1 NP_777250.1, cattle2 NP_777251.1, cattle3 NP_776426.1, cattle4 NP_777252.1, cattle5 NP_776935.1, cattle6 NP_777257.1, cattle7 NP_777256.1.

For the CD163 family we see 10 genes, as opposed to 1 or 2 in eutherians. *CD163*

plays a key role in endocytosis of hemoglobin-haptoglobin complexes in plasma[34]. The unusually high level of hemoglobin of platypus and its higher oxygen capacity have been interpreted as an adaptive response to the hypoxic conditions in their burrows[35], in which half their day and extended periods after hatching are spent. Taken together, therefore, it is possible that the increased repertoire of *CD163* molecules also reflects a monotreme adaptation to hypoxic conditions, either in burrows, underwater, or, as seen for echidnas, during hibernation of ancestral species[36].

Phylogenetic position of platypus S19. We applied three independent computational techniques to unravel the monotreme-marsupial-eutherian relationship: (1) maximum-likelihood-based phylogenetic reconstruction and the taxonomic distributions of both; (2) rare insertions or deletions in protein-coding regions and (3) insertions of interspersed repeats. Detailed reports follow for each of the three.

We extended the basic data-sampling approach described previously[37] to protein-coding genes. Genes annotated as orthologs in Ensembl release 43 for chicken, platypus, opossum, mouse, rat, dog, and human were extracted. A multiple sequence alignment of the DNA sequences was generated by translation into amino acids, alignment using MUSCLE[21], and reintroduction of gaps into the original DNA sequences. As the evolutionary models of sequence change we used represent point substitution processes only, all alignment columns containing gaps or ambiguous characters (such as N's) were excluded. Alignment segments of exactly 750 nucleotides were saved for later use in analyses. We selected 750 nucleotides as a trade-off between ensuring each alignment block had sufficient positions for reliable parameter estimation in the subsequent likelihood analyses and sufficient genomic regions were sampled to ensure no one region biased the results.

The primary phylogenetic question concerned whether analysis of substantial genomic data supported either the Theria (monotreme, marsupial, eutherian) or Marsupionta (monotreme, marsupial), eutherian tree topologies. We addressed these competing hypotheses, fitting each distinct topology to each alignment separately for each of a number of different models of substitution by maximizing the likelihood. As the eutherian sub-tree involving these species remains controversial, with support for a rodent-first, or carnivore-first topology being sensitive to the sampled lineages or types of molecular markers, for the current case we consider both possible sub-tree's. Likewise, because inference by the likelihood approach can be sensitive to model choice we considered a selection of models designed to differ in a manner previously demonstrated to affect phylogenetic inference. The set of substitution models considered were: an empirical amino acid substitution model[38]; a codon substitution model[39]; the general time reversible nucleotide substitution model[40]; and a purine/pyrimidine (RY) substitution model employed to address the violation of compositional stationarity[37]. Gamma distributed rate-heterogeneity variants for most of these models were also used where practical. All models were implemented using the PyEvolve package[41] with numerical optimizations conducted using PyEvolve's built-in Powell optimizer at default settings.

This phylogenetic inference using likelihood provided consistent support for the Theria hypothesis across all substitution models. The percentage of alignment

segments supporting the Theria topology ranged from 69-81% (RY + gamma and GTR substitution models respectively[37]. Both percentages are significantly greater than 50% (both $P < 10^{-15}$).

For protein-coding indels we searched the pairwise alignments of human with each of opossum, platypus and chicken, downloaded from the UCSC Browser. Our goal was to find protein-coding exons with the following properties: (1) in each of the three pairwise alignments, the exon is covered by a single local alignment, (2) there is no gap in the exon's alignment for one of the three non-human species, and (3) alignments to the other two non-human species have gaps of length divisible by three in precisely the same place. Notice that when chicken is taken as the first non-human species, then the indel argues for the Marsupionta hypothesis[42], according to which the eutherian lineage diverged from the common ancestor of the monotreme and marsupial lineages. When opossum is taken as the first species, indels support the Theria hypothesis. At these (not very restrictive) conditions, we found 84 examples that putatively support Theria, but only 26 that might argue for Marsupionta.

We then applied a manual-curation pipeline[43] to enforce a variety of additional criteria that help eliminate cases of homoplasy and/or non-orthologous matches. In particular, we looked for GenBank sequence data from other vertebrates that cover the exon in question. All of the compelling examples that we detected this way supported the Theria hypothesis. An alignment of 43 vertebrate sequences for an exon of the *PTPN4* gene (Table A15) is a representative example. The extreme conservation of the exon, the size of the indel (3 amino acids) and consequent low potential for homoplasic reversion, and absence of potential confusion from close paralogs or processed pseudogenes make this example compelling. Not surprisingly, orthologous sequence could not be recovered from a few low-coverage genome projects — as of 29 July 2007 these consisted of *Otolemur garnettii* (bushbaby), *Cynocephalus volans* (flying lemur), *Erinaceus europaeus* (hedgehog), *Sorex araneus* (shrew), and *Loxodonta africana* (elephant). We predict that all five of these species will exhibit the same three-residue deletion. Conversely, though echidna sequences are not currently available, as sister group to platypus, we predict *Tachyglossidae* will not contain the deletion (in the absence of lineage sorting).

Our third approach consisted of a three-directional screening for retrotransposon presence and absence[44,45] to resolve the root of the mammalian tree. Three different evolutionary relationships are theoretically possible (1) Theria (placentals + marsupials) to the exclusion of monotremes, (2) Marsupionta (marsupials + monotremes) to the exclusion of placentals, or the hypothetical relationship of (3) monotremes + placentals to the exclusion of marsupials. Despite the very ancient divergence points, from ~90,000 inspected mammalian loci we identified the presence of three retroposed elements in both human and opossum that were clearly absent in platypus, providing significant support for the Therian hypothesis supported by sequence-based reconstructions. Not a single element was found to support the other potential relationships. Interestingly all three retrotransposed markers are MIR3 elements, which were thought to have been extinct before the mammalian divergence. The presence/absence data support their continuous activity in the common ancestor of therians.

To find informative retrotransposed elements and to test the above hypotheses, we applied a novel strategy screening two, three-way whole-genome alignments of 1,476 and 1,570Gb each, respectively, in MAF format (UCSC - University of California Santa Cruz; human-opossum-platypus; opossum-platypus-human). The alignments were downloaded from the UCSC web server (http://hgdownload.cse.ucsc.edu/downloads.html).

Using a novel C-language script, we extracted from each MAF-file all sets of three continuous blocks of alignments that contained an embedded two-species alignment (block 2). In all, we extracted 89,170 such triple blocks. The embedded two-species blocks were size-restricted from 75 to 2000 nt, based on the coordinates of the MAF-alignment (with respect to the nt-position on a given chromosome), and all triple blocks were converted into continuous FASTA sequences using another C-language script. The FASTA sequences were screened for non-lineage-specific SINEs, LINEs, and LTRs with the local version of RepeatMasker (www.reapeatmasker.org) using the mammalian repeat library settings. From this output file we selected cases in which at least 50% of block 2 was composed of a recognizable retrotransposed sequence (human-opossum: 1091 blocks, human-platypus: 342 blocks and opossum-platypus: 1571 blocks). Links to the Genome Browser (http://genome.ucsc.edu/cgi-bin/hgBlat) were generated and used to filter out all loci with less than ~70% similarities among the alignments, as well as those with retrotransposed elements that overlapped the flanking regions of the neighboring blocks (blocks 1 and 3). The resulting data set contained 25 blocks of human-opossum, 16 blocks of human-platypus, and 23 blocks of opossum-platypus alignments. To obtain a more complete picture of each retrotransposon alignment, sequences from MAF alignments and from trace archives from additional species in the given lineages (eutherians: elephant and armadillo; marsupials: wallaby; and outgroup: chicken) were added to the remaining candidate loci (human-opossum: 25, human-platypus: 16 blocks and opossum-platypus: 23 blocks) and realigned by hand. The candidate loci were screened again for the non-lineage-specific SINEs, LINEs, and LTRs using RepeatMasker by aligning the entire retroelements taken from the Repbase (www.girinst.org/repbase) to the corresponding loci. Each of the insertion sites was inspected to determine their exact borders and validate their orthology. The non-biased screening presented in Accessory Fig. 14A was designed to find potential support for any of the three possible tree topologies shown in Accessory Fig. 14B. Three independent insertions of MIR3 SINEs were found in both human and opossum that were not present in platypus, providing significant support for the Theria hypothesis. No retroposed elements were found to support either of the other two possibilities. A section of a representative alignment is shown in Accessory Fig. 14C. The boxed area contains the MIR3 repeat region. Complete alignments are available in FASTA format.

Interspersed repeats S20. We used a modified version of the clustering software developed Price *et al*.[46] to analyze the LINE2 subfamily structure in the platypus genome. The software performs best on regions lacking indel differences between subfamilies. Therefore, we analyzed a 500 bp region at the 3' end of the second ORF2, which is both conserved between subfamilies, especially with respect to indels, and relatively abundant. However, most LINE2 copies are so 5' truncated that they don't extend into the ORF (the median length of LINE2 copies is a mere 131 bp). To reduce the bias of ascertainment towards younger copies and increase the number of old copies including this region, younger insertions were

clipped out of older LINE2 copies. Still, only 4025 of the 1.5 million LINE2 copies contained this region in full. The Supplementary Fig. 4 shows the result of the analysis varying the minimum number of copies in a subfamily (M). The size of the circle corresponds to the (binned) sizes of the subfamilies. Dark colored circles represent subfamilies distinguished in the first steps of the analysis by co-segregation of multiple diagnostic substitutions at different sites.

At least 8 LINE families were once active in parallel in the monotreme genome, most since before the speciation from therians, all but the LINE2 element has become extinct. Likewise, only LINE1 has survived in therians. As predicted[47], no LINE1 copies, not even ancient ones, are observed in platypus. Since we can find many copies of transposable elements active in the ancestor of therians and monotremes in the platypus (by itself suggestive of low neutral substitution and/or deletion rates compared to therians), perhaps LINE1 was introduced in the germline of a therian ancestor.

The software creates a phylogenetic tree based on the subfamily consensus sequences alone and it is unaware of the average substitution level of the subfamily copies. For each M, the tree is completely consistent with a linear progression from the oldest subfamilies on the left to the youngest on the right. Most remarkably, even when 70 subfamilies are created with M=10, about 90% of the copies belong to subfamilies that lie on the main branch, while consensus sequences for subfamilies on side branches do not diverge for more than 10% from those on the main branch.

Population structure analysis S21. Genomic locations containing the youngest retrotransposon subfamilies of Mon1 and LINE2 were selected for this phylogenetic and population genetic study. Each locus was extracted from the platypus genome assembly (OrnAna1 UCSC v 5.0.1) with 3' and 5' flanking sequence, and repeatmasked with a local installation of RepeatMasker using a custom library (RepeatMasker Open-3.0. at http://repeatmasker.org). Mon1 loci were randomly selected using two different approaches. In the first scenario, only the Mon1 subfamilies predicted to be the youngest were included and loci with less than 2% divergence from each consensus sequence were selected for further analysis. In the other approach, Mon1 sequences with 95% identity to one another were selected from all available platypus "target" data without identification of the subfamily. The Mon1 loci were selected from all chromosomes, contigs, and ultra sequence available. L2 loci of the youngest subfamily L2_plat1a were randomly selected from the chromosomes 1, 10, 11, and 12.

Primers were designed with a locally installed version of Primer3[48]. Each primer was checked with BLAT-The BLAST-Like Alignment Tool[49] and a virtual PCR (http://genome.ucsc.edu) was performed for each primer combination in order to investigate if each primer combination would likely result in a single PCR product. The sequences of each primer, the PCR sizes of filled and empty amplicons, the annealing temperature for each primer combination, and the genomic location of each locus are provided at this web address: (http://batzerlab.lsu.edu).
PCR amplifications were performed in 25 μl reactions containing 10 ng of template DNA (platypus, echidna, possum, wallaby or human); 200 nM of each oligonucleotide primer; 1.5 mM $MgCl_2$; 10 X PCR buffer (50mM KCl; 10mM Tris-HCl, pH 8.4); 0.2 mM dNTPs; and 1.25 U *Taq* DNA polymerase. Each PCR reaction

was performed under the following conditions: An initial denaturation at 94 °C for 60 seconds (s) was followed by 32 cycles of denaturation at 94 °C for 15 to 30 s, 30 s at annealing temperature (57 °C), and extension at 72 °C for 30 s. The PCR was concluded by a final extension step at 72 °C for 2 min. 20 µl of each PCR product were fractionated in a horizontal gel chamber on a 2% agarose gel containing ~ 0.1 µg/ml ethidium bromide for 50-60 min at 175-200 V. The DNA fragments were visualized with UV-fluorescence.

To investigate the population structure of 90 platypuses from different regions in Australia including Tasmania a Structure analysis was performed using Structure software v2.1[50](Accessory Fig. A15). This software package performs model-based clustering using genotypic data from unlinked markers to infer population structure. For this analysis all 57 polymorphic retrotransposon loci were included but information about the origin of the samples was omitted. Initially, K (number of population clusters) was set from 1 to 6 to allow the software to determine the most likely value of K clusters. The initial burn-in period was set at 10,000 iterations and followed by a run of 10,000 replications. For the highest likelihood of K (here 4) a run with 25 replications was used. All runs were performed on a desktop computer.

Microsatellites analyses S22. Microsatellites were identified across the platypus genome (ornAna1) combining two programs: Tandem Repeat Finder (TRF)[15] and Sputnik[51] (http://wheat.pw.usda.gov/ITMI/EST-SSR/LaRota/). The assembled chromosomes and the unassembled contigs and ultracontigs were analyzed separately. Microsatellites are usually defined as perfect repetitions of 1-6 nucleotide motifs, but can also be interrupted by point mutations or possess a mixture of different motifs within the same locus. For our analyses the minimum length for a microsatellite was set to fifteen nucleotides and independent searches were run with stringent and relaxed parameters to find perfect and imperfect microsatellites, respectively. The parameters for each programme were as follows. TRF: perfect repeats (2, 7, 7, 80, 10, 30, 6); imperfect repeats (2, 3, 5, 80, 10, 30, 6). Sputnik: perfect repeats (-v 1 -u 5 -s 11 -p -r 0 -L 15 -l -1); Long imperfect repeats (-v 4 -u 5 -m 2 -n -6 -s 24 -A -p -L 16 -l -1); Short imperfect repeats (-v 1 -u 3 -m 2 -n -6 -s 16 -A -p -L 16 -l -1). Under the chosen parameters, using both Sputnik and TRF, improves search sensitivity by an additional 10% over searches employing TRF alone (Vargas et al., unpublished).

Inter- and intra-genomic comparisons were carried out with microsatellite datasets for human (hg18), dog (canFam2), mouse (mm8), opossum (monDom4), chicken (galGal3) and lizard (anoCar1) genomes (obtained from the UCSC Genome Browser) using the parameters described above. To obtain a thorough but non-redundant estimate of microsatellite density the results for perfect and imperfect microsatellites were merged and analyzed using Visual Basic and Java scripts. Density, shown as percentage coverage, was calculated as the total length of microsatellites for each 10kb non-overlapping window of genomic sequence. Only windows uninterrupted by gaps were analyzed. The density plots from these analyses are available on request.

The characterisation of microsatellites in terms of array length, AT-content, motif and size class preference were performed using just TRF data for all genomes analysed but for dog. The set was filtered for redundant microsatellites (i.e. those occupying the same genomic position) retaining the longest array and for microsatellites with more

than three repeats using Visual Basic scripts. The subsequent were done using R, version 2.2.1 (http://www.R-project.org).

Platypus' mean microsatellite length is 22.31nt for perfect repeats and 31.42nt for imperfect repeats; the shortest mean array length within the compared genomes (Supplementary Table 9). Mono- and dinucleotide repeats common to the other mammalian genomes are rare in the platypus genome (Supplementary Fig. 8). Microsatellite frequencies were 389.25 loci/Mbp for perfect repeats and 719.412 loci/Mbp for imperfect microsatellites, putting platypus in third place for the lowest microsatellite frequency after chicken and opossum (Supplementary Table 9). This likely reflects a genuine composition difference among the genomes, potentially strengthening the similarities between platypus and the non-mammalian genomes, but may be due to difficulties sequencing this fraction of the platypus genome. Analyses of motif preference also support this relationship. The three most common motifs in platypus are ATT (12.9%), TAA (7.6%) and TGAA (6.6%); highly similar to the motif usage in lizard.

Differences in microsatellite content observed among genomes are not due to genome size differences, because genome size (nucleotides of available sequence data) does not correlate with microsatellite coverage. At a chromosome specific level the distribution of microsatellites is homogeneous with no significant relationship between microsatellite coverage and chromosome size observed in all species (data not shown).

The platypus G+C nucleotide composition (45.5%) is higher than that found in other mammals (e.g. human, 40.7%). However, microsatellite nucleotide composition differs from the overall genomic values in all genomes analysed, with microsatellites having a higher A+T content in both perfect and imperfect repeats and G+C microsatellites exceedingly rare (Supplementary Figure 9). The analysis of microsatellite abundance by A+T content identifies four major peaks, corresponding to approximately 0, 50, 70 and 90% A+T. In comparison to other genomes, platypus has fewer microsatellites with ~50% A+T and more with ~70% A+T, leading to an abundance distribution that has more in common with chicken and lizard than with mammals.

Microsatellite sequences are predominantly non-coding, evolve neutrally, and are generally considered to be highly liable in an evolutionary sense[52], with most loci conserved only among closely related species (e.g. human-chimpanzee). Our null expectation therefore was that very few, if any, microsatellites would be conserved across the evolutionary timescale separating the monotremes from the other mammals. However, we found that of 352,034 platypus microsatellites identified in the whole genome alignment, the percentage of these loci conserved in other species was 0.77% in lizard, 1.19% in chicken, 1.81% in mouse, 1.85% in human and 2.55% in opossum.

Microsatellites were searched for in ungapped sequences extracted from the multiple alignment of the platypus genome (ornAna1) against those of lizard (anoCar1), chicken (galGal3), human (hg18), mouse (mm8) and opossum (monDom4) available at the UCSC Genome Browser[53]. FASTA-formatted sequences were extracted using Galaxy[54] gaps were removed using the module degapseq from the EMBOSS 5.0

package[55], and perfect and imperfect microsatellites were searched using SciRoKo 3.1 with fixed penalty parameters: 12, 4, 3, 3, 3[56]. Duplicated microsatellites and microsatellites that had any overlap with repeats other than simple and low-complexity repeats were discarded. Genomic positions of non-platypus microsatellites were converted to the ornAna1 platypus genome assembly using the LiftOver utility and chain files available at the UCSC Genome Browser. The fraction of platypus microsatellite positions that overlapped with any of the converted microsatellite positions indicated conserved sites. The genomic locations of conserved microsatellites were determined using the Ensembl gene annotation, using an overlap threshold of 0.50001% applied to avoid any duplicated results.

Over 95% of the conserved microsatellites identified are in non-coding or unannotated sequences (74.7% in IGRs and 20.7% in introns). This value compares favourably with the distribution of similar elements in the human genome, in which 94% of conserved microsatellites were found in non-coding regions (52% in IGRS, 42% in introns). Of the 919 conserved microsatellites localized to coding regions, the majority, 779, are trinucleotide repeats, consistent with the heightened abundance of this class of microsatellite in the coding regions of other genomes[57].

Most platypus microsatellites are conserved in one species, with decreasing numbers of loci conserved as the number of species increases (Supplementary Fig. 10). As expected, more platypus microsatellite loci are conserved in mammals than chicken and lizard. Curiously, more platypus microsatellites are conserved in opossum than the other mammals, suggesting a closer relationship between monotremes and marsupials. However, the result may be influenced by the incomplete nature of the whole-genome alignments, with the current alignment representing ~34% of the assembled genome.

High heterogeneity in microsatellite conservation exists among chromosomes, which is unexplained by microsatellite coverage, with the exception of chromosome 17. Chromosomes 6, 7, X1–3, and the grouped contigs and ultracontigs have 5% or fewer conserved microsatellites, while chromosomes 11 and 17 show high conservation of microsatellites, with ~25% of loci showing conservation among the species examined. While the extent of microsatellite conservation per chromosome is as yet not fully understood in relation to the features each chromosome possesses, it is noteworthy that platypus chromosome 6 is homologous to human X chromosome, which also has fewer conserved microsatellite loci than expected (Buschiazzo *et al*., unpublished), suggesting that microsatellite conservation may be useful for examining chromosomal conservation and potentially synteny.

Contrary to expectations that few, if any, microsatellites would be conserved across the evolutionary timescale separating the monotremes from the other vertebrates we found that the percentage of these microsatellite loci conserved in other species ranged from 0.77% in lizard to 2.55% in opossum, with fewer loci conserved as the number of species increases (Supplementary Fig. 10). Curiously, more platypus microsatellites are conserved in opossum than the other mammals, suggesting a closer relationship between monotremes and marsupials, at least for this segment of the genome architecture.

G+C fraction in various mammalian species S23. Supplementary Fig. 12 shows the distribution of this fraction. Preliminary data suggests that chromosomes with a high G+C fraction tend to be short (Supplementary Fig. 13. However, some platypus chromosomes currently have very little sequence data assigned to them, so the current analysis cannot be considered definitive.

CpGs at promoters and other regulatory elements S24. In addition to papers cited in the main paper, the following data sources were used for Supplementary Fig. 14: CTCF binding sites[58], PRPs[59] and 93 known regulatory regions[60]. These putatively functional regions are compared to the non-coding, non-repetitive regions of the genome, denoted NCNR in Supplementary Fig.14.

In eutherians, a high local concentration of CpG dinucleotides often coincides with a gene promoter, i.e., a start site of transcription[61]. We investigated whether this phenomenon can be observed in genomes with dramatically different G+C fractions. Putative promoters and regulatory regions in human were mapped to mouse, opossum[10], and platypus using whole-genome alignments, treating the homologous sequences as likely promoters and regulatory regions. Only regions that could be mapped to all three species were considered for further analysis, and CpG fractions were calculated for the human promoters and the regions to which they mapped in companion species (Supplementary Tables 10,11). Promoters predicted by the presence of at least 100 CAGE tags were separated into four classes as in Carninci *et al.* [62]. The classes are SP (sharp peak), PB (broad but with dominant peak), MU (multimodal peaks), and BR (broad with no dominant peaks). While CpG fractions of all classes of CAGE promoters are considerably higher than that of the bulk genome (NCNR=non-coding, non-repetitive) for all species examined (Supplementary Fig. 14), it was higher in human than in platypus, despite the elevated G+C content for the platypus genome. Interestingly, for the other putative regulatory region25s pictured in Supplementary Fig. 14, the homologous DNA in platypus does tend to show the expected increase in CpG content relative to human, but the CpG fraction is much lower than for the CAGE promoters. One possible explanation for these observations is that the CpG content for promoters has a maximum. Perhaps the thermodynamics of strand separation at initiation become unfavorable at higher G+C and CpG content, or the bias in base composition precludes the existence of needed transcription-factor binding sites.

We observed that CpG fractions for all classes of CAGE promoters are higher in human than their putative orthologs in platypus, where we considered only the CAGE promoters that we could map (via multi-species alignments downloaded from the UCSC Browser) to all of mouse, opossum and platypus. One possible explanation for our observation may be occasional failure of our implicit assumption that human promoters and regulatory regions map to platypus regions with the same function. Interestingly, the strongest evidence that this effect is not dominating our results is provided by data for the marsupial *Monodelphis domestica*. There, the CpG fractions of some classes of promoters are 10 times higher than the genome average, which would be unlikely to happen if many human promoters were mapped to non-functional DNA. In platypus, the difference in CpG fraction between mapped promoters and background averages about four-fold. Data of Frith *et al.*[63] suggest that for 20% of human promoters the orthologous mouse interval is not a promoter, and

the rate of promoter turnover between human and Monodelphis or platypus could well be higher. However, the analysis described here should ignore most of those cases, because the region orthologous to the human promoter is evolving neutrally and has accumulated too many mutations to align over a large evolutionary distance.

Based on these data, we speculated that G+C and/or CpG fractions of promoters might have practical limits, which have been attained in human and cannot be exceeded even in mammalian genomes with higher overall G+C fraction (i.e., platypus). Since our observation was based on the average fraction for each tested class of promoters and regulatory elements, we investigated the high end of the distribution of G+C and CpG fractions. Results are given in Supplementary Tables 10 and 11. In the majority of cases, the human fraction at the 95-th percentile is higher than the platypus fraction at the 95-th percentile, again consistent with the existence of maximum G+C and CpG fractions in mammalian promoters.

2. Supplementary Tables

Table 1. Sequence coverage of the platypus genome

Insert size (kb)	Read number (thousand)			Read bases (million)		Sequence Coverage	Physical Coverage
	Input	Assembled	Paired	Assembled	Paired	Assembled	Paired
4	25894	23881	20967	16395	14606	6.09	12.17
40	653	494	311	304	196	0.08	1.63
150	408	380	297	270	214	0.09	6.69
Total	26954	24754	21576	16969	15016	6.26	20.49

Table 2. Whole genome assembly statistics

Assembly Feature	>2kb number	N50 length (kb)	N50 number	Largest (kb)
Contigs	177,028	12	39,589	246
Supercontigs	61,239	967	298	14,341

Contigs are contiguous sequences not interrupted by gaps, and supercontigs are ordered and oriented contigs including estimated gap sizes. The N50 statistic is defined as the largest length L such that 50% of all nucleotides are contained in contigs of size at least L. A total of 24,754,112 reads were included in the final assembly; eight percent of the total sequencing reads presented to the assembler were not used in the final assembly. The final integrated assembly was composed of 205,534 supercontigs (contigs ordered and oriented by read-pairing data); of those, 4,197 supercontigs were organized into 689 ultracontigs.

Table 3. Grouping of experimentally- (Oa) and BLAST- (bOa) identified non-protein coding (npc) RNAs in platypus. (I) Vertebrate-wide conserved npcRNAs, (II) conserved in mammals, and (III) Platypus-specific npcRNAs. snoRNAs were categorized as either C/D-box or H/ACA-box snoRNAs and were found in intronic, intergenic or unidentified (not analyzed - na) regions. Owing to the large number of spliceosomal RNA paralogs, we were not able to identify the true orthologs of the respective U1–U13 snRNAs in platypus. RPG, ribosomal protein genes; U1–U6, U7-8, U11-13, spliceosomal RNAs; snoRTE HACA, HACAs distributed by retroposition; Others, additional, known npcRNAs. The U1-13 and "Others" groups were experimentally detected but not further analyzed.

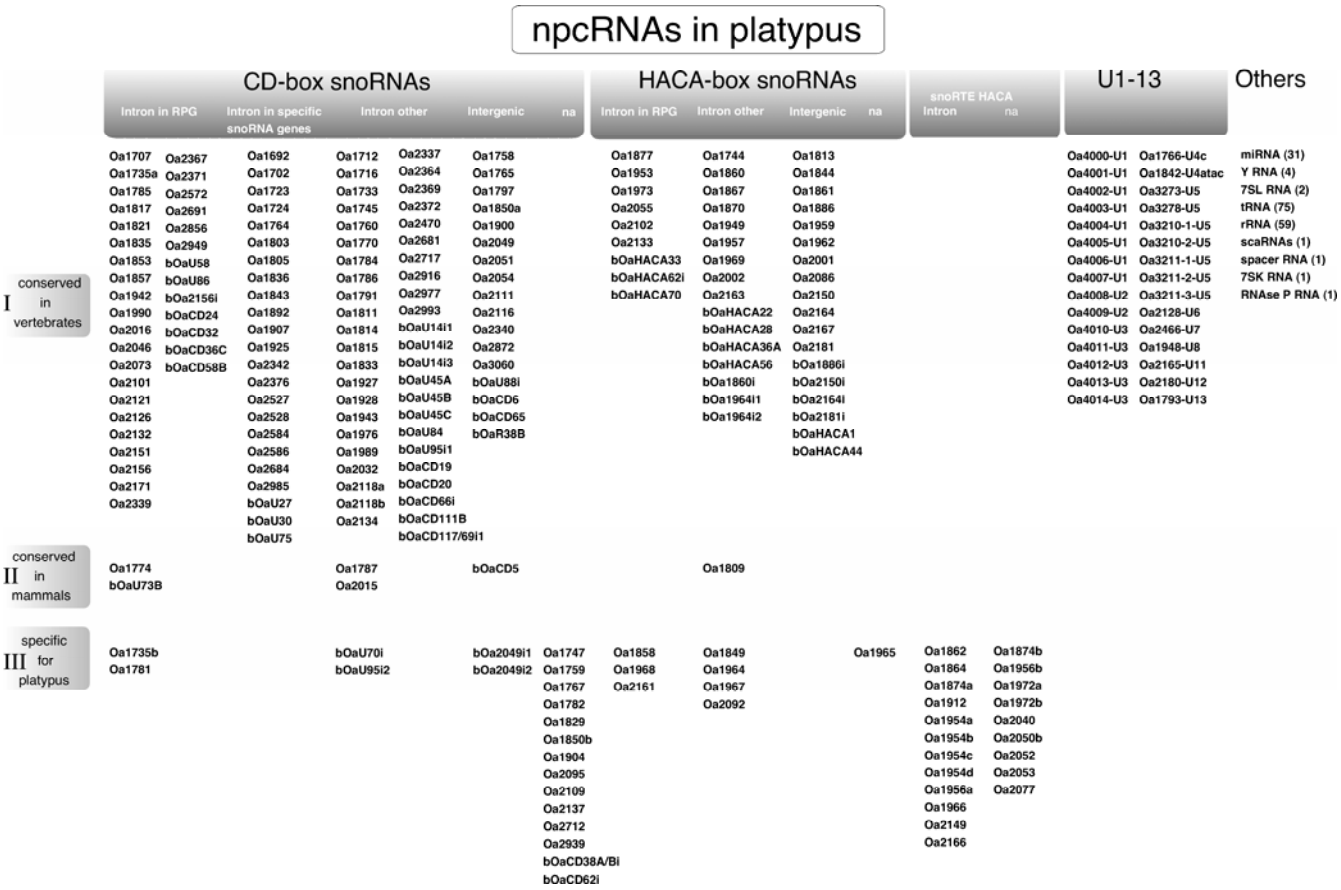


Table 4. Gene Prediction Identification numbers for genes involved in the RNA interference pathway in platypus.

Gene Name	Representative Ensembl Gene ID
Dicer	ENSOANT00000019065
Drosha	ENSOANT00000024819
Argonaute1	ENSOANT00000003300
Argonaute2	ENSOANT00000011337
Argonaute3	ENSOANT00000003299
Argonaute4	ENSOANT00000003308
PiwiL1	ENSOANT00000018940
PiwiL2	ENSOANT00000017089

Table 5: Recall of genes after orthology assignment. The human gene set does not include mitochondrial genes.

Species	Genes	Orthologs		Orphans	
H. sapiens	22,611	19,339	86%	3,272	14%
M. musculus	24,442	20,758	85%	3,684	15%
C. familiaris	19,314	18,066	94%	1,248	6%
M. domestica	19,597	18,123	92%	1,474	8%
O. anatinus	18,596	15,312	82%	3,284	18%
G. gallus	16,715	13,893	83%	2,822	17%

Table 6. Significant over-representations of Gene Ontology (GO) annotations for simple 1:1 orthologues that have been conserved without duplication, deletion or non-functionalisation since the common ancestor of five mammalian species.

Biological process Accession	GO Term	Fold	P-Value	P-Value
GO:0009790	embryonic development	1.37	2.2 x 10-10	2.17E-10
GO:0009653	morphogenesis	1.31	5.9 x 10-22	5.86E-22
GO:0016043	cell organization and biogenesis	1.27	1.9 x 10-17	1.86E-17
GO:0006519	amino acid and derivative metabolism	1.25	4.2 x 10-6	4.22E-06
GO:0006464	protein modification	1.25	6.8 x 10-24	6.78E-24
GO:0015031	protein transport	1.24	1.6 x 10-10	1.58E-10
GO:0007010	cytoskeleton organization and biogenesis	1.22	8.1 x 10-7	8.14E-07
GO:0007275	development	1.21	8.7 x 10-18	8.73E-18
GO:0030154	cell differentiation	1.21	2.4 x 10-11	2.35E-11
GO:0007049	cell cycle	1.2	1.1 x 10-8	1.09E-08
GO:0006810	transport	1.18	8.9 x 10-13	8.87E-13
GO:0009058	biosynthesis	1.18	9.2 x 10-7	9.23E-07
GO:0006811	ion transport	1.18	1.7 x 10-7	1.71E-07
GO:0008283	cell proliferation	1.15	2.8 x 10-5	2.78E-05
GO:0007267	cell-cell signaling	1.15	4.6 x 10-5	4.63E-05
GO:0009056	catabolism	1.13	6.8 x 10-4	6.84E-04
GO:0006139	nucleobase, nucleoside, nucleotide and nucleic acid metabolism	1.13	6.2 x 10-5	6.24E-05
GO:0006629	lipid metabolism	1.12	6.4 x 10-4	6.38E-04

Molecular function Accession	GO Term	Fold	P-Value	P-Value
GO:0008092	cytoskeletal protein binding	1.42	2.2 x 10-7	2.22E-07
GO:0004672	protein kinase activity	1.31	9.2 x 10-15	9.21E-15
GO:0003779	actin binding	1.3	7 x 10-7	7.01E-07
GO:0005216	ion channel activity	1.25	8 x 10-7	8.04E-07
GO:0016301	kinase activity	1.24	1.6 x 10-5	1.56E-05
GO:0016740	transferase activity	1.2	4.6 x 10-19	4.58E-19
GO:0004871	signal transducer activity	1.19	1.6 x 10-5	1.55E-05

GO:0005515	protein binding	1.17	3.6 x 10-48	3.64E-48
GO:0005509	calcium ion binding	1.17	1.4 x 10-7	1.37E-07
GO:0000166	nucleotide binding	1.16	7.7 x 10-16	7.72E-16
GO:0030528	transcription regulator activity	1.15	1.8 x 10-5	1.76E-05
GO:0016787	hydrolase activity	1.13	8 x 10-8	7.95E-08
GO:0030234	enzyme regulator activity	1.13	1.6 x 10-4	1.55E-04
GO:0005215	transporter activity	1.12	8.9 x 10-5	8.86E-05
GO:0003824	catalytic activity	1.1	1.4 x 10-5	1.37E-05

Location

Accession	GO Term	Fold	P-Value	P-Value
GO:0005815	microtubule organizing center	1.38	4 x 10-4	4.03E-04
GO:0005768	endosome	1.34	3.1 x 10-4	3.13E-04
GO:0005794	Golgi apparatus	1.33	4.1 x 10-10	4.13E-10
GO:0005635	nuclear envelope	1.31	3 x 10-4	2.99E-04
GO:0005578	extracellular matrix (sensu Metazoa)	1.31	5.5 x 10-8	5.50E-08
GO:0005654	nucleoplasm	1.26	1.2 x 10-7	1.15E-07
GO:0016023	cytoplasmic membrane-bound vesicle	1.25	1.1 x 10-4	1.09E-04
GO:0005737	cytoplasm	1.18	1.3 x 10-16	1.28E-16
GO:0005783	endoplasmic reticulum	1.15	3.8 x 10-5	3.81E-05
GO:0005856	cytoskeleton	1.13	1.5 x 10-5	1.51E-05
GO:0005886	plasma membrane	1.12	3.2 x 10-9	3.20E-09
GO:0005615	extracellular space	1.11	1.5 x 10-6	1.54E-06
GO:0043234	protein complex	1.11	2.6 x 10-7	2.56E-07
GO:0005739	mitochondrion	1.11	6.9 x 10-4	6.92E-04

Table 7. Number of copies and fraction of genome for interspersed repeats.

(RepeatMasker library version 3.1.8)

	Number of copies (x1000)	Total number of bases in the draft genome sequence (Mb)	Fraction of the draft genome sequence (%)	Number of families (subfamilies)
SINEs				
	2275.10	414.95	22.43%	
Mon1	2145.38	394.04	21.30%	20
RTE-SINE	52.88	12.12	0.66%	2
MIR	35.26	3.09	0.17%	3
MIR3	2.56	0.24	0.01%	1
Other	39.02	5.44	0.29%	4
LINEs				
	2050.23	389.17	21.04%	
L1	0.06	0.01	0.00%	6
L2	1910.97	360.23	19.47%	21
L3	8.04	0.80	0.04%	2
CR1	43.76	8.56	0.46%	4
RTE	85.69	19.17	1.04%	3
Dong-R4	1.70	0.41	0.02%	1
LTR elements				
	5.79	2.72	0.15%	
ERV-class I	1.95	0.91	0.05%	11
ERV(K)-class II	0.41	0.32	0.02%	5
ERV(L)-class III	0.02	>0.01	>0.01%	4
Gypsy	0.04	>0.01	>0.01%	8
Other	3.38	1.49	0.08%	3
DNA elements				
	58.13	10.27	0.56%	
Tigger	49.76	8.91	0.48%	15
hAT	7.25	1.05	0.06%	1
Charlie	0.42	0.07	>0.01%	4
AcHobo	0.40	0.22	0.01%	2
MER1	0.30	0.02	>0.01%	7
Tip100	0.01	>0.01	>0.01%	4
Unclassified				
	63.83	8.235	0.45%	3

Table 8. Platypus RepeatMasker output statistics (library version 3.1.8)

Repeat Name	Family	Number
AmnSINE1	SINE	1975
BovB_Plat	LINE/RTE-BovB	9048
CR1_Mam	LINE/CR1	283
L1M5	LINE/L1	22
L1M6	LINE/L1	7
L1M7	LINE/L1	2
L1ME5	LINE/L1	16
L1MEf	LINE/L1	5
L1MEg	LINE/L1	6
L2	LINE/L2	1410
L2_Plat1a	LINE/L2	23022
L2_Plat1b	LINE/L2	28748
L2_Plat1c	LINE/L2	11445
L2_Plat1d	LINE/L2	15113
L2_Plat1e	LINE/L2	37643
L2_Plat1f	LINE/L2	31941
L2_Plat1g	LINE/L2	57870
L2_Plat1h	LINE/L2	76955
L2_Plat1i	LINE/L2	190532
L2_Plat1m	LINE/L2	566041
L2_Plat1n	LINE/L2	136675
L2_Plat1o	LINE/L2	131783
L2_Plat1q	LINE/L2	64066
L2_Plat1r	LINE/L2	57143
L2_Plat1s	LINE/L2	57052
L2_Plat1t	LINE/L2	39841
L2_Plat1u	LINE/L2	52015
L2a	LINE/L2	3296

Repeat Name	Family	Number
MIRb	SINE/MIR	6131
MIRc	SINE/MIR	17385
MamSINE1	SINE/tRNA	532
Mon1a1	SINE/MIR	139078
Mon1a2	SINE/MIR	142139
Mon1a3	SINE/MIR	25993
Mon1a4	SINE/MIR	56169
Mon1a5	SINE/MIR	62377
Mon1a6	SINE/MIR	60012
Mon1a7	SINE/MIR	72178
Mon1d	SINE/MIR	198047
Mon1e	SINE/MIR	70374
Mon1f-1	SINE/MIR	28983
Mon1f-2	SINE/MIR	19592
Mon1f0	SINE/MIR	4433
Mon1f1	SINE/MIR	4402
Mon1f2	SINE/MIR	44356
Mon1f3	SINE/MIR	215639
Mon1f4	SINE/MIR	64475
Mon1f5	SINE/MIR	277811
Mon1g0	SINE/MIR	60692
Mon1g1	SINE/MIR	327919
Mon1g3	SINE/MIR	150701
MonoRep87A	SINE	15458
MonoRep87B	SINE	13715
PlatCR1	LINE/CR1	28173
PlatCR1_old1	LINE/CR1	2090
PlatCR1_old2	LINE/CR1	2045

L2b	LINE/L2	4634
L2c	LINE/L2	8421
L3	LINE/CR1	2674
L3b	LINE/CR1	5066
L4	LINE/RTE	903
MIR3	SINE/MIR	2521

PlatSINE1	SINE/CORE	38919
Plat_L3	LINE/CR1	14028
Plat_L3b	LINE/CR1	13843
Plat_R4	LINE/Dong-R4	1119
Plat_RTE1	LINE/RTE-BovB	50533
Plat_RTE1_SINE	SINE	8673

TinT analysis was restricted to platypus specific non-LTR retroposons
(all gray shaded elements were omitted). For additional analyses see Accessory Fig. A16.

Table 9. Number, length, density and frequency of perfect and imperfect microsatellites in representative vertebrate genomes

Perfect Loci

	Total count	Mean length (nt)	Density (nt/Mbp)	Frequency (loci/Mbp)	Genome size (Mbp)
Platypus	159471	22.31	8685.54	389.25	409.691489
Opossum	1105718	44.68	14109.82	315.77	3501.643220
Mouse	1654245	40.87	26478.73	647.92	2553.156572
Human	1366262	29.55	14128.43	478.044	2858.023193
Chicken	289384	26.10	2946.09	293.832	984.860953
Lizard (incl. redmsats)	825796	30.62	14520.39	474.19	1741.478929

Imperfect Loci

	Total count	Mean length	Density (nt/Mbp)	Frequency (loci/Mbp)	Genome size (Mbp)
Platypus	294737	31.43	22608.14	719.41	409.691489
Opossum	2165023	51.01	31541.29	618.29	3501.643220
Mouse	2243790	51.43	45197.77	878.83	2553.156572
Human	2073146	38.62	28013.06	725.38	2858.023193
Chicken	515066	33.92	6842.90	522.98	984.860953
Lizard (incl. redmsats)	1327057	42.13	32100.83	762.03	1741.478929

Table 10. Distribution of CpG content by feature class / species.

Feature	Species	Min	5%	25%	median	75%	95%	Max	p(CpG)*	p(dCpG)**
cage_promoters_BR	hg18	0	0.0417	0.0874	0.1181	0.1481	0.1782	0.2432	1.88E-242	N/A
cage_promoters_BR	mm8	0	0.0250	0.0759	0.1082	0.1395	0.1780	0.2405	7.74E-207	1.27E-11
cage_promoters_BR	monDom4	0	0.0158	0.0542	0.0861	0.1135	0.1576	0.2237	3.63E-177	6.50E-40
cage_promoters_BR	ornAna1	0	0.0198	0.0684	0.1000	0.1358	0.1818	0.4000	1.71E-154	2.65E-22
cage_promoters_MU	hg18	0	0.0219	0.0832	0.1261	0.1553	0.1918	0.2255	1.74E-199	N/A
cage_promoters_MU	mm8	0	0.0162	0.0736	0.1172	0.1474	0.1857	0.2245	9.76E-186	3.20E-10
cage_promoters_MU	monDom4	0	0.0110	0.0459	0.0767	0.1071	0.1487	0.2073	9.21E-158	1.30E-59
cage_promoters_MU	ornAna1	0	0.0176	0.0562	0.0940	0.1261	0.1654	0.2133	1.04E-136	1.17E-44
cage_promoters_PB	hg18	0	0.0280	0.0868	0.1215	0.1481	0.1858	0.2211	3.27E-178	N/A
cage_promoters_PB	mm8	0	0.0162	0.0741	0.1111	0.1429	0.1795	0.2234	2.25E-162	2.20E-08
cage_promoters_PB	monDom4	0	0.0112	0.0499	0.0804	0.1087	0.1525	0.2273	8.46E-138	3.33E-38
cage_promoters_PB	ornAna1	0	0.0163	0.0605	0.0951	0.1228	0.1768	1.0000	6.34E-86	1.19E-20
cage_promoters_SP	hg18	0	0.0000	0.0538	0.1053	0.1522	0.2174	0.2917	5.51E-141	N/A
cage_promoters_SP	mm8	0	0.0000	0.0427	0.0874	0.1311	0.1949	0.4000	3.12E-118	1.16E-14
cage_promoters_SP	monDom4	0	0.0000	0.0239	0.0599	0.1000	0.1571	0.2973	2.51E-101	7.04E-40
cage_promoters_SP	ornAna1	0	0.0000	0.0370	0.0744	0.1186	0.1766	0.2692	3.09E-86	4.99E-35
CTCF	hg18	0	0.0050	0.0115	0.0185	0.0315	0.0682	0.1626	4.97E-182	N/A
CTCF	mm8	0	0.0040	0.0098	0.0155	0.0252	0.0573	0.1956	9.02E-84	2.67E-42
CTCF	monDom4	0	0.0000	0.0043	0.0096	0.0235	0.0703	0.1631	1.25E-90	5.13E-11
CTCF	ornAna1	0	0.0000	0.0149	0.0309	0.0562	0.1090	0.2857	5.39E-106	1.10E-07
PRPv2	hg18	0	0.0000	0.0047	0.0098	0.0185	0.0722	0.2237	1.70E-183	N/A
PRPv2	mm8	0	0.0000	0.0065	0.0123	0.0209	0.0583	0.2407	7.22E-204	0.6343554
PRPv2	monDom4	0	0.0000	0.0025	0.0068	0.0143	0.0473	0.5000	2.01E-129	5.50E-22
PRPv2	ornAna1	0	0.0000	0.0061	0.0135	0.0283	0.0802	0.5000	5.06E-06	3.46E-151
known_regulatory_93	hg18	0	0.0042	0.0097	0.0164	0.0318	0.0747	0.1789	0.002616081	N/A
known_regulatory_93	mm8	0	0.0000	0.0079	0.0117	0.0186	0.0503	0.1840	0.1411141	2.94E-05
known_regulatory_93	monDom4	0	0.0000	0.0057	0.0114	0.0290	0.0484	0.1093	0.0009376944	0.2516276
known_regulatory_93	ornAna1	0	0.0043	0.0134	0.0320	0.0581	0.1020	0.1094	0.00344834	0.9656966
ptrr	hg18	0	0.0000	0.0101	0.0210	0.0529	0.1212	0.1616	8.41E-17	N/A
ptrr	mm8	0	0.0000	0.0070	0.0195	0.0405	0.1053	0.2048	8.21E-11	0.0002408794
ptrr	monDom4	0	0.0000	0.0000	0.0177	0.0409	0.0924	0.1967	1.18E-14	0.008035623
ptrr	ornAna1	0	0.0000	0.0172	0.0361	0.0789	0.1375	0.2000	1.07E-14	0.3144259
ncnr-random	hg18	0	0.0000	0.0030	0.0063	0.0122	0.0431	0.5000	1	N/A
ncnr-random	mm8	0	0.0000	0.0037	0.0081	0.0144	0.0379	0.5000	1	1
ncnr-random	monDom4	0	0.0000	0.0000	0.0026	0.0075	0.0417	0.5000	1	1
ncnr-random	ornAna1	0	0.0000	0.0019	0.0118	0.0313	0.0946	0.5000	1	1

* p(CpG): p-value for test of whether mean of CpG within feature class is the same as for background (ncnr-random) by a two sided t-test

** p(dCpG): p-value for test of whether mean of difference in CpG from human within feature class is the same as for background (ncnr-random) by a two sided t-test

Table 11. Distribution of GC content by feature class / species.

Feature	Species	Min	5%	25%	median	75%	95%	Max	p(GC)*	p(dGC)**
cage_promoters_BR	hg18	0.3500	0.5658	0.6667	0.7220	0.7755	0.8319	0.9032	~0	N/A
cage_promoters_BR	mm8	0.3838	0.5503	0.6447	0.6979	0.7500	0.8215	0.9211	3.80E-299	3.80E-299
cage_promoters_BR	monDom4	0.3471	0.5000	0.6189	0.6779	0.7262	0.8011	0.9091	2.27E-296	2.27E-296
cage_promoters_BR	ornAna1	0.3273	0.5478	0.6638	0.7206	0.7658	0.8329	1.0000	5.02E-280	5.02E-280
cage_promoters_MU	hg18	0.3304	0.5170	0.6577	0.7346	0.7931	0.8394	0.9004	2.02E-279	N/A
cage_promoters_MU	mm8	0.3019	0.5226	0.6362	0.7172	0.7684	0.8220	0.8953	5.71E-255	5.71E-255
cage_promoters_MU	monDom4	0.2807	0.4852	0.5939	0.6667	0.7164	0.7916	0.8533	2.64E-263	2.64E-263
cage_promoters_MU	ornAna1	0.2805	0.4649	0.6316	0.7062	0.7588	0.8141	1.0000	2.79E-208	2.79E-208
cage_promoters_PB	hg18	0.3900	0.5329	0.6584	0.7270	0.7778	0.8358	0.8926	2.64E-252	N/A
cage_promoters_PB	mm8	0.3737	0.5241	0.6355	0.7093	0.7632	0.8271	0.9000	3.12E-220	3.12E-220
cage_promoters_PB	monDom4	0.3535	0.4785	0.6132	0.6667	0.7225	0.7820	1.0000	3.93E-237	3.93E-237
cage_promoters_PB	ornAna1	0.3511	0.5126	0.6528	0.7066	0.7505	0.8170	1.0000	1.08E-212	1.08E-212
cage_promoters_SP	hg18	0.2529	0.4500	0.6113	0.7009	0.7818	0.9091	1.0000	1.06E-209	N/A
cage_promoters_SP	mm8	0.2000	0.4437	0.5860	0.6607	0.7329	0.8575	1.0000	1.58E-178	1.58E-178
cage_promoters_SP	monDom4	0.2000	0.3837	0.5455	0.6220	0.7000	0.8011	1.0000	2.35E-176	2.35E-176
cage_promoters_SP	ornAna1	0.1818	0.3996	0.5724	0.6765	0.7500	0.8336	1.0000	2.19E-151	2.19E-151
CTCF	hg18	0.2743	0.3717	0.4527	0.5068	0.5665	0.6375	0.8111	~0	N/A
CTCF	mm8	0.1556	0.3919	0.4580	0.5000	0.5377	0.6038	0.8269	1.92E-304	1.92E-304
CTCF	monDom4	0.1667	0.3227	0.3985	0.4577	0.5365	0.6637	0.8571	8.37E-234	8.37E-234
CTCF	ornAna1	0.1000	0.3559	0.4381	0.5243	0.6221	0.7333	0.8750	1.91E-170	1.91E-170
PRPv2	hg18	0.1400	0.3082	0.3625	0.4078	0.4719	0.6410	0.9063	4.61E-48	N/A
PRPv2	mm8	0.1481	0.3204	0.3813	0.4280	0.4859	0.6085	1.0000	3.83E-06	3.83E-06
PRPv2	monDom4	0.0000	0.2993	0.3551	0.3980	0.4524	0.5984	1.0000	2.57E-110	2.57E-110
PRPv2	ornAna1	0.0000	0.3106	0.3658	0.4125	0.4792	0.6704	1.0000	3.39E-73	3.39E-73
known_regulatory_93	hg18	0.3366	0.3607	0.5051	0.5663	0.6214	0.6906	0.8377	4.18E-11	N/A
known_regulatory_93	mm8	0.3991	0.4123	0.4779	0.5190	0.5714	0.6641	0.8287	5.32E-10	5.32E-10
known_regulatory_93	monDom4	0.3238	0.3698	0.4900	0.5380	0.6201	0.6513	0.7661	6.69E-14	6.69E-14
known_regulatory_93	ornAna1	0.3438	0.3910	0.5256	0.5934	0.6665	0.7549	0.7778	5.80E-10	5.80E-10
ptrr	hg18	0.2500	0.3699	0.4483	0.5400	0.6115	0.7100	0.7700	5.90E-35	N/A
ptrr	mm8	0.3000	0.3731	0.4559	0.5140	0.5690	0.7011	0.8333	3.04E-25	3.04E-25
ptrr	monDom4	0.0000	0.3207	0.4197	0.4987	0.5947	0.6984	0.8462	1.44E-27	1.44E-27
ptrr	ornAna1	0.0000	0.3405	0.4590	0.5543	0.6826	0.8057	0.9091	7.51E-19	7.51E-19
ncnr-random	hg18	0.0833	0.2745	0.3333	0.3854	0.4819	0.6434	1.0000	1	N/A
ncnr-random	mm8	0.0000	0.2947	0.3730	0.4284	0.4943	0.6036	1.0000	1	1
ncnr-random	monDom4	0.0000	0.2457	0.3194	0.3710	0.4444	0.6286	1.0000	1	1
ncnr-random	ornAna1	0.0000	0.2857	0.3623	0.4234	0.5351	0.7200	1.0000	1	1

* p(GC): p-value for test of whether mean of GC within feature class is the same as for background (ncnr-random) by a two sided t-test

** p(dGC): p-value for test of whether mean of difference in GC from human within feature class is the same as for background (ncnr-random) by a two sided t-test

3. Supplementary Figures

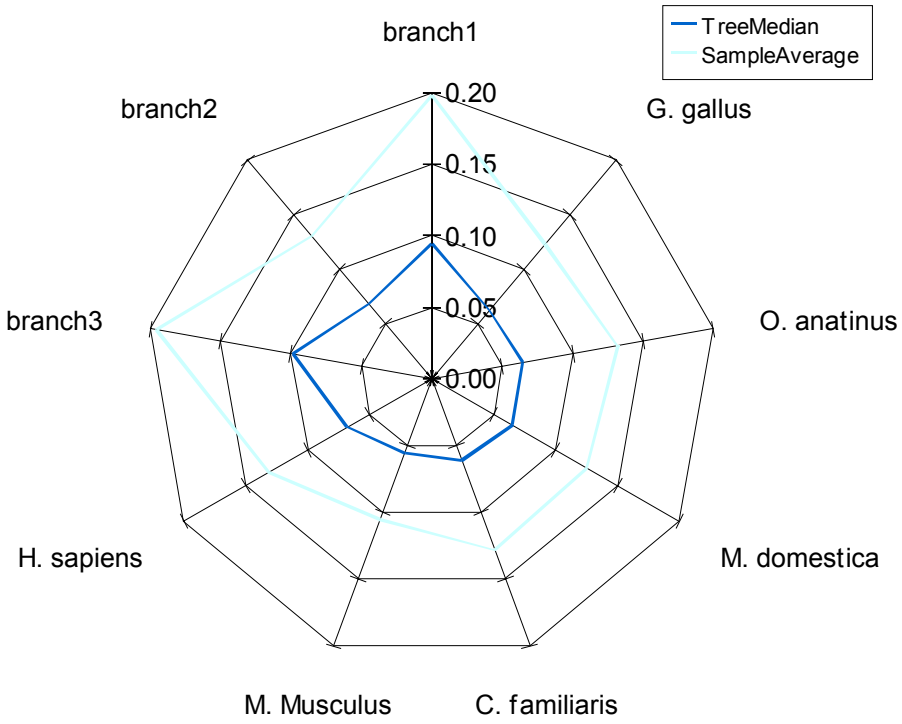


Figure 1. Branch specific d_N/d_S ratios estimated with two different methods (see Accesssory Table A12). The internal branches are: branch1: dog/human/mouse/opossum; branch2: dog/human/mouse; branch3: human/mouse. While the magnitude of this ratio differs, both methods show similar trends.

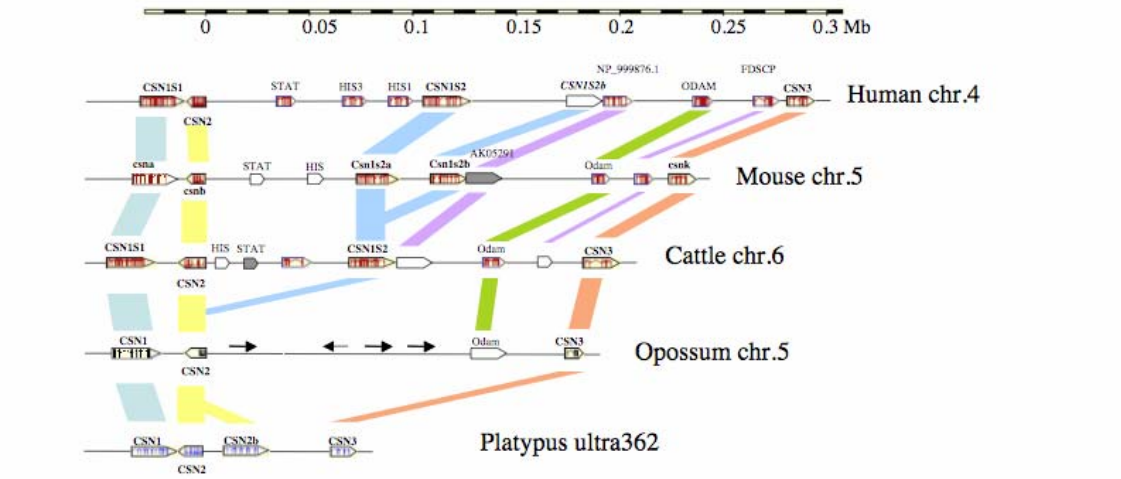


Figure 2. A comparison of the casein loci organization. The casein loci in platypus, opossum, cattle, mouse and human genomes are drawn approximately to scale and aligned on the beta-casein gene[64]. Genes are each represented by a box with an arrow pointing in the direction of gene transcription. Gene models for confirmed genes were generated (platypus) or retrieved from Ensembl (others) when available. Blank boxes represent putative genes based on similarity, whereas grey boxes represent genes with observed expression. The opossum locus, there is no casein duplication and the spacing region contains several copies of an invading repetitive element (black arrows).

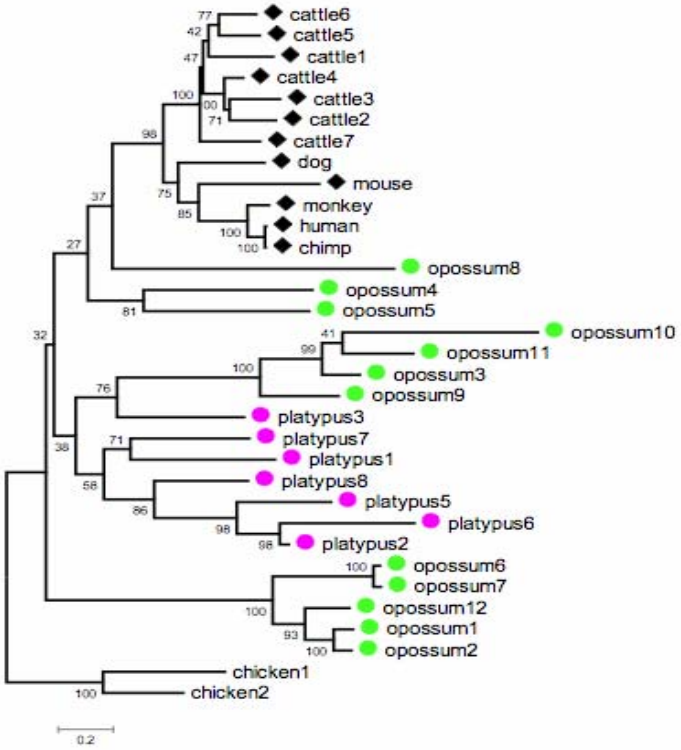
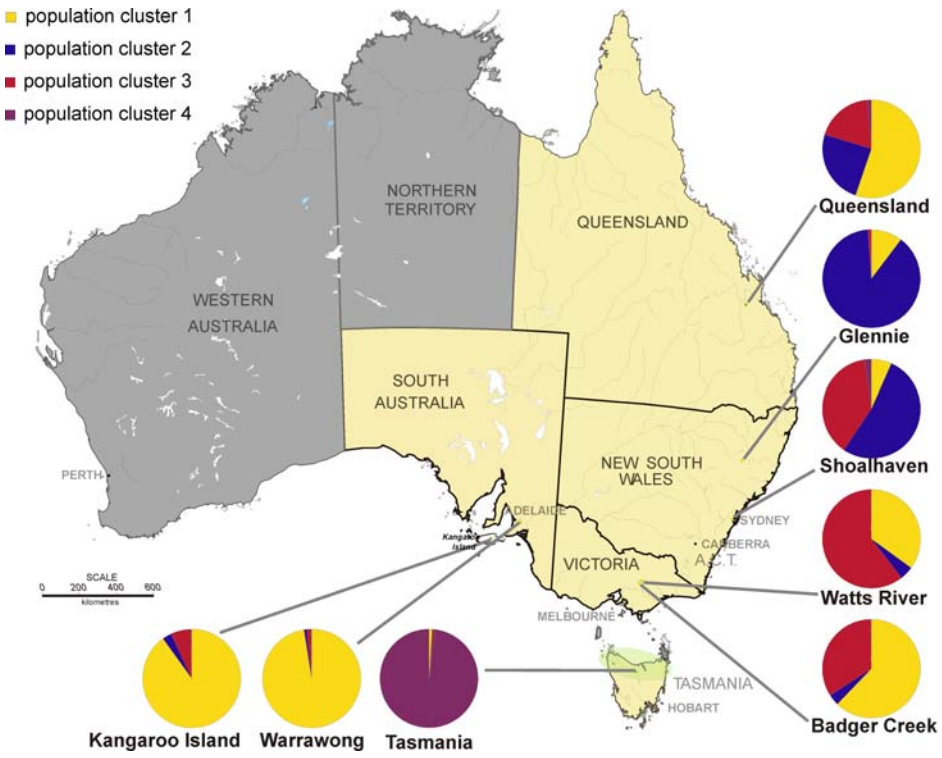


Figure 3. The cathelicidin antimicrobial peptide gene family. has expanded and is highly heterogeneous in the platypus (pink) and opossum (green).These divergent peptides may provide marsupials and monotremes with a unique mechanism for protecting immunologically naïve young from pathogens (See Supplementary Notes S19).

a


population cluster 1
population cluster 2
population cluster 3
population cluster 4
NORTHERN TERRITORY
QUEENSLAND
WESTERN AUSTRALIA
SOUTH AUSTRALIA
NEW SOUTH WALES
VICTORIA
TASMANIA
PERTH
ADELAIDE
SYDNEY
CANBERRA A.C.T.
MELBOURNE
HOBART
SCALE
0 200 400 600
kilometres
Queensland
Glennie
Shoalhaven
Watts River
Badger Creek
Kangaroo Island
Warrawong
Tasmania

b

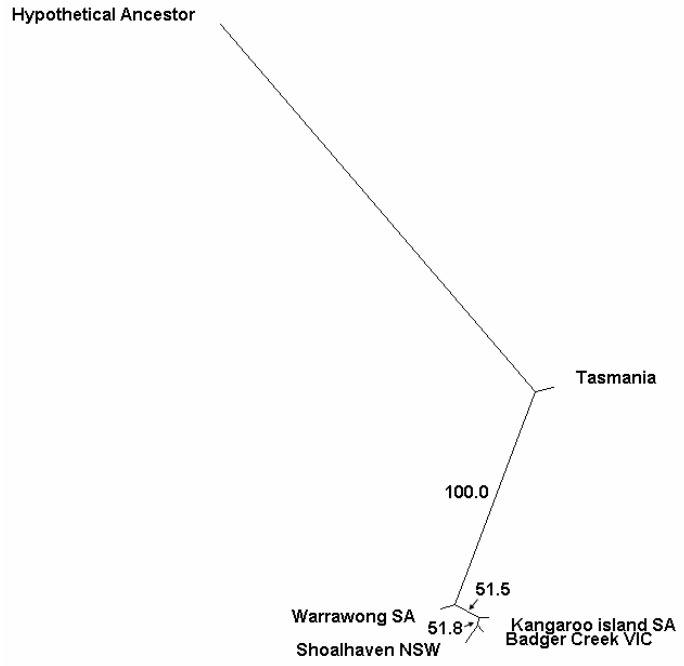

Hypothetical Ancestor
Tasmania
100.0
51.5
Warrawong SA
51.8
Kangaroo island SA
Badger Creek VIC
Shoalhaven NSW

Figure 4. Platypus population structure analysis. Using Structure v2.1 software[50], an analysis was performed with 57 polymorphic retrotransposon loci (LINE2 and Mon1) with 90 platypus samples from various regions in Australia including Tasmania were included in this study. **a,** Map of Australia showing inferred population structure for 90 platypus DNA samples. Grey lines point to the sample geographic origins by name (the precise location within Queensland is unknown). Platypus samples from Tasmania were collected from the northern portion, shaded in green. Structure analyses revealed four distinct genetic clusters, shown in red, green, purple and yellow. Pie charts illustrate the distribution of the four clusters for each platypus population, with samples from Tasmania and Warrawong, showing near single cluster affiliation, while the remaining populations show varying degrees of genetic admixture. **b,** A neighbor-joining tree of platypus population relationships. Genetic distances were calculated using Gendist and Nei's standard genetic distance indicating that the Tasmanian population is distinct from the Australian mainland population and the two South Australian populations cluster together.

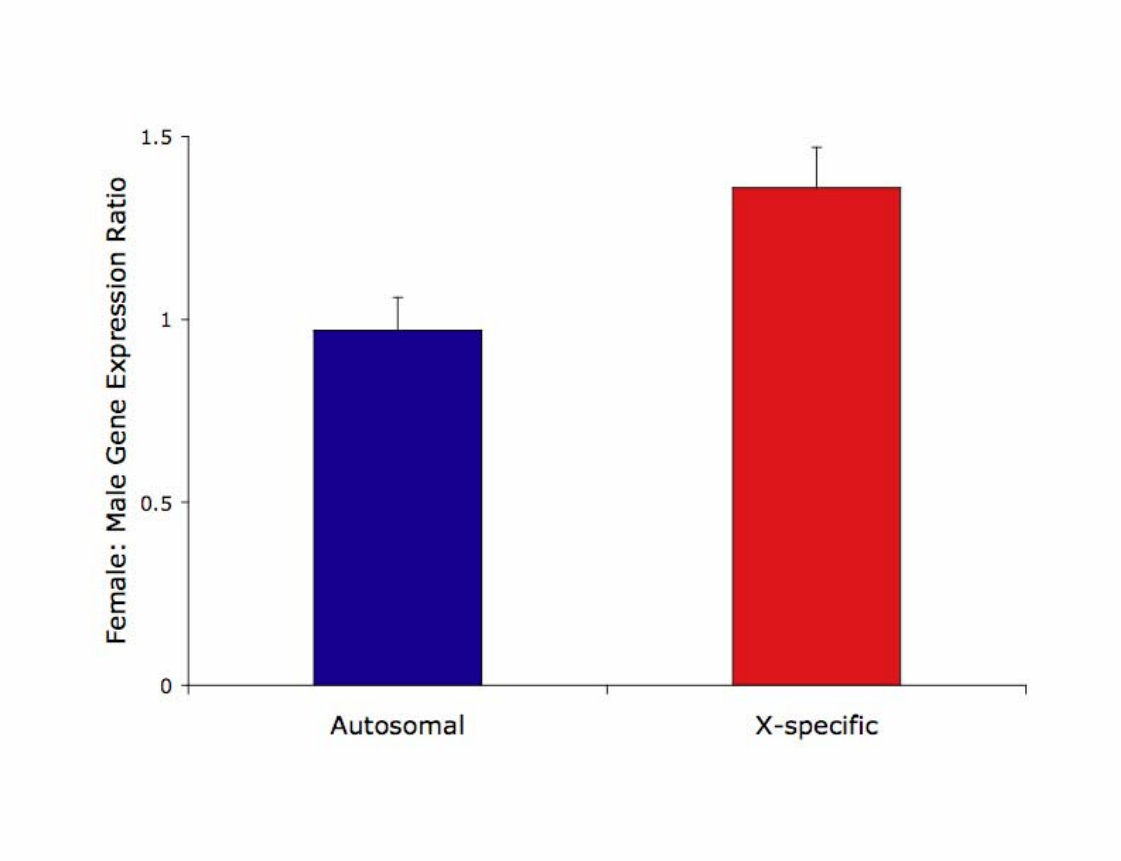


Figure 5. A comparision of female:male ratios for expression of platypus autosomal and X-specific genes in fibroblast cells. Real-time RT-PCR gene estimates were normalized to the autosomal ACTB housekeeping gene. Female:male ratios for autosomal genes was close to 1.0, whereas X-specific gene ratios ranged from near 2.0 (indicative of no dosage compensation), to close to 1.0 (expected if there is complete dosage compensation). Error bars indicate S.E.M.

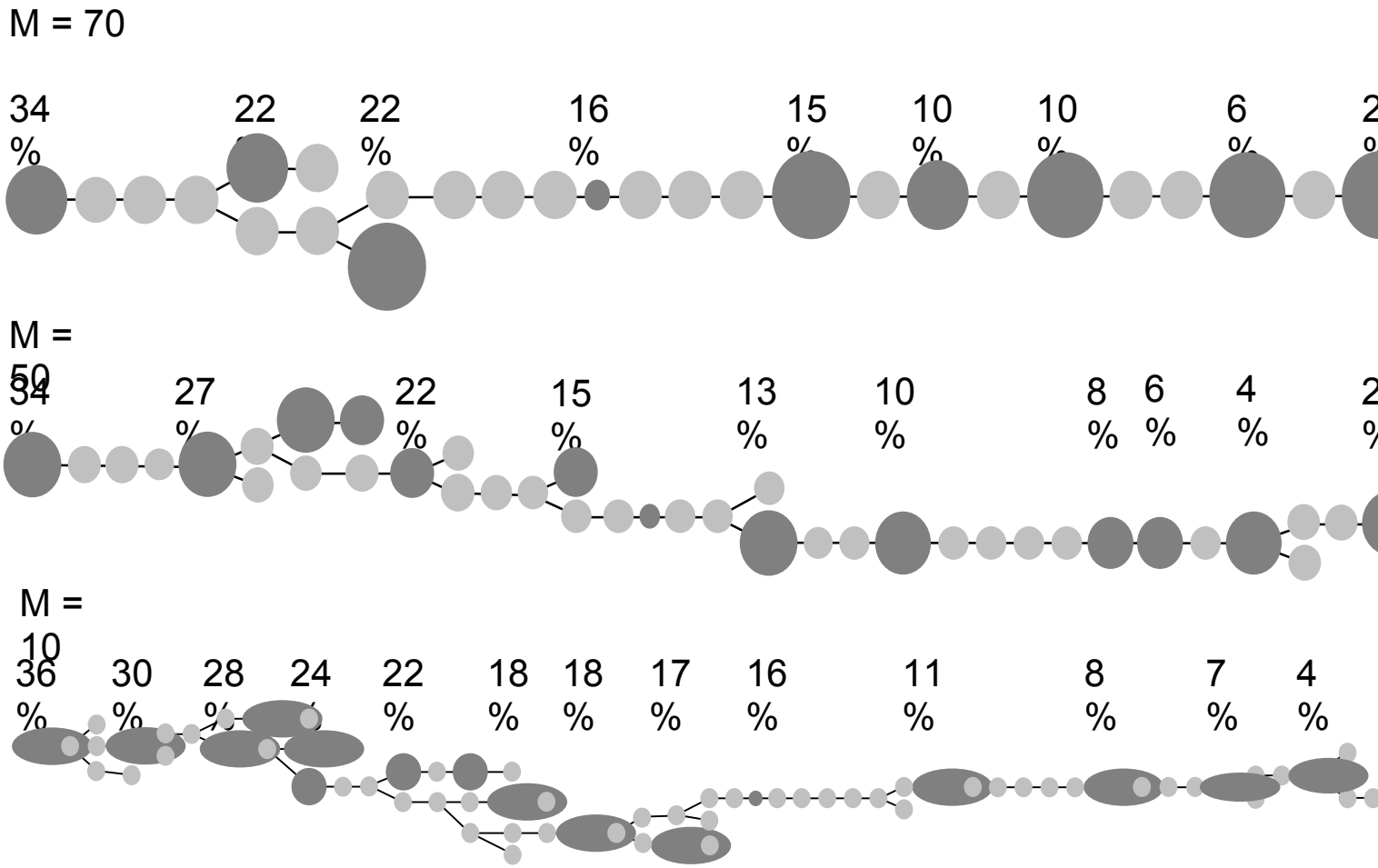
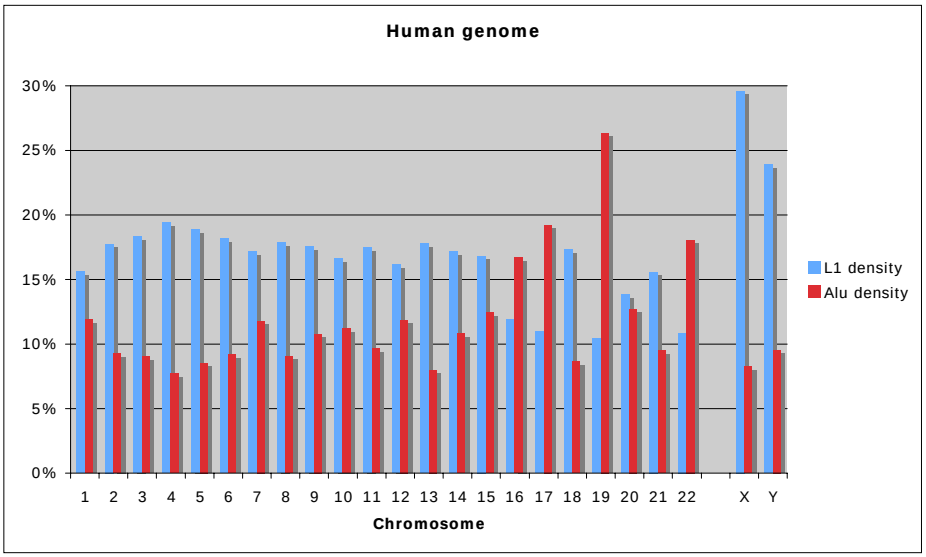
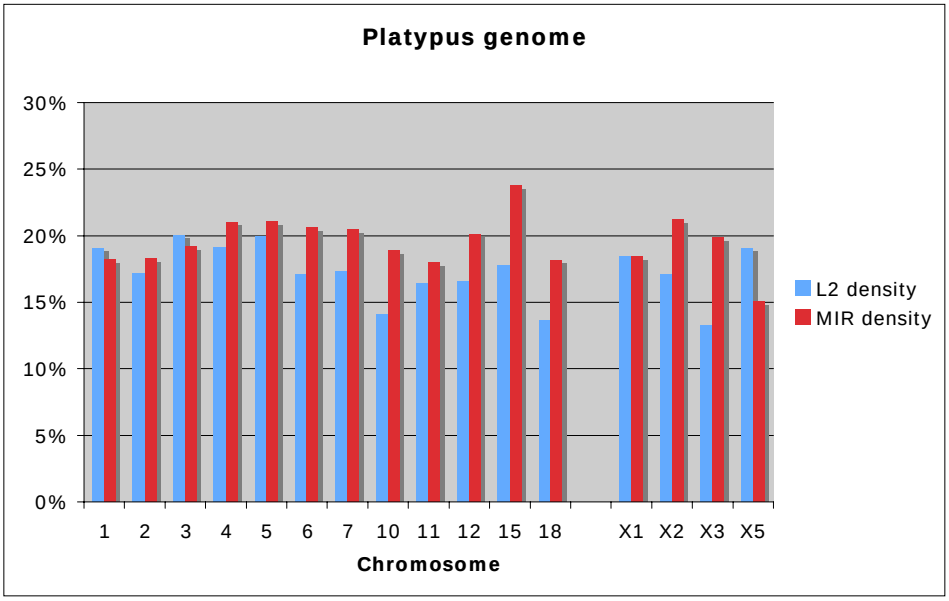


Figure 6. LINE2 evolution in the platypus. (see Supplementary Notes S20).

a


Human genome
30%
25%
20%
15%
10%
5%
0%
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 X Y
Chromosome
L1 density
Alu density

b


Platypus genome
30%
25%
20%
15%
10%
5%
0%
1 2 3 4 5 6 7 10 11 12 15 18 X1 X2 X3 X5
Chromosome
L2 density
MIR density

c

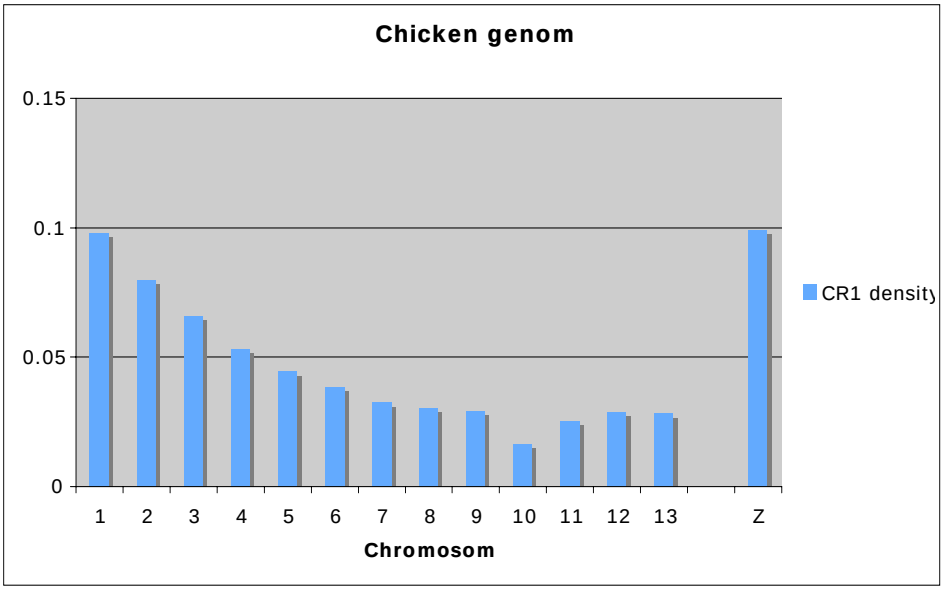


Figure 7. LINE2 density across human, platypus and chicken chromosomes. A, human chromosomal distribution, b, platypus mapped chromosome distribution and c, chicken chromosomal distribution. Neither LINE2 nor MIR/Mon-1 density varies much between platypus chromosomes.

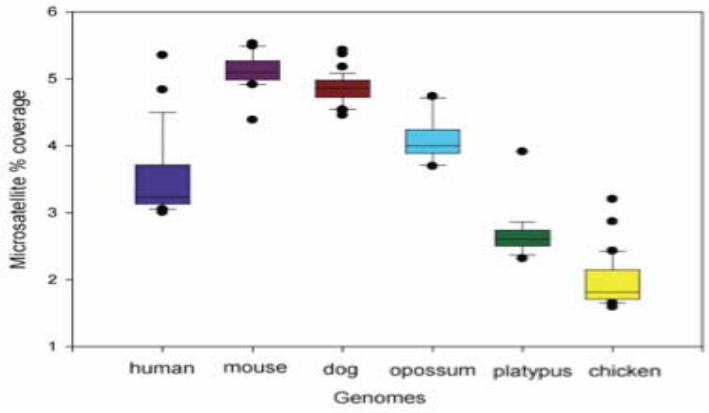
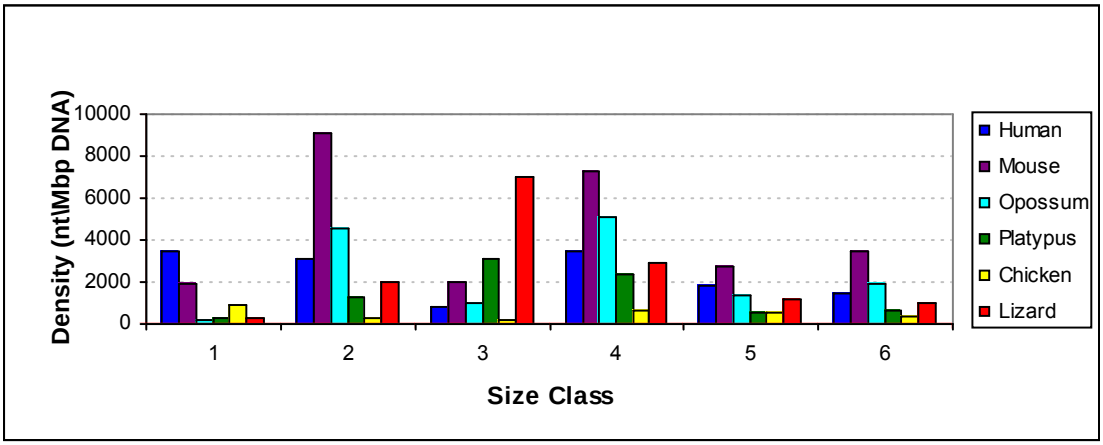


Figure 8. Platypus whole genome microsatellites. Coverage was compared across representative mammalian and avian genomes. For each species, the variation in microsatellite coverage by chromosome is represented by the box plot.

a



b

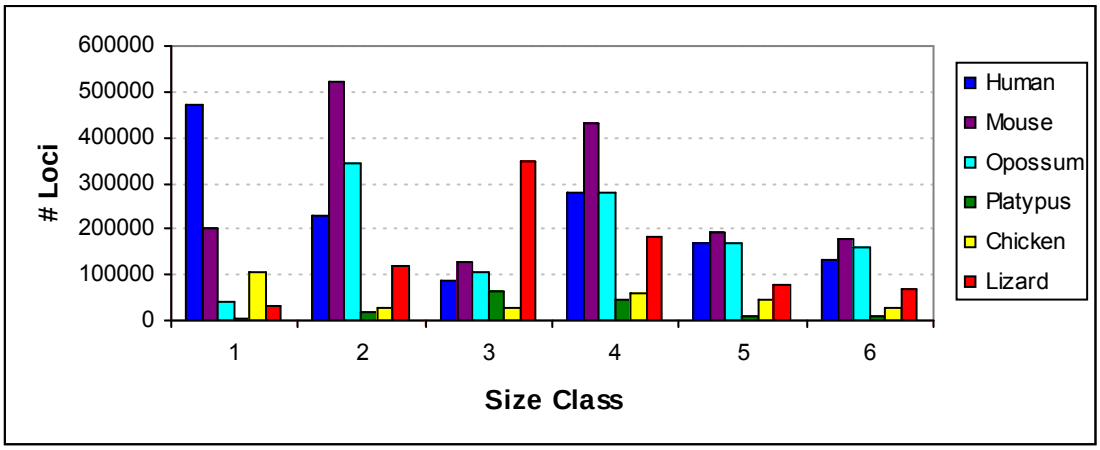


Figure 9. Number and density of microsatellite size classes in each genome. a, density of microsatellites by size class and b, number of microsatellite loci by size class.

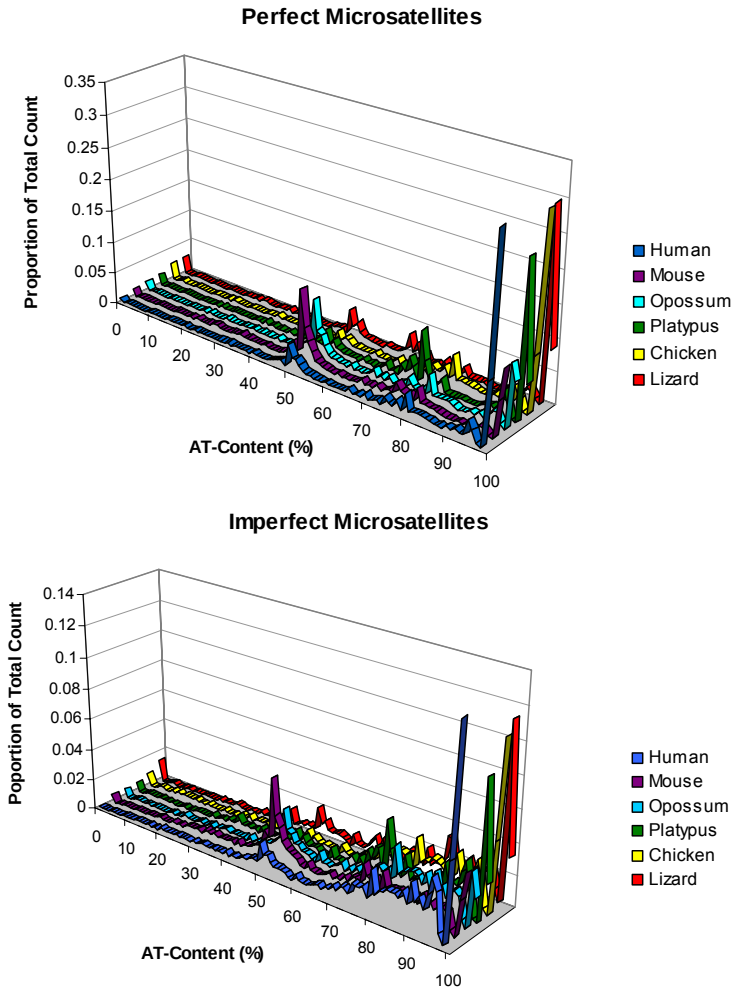


Figure 10. Microsatellite sequence composition measured by percentage A+T of perfect and imperfect microsatellites across different genomes. Despite the platypus genome being enriched for G+C, the three most common motifs in platypus are ATT (12.9%), TAA (7.6%) and TGAA (6.6%); highly similar to the motif usage in lizard.

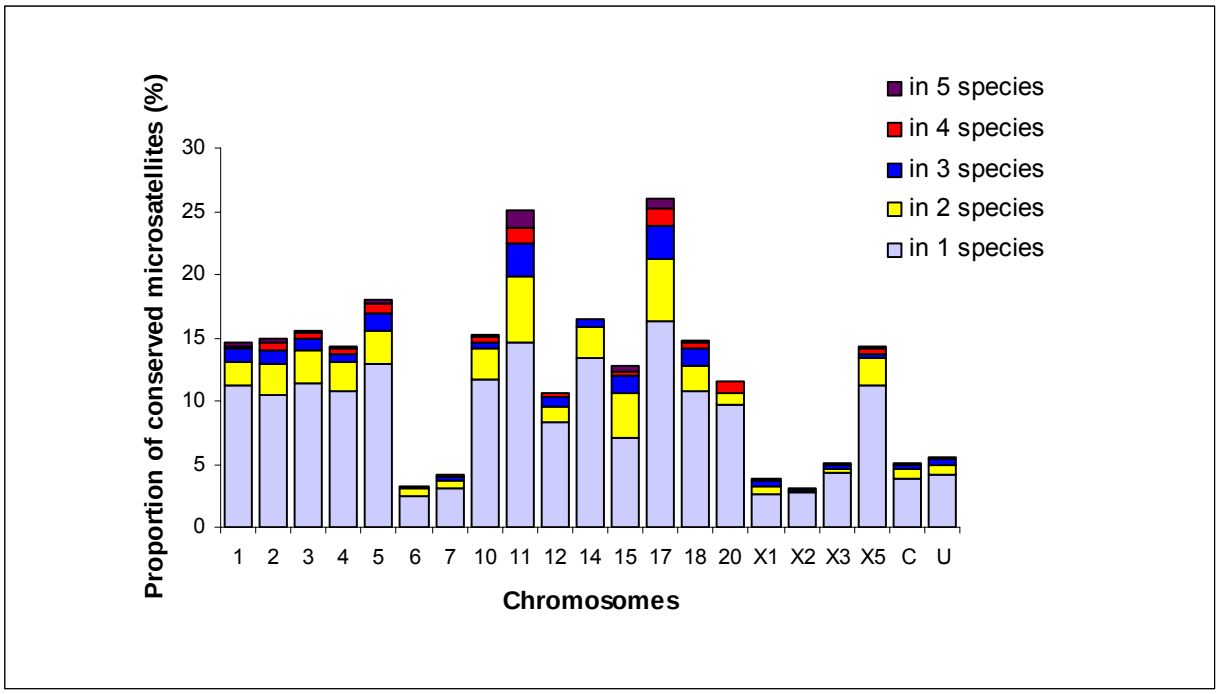


Figure 11. Proportion and distribution of platypus microsatellites conserved in one or more species. We found that of 352,034 platypus microsatellites identified in the whole genome alignment, the percentage of these loci conserved in other species was 0.77% in lizard, 1.19% in chicken, 1.81% in mouse, 1.85% in human and 2.55% in opossum. Most platypus microsatellites are conserved in one species, with decreasing numbers of loci conserved as the number of species increases.

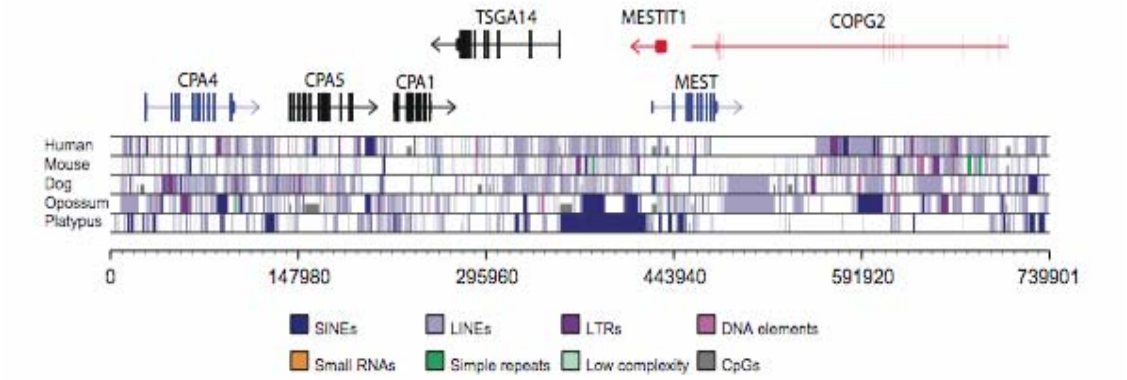


Figure 12. Repeat distribution plots for each mammalian species across the *PEG1/MEST* eutherian imprinted gene cluster. Gene structure is shown by a line (introns) connecting boxes (exons), transcripts in blue are paternally imprinted, red are maternally imprinted and black represents unknown or non-imprinted genes within the cluster (taken from the mouse). There is a dramatic difference in the number and distribution of repeat elements between the platypus and other mammalian species.

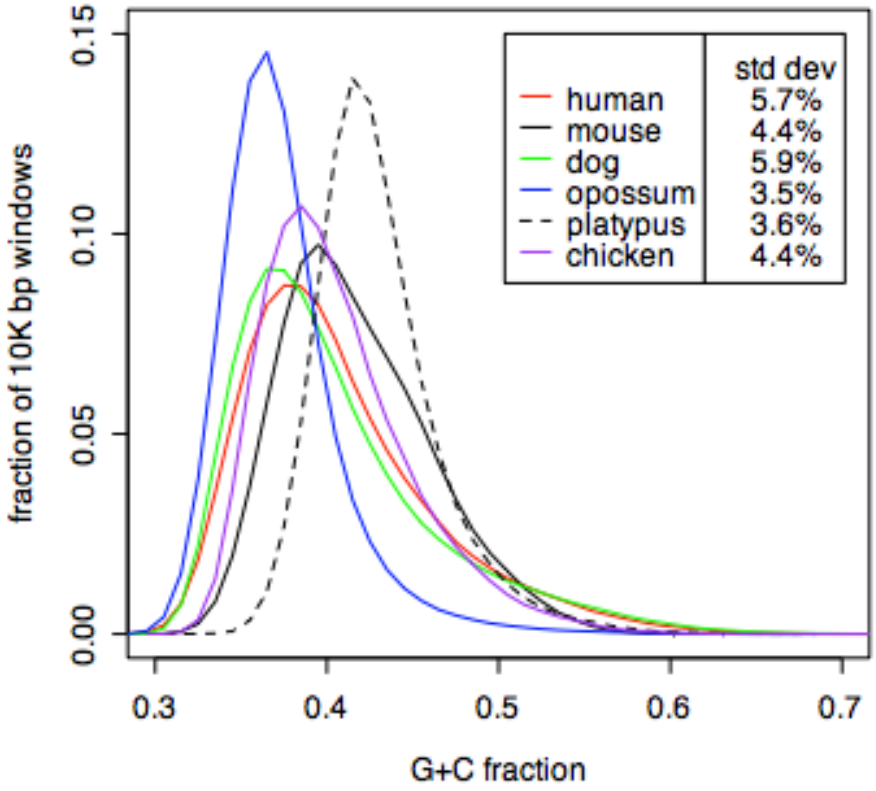


Figure 13. Distribution of G+C fractions in 10-kb windows. All platypus genomic sequences over 10 kb in length were used; for other species, only sequences assigned to chromosomes were examined.

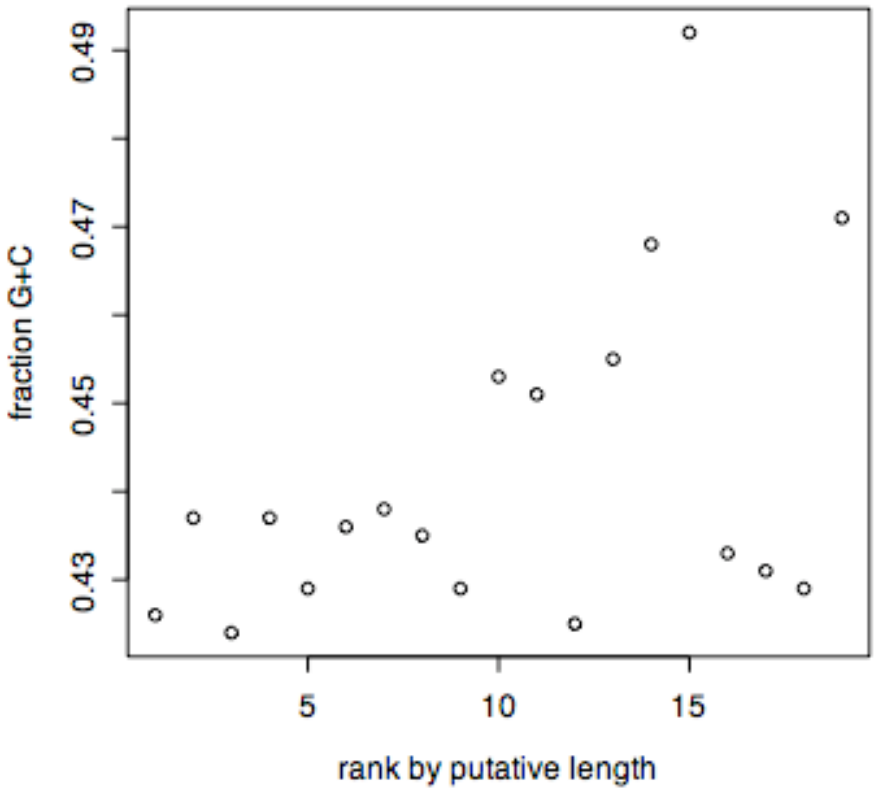


Figure 14. Plot of relative length vs. G+C fraction for platypus chromosomes.

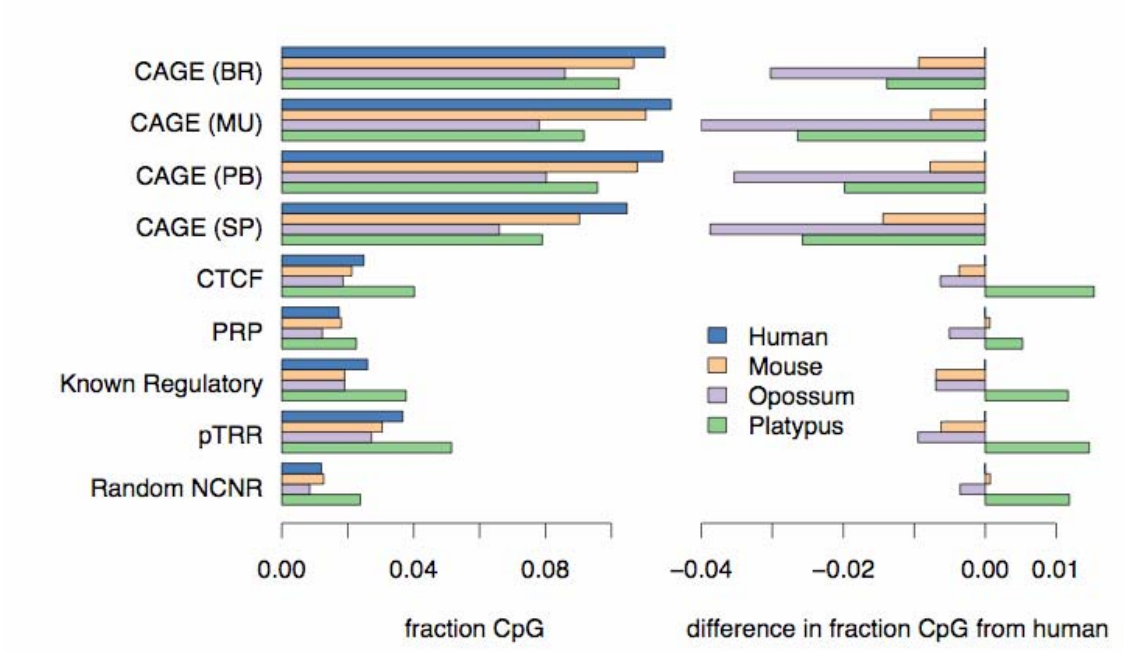


Figure 15. CpG content of putative promoters and other regulatory elements. CpG content and change in CpG content relative to human are shown for several sets of DNA sequences implicated in gene regulation in human. These are the four classes of promoters defined by the CAGE-tag clusters from the FANTOM consortium, binding sites for CTCF from ChIP-chip data, PRPs, which are predicted regulatory regions from the intersection of two methods based on genome comparisons, a set of 93 known regulatory regions, and PTRRs, which are a set of regions identified as bound by sequence-specific transcription factors and supported by chromatin alteration data from the ENCODE project[65].

4. Supplementary References

1. Bick, Y.A.E. & Sharman, G.B. The chromosomes of the platypus (*Ornithorynchus*: Monotremata). *Cytobios* **14**, 17-28 (1975).
2. Margulies, E. H. *et al*.. Comparative sequencing provides insights about the structure and conservation of marsupial and monotreme genomes. *Proc Natl Acad Sci USA* **102**, 3354-3359 (2005).
3. Trask, B., Van den Engh, G., Mayall, B. & Gray, J.W. Chromosome heteromorphism quantified by high-resolution bivariate flow karyotyping. *Am J Hum Genet* **45**, 739-752 (1989).
4. Huang, X. *et al*. Application of a superword array in genome assembly. *Nucleic Acids Res* **34**, 201-205 (2006).
5. Wallis, J. W. *et al.* A physical map of the chicken genome. *Nature* **432**, 761-764 (2004).

6. Gibbs, R. A. *et al.* Evolutionary and biomedical insights from the rhesus macaque genome. *Science* **316**, 222-34 (2007).
7. Hillier, L. W. *et al*. Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature* **9**, 695-718 (2004).
8. McMillan, D. *et al.* Characterizing the chromosomes of the platypus (Ornithorhynchus anatinus). *Chromosome Res* **15**, 961-974 (2007).
9. Collins, F. S. *et al.* Finishing the euchromatic sequence of the human genome. Nature **431**, 931-945 (2004).
10. Mikkelsen, T. S. *et al*. Genome of the marsupial Monodelphis domestica reveals innovation in non-coding sequences. *Nature* **447**, 167-177 (2007).
11. Lindblad-Toh, K. *et al.* Genome sequence, comparative analysis and haplotype structure of the domestic dog. *Nature* **438**, 803-819 (2005).
12. Bailey, J. A., Yavor, A. M., Massa, H. F., Trask, B. J. & Eichler, E. E. Segmental duplications: organization and impact within the current human genome project assembly. *Genome Res* **11**, 1005-17 (2001).
13. Bailey, J. A. *et al*. Recent segmental duplications in the human genome. *Science* **297**, 1003-1007 (2002).
14. She, X. *et al*. Shotgun sequence assembly and recent segmental duplications within the human genome. *Nature* **431**, 927-930 (2004).
15. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res* **27**, 573-80 (1999).
16. Curwen, V. *et al.* The Ensembl automatic gene annotation system. *Genome Res* **14**, 942-50 (2004).
17. Goodstadt, L. & Ponting, C. P. Phylogenetic reconstruction of orthology, paralogy, and conserved synteny for dog and human. *PLoS Comput Biol* **2**, e133 (2006).
18. Goodstadt, L., Heger, A., Webber, C. & Ponting, C. P. An analysis of the gene complement of a marsupial, Monodelphis domestica: evolution of lineage-specific genes and giant chromosomes. *Genome Res* **17**, 969-981 (2007).
19. Heger, A. P. & Ponting, C.P. Evolutionary rate analyses of orthologues and paralogues from twelve Drosophila genomes. *Genome Res* **17,** 1837-1849 (2007).
20. Altschul, S. F. *et al*. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25**, 3389-3402 (1997).
21. Edgar, R. C. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatic*s **5**, 113 (2004).
22. Hedges, S. B. The origin and evolution of model organisms. *Nat Rev Genet* **3**, 838-849 (2002).
23. Yang, Z. PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput Appl Biosci* **13**, 555-556 (1997).
24. Goldman, N. & Yang, Z. A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Mol Biol Evol* **11**, 725-736 (1994).
25. Castresana, J. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol Biol Evol* **17**, 540-552 (2000).
26. Bernadette, L., Mondy, A. & Keenan, T.W. . Butyrophilin and xanthine oxidase occur in constant molar proportions in milk lipid globule membrane but vary in amount with breed and stage of lactation. *Protoplasma* **177**, 32-36 (1993).

27. McManaman, J. L., Neville, M. C. & Wright, R. M. Mouse mammary gland xanthine oxidoreductase: purification, characterization, and regulation. *Arch Biochem Biophys* **371**, 308-316 (1999).
28. Pettigrew, J. D. Electroreception in monotremes. *J Exp Biol* **202**, 1447-1454 (1999).
29. Proske, U., Gregory, J. E. & Iggo, A. Sensory receptors in monotremes. *Philos Trans R Soc Lond B Biol Sci* **353**, 1187-98 (1998).
30. Davies, W. L. *et al*. Visual pigments of the platypus: a novel route to mammalian colour vision. *Curr Biol* **17**, R161-163 (2007).
31. Wakefield, M. J. *et al.* Cone visual pigments of monotremes: filling the phylogenetic gap. *Visual Neuroscience* (2008) in press.
32. Amsterdam, A. *et al*. A large-scale insertional mutagenesis screen in zebrafish. *Genes Dev* **13**, 2713-2724 (1999).
33. Gross, J. M. *et al.* Identification of zebrafish insertional mutants with defects in visual system development and function. *Genetics* **170**, 245-261 (2005).
34. Ascenzi, P. *et al*. Hemoglobin and heme scavenging. *IUBMB Life* **57**, 749-759 (2005).
35. Grant, T. *The platypus: a unique mammal* (UNSW Press, Sydney, 1995).
36. Andersen, N. A., Mesch, U., Lovell, D. J. & Nicol, S. C. The effects of sex, season and hybernation on haemotology and blood viscosity of free-ranging echidnas (*Tachyglossus aculeatus*). *Canadian Journal of Zoology* **78**, 174-181 (2000).
37. Huttley, G. A., Wakefield, M. J. & Easteal, S. Rates of genome evolution and branching order from whole genome analysis. *Mol Biol Evol* **24**, 1722-1730 (2007).
38. Jones, D. T., Taylor, W. R. & Thornton, J.M. The rapid generation of mutation data matrices from protein sequences. *Comput Appl Biosci* **8**, 275-282 (1992).
39. Huttley, G. A. Modeling the impact of DNA methylation on the evolution of BRCA1 in mammals. *Mol Biol Evol* **21**, 1760-1768 (2004).
40. Tavare, S. Some probabilistic and statistical problems in the analysis of DNA sequences. *Lec. Math. Life Sci.* **17** (1986).
41. Butterfield, A. *et al.* PyEvolve: a toolkit for statistical modelling of molecular evolution. *BMC Bioinformatics* **5**, 1 (2004).
42. Janke, A., Gemmell, N. J., Feldmaier-Fuchs, G., von Haeseler, A. & Paabo, S. The mitochondrial genome of a monotreme--the platypus (Ornithorhynchus anatinus). *J Mol Evol* **42**, 153-159 (1996).
43. Murphy, W. J., Pringle, T. H., Crider, T. A., Springer, M. S. & Miller, W. Using genomic data to unravel the root of the placental mammal phylogeny. *Genome Res* **17**, 413-21 (2007).
44. Schmitz, J., Ohme, M. & Zischler, H. SINE insertions in cladistic analyses and the phylogenetic affiliations of Tarsius bancanus to other primates. *Genetics* **157**, 777-784 (2001).
45. Kriegs, J. O. *et al.* Retroposed elements as archives for the evolutionary history of placental mammals. *PLoS Biol* **4**, e91 (2006).

46. Price, A. L., Eskin, E. & Pevzner, P.A. Whole-genome analysis of Alu repeat elements reveals complex evolutionary history. *Genome Res* **14**, 2245-2252 (2004).
47. Kordis, D., Lovsin, N. & Gubensek, F. Phylogenomic analysis of the L1 retrotransposons in Deuterostomia. *Syst Biol* **55**, 886-901 (2006).
48. Rozen, S. & Skaletsky, H. Primer3 on the WWW for general users and for biologist programmers. *Methods Mol Biol* **132**, 365-386 (2000).
49. Kent, W. J. BLAT--the BLAST-like alignment tool. *Genome Res* **12**, 656-664 (2002).
50. Pritchard, J. K., Stephens, M. & Donnelly, P. Inference of population structure using multilocus genotype data. *Genetics* **155**, 945-959 (2000).
51. La Rota, M., Kantety, R.V., Yu, J.K. & Sorrells, M.E. Nonrandom distribution and frequencies of genomic and EST-derived microsatellite markers in rice, wheat, and barley. *BMC Genomics* **6**, 23-29 (2005).
52. Buschiazzo, E. & Gemmell, N. J. The rise, fall and renaissance of microsatellites in eukaryotic genomes. *Bioessays* **28**, 1040-1050 (2006).
53. Karolchik, D. *et al.* The UCSC Genome Browser Database. *Nucleic Acids Res* **31,** 51-54 (2003).
54. Giardine, B. *et al.* Galaxy: a platform for interactive large-scale genome analysis. *Genome Res* **15**, 1451-1455 (2005).
55. Rice, P., Longden, I. & Bleasby, A. EMBOSS: the European Molecular Biology Open Software Suite. *Trends Gene*t **16**, 276-287 (2000).
56. Kofler, R., Schlotterer, C. & Lelley, T. SciRoKo: a new tool for whole genome microsatellite search and investigation. *Bioinformatics* **23**, 1683-5 (2007).
57. Toth, G., Gaspari, Z. & Jurka, J. Microsatellites in different eukaryotic genomes: survey and analysis. *Genome Res* **10,** 967-81 (2000).
58. Kim, T. H. *et al.* Analysis of the vertebrate insulator protein CTCF-binding sites in the human genome. *Cell* **128,** 1231-1245 (2007).
59. Miller, W. *et al*. 28-Way vertebrate alignment and conservation track in the UCSC Genome Browser. *Genome Res* **17**, 1797-1808 (2007).
60. Elnitski, L., Hardison, R.C., Li, J., Yang, S., Kolbe, D., Eswara, P., O'Connor, M.J., Schwartz, S., Miller, W. & Chiaromonte, F. Distinguishing regulatory DNA from neutral sites. *Genome Res* **13**, 64-72 (2003).
61. Antequera, F. Structure, function and evolution of CpG island promoters. *Cell Mol Life Sci* **60**, 1647-1658 (2003).
62. Carninci, P. *et al*. Genome-wide analysis of mammalian promoter architecture and evolution. *Nat Genet* **38**, 626-635 (2006).
63. Frith, M. C. et al. Evolutionary turnover of mammalian transcription start sites. *Genome Res* **16**, 713-22 (2006).
64. Rijnkels, M. Multispecies comparison of the casein gene loci and evolution of casein gene family. *Journal of Mammary Gland Biology and Neoplasia* **7**, 327-345 (2002).
65. King, D. C. *et al*. Finding cis-regulatory elements using comparative genomics: some lessons from ENCODE data. *Genome Res* **17**, 775-786 (2007).