

**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<sup>1</sup> 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  $C \cdot N_A (\text{Avogadro's num}) \cdot (\text{mean molar mass of a basepair})^{-1}$ . 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 =  $3.0 \cdot 10^{-12} \cdot 6.022 \cdot 10^{23} / 660 = 2.7\text{Gb}$ .

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A more recent genome size estimate (J.S. Johnston, unpublished data) is 2.35Gb.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<sup>2</sup>. 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. O'Brien, W. Rens and M. Ferguson-Smith, unpublished data) using the method reported by Trask *et al.*<sup>3</sup>, 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 (Accessory Table A2).

The combined sequence reads were assembled using the PCAP software<sup>4</sup>, 300 parallel PCAP jobs ran on a cluster of AMD® opteron<sup>tm</sup> 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 consensus 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 1<sup>st</sup> 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 *Theileria* 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<sup>5</sup>. There were a total of 379,614 BES that could be used for this purpose. The predicted *HindIII* 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<sup>6,7</sup> 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<sup>8</sup>.

**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)<sup>9</sup>, chicken (galGal3)<sup>7</sup>, opossum (monDom4)<sup>10</sup> and dog (canFam2)<sup>11</sup> 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<sup>1,2</sup> (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<sup>2</sup> 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<sup>12,13</sup> 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<sup>14</sup>, 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<sup>4</sup>; 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”<sup>15</sup>. 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 chimera 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 OrnAnal. 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 II restriction enzyme sites (*MmeI*). 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 *MmeI* 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' termini of the cDNA molecules contained *MmeI* 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 *MmeI*-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<sup>16</sup>. 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<sup>16</sup>. 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<sup>17</sup> 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<sup>17,18</sup>. 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<sup>19</sup>. Orthologs were predicted in three steps: (1) orthologs were predicted between pairs of genomes using PhyOP<sup>17</sup>, (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<sup>20</sup>. 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<sup>21</sup>. 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<sup>22</sup> 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)<sup>n</sup> simple ortholog sets and orthologous groups with duplications and deletions. As a consequence of using NJTree, (1:1)<sup>n</sup> simple orthology gene trees almost always correspond to the species phylogeny.

Rate estimation was performed with PAML<sup>23</sup> using the Goldman & Yang (1994) model<sup>24</sup>. 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 ( $\kappa$ ), synonymous substitution rate ( $d_S$ ), non-synonymous substitution rate ( $d_N$ ) and the ratio of synonymous to non-synonymous substitutions ( $\omega$ ). 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<sup>25</sup>, 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  $d_S$ ,  $d_N$  and  $d_N/d_S$  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<sup>26</sup>. As its protein level correlates with the maturation of mouse mammary tissue in pregnancy<sup>27</sup> 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<sup>28,29</sup>. 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<sup>30</sup>. 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<sup>30,31</sup> 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<sup>32,33</sup>.

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:

Chicken1 NP\_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<sup>34</sup>. 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<sup>35</sup>, 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<sup>36</sup>.

**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<sup>37</sup> 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<sup>21</sup>, 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<sup>38</sup>; a codon substitution model<sup>39</sup>; the general time reversible nucleotide substitution model<sup>40</sup>; and a purine/pyrimidine (RY) substitution model employed to address the violation of compositional stationarity<sup>37</sup>. Gamma distributed rate-heterogeneity variants for most of these models were also used where practical. All models were implemented using the PyEvolve package<sup>41</sup> 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<sup>37</sup>). 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<sup>42</sup>, 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<sup>43</sup> 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 homoplastic 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<sup>44,45</sup> 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.repeatmasker.org](http://www.repeatmasker.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/replib](http://www.girinst.org/replib)) 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.*<sup>46</sup> 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<sup>47</sup>, 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<sup>48</sup>. Each primer was checked with BLAT-The BLAST-Like Alignment Tool<sup>49</sup> 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  $\mu$ l reactions containing 10 ng of template DNA (platypus, echidna, possum, wallaby or human); 200 nM of each oligonucleotide primer; 1.5 mM MgCl<sub>2</sub>; 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<sup>50</sup> (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)<sup>15</sup> and Sputnik<sup>51</sup> (<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<sup>52</sup>, 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<sup>53</sup>. FASTA-formatted sequences were extracted using Galaxy<sup>54</sup> gaps were removed using the module degapseq from the EMBOSS 5.0

package<sup>55</sup>, and perfect and imperfect microsatellites were searched using SciRoKo 3.1 with fixed penalty parameters: 12, 4, 3, 3, 3<sup>56</sup>. 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<sup>57</sup>.

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 (Buschiazzi *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<sup>58</sup>, PRPs<sup>59</sup> and 93 known regulatory regions<sup>60</sup>. 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<sup>61</sup>. 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<sup>10</sup>, 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.*<sup>62</sup>. 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.*<sup>63</sup> 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.

npcRNAs in platypus															
	CD-box snoRNAs					HACA-box snoRNAs				snoRTE HACA		U1-13	Others		
	Intron in RPG	Intron in specific snoRNA genes	Intron other	Intergenic	na	Intron in RPG	Intron other	Intergenic	na	Intron	na				
I conserved in vertebrates	Oa1707	Oa2367	Oa1692	Oa1712	Oa2337	Oa1758	Oa1877	Oa1744	Oa1813			Oa4000-U1	Oa1766-U4c	miRNA (31)	
	Oa1735a	Oa2371	Oa1702	Oa1716	Oa2364	Oa1765	Oa1953	Oa1860	Oa1844			Oa4001-U1	Oa1842-U4atac	Y RNA (4)	
	Oa1785	Oa2572	Oa1723	Oa1733	Oa2369	Oa1797	Oa1973	Oa1867	Oa1861			Oa4002-U1	Oa3273-U5	7SL RNA (2)	
	Oa1817	Oa2691	Oa1724	Oa1745	Oa2372	Oa1850a	Oa2055	Oa1870	Oa1886			Oa4003-U1	Oa3278-U5	tRNA (75)	
	Oa1821	Oa2856	Oa1764	Oa1760	Oa2470	Oa1900	Oa2102	Oa1949	Oa1959			Oa4004-U1	Oa3210-1-U5	rRNA (59)	
	Oa1835	Oa2949	Oa1803	Oa1770	Oa2681	Oa2049	Oa2133	Oa1957	Oa1962			Oa4005-U1	Oa3210-2-U5	scaRNAs (1)	
	Oa1853	bOaU98	Oa1805	Oa1784	Oa2717	Oa2051	bOaHACA33	Oa1969	Oa2001			Oa4006-U1	Oa3211-1-U5	spacer RNA (1)	
	Oa1857	bOaU86	Oa1836	Oa1786	Oa2916	Oa2054	bOaHACA62i	Oa2002	Oa2086			Oa4007-U1	Oa3211-2-U5	7SK RNA (1)	
	Oa1942	bOa2156i	Oa1843	Oa1791	Oa2977	Oa2111	bOaHACA70	Oa2163	Oa2150			Oa4008-U2	Oa3211-3-U5	RNase P RNA (1)	
	Oa1990	bOaCD24	Oa1892	Oa1811	Oa2993	Oa2116		bOaHACA22	Oa2164			Oa4009-U2	Oa2128-U6		
	Oa2016	bOaCD32	Oa1907	Oa1814	bOaU141i	Oa2340		bOaHACA28	Oa2167			Oa4010-U3	Oa2466-U7		
	Oa2046	bOaCD36C	Oa1925	Oa1815	bOaU1412	Oa2872		bOaHACA36A	Oa2181			Oa4011-U3	Oa1948-U8		
	Oa2073	bOaCD58B	Oa2342	Oa1833	bOaU1413	Oa3060		bOaHACA56	bOa1886i			Oa4012-U3	Oa2165-U11		
	Oa2101		Oa2376	Oa1927	bOaU45A	bOaU88i		bOa1860i	bOa2150i			Oa4013-U3	Oa2180-U12		
	Oa2121		Oa2527	Oa1928	bOaU45B	bOaCD6		bOa19641i	bOa2164i			Oa4014-U3	Oa1793-U13		
	Oa2126		Oa2528	Oa1943	bOaU45C	bOaCD65		bOa196412	bOa2181i						
	Oa2132		Oa2584	Oa1976	bOaU84	bOaR38B			bOaHACA1						
	Oa2151		Oa2586	Oa1989	bOaU951i				bOaHACA44						
	Oa2156		Oa2684	Oa2032	bOaCD19										
	Oa2171		Oa2985	Oa2118a	bOaCD20										
	Oa2339		bOaU27	Oa2118b	bOaCD66i										
			bOaU30	Oa2134	bOaCD111B										
			bOaU75		bOaCD117/691i										
	II conserved in mammals	Oa1774		Oa1787		bOaCD5									
bOaU73B			Oa2015				Oa1809								
III specific for platypus	Oa1735b		bOaU70i		bOa20491i	Oa1747	Oa1858	Oa1849		Oa1965	Oa1862	Oa1874b			
	Oa1781		bOaU95i2		bOa2049i2	Oa1759	Oa1968	Oa1964			Oa1864	Oa1956b			
						Oa1767	Oa2161	Oa1967			Oa1874a	Oa1972a			
						Oa1782		Oa2092			Oa1912	Oa1972b			
						Oa1829					Oa1954a	Oa2040			
						Oa1850b					Oa1954b	Oa2050b			
						Oa1904					Oa1954c	Oa2052			
						Oa2095					Oa1954d	Oa2053			
						Oa2109					Oa1956a	Oa2077			
						Oa2137					Oa1966				
						Oa2712					Oa2149				
						Oa2939					Oa2166				
						bOaCD38A/BI									
						bOaCD62i									

**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	
<i>H. sapiens</i>	22,611	19,339	86%	3,272	14%
<i>M. musculus</i>	24,442	20,758	85%	3,684	15%
<i>C. familiaris</i>	19,314	18,066	94%	1,248	6%
<i>M. domestica</i>	19,597	18,123	92%	1,474	8%
<i>O. anatinus</i>	18,596	15,312	82%	3,284	18%
<i>G. gallus</i>	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 <sup>-10</sup>	2.17E-10	
GO:0009653	morphogenesis	1.31	5.9 x 10 <sup>-22</sup>	5.86E-22	
GO:0016043	cell organization and biogenesis	1.27	1.9 x 10 <sup>-17</sup>	1.86E-17	
GO:0006519	amino acid and derivative metabolism	1.25	4.2 x 10 <sup>-6</sup>	4.22E-06	
GO:0006464	protein modification	1.25	6.8 x 10 <sup>-24</sup>	6.78E-24	
GO:0015031	protein transport	1.24	1.6 x 10 <sup>-10</sup>	1.58E-10	
GO:0007010	cytoskeleton organization and biogenesis	1.22	8.1 x 10 <sup>-7</sup>	8.14E-07	
GO:0007275	development	1.21	8.7 x 10 <sup>-18</sup>	8.73E-18	
GO:0030154	cell differentiation	1.21	2.4 x 10 <sup>-11</sup>	2.35E-11	
GO:0007049	cell cycle	1.2	1.1 x 10 <sup>-8</sup>	1.09E-08	
GO:0006810	transport	1.18	8.9 x 10 <sup>-13</sup>	8.87E-13	
GO:0009058	biosynthesis	1.18	9.2 x 10 <sup>-7</sup>	9.23E-07	
GO:0006811	ion transport	1.18	1.7 x 10 <sup>-7</sup>	1.71E-07	
GO:0008283	cell proliferation	1.15	2.8 x 10 <sup>-5</sup>	2.78E-05	
GO:0007267	cell-cell signaling	1.15	4.6 x 10 <sup>-5</sup>	4.63E-05	
GO:0009056	catabolism	1.13	6.8 x 10 <sup>-4</sup>	6.84E-04	
GO:0006139	nucleobase, nucleoside, nucleotide and nucleic acid metabolism	1.13	6.2 x 10 <sup>-5</sup>	6.24E-05	
GO:0006629	lipid metabolism	1.12	6.4 x 10 <sup>-4</sup>	6.38E-04	
Molecular function					
Accession	GO Term	Fold	P-Value	P-Value	
GO:0008092	cytoskeletal protein binding	1.42	2.2 x 10 <sup>-7</sup>	2.22E-07	
GO:0004672	protein kinase activity	1.31	9.2 x 10 <sup>-15</sup>	9.21E-15	
GO:0003779	actin binding	1.3	7 x 10 <sup>-7</sup>	7.01E-07	
GO:0005216	ion channel activity	1.25	8 x 10 <sup>-7</sup>	8.04E-07	
GO:0016301	kinase activity	1.24	1.6 x 10 <sup>-5</sup>	1.56E-05	
GO:0016740	transferase activity	1.2	4.6 x 10 <sup>-19</sup>	4.58E-19	
GO:0004871	signal transducer activity	1.19	1.6 x 10 <sup>-5</sup>	1.55E-05	

GO:0005515	protein binding	1.17	3.6 x 10 <sup>-48</sup>	3.64E-48
GO:0005509	calcium ion binding	1.17	1.4 x 10 <sup>-7</sup>	1.37E-07
GO:0000166	nucleotide binding	1.16	7.7 x 10 <sup>-16</sup>	7.72E-16
GO:0030528	transcription regulator activity	1.15	1.8 x 10 <sup>-5</sup>	1.76E-05
GO:0016787	hydrolase activity	1.13	8 x 10 <sup>-8</sup>	7.95E-08
GO:0030234	enzyme regulator activity	1.13	1.6 x 10 <sup>-4</sup>	1.55E-04
GO:0005215	transporter activity	1.12	8.9 x 10 <sup>-5</sup>	8.86E-05
GO:0003824	catalytic activity	1.1	1.4 x 10 <sup>-5</sup>	1.37E-05
Location				
Accession	GO Term	Fold	P-Value	P-Value
GO:0005815	microtubule organizing center	1.38	4 x 10 <sup>-4</sup>	4.03E-04
GO:0005768	endosome	1.34	3.1 x 10 <sup>-4</sup>	3.13E-04
GO:0005794	Golgi apparatus	1.33	4.1 x 10 <sup>-10</sup>	4.13E-10
GO:0005635	nuclear envelope	1.31	3 x 10 <sup>-4</sup>	2.99E-04
GO:0005578	extracellular matrix (sensu Metazoa)	1.31	5.5 x 10 <sup>-8</sup>	5.50E-08
GO:0005654	nucleoplasm	1.26	1.2 x 10 <sup>-7</sup>	1.15E-07
GO:0016023	cytoplasmic membrane-bound vesicle	1.25	1.1 x 10 <sup>-4</sup>	1.09E-04
GO:0005737	cytoplasm	1.18	1.3 x 10 <sup>-16</sup>	1.28E-16
GO:0005783	endoplasmic reticulum	1.15	3.8 x 10 <sup>-5</sup>	3.81E-05
GO:0005856	cytoskeleton	1.13	1.5 x 10 <sup>-5</sup>	1.51E-05
GO:0005886	plasma membrane	1.12	3.2 x 10 <sup>-9</sup>	3.20E-09
GO:0005615	extracellular space	1.11	1.5 x 10 <sup>-6</sup>	1.54E-06
GO:0043234	protein complex	1.11	2.6 x 10 <sup>-7</sup>	2.56E-07
GO:0005739	mitochondrion	1.11	6.9 x 10 <sup>-4</sup>	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)
<b>SINEs</b>				
	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
<b>LINEs</b>				
	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
<b>LTR elements</b>				
	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
<b>DNA elements</b>				
	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
<b>Unclassified</b>				
	63.83	8.235	0.45%	3

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

Repeat Name	Family	Number	Repeat Name	Family	Number
AmnSINE1	SINE	1975	MIRb	SINE/MIR	6131
BovB_Plat	LINE/RTE-BovB	9048	MIRc	SINE/MIR	17385
CR1_Mam	LINE/CR1	283	MamSINE1	SINE/tRNA	532
L1M5	LINE/L1	22	Mon1a1	SINE/MIR	139078
L1M6	LINE/L1	7	Mon1a2	SINE/MIR	142139
L1M7	LINE/L1	2	Mon1a3	SINE/MIR	25993
L1ME5	LINE/L1	16	Mon1a4	SINE/MIR	56169
L1MEf	LINE/L1	5	Mon1a5	SINE/MIR	62377
L1MEg	LINE/L1	6	Mon1a6	SINE/MIR	60012
L2	LINE/L2	1410	Mon1a7	SINE/MIR	72178
L2_Plat1a	LINE/L2	23022	Mon1d	SINE/MIR	198047
L2_Plat1b	LINE/L2	28748	Mon1e	SINE/MIR	70374
L2_Plat1c	LINE/L2	11445	Mon1f-1	SINE/MIR	28983
L2_Plat1d	LINE/L2	15113	Mon1f-2	SINE/MIR	19592
L2_Plat1e	LINE/L2	37643	Mon1f0	SINE/MIR	4433
L2_Plat1f	LINE/L2	31941	Mon1f1	SINE/MIR	4402
L2_Plat1g	LINE/L2	57870	Mon1f2	SINE/MIR	44356
L2_Plat1h	LINE/L2	76955	Mon1f3	SINE/MIR	215639
L2_Plat1i	LINE/L2	190532	Mon1f4	SINE/MIR	64475
L2_Plat1m	LINE/L2	566041	Mon1f5	SINE/MIR	277811
L2_Plat1n	LINE/L2	136675	Mon1g0	SINE/MIR	60692
L2_Plat1o	LINE/L2	131783	Mon1g1	SINE/MIR	327919
L2_Plat1q	LINE/L2	64066	Mon1g3	SINE/MIR	150701
L2_Plat1r	LINE/L2	57143	MonoRep87A	SINE	15458
L2_Plat1s	LINE/L2	57052	MonoRep87B	SINE	13715
L2_Plat1t	LINE/L2	39841	PlatCR1	LINE/CR1	28173
L2_Plat1u	LINE/L2	52015	PlatCR1_old1	LINE/CR1	2090
L2a	LINE/L2	3296	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

<b>Perfect Loci</b>					
	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
<b>Imperfect Loci</b>					
	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

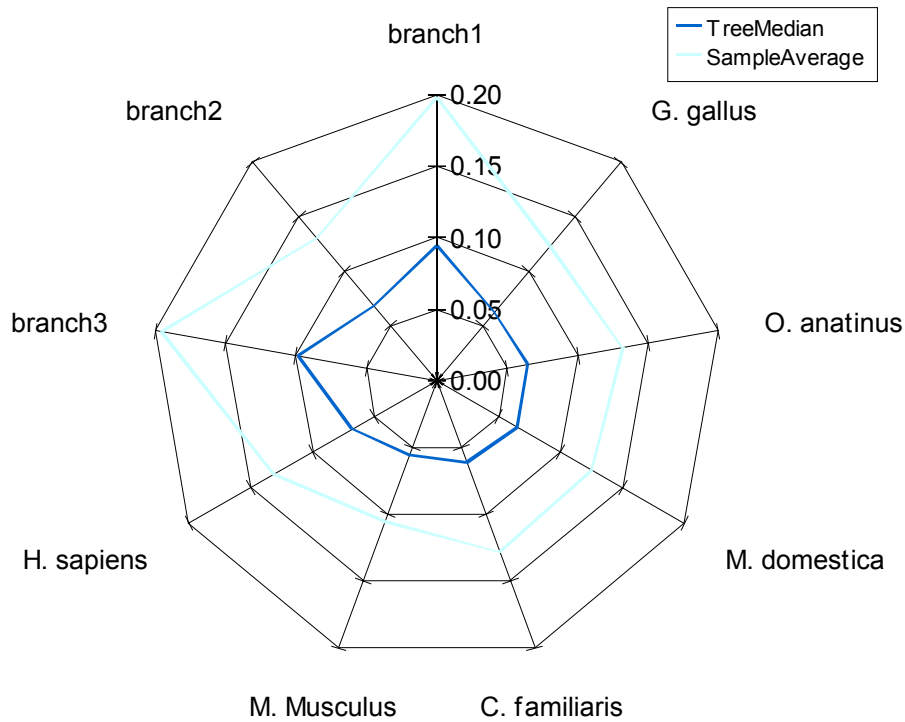
**Table 11.** Distribution of GC content by feature class / species.

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

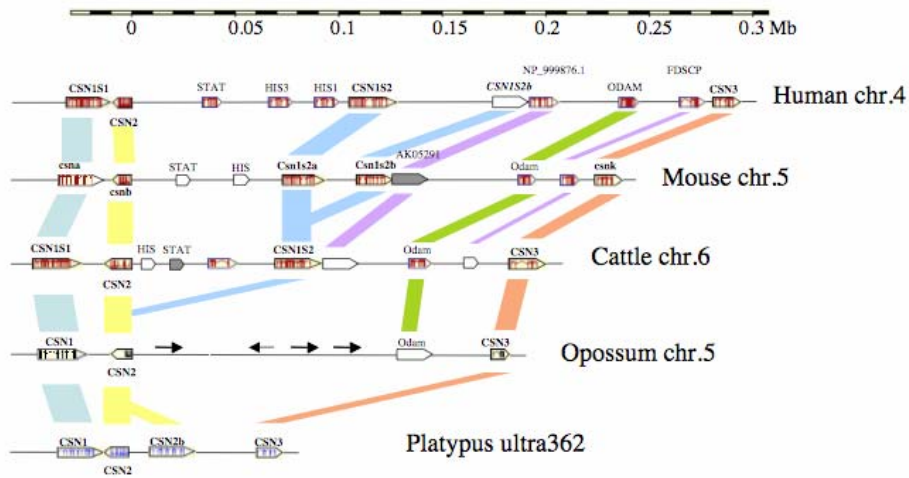
\* p(GC): p-value for test of whether mean of GC within feature class is the same as for background (ncnr-random) by a two sided t-test

\*\* 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

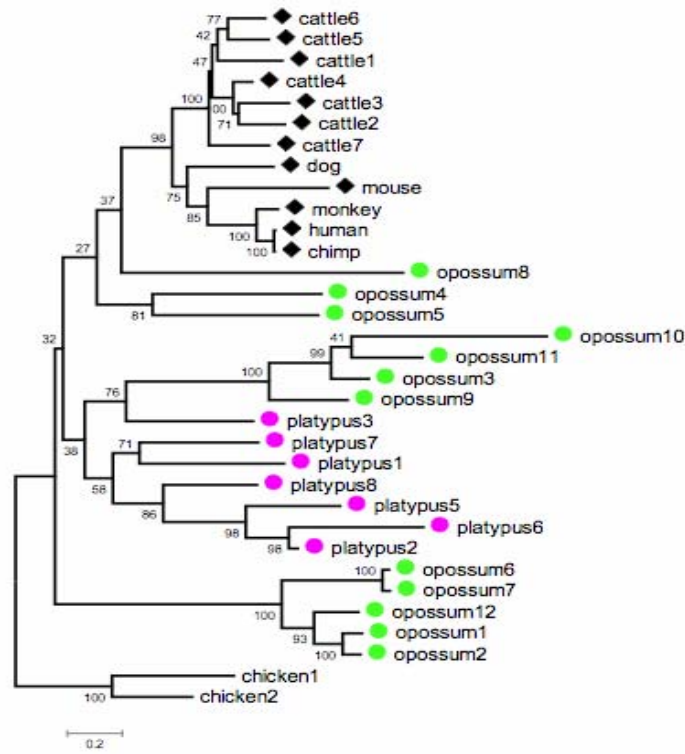
### 3. Supplementary Figures



**Figure 1.** Branch specific  $d_N/d_S$  ratios estimated with two different methods (see Accessory 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.

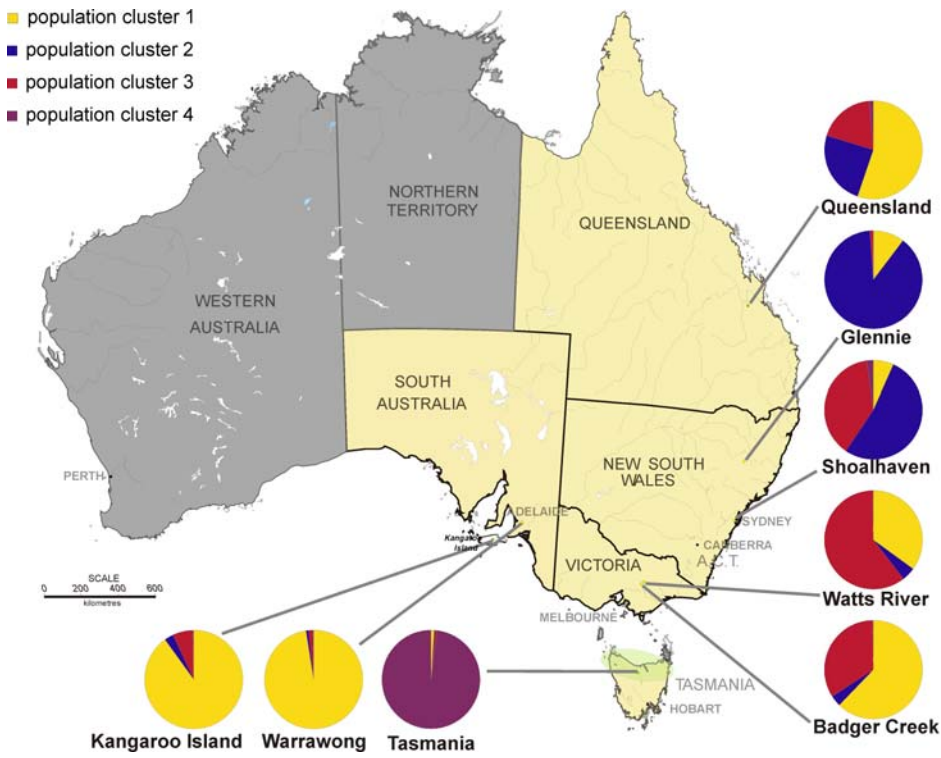


**Figure 2.** A comparison of the casein loci organization. The casein loci in platypus, opossum, cattle, mouse and human genomes are drawn approximately to scale and aligned on the beta-casein gene<sup>64</sup>. Genes are each represented by a box with an arrow pointing in the direction of gene transcription. 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).

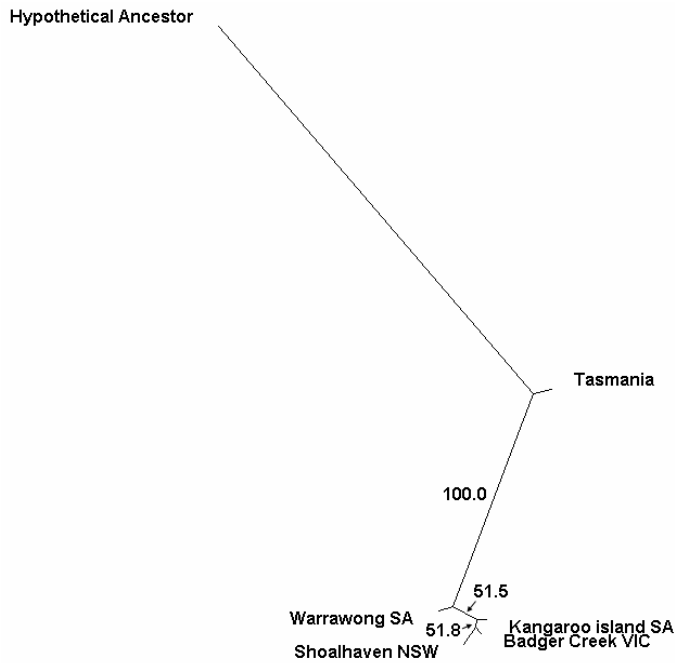


**Figure 3.** The cathelicidin antimicrobial peptide gene family has expanded and is highly heterogeneous in the platypus (pink) and opossum (green). These divergent peptides may provide marsupials and monotremes with a unique mechanism for protecting immunologically naïve young from pathogens (See Supplementary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

a

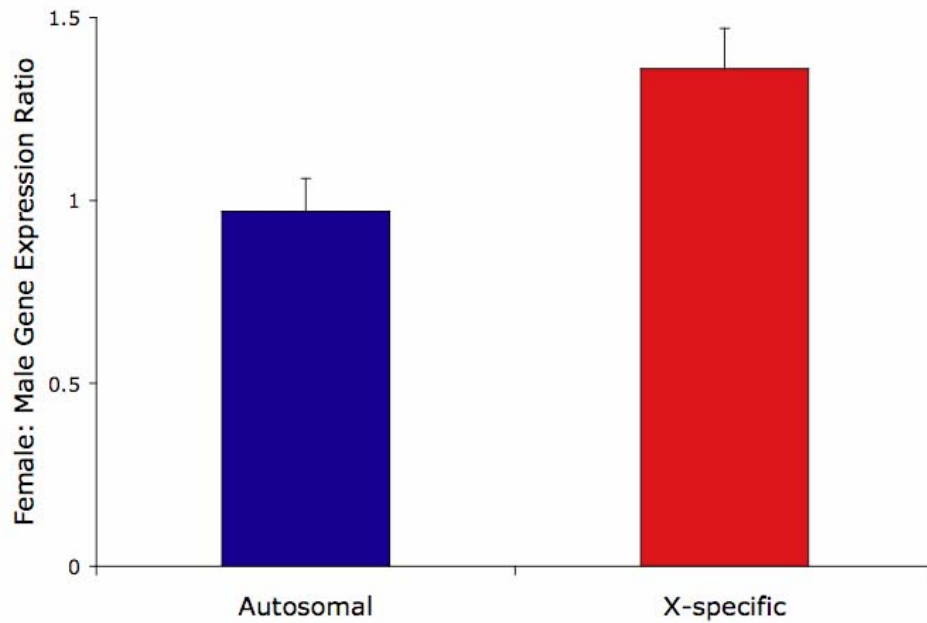


b



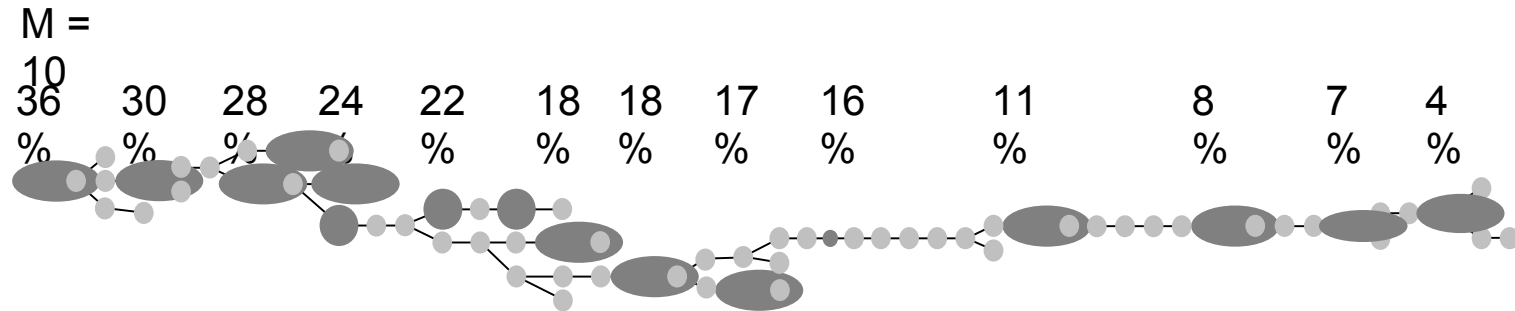
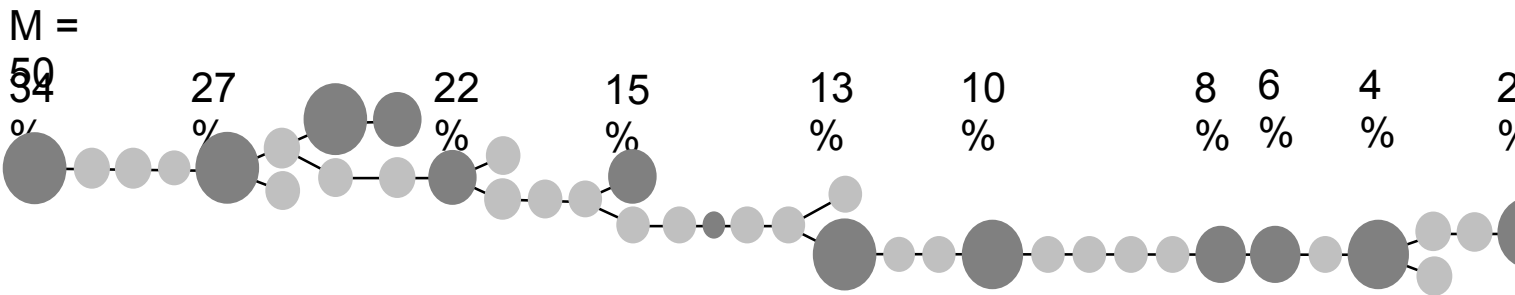
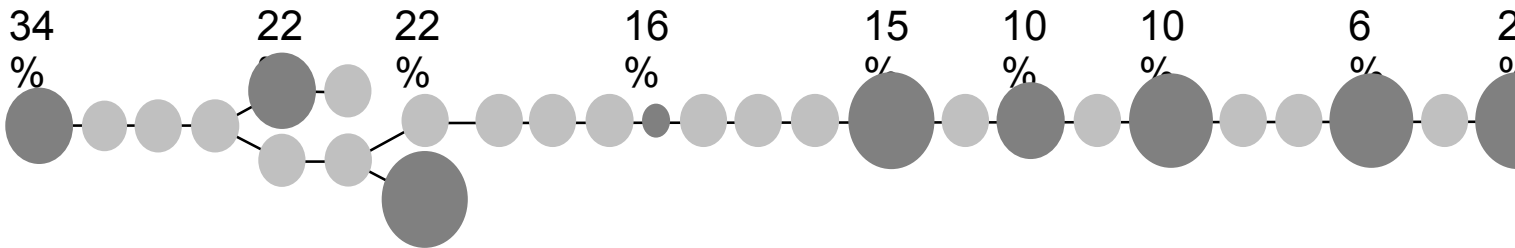
**Figure 4.** Platypus population structure analysis. Using Structure v2.1 software<sup>50</sup>, an analysis was performed with 57 polymorphic retrotransposon loci (LINE2 and Mon1) with 90 platypus samples from various regions in Australia including Tasmania were included in this study. **a**, Map of 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.



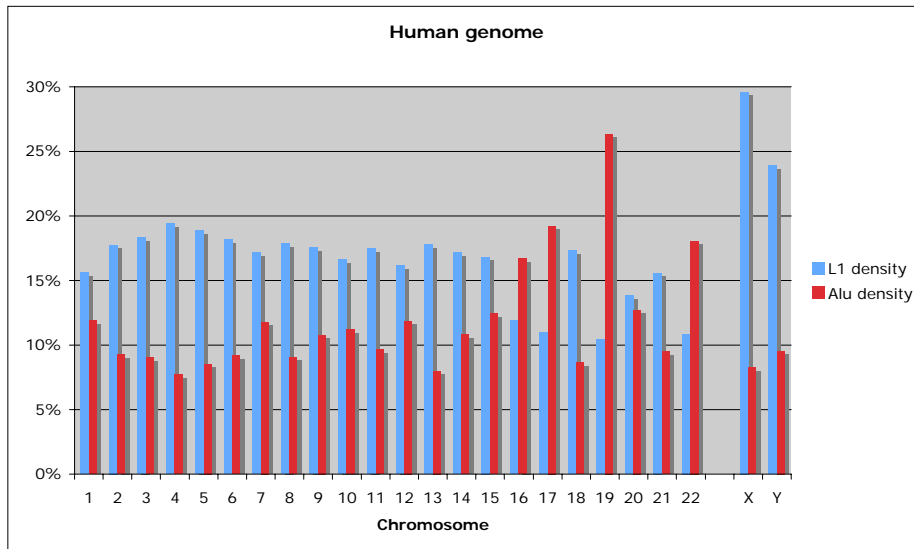
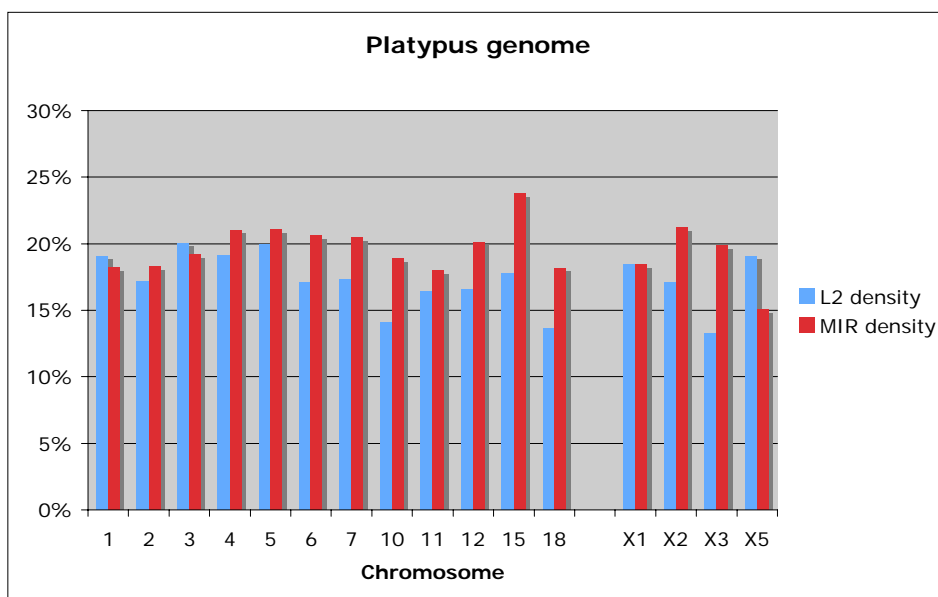


**Figure 5.** A comparison of female:male ratios for expression of platypus autosomal and X-specific genes in fibroblast cells. Real-time RT-PCR gene estimates were normalized to the autosomal ACTB housekeeping gene. Female:male ratios for autosomal genes was close to 1.0, whereas X-specific gene ratios 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.

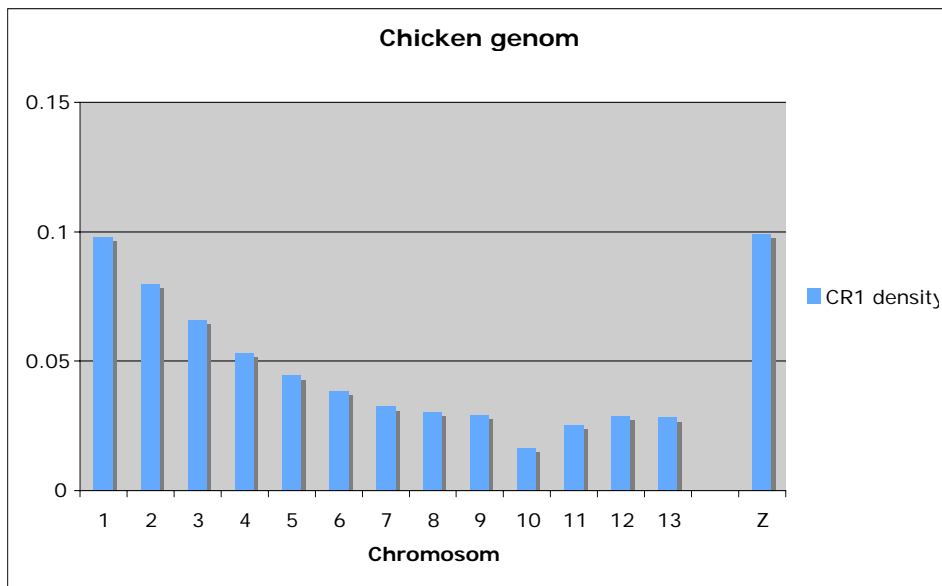
M = 70



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

**a****b**

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.

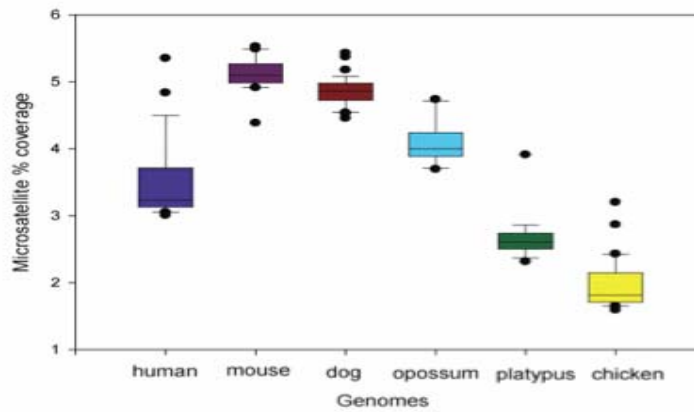
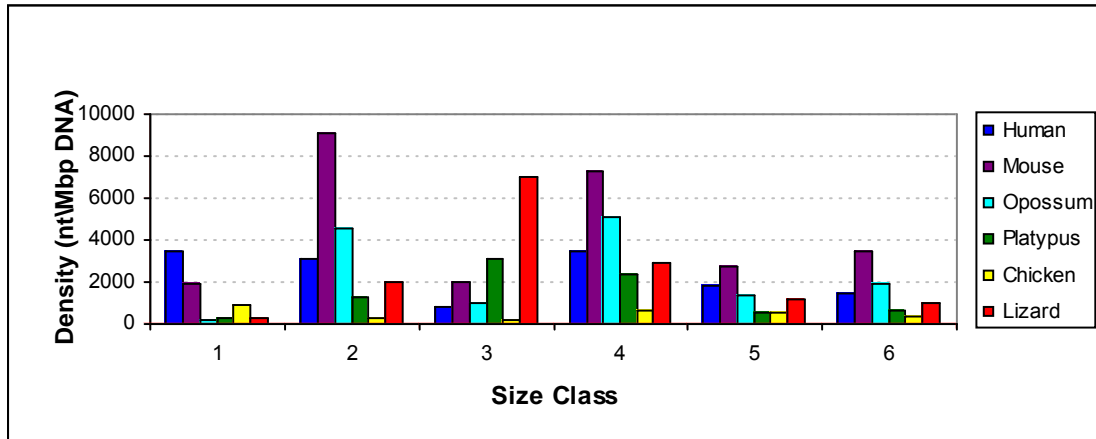
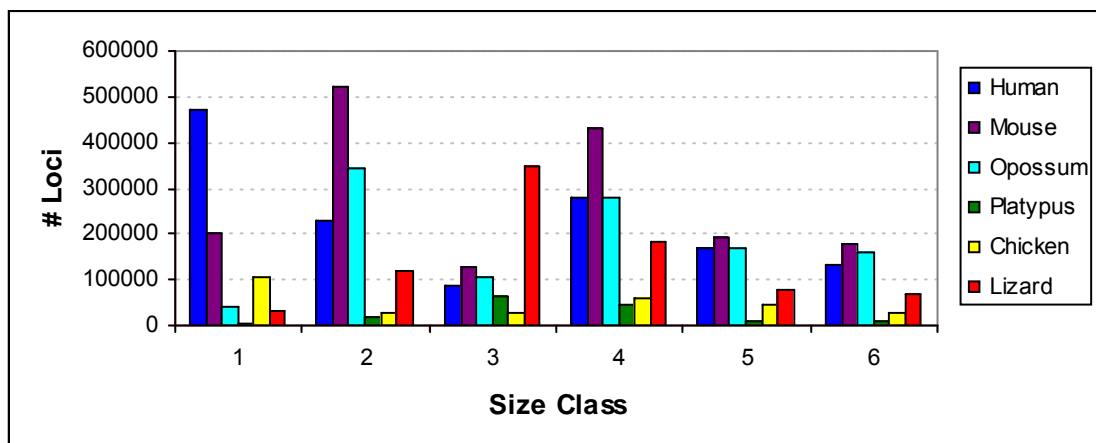


Figure 8. Platypus whole genome microsatellites. Coverage was compared across representative mammalian and avian genomes. For each species, the variation in microsatellite coverage by chromosome is represented by the box plot.

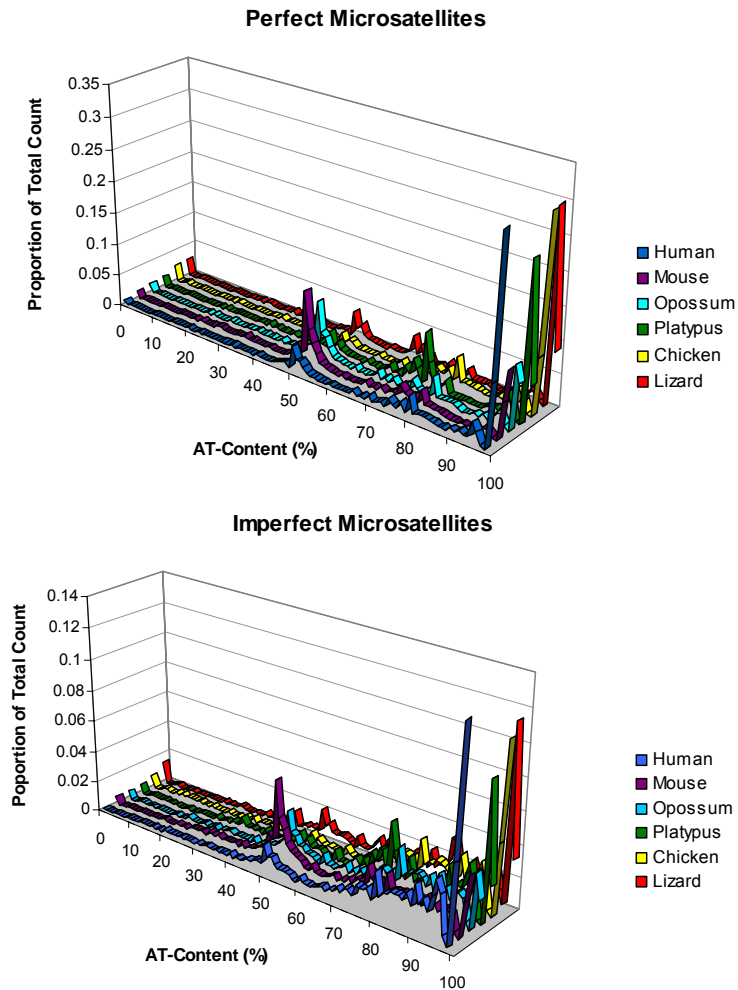
a



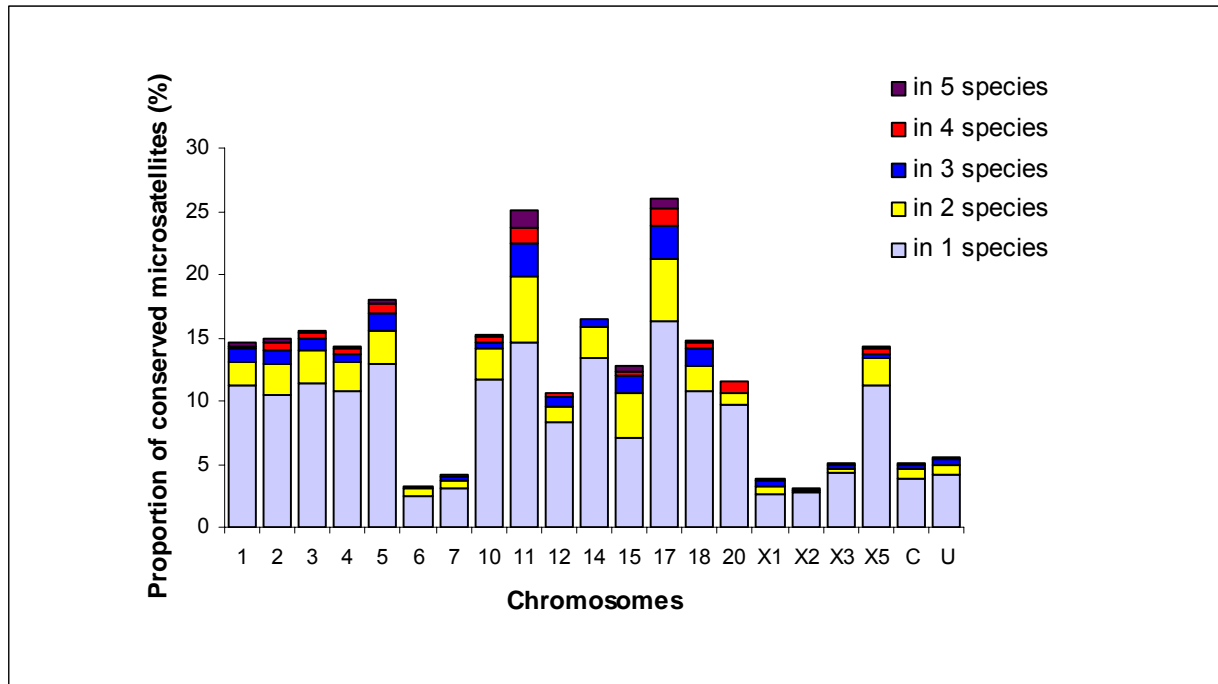
b



**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.



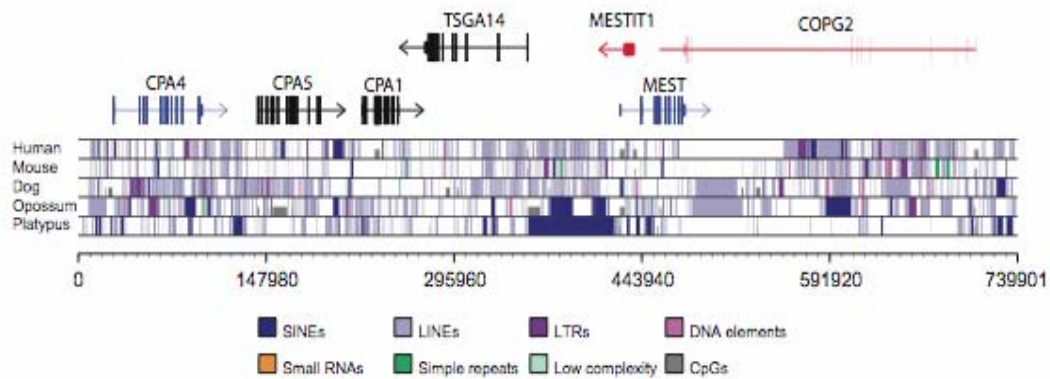
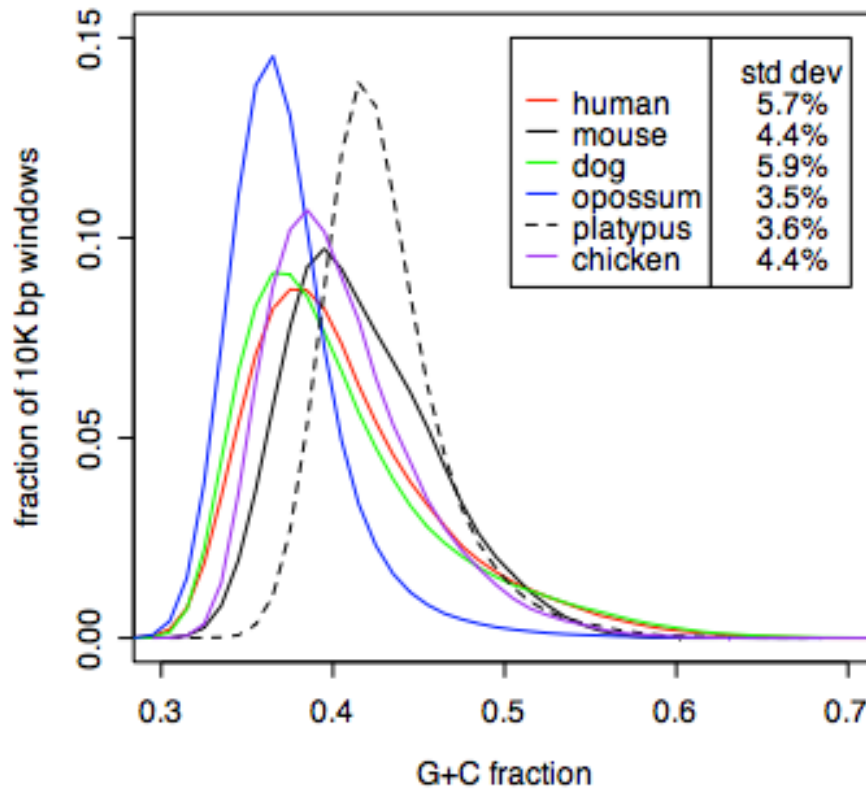
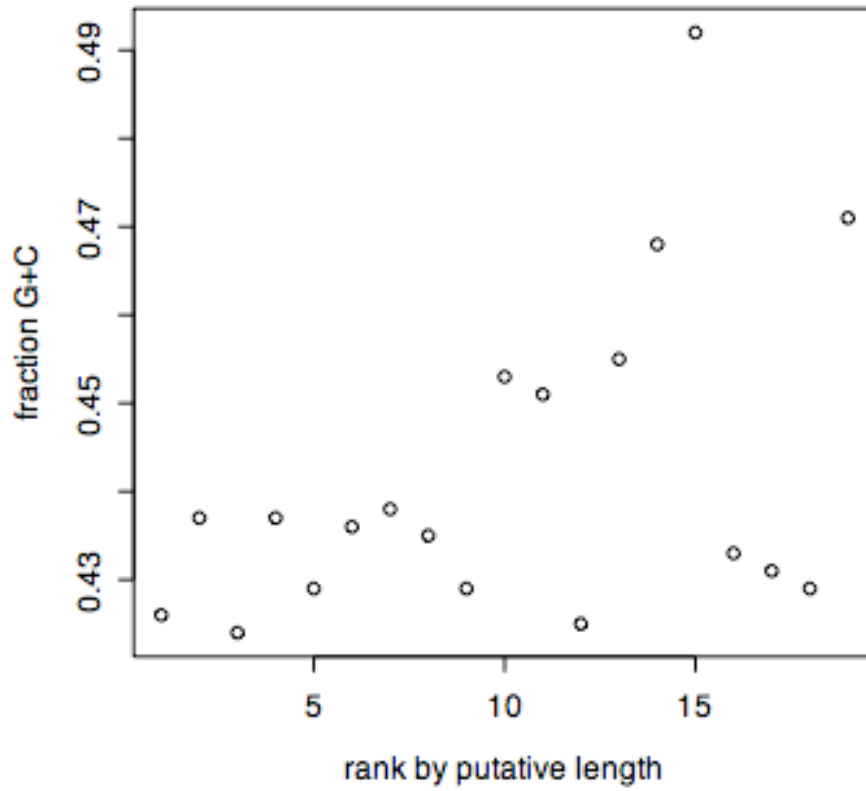


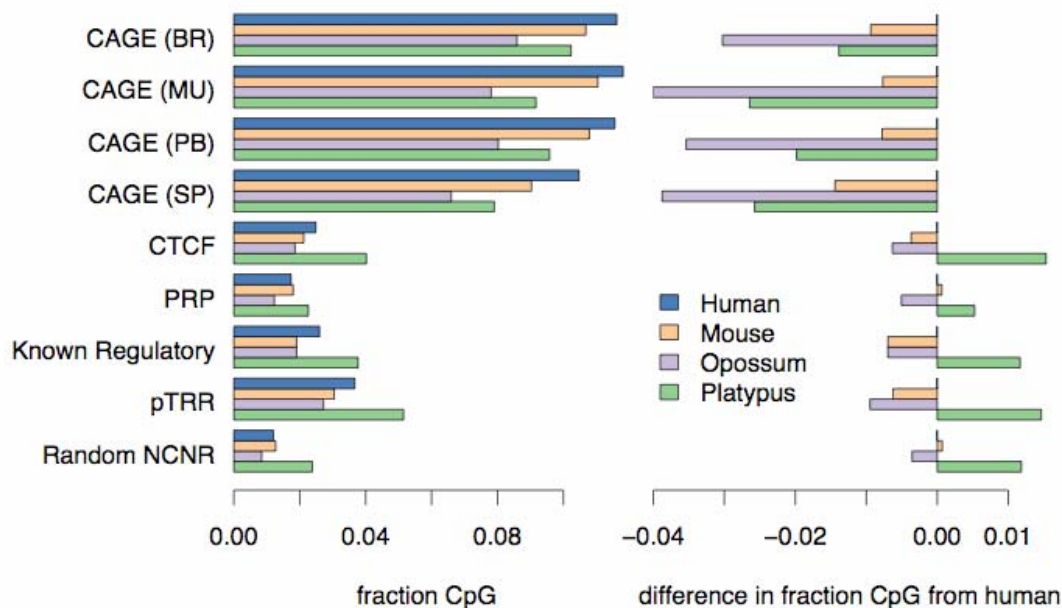
Figure 12. Repeat distribution plots for each mammalian species across the *PEG1/MEST* eutherian imprinted gene cluster. Gene structure is shown by a line (introns) connecting boxes (exons), transcripts in blue are paternally imprinted, red are maternally imprinted and black represents unknown or non-imprinted 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<sup>65</sup>.

#### 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 Bioinformatics* **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 hibernation on haematology 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. Tavaré, 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. Buschiazio, 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 Genet* **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).

## Accessory Tables

**Table A1.** C-values vs genome size.

Species	C value	Genome size (Gb)
Mouse	3.3	2.5
Opossum	4.3	3.4
Chimpanzee	3.76	3.0
Chicken	1.25	1.0

**Table A2.** Sequencing data by insert size.

Insert Size(kb)	Count number	Assembled Num	Trimmed bases	Paired read num	Paired bases
1.5-2.0	13688	12769	7864169	11266	7147302
2.0-2.5	30860	27868	16566633	23460	14456677
2.0-3.0	62952	57781	37405066	51330	33991476
2.0-4.0	123069	114217	78470191	104279	72503023
2.5-3.0	31829	28639	17244301	24360	15097909
3.0-3.5	15760	14514	8987717	12686	8086119
3.0-4.0	1298245	1166093	681919515	961656	578014809
3.5-4.0	573	321	132130	270	108856
4.0-4.3	11964296	11048205	7422999427	9907029	6759757330
4.3-6.5	7564965	7058321	5164607068	6203560	4585400739
6.5-9.5	4787407	4352061	2959216736	3667469	2531831207
40	652982	493809	304165771	311447	196298750
147	407649	379614	269859963	297082	213731642
total	26954275	24754212	16969438687	21575894	15016425839

**Table A3.** Sequence contiguity by species.

Platypus Contiguity stat	Ctg N50 len	Sctg N50 Len	Ctg N50 Num	Sctg N50 Num
platypus_051215(1.7G)	13,026	365,576	33,796	736
Mouse 7X(2.5G)	24,775	11,479,562	29,556	66
Chimp050130(2.9G) 6.7X	28,231	13,733,919	29,244	60
Chicken040612(1G)6.6X	38,867	10,918,580	6,835	26
chimp050412(no fosmid 3.0G)	28,238	3,323,878	29,362	211
Platypus Contiguity stat	Largest Ctg	Ave. Ctg	Largest Sctg	Ave. Sctg
platypus_051215(1.7G)	245,967	6,278	14,341,343	28,661
Mouse 7X(2.5G)	225,618	11,576	44,729,073	78,832
Chimp050130(2.9G) 6.7X	380,600	9,425	59,100,486	29,107
Chicken040612(1G)6.6X	441,790	12,007	50,876,398	42,110
chimp050412(no fosmid 3.0G)	288,304	8,765	25,457,179	21,251



**Table A4.** Total sequence assigned to platypus chromosomes.

Sequence per  
Chromosome

Chr	Sequence*
1	47594283
2	54797317
3	59581953
4	58987262
5	24609220
6	16302927
7	40039088
10	11243762
11	6809224
12	15872666
14	2696122
15	3786880
17	1399469
18	6611290
20	1816412
X1	45541551
X2	5652501
X3	5951358
X5	27786739
Un	1559714169

\*Sequence lengths include gap size estimates

**Table A5.** G+C content distribution of finished BACs: “Covered” vs “Not covered” regions.

%G+C	Not covered		Covered	
	num	%age	num	%age
10-19	2	0.1	0	0
20-29	8	0.6	0	0
30-39	66	4.9	54	3.7
40-49	592	44	1045	70.7
50-59	438	32.6	283	19.1
60-69	181	13.5	85	5.8
70-79	51	3.8	11	0.7
80-89	6	0.4	0	0

**Table A6.** Repeat content based on the platypus RepeatMasker library.**Covered**

Covered regions of finished platypus BACs aligning to the WGS sequence

total_length:	11687519		
GC_level:	44.38		
bases_masked:		5268945	45.08
SINEs:	13041	2516541	21.53
LINEs:	12251	2479459	21.21
LTR_elements:	25	13698	0.12
DNA_elements:	354	72473	0.62
Unclassified:	443	51518	0.44
---			
Total_interspersed_repeats:	NA	5133689	43.92
Small_RNA:	0	0	0

Satellites:	71	7117	0.06
Simple_repeats:	1907	77311	0.66
Low_complexity:	795	51330	0.44

**Not Covered (Gaps)**

(Regions of finished platypus BACs not aligning to the WGS sequence)

Total_length:	1213497		
GC_level:	49.85		
Bases_masked:		765383	63.07
SINES:	1635	271373	22.36
LINEs:	1463	454461	37.45
LTR_elements:	5	2028	0.17
DNA_elements:	23	4300	0.35
Unclassified:	36	2948	0.24
---			
Total_interspersed_repeats:	NA	735110	60.58
Small_RNA:	0	0	0
Satellites:	4	613	0.05
Simple_repeats:	279	12463	1.03
Low_complexity:	215	17256	1.42

**Table A7.** Platypus segmental duplication analysis by chromosome assignment.

chrom	size	size (gap-free)	All WGAC (>90% > 1 kb)	WGAC (<94% <10 kb)	WGAC (>94% > 10 kb)	WSSD	WGAC+WSSD	Total SD	% SD
	47,594,283	44,500,396	2,409,545	44,445	55,609	0	0	44,445	0.10%
chr2	54,797,317	50,943,648	2,947,515	61,079	107,719	74,917	53,300	114,379	0.22%
chr3	59,581,953	55,753,503	3,970,872	157,160	254,692	170,811	130,311	287,471	0.52%
chr4	58,987,262	55,339,946	3,373,590	101,053	126,600	0	0	101,053	0.18%
chr5	24,609,220	22,751,819	1,051,886	19,770	23,747	0	0	19,770	0.09%
chr6	16,302,927	15,302,526	1,075,579	31,053	63,676	73,252	58,449	89,502	0.58%
chr7	40,039,088	37,716,602	2,428,143	70,516	120,910	78,862	55,796	126,312	0.33%
chr10	11,243,762	10,686,721	728,385	16,340	11,036	0	0	16,340	0.15%
chr11	6,809,224	6,329,107	319,445	6,126	0	0	0	6,126	0.10%
chr12	15,872,666	15,079,184	1,186,676	48,638	128,452	72,624	48,514	97,152	0.64%
chr14	2,696,122	2,492,320	38,725	0	10,912	15,000	0	0	0.00%
chr15	3,786,880	3,543,753	154,816	5,858	0	0	0	5,858	0.17%
chr17	1,399,469	1,274,415	32,044	2,187	0	0	0	2,187	0.17%
chr18	6,611,290	6,314,477	553,310	2,073	21,649	0	0	2,073	0.03%
chr20	1,816,412	1,654,293	92,022	0	0	0	0	0	0.00%
chrX1	45,541,551	42,774,007	2,477,972	120,243	122,529	0	0	120,243	0.28%
chrX2	5,652,501	5,467,671	414,593	0	0	23,373	0	0	0.00%
chrX3	5,951,358	5,824,245	507,442	1,854	0	0	0	1,854	0.03%
chrX5	27,786,739	25,788,079	199,997	14,873	0	10,133	0	14,873	0.06%
subtotal	437,080,024	409,536,712	23,962,557	703,268	1,047,531	518,972	346,370	1,049,638	0.26%
all contigs	894,359,526	786,450,919	121,274,573	24,988,227	14,207,330	12,907,044	7,460,009	32,448,236	4.13%
all ultras	689,150,142	646,450,220	45,273,432	2,250,071	3,692,305	3,145,198	2,219,933	4,470,004	0.69%
total	2,020,589,692	1,842,437,851	190,510,562	27,941,566	18,947,166	16,571,214	10,026,312	37,967,878	2.06%

Total SD=total duplication content based on all WGAC duplications <94% (column 6) and all duplications >94% and > 10

kb that are confirmed by both WGAC and WSSD. % SD represents a lower bound estimate of duplication content.

**Table A8.** Chimeric clones in chimp and platypus assemblies.

	Chimp	Platypus	Platypus/Chimp
Chimera	83,570	196,953	2.36
Contig_Ends	1,083,550	1,092,049	1.01
Both	2,395	5,213	2.18

\*both=chimeric reads at both ends of a given contig

**Table A9.** Orientation of reads at contig ends.

count	platy_start	percentage
119,456	0_0_0	23.46
56,575	0_0_1	11.11
76,137	0_1_0	14.95
21,594	0_1_1	4.24
41,472	1_0_0	8.14
13,198	1_0_1	2.59
139,724	1_1_0	27.44
41,127	1_1_1	8.08
509,283	total	100.00

count	chimp_start	percentage
111,557	0_0_0	22.12
44,473	0_0_1	8.82
97,379	0_1_0	19.31
32,511	0_1_1	6.45
58,706	1_0_0	11.64
23,660	1_0_1	4.69
98,851	1_1_0	19.60
37,155	1_1_1	7.37
504,292	total	100.00

count	platy_end	percentage
41,391	0_0_0	8.13
138,810	0_0_1	27.26
13,440	0_1_0	2.64
42,258	0_1_1	8.30
123,856	1_0_0	24.32
71,021	1_0_1	13.95
57,381	1_1_0	11.27
21,126	1_1_1	4.15
509,283	total	100.00

count	chimp_end	percentage
56,761	0_0_0	11.26
94,667	0_0_1	18.77
23,784	0_1_0	4.72
36,854	0_1_1	7.31
117,301	1_0_0	23.26
97,954	1_0_1	19.42
44,467	1_1_0	8.82
32,504	1_1_1	6.45
504,292	total	100.00

**Table A10.** Ensembl gene sets used for orthology analysis

Species	Genes	Transcripts	Assembly
<i>H. sapiens</i>	22810	43605 ncbi36 (hg18)	
<i>M. musculus</i>	24442	32078 ncbi36 (m36)	
<i>C. familiaris</i>	19314	25568 canFam2, Broad Institute (May 2005)	
<i>M. domestica</i>	19597	32690 monDom4, Broad Institute	
<i>O. anatinus</i>	18597	27557 WashU assembly	
<i>G. gallus</i>	16715	22186 galGal3, Chicken genome sequencing consortium (May 2006)	

**Table A11.** Simple 1:1 orthologs across the amniota. Each set contains exactly one gene from each species.

Species	(1:1)n ortholog sets	
<i>H.sapiens</i>	22,611	100%
+ <i>M. musculus</i>	14,905	66%
+ <i>C. familiaris</i>	13,788	61%
+ <i>M. domestica</i>	12,042	53%
+ <i>O. anatinus</i>	9,264	41%
+ <i>G. gallus</i>	7,587	34%

**Table A12.** Different ways to estimate genome-wide substitution rates.

<u>Method</u>	<u>Aggregation</u>	<u>Ortholog sets</u>	<u>Fit on tree</u>
Method 1 / Pairwise	Median over gene pairs	Pairwise	No
Method 2 / PairwiseStrict	Median over gene pairs	Multiple	No
Method 3 / TreeMedian	Median over gene trees	Multiple	Yes
Method 4 / Sample Average	20 samples of 200 concatenated gene alignments	Multiple	Yes

**Table A13.** Estimated genome-wide  $d_S$  and  $d_N$  for different methods (Supplementary Notes S18). Pairwise rates between genomes have been read off the tree for tree based methods.

<u>Species</u>	<u>Species</u>	<u>Pairwise</u>		<u>PairwiseStrict</u>		<u>TreeMedian</u>		<u>SampleAverage</u>	
		<u>dS</u>	<u>dN</u>	<u>dS</u>	<u>dN</u>	<u>dS</u>	<u>dN</u>	<u>dS</u>	<u>dN</u>
<i>C. familiaris</i>	<i>G. gallus</i>	1.741	0.053	1.547	0.043	1.411	0.032	1.314	0.056
<i>C. familiaris</i>	<i>H. sapiens</i>	0.369	0.041	0.338	0.032	0.315	0.022	0.344	0.044
<i>C. familiaris</i>	<i>M. domestica</i>	1.113	0.063	0.986	0.051	0.867	0.036	0.879	0.067
<i>C. familiaris</i>	<i>M. musculus</i>	0.667	0.088	0.637	0.074	0.596	0.062	0.605	0.106
<i>C. familiaris</i>	<i>O. anatinus</i>	1.461	0.103	1.277	0.087	1.126	0.076	1.121	0.129
<i>G. gallus</i>	<i>H. sapiens</i>	1.642	0.094	1.469	0.078	1.381	0.062	1.292	0.107
<i>G. gallus</i>	<i>M. domestica</i>	1.586	0.099	1.426	0.083	1.345	0.088	1.254	0.145
<i>G. gallus</i>	<i>M. musculus</i>	1.891	0.111	1.687	0.092	1.662	0.102	1.553	0.167
<i>G. gallus</i>	<i>O. anatinus</i>	1.707	0.102	1.497	0.086	1.442	0.088	1.311	0.145
<i>H. sapiens</i>	<i>M. domestica</i>	1.037	0.095	0.918	0.080	0.837	0.082	0.858	0.136
<i>H. sapiens</i>	<i>M. musculus</i>	0.575	0.112	0.543	0.097	0.534	0.105	0.530	0.163
<i>H. sapiens</i>	<i>O. anatinus</i>	1.406	0.124	1.220	0.107	1.096	0.119	1.100	0.185
<i>M. domestica</i>	<i>M. musculus</i>	1.305	0.119	1.199	0.103	1.118	0.105	1.119	0.163
<i>M. domestica</i>	<i>O. anatinus</i>	1.308	0.111	1.152	0.097	1.061	0.099	1.062	0.154
<i>M. musculus</i>	<i>O. anatinus</i>	1.696	0.107	1.454	0.093	1.377	0.105	1.361	0.158

**Table A14.** Branch-specific genome-wide  $d_N/d_S$  values across the amniota for two different methods (Methods 3 and 4; see **Table A13**).

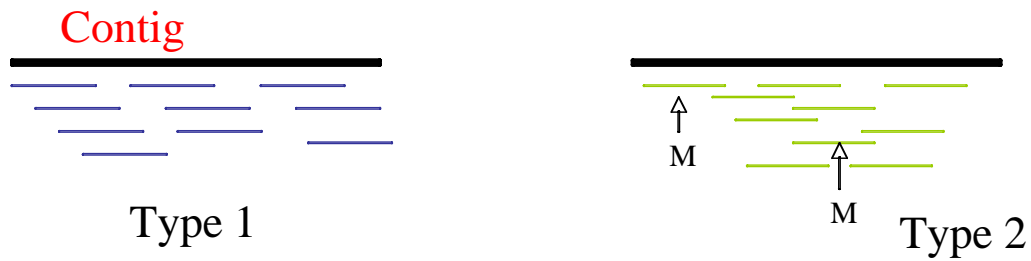
<i>Branch</i>	<i>TreeMedian</i>		<i>SampleAverage</i>	
	<b>Median</b>	<b>Stddev</b>	<b>Mean</b>	<b>Stddev</b>
branch1	0.095	76.03	0.199	0.020
branch2	0.068	22.58	0.130	0.010
branch3	0.098	69.73	0.195	0.024
<i>H. sapiens</i>	0.068	15.23	0.132	0.010
<i>M. Musculus</i>	0.056	12.5	0.105	0.008
<i>C. familiaris</i>	0.061	21.68	0.128	0.009
<i>M. domestica</i>	0.065	19.96	0.125	0.008
<i>O. anatinus</i>	0.065	13.92	0.132	0.010
<i>G. gallus</i>	0.062	10.86	0.123	0.008



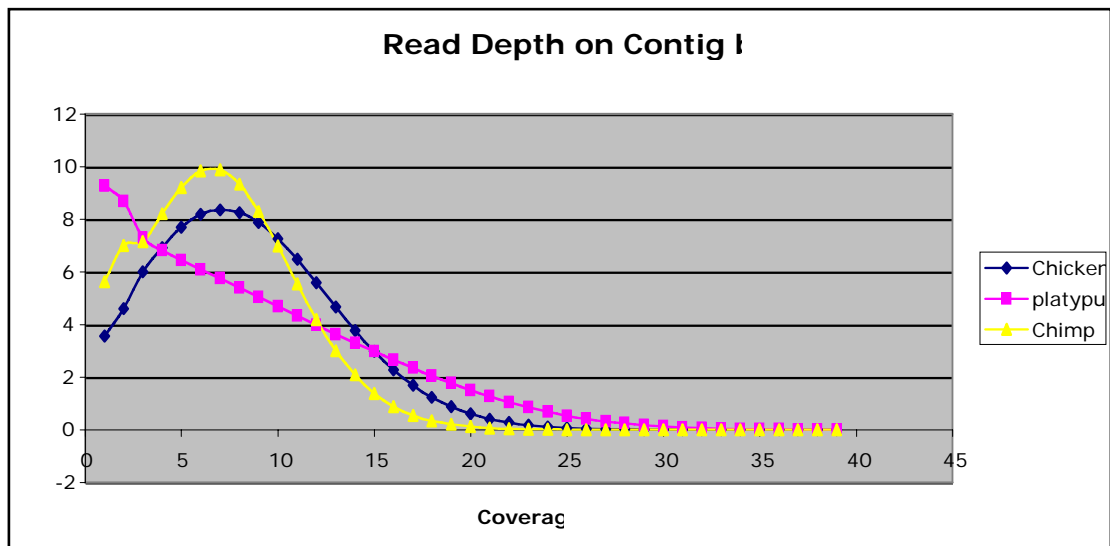
**Table A15.** Alignment of a region of the *PTPN4* gene showing a three-amino-acid deletion believed to have occurred after marsupials and eutherians diverged from monotremes.

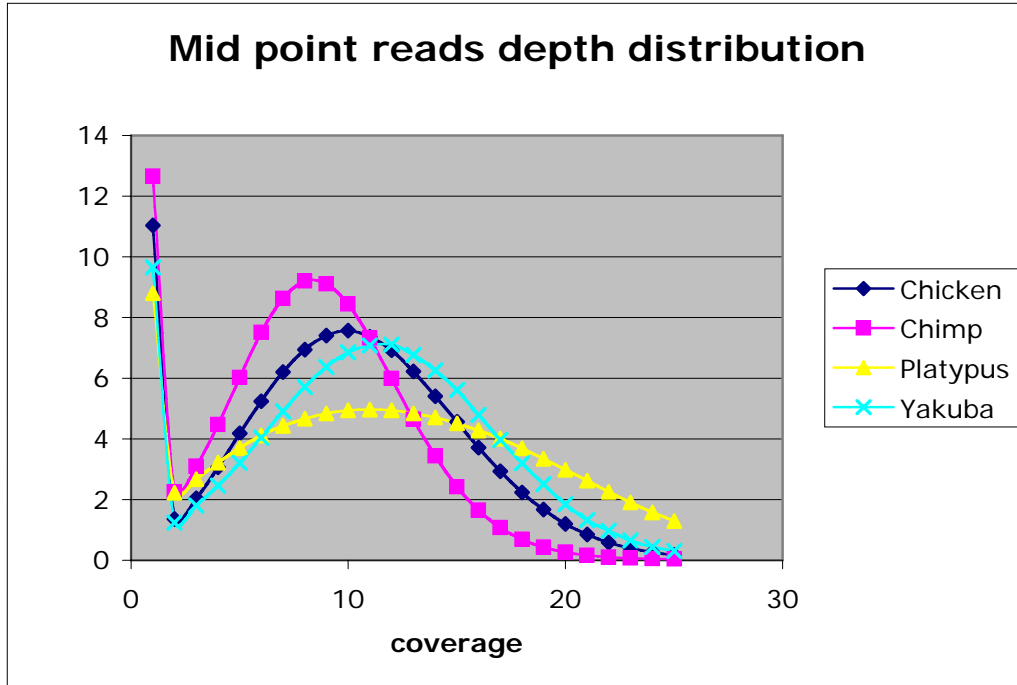
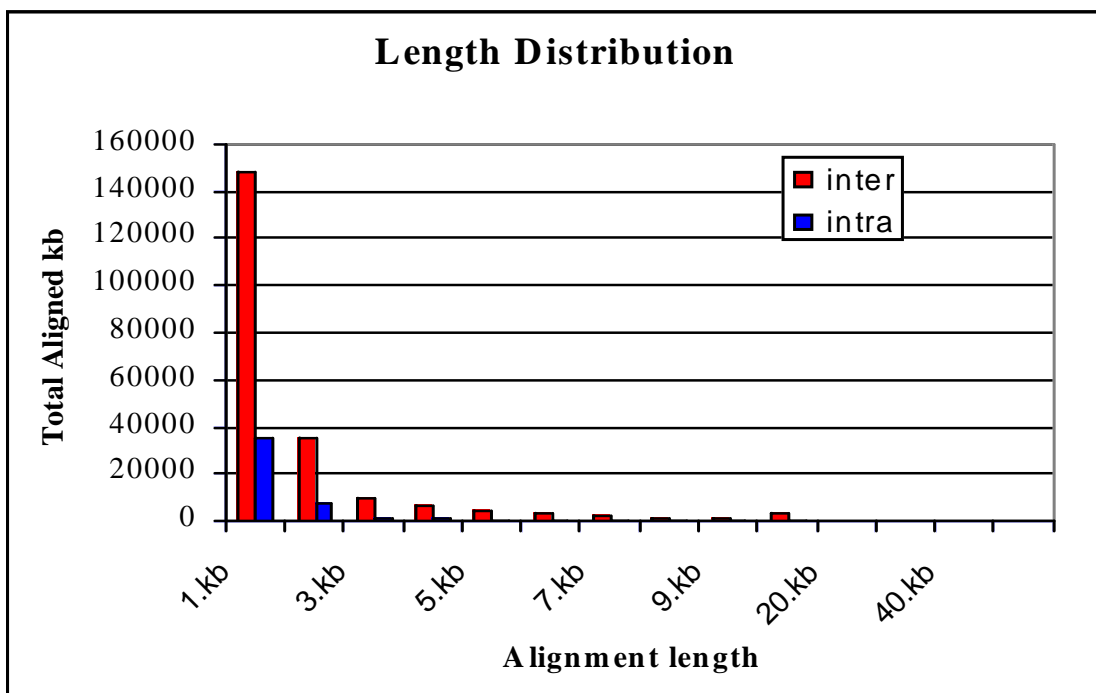
SPSKPLARKLM---DWEVSRNSISDDRLETQSLPSRSPPGTPN	<i>sapiens</i> (human)
SPSKPLARKLM---DWEVSRNSISDDRLETQSLPSRSPPGTPN	<i>trogodytes</i> (chimp)
SPSKPLARKLM---DWEVSRNSISDDRLETQSLPSRSPPGTPN	<i>gorilla</i> (gorilla)
SPSKPLARKLM---DWEVSRNSISDDRLETQSLPSRSPPGTPN	<i>pygmaeus</i> (orang_sumatran)
SPSKPLARKLM---DWEVSRNSISDDRLETQSLPSRSPPGTPN	<i>leucogenys</i> (gibbon)
SPSKPLARKLM---DWEVSRNSISDDRLETQSLPSRSPPGTPN	<i>mulatta</i> (rhesus)
SPSKPLARKLM---DWEVSRNSISDDRLETQSLPSRSPPGTPN	<i>jacchus</i> (marmoset_nwm)
SPSKPLARKLM---DWEVSRNSISDDRLETQSLPSRSPPGTPN	<i>syrichta</i> (tarsier)
SPSKPLARKLM---DWEVSRNSVSDDRLETQSLPSRSPPGTPN	<i>murinus</i> (mouse_lemur)
SPSKPLVRKLM---DWEVSRNSVSDDRLETQSLPSRSPPGTPN	<i>belangeri</i> (tree_shrew)
SPSKPLARKLM---DWEVSRNSISDDRLETQSLPSRSPPGTPN	<i>musculus</i> (mouse)
SPSKPLARKLM---DWEVSRNSISDDRLETQSLPSRSPPGTPN	<i>norvegicus</i> (rat)
SPSKPLARKLM---DWEVSRNSISDDRLETQSLPSRSPPGTPN	<i>porcellus</i> (guinea_pig)
SPSKPLARKLM---DWEVSRNSISDDRLETQSLPSRSPPGTPN	<i>tridecemlineatus</i> (ground_squirrel)
SPSKPLARKLM---DWEVSRNSVSDDRLETQSLPSRSPPGTPN	<i>ordii</i> (kangaroo_rat)
SPSKPLARKLM---DWEVSRNSISDDRLETQSLPSRSPPGTPN	<i>cuniculus</i> (rabbit)
SPSKPLARKLM---DWEVSRNSISDDRLETQSLPSRSPPGTPN	<i>princeps</i> (pika)
SPSKPLARKLM---DWEVSRNSISDDRLETQSLPSRSPPGTPN	<i>familiaris</i> (dog)
SPSKPLARKLM---DWEVSRNSISDDRLETQSLPSRSPPGTPN	<i>catus</i> (cat)
SPSKPLARKLM---DWEVSRNSISDDRLETQSLPSRSPPGTPN	<i>caballus</i> (horse)
SPSKPLARKLM---DWEVSRNSVSDDRLETQSLPSRSPPGTPN	<i>lucifugus</i> (microbat)
SPSKPLARKLM---DWEVSRNSISDDRLEnQSLPSRSPPGTPN	<i>vampyrus</i> (macrobat)
SPSKPLARKLM---DWEVSRNSISDDRLETQSLPSRSPPGTPN	<i>taurus</i> (cow)
SPSKPLARKLM---DWEVSRNSISDDRLETQSLPSRSPPGTPN	<i>aries</i> (sheep)
SPSKPLARKLM---DWEVSRNSISDDRLETQSLPSRSPPGTPN	<i>truncatus</i> (dolphin)
SPSKPLARKLM---DWEVSRNSISDDRLETQSLPSRSPPGTPN	<i>scrofa</i> (pig)
SPSKPLARKLM---DWEVSRNSISDDRLETQSLPSRSPPGTPN	<i>novemcinctus</i> (armadillo)
SPSKPLARKLM---DWEVSRNSISDDRLETQSLPSRSPPGTPN	<i>hoffmanni</i> (sloth)
SPSKPLARKLM---DWEVSRNSISDDRLETQSLPSRSPPGTPN	<i>capensis</i> (hyrax)
SPSKPLARKLM---DWEIVSRISLSDDRLETQSLPSRSPPGTPN	<i>telfairi</i> (tenrec)
SPSKPLARKLM---DWEVSRNSLSDERLETQSLPSRSPPGTPN	<i>domestica</i> (opossum)
SPSKPLARKLM---DWEVSRNSLSDERLETQSLPSRSPPGTPN	<i>eugenii</i> (wallaby)
SPSKPLARKLMGMDWEVSRNSISDDRLETQSLPSRSPPGTPN	<i>anatinus</i> (platypus)
SPSKPLARKLMGMDWEVSRNSISDDRLETQSLPSRSPPGTPN	<i>gallus</i> (chicken)
SPSKPLARKLMGMDWEVSRNSISDDRLETQSLPSRSPPGTPN	<i>platyrhynchos</i> (duck)
SPSKPLARKVMGMDWEVSRNSISDDRLETQSLPSRSPPGTPN	<i>guttata</i> (finch)
SPSKPLARKLISGMDWEVSRNSISDDRLETQSLPSRSPPGTPN	<i>carolinensis</i> (lizard)
SPSKPLSRKLVSEMDWEVSRNSLSNDRLETQSLPSRSPPGTPN	<i>tropicalis</i> (frog)
SPSKPLARKLLGGTDWESVSRNSLSDERLETQSLPTRSPPGTPN	<i>rubripes</i> (fugu)
SPSKPLARKLLGGTDWESVSRNSLSDERLETQSLPTRSPPGTPN	<i>nigroviridis</i> (pufferfish)
SPSKPLARKLMGGTDWESVSRNSLSDERLETQSLPTRSPPGTPS	<i>latipes</i> (medaka)
SPSKPLARKLMGVTDWEAVSRNSLSEERLETRSLPTRSPPGTPN	<i>rerio</i> (zebrafish)
SPSKPLARKLVGGTDWESVSRNSLSDERLETQSLPTRSPPGTPN	<i>aculeatus</i> (stickleback)

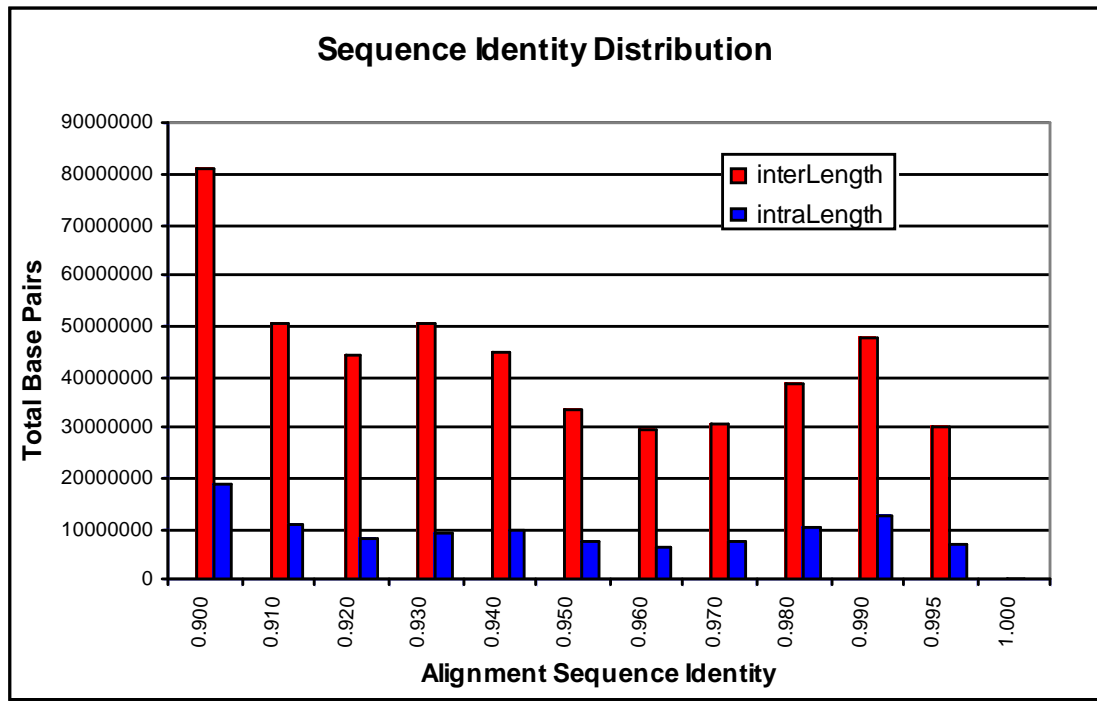
**Figure A1.** Describes format followed for read depth distribution.

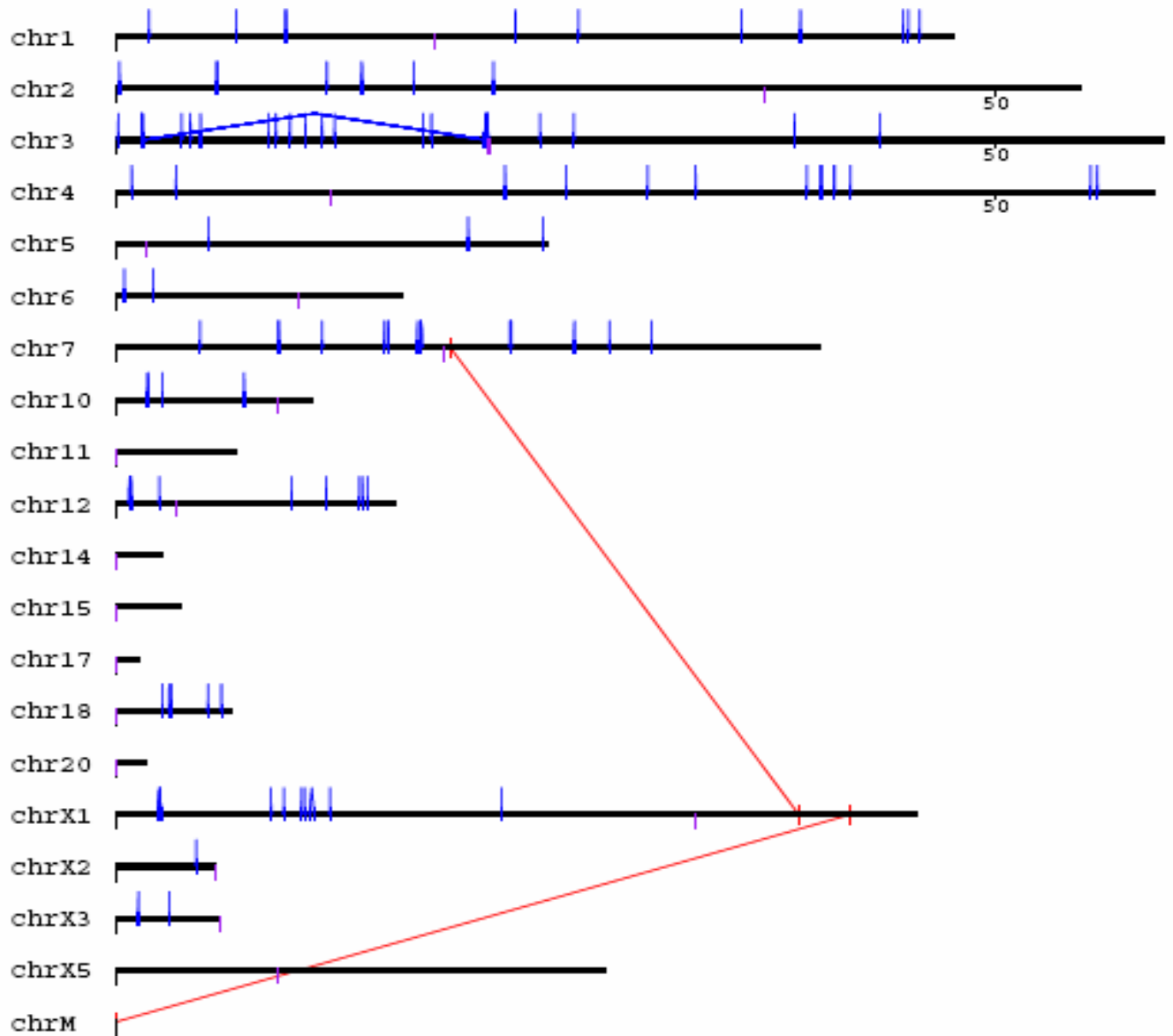


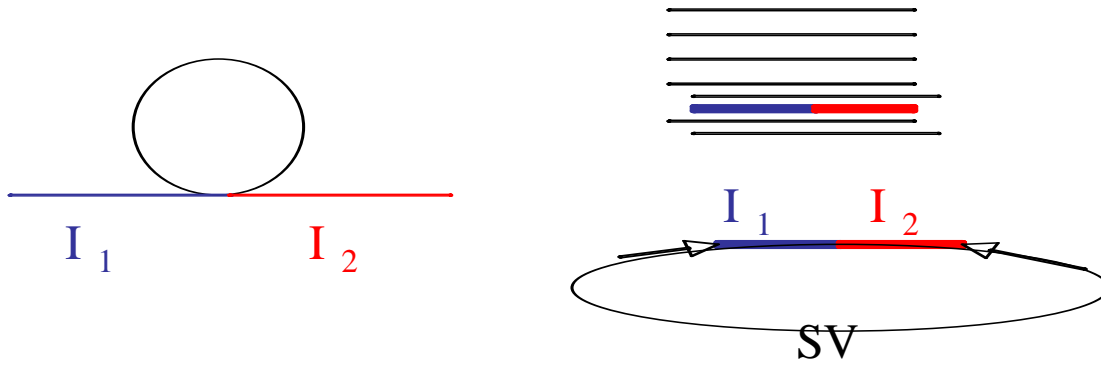
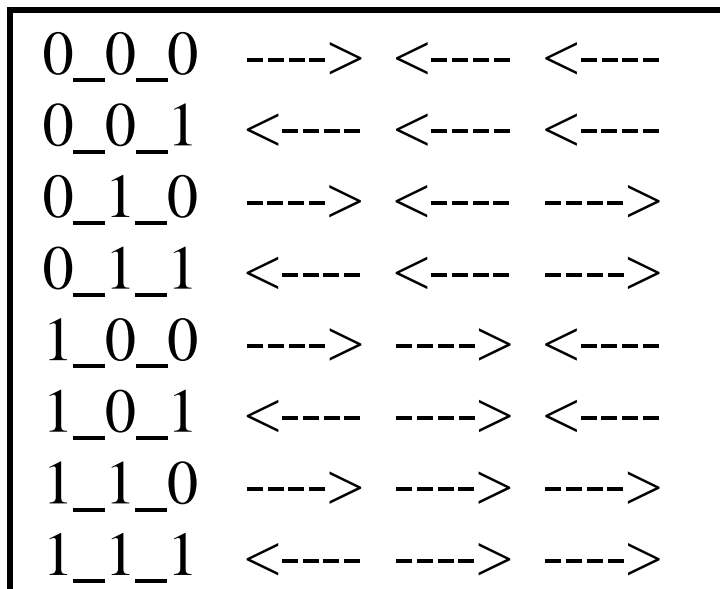
**Figure A2.** Read depth on contig base pairs.

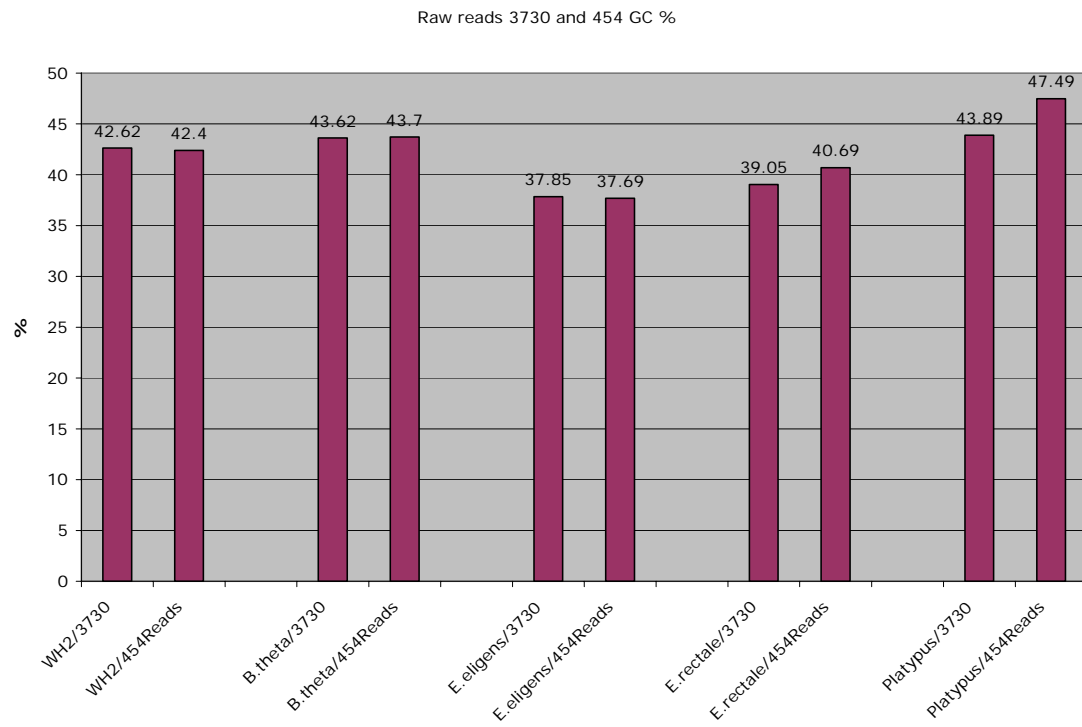


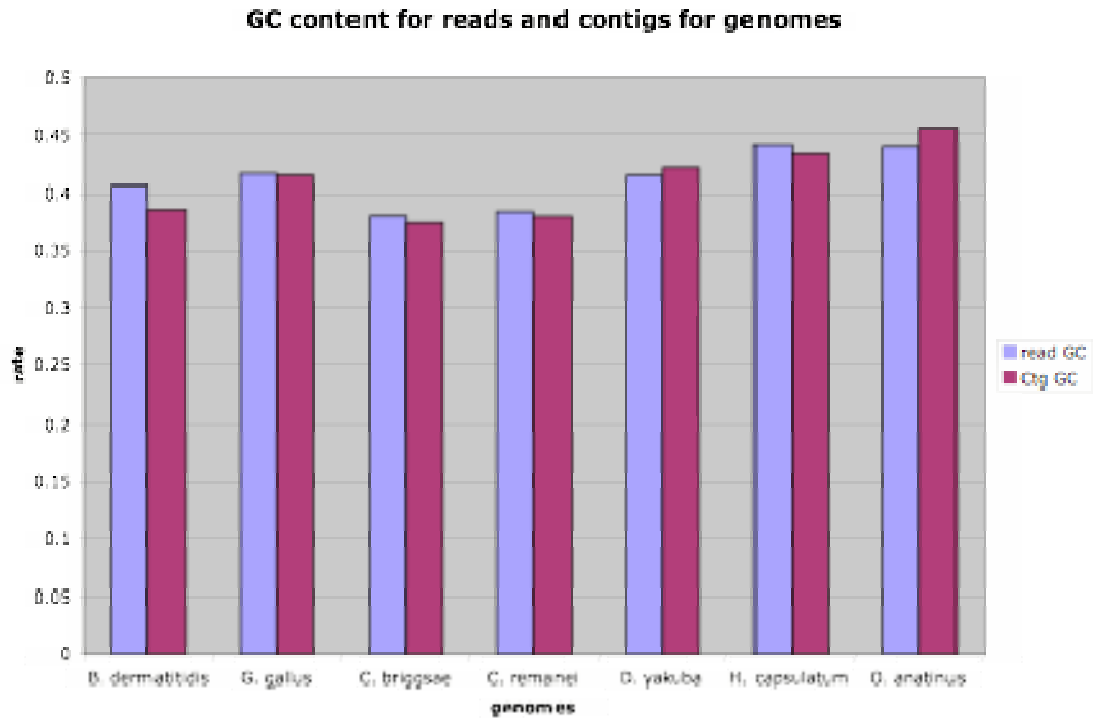
**Figure A3.** Distribution of read depths.**Figure A4.** Platypus segmental duplication by alignment length for intrachromosomal and interchromosomal location.

**Figure A5.** Platypus segmental duplication by alignment identity.

**Figure A6.** Segmental duplication sequences by chromosome (>5kb at 90% identity).

**Figure A7.** Chimeric read types.**Figure A8.** Combinations of reads at the end of contigs.

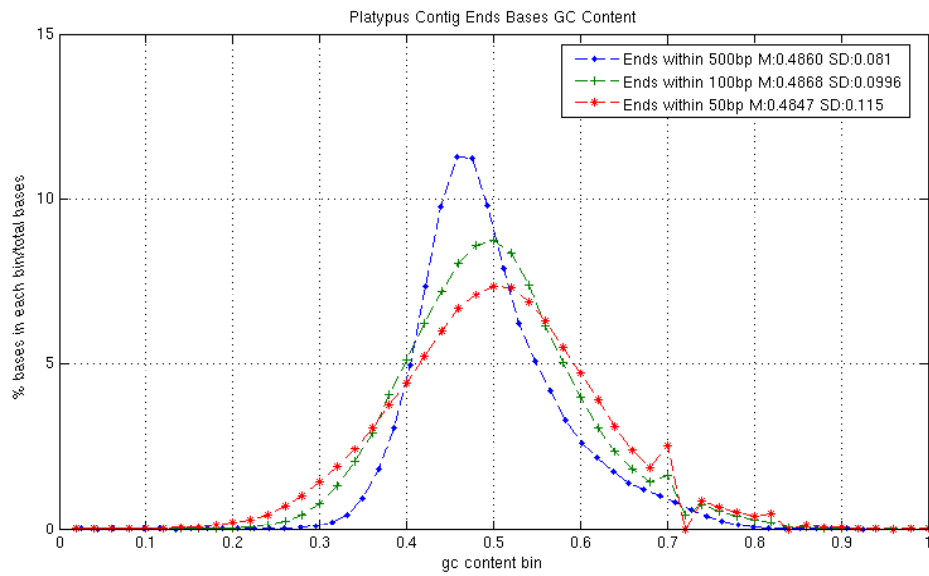
**Figure A9.** GC content for ABI3730 and 454 read types.

**Figure A10.** Multiple species GC content for reads and contigs.

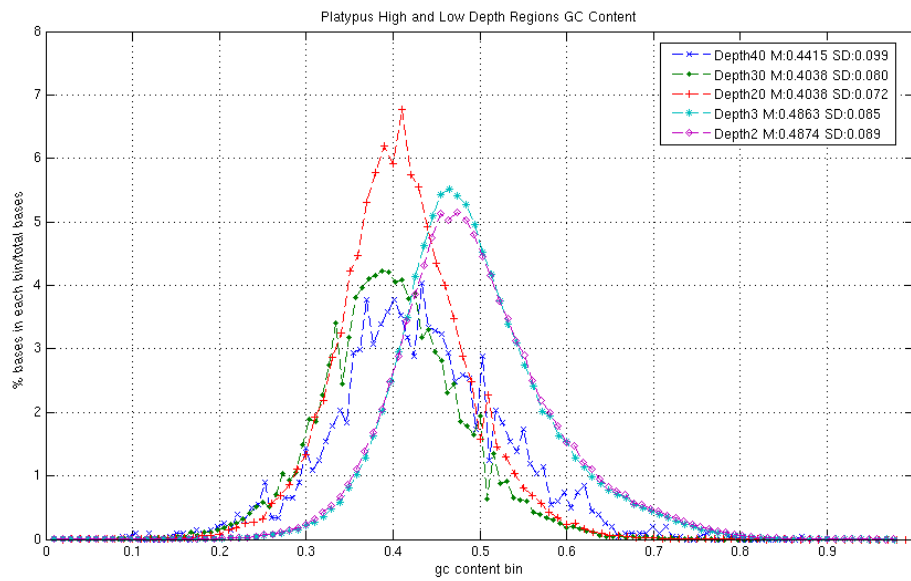


**Figure A11.** GC content by contig ends and read depth a) platypus contig end GC content b) platypus GC content in high and low read depth regions.

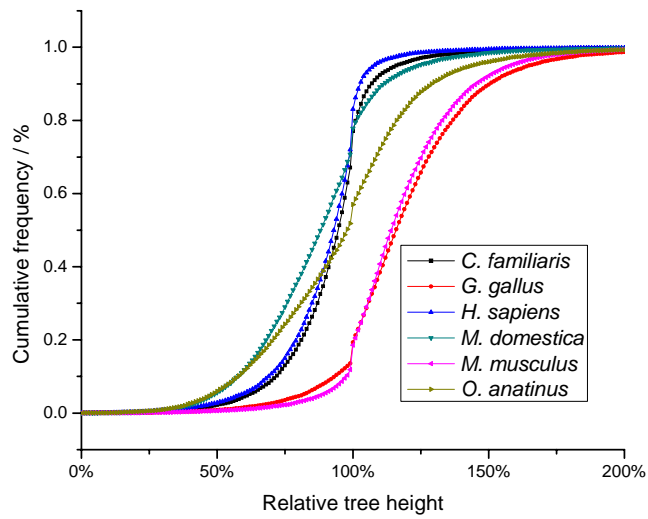
**a**



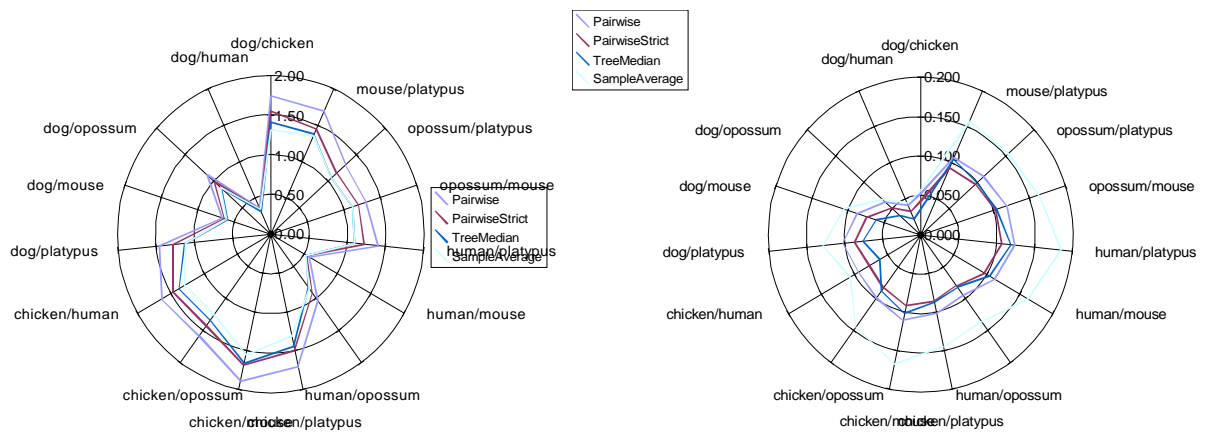
**b**



**Figure A12.** Lineages leading to chicken or mouse have a larger distance to root than other lineages. Shown are cumulative distributions of the distance to root for each species normalized by the median distance to root of each tree.

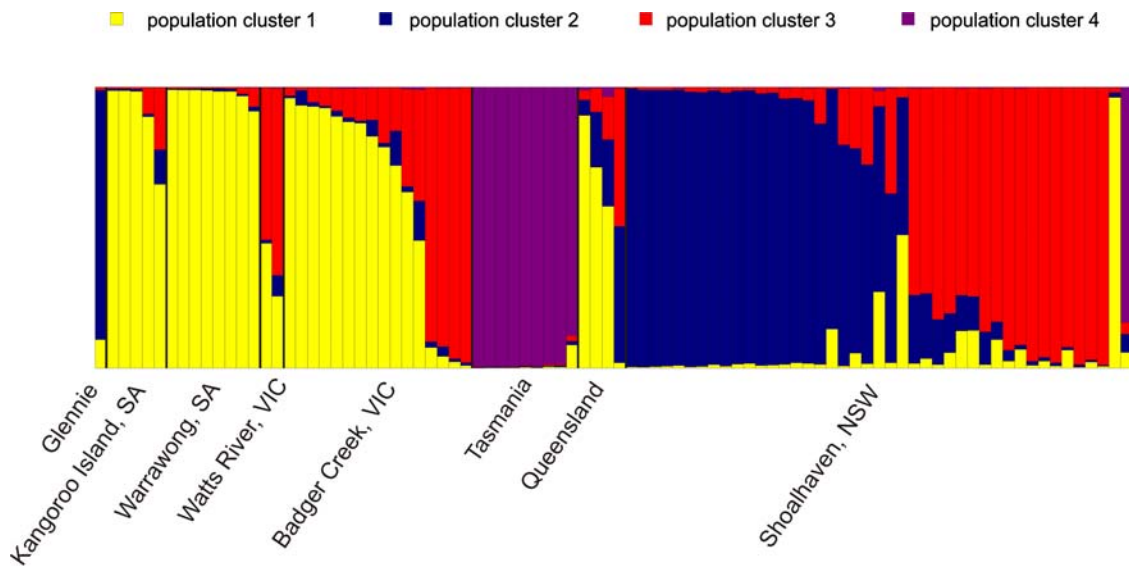


**Figure A13.** Genome-wide  $d_S$  (left) and  $d_N$  (right) values for pairs of species computed with different methods. Each spoke in the wheel corresponds to a species pair (see Accessory Table A13 for values). Method 1 (Pairwise) provides the median values from pairwise comparisons of all 1:1 orthologues. Method 2 (PairwiseStrict) provides the median values from pairwise comparisons of 1:1 orthologues drawn only from the simple (1:1)<sup>n</sup> orthologue set. Method 3 (TreeMedian) provides the median values from rate estimates inferred across the species phylogeny for each simple (1:1)<sup>n</sup> orthologue set. Method 4 (SampleAverage) provides values inferred from 20 samples of 200 concatenated multiple alignments from the simple (1:1)<sup>n</sup> orthologue set (See Accessory Table A12).

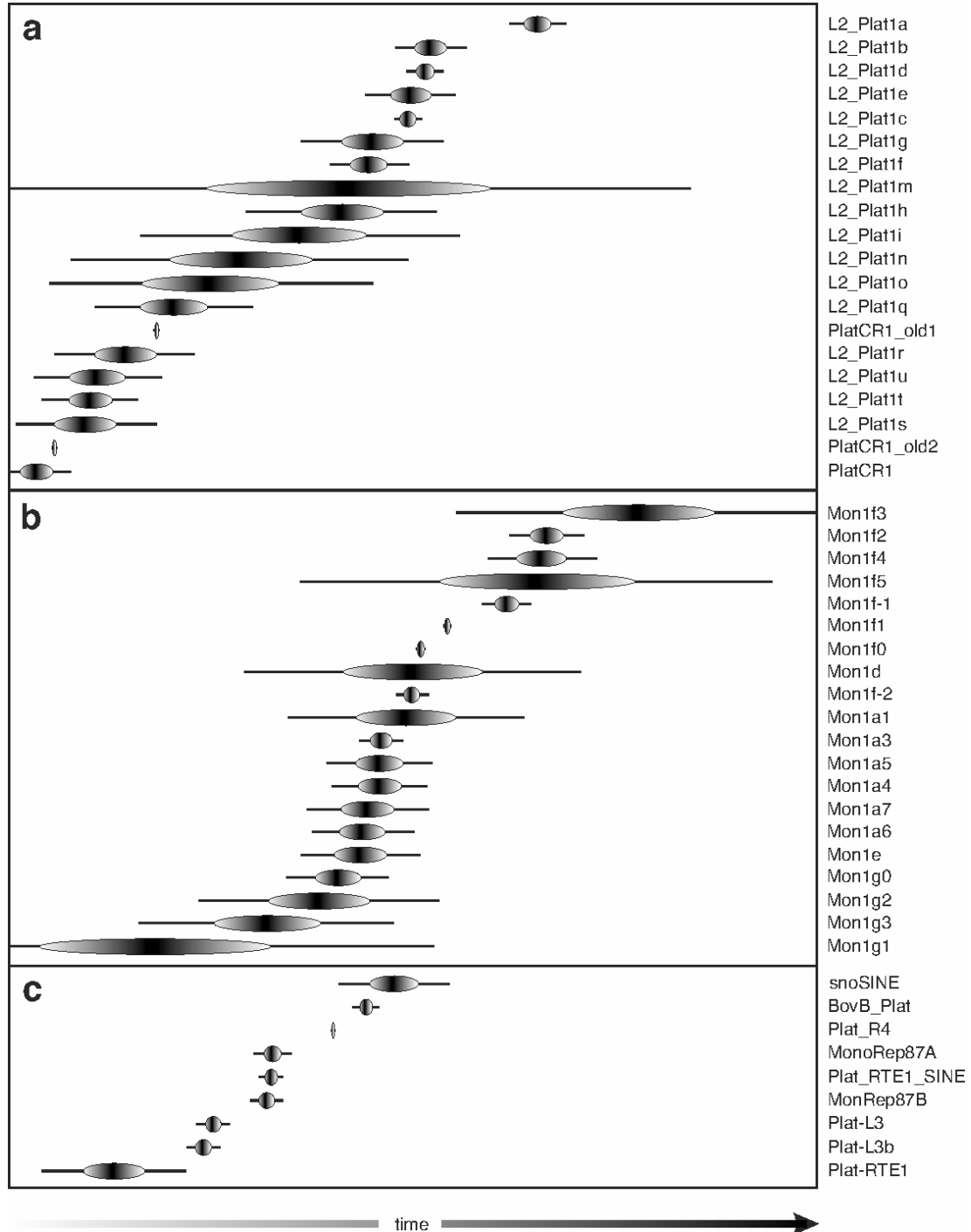




**Figure A15.** A platypus population structure plot for 90 individuals. Each platypus individual is represented by a vertical bar. The color composition of each vertical bar represents the probability of assignment for each individual to each of the four inferred population clusters as estimated using Structure 2.1.



**Figure A16.** TinT distribution: Activity pattern of platypus specific non-LTR retroposed elements over evolutionary time.



**Table R1. Genomic locations, population diversity, oligonucleotide primers**

	<b>Locus name</b>	<b>Forward Primer (5' to 3')</b>
1	DP000299.1(1)	GCAAGTCATTTGGCCTCTCT
2	DP000299.1(2)	AGTGTCCACGGACATTCTCT
3	DP000299.1(3)	GGCTCTGTTTCAGAGTCTTCCA
4	DP000299.1(4)	AGCTCTGCCATTCATCTGCT
5	DP000299.1(5)	TTGTATTTATTGAGCACTTACTGCAT
6	DP000260.1(1)	AAGGTACGAAATCCCATCCA
7	DP000260.1(2)	TTCTTATAATTGTTAAGGGAGGCAAT
8	DP000260.1(4)	CGTTTCTCTCACACGGAAAA
9	DP000260.1(5)	ACAGGGAACGTGACAACCAT
10	DP000260.1(6)	TGGCTAGGCCAGCTATTGTA
11	DP000260.1(7)	TGTCCTTTTGCCCAGAATTA
12	DP000252.1(1)	TCGTTGTCACCACTCCAAAA
13	DP000252.1(2)	TGGAATTCAGTGTCCACAGG
14	DP000252.1(3)	GGCTCATGGAATATCGAGGT
15	DP000252.1(4)	CTTCACCTACGGCCAATCAT
16	DP000252.1(5)	GATCTCGGGCATGTCACTTC
17	DP000252.1(6)	CTGTGCGCAGTACACAGTACTAA
18	DP000252.1(7)	CCTGCCACAGGAGCTTAC
19	DP000252.1(8)	CCAAGCCTTGACTTGGAGAC
20	DP000252.1(9)	TCGGTCTCTTTCCCTCAGTCG
21	DP000252.1(10)	GTTTCCCGAGTGCTGACTGT
22	DP000252.1(12)	CCATCTTCCCCTCTAGTCTGC
23	DP000252.1(13)	CTTGCTGGCTGTGTGATCTT
24	DP000252.1(14)	CACGGAGAGGAAGGTACCAA
25	DP000252.1(15)	TCCATCAGTGGTATTTACTGAGC
26	DP000252.1(16)	AAATGGCCACCCTTTTCC
27	DP000224.1(1)	AAACGGTGGGAGTCACAGAT
28	DP000224.1(2)	ACGGCCTTCCACTTGCTAC
29	DP000224.1(3)	GGGTCTCTAAAAATCTCCTAGTTGG
30	DP000211.1(2)	CGTTAATAGAAATAAACAGAATTGGAA
31	DP000202.1(1)	TAGGCCTGCTTCTTGTCTC
32	DP000202.1(3)	CAAGCTAATCAGTTTGGACACAG
33	DP000202.1(4)	CCTACAGGGGAGATAGGTTTAGA
34	DP000185.1(2)	GGTCTTTGCCACCTGATCTC
35	DP000185.1(3)	GATGAAGACACCATATCCTCTGG
36	DP000185.1(4)	CCTGATTCCCATTCTATGC
37	DP000185.1(5)	ATTCCTGTGCAGGGAAAGGT
38	DP000185.1(7)	GGCTGTGTCTCTTTTTTCCA
39	DP000185.1(8)	GTTTCCCAGACAGGGATTCA
40	DP000185.1(9)	TTTCAGTGAATCAGTGGCATT
41	DP000185.1(10)	TTTTAAATGAAATTCCTTCTACCC
42	DP000185.1(11)	GTGCTTGGCCCATAATAAGC
43	DP000185.1(12)	TGTGGAGGACACTGCACTAGA
44	DP000185.1(13)	GCACTGGGCTAGATACAAGGA
45	DP000119.1(1)	AAGCATCCACTATGGGCAAA
46	DP000119.1(2)	AAAAACTTCTCATCATTGGCTTT
47	DP000077.1(1)	TTAACTTTGGGCCATTTCAGG
48	DP000077.1(3)	CAAGAAAATCAGGCTGGACA
49	DP000077.1(4)	AGCGTGGCTCAGTGGAAC
50	C1-001-2439995	GCTGCTTACCCAGAACAAGC
51	C1-002-4623259	CAAATGCATGCCAGCTCTC
52	C1-003-6186795	CAGGGAATGTGTTTCATGCTT
53	C1-005-7369364	GACGGCACAATCAGTAAGCTC
54	C1-008-13131414	CCCTCTCTCCCTCCACCTT
55	C1-009-13439305	GGGACAATCTGAAGACCCTGT
56	C1-011-13897057	GGGGACCTGGAGATAAGCTG
57	C1-013-16935581	CATGGCCTTTGACCATAATGA

58	C1-017-20138644	CTTAGTGCCCAATGTGTGAA
59	C1-020-25913344	TCCCAAGCAGTTGGTAAAGTG
60	C1-022-27664176	CCTGCTGGTGTGTGCATAG
61	C1-023-30282273	ATGCCTAGAAACCCCATTC
62	C1-024-30863128	TGGTTTCCATTCACTAAACCAG
63	C1-025-31191832	CTGGCCCTGGCAGTCTTAT
64	C1-026-32191881	TGCCCTCAAACACTTCTGC
65	C1-027-32722260	TGCACGCACTTCTAATCATCA
66	C1-028-33293636	TCCCTGTCCCTAGACACTTCA
67	C1-031-34717813	CATCTCCTCCAAACCCTTCA
68	C1-032-35009282	TGCTTTGATGGACATGTTGAG
69	C1-033-35063550	GGTTAAAATCCAGACCATGGAA
70	C1-034-35835046	TGGAAGGTCAGGTTGATTTTCT
71	C1-036-41231888	TAGCTGTCTATGGGGCCAAC
72	C1-039-45450457	TCAACCCAAATTGACAGATGA
73	C1-040-46330980	TTCCTCACTAAGGTGCAATGG
74	C1-041-46395538	CAATGGTATGCGTGGAGATG
75	C10-042-255284	ACATTAGGCTGCTCGTTTGG
76	C10-043-5494470	GCGTAGCCTAGGAGAAAGCA
77	C10-044-5574755	TTGGCTTCAAGGCTTTCTGT
78	C10-046-7484318	GCCGAGAACCCCTAGAACTC
79	C12-053-1053642	GGCAGCGATTGTGTCTATGA
80	C12-054-2044985	GGGAAAAGGTCTTGAATTTG
81	C12-055-4932432	TCTTGGATGTGTCGCTCAAC
82	C12-056-5784023	TCCACAGTCCAGGGAAAGAC
83	C12-057-6499232	GGAAGGGGAATAGAGGACCA
84	C12-058-6899752	TCAAGAAGAGATAAATAAGGCAAAA
85	C12-059-7175632	AAAATGGTGGTTGTGGATCA
86	C12-060-7452902	CACACAGAAGTTACAACCAGCTAGA
87	C12-061-7801849	TCCTAGCACTCACTTGCTCAAC
88	C12-062-9172697	GAGCTGCACATGTGATTCAT
89	C14-065-137579	TAACCAGGACCCCAACTCTG
90	C14-067-1307932	CTATCCCACCCACCCTTAG
91	C17-069-566158	TGGAGGGCTGCTTGAATATC
92	C18-072-861774	TGTGTGAATTTGGGCATCAC
93	C18-073-2130927	CCATGTGAGACAGGGACCTT
94	C18-074-2256902	CAAGCAGCTGTGGAATGGTA
95	C18-075-2515658	TTTGGGGATAGGGAGGAAAG
96	C18-076-3138592	CTCGTTCTCCCGAGGAACAT
97	C18-077-4359820	AATGTGGGCAGTGAACGAGT
98	C2-084-8397014	TGAACATCAACCCTTCATGG
99	C2-085-8660181	TTCTCCGAATTCTGCCATCT
100	C2-088-11466521	CATAGCAGAAACACGGCAAA
101	C2-091-14029452	GCTCATCCCAGGGAGAAAAT
102	C2-092-14290900	AGATCCAGGCTATGCGGTTT
103	C2-094-15767493	TCGGTTGTAATGGGTCTGTTC
104	C2-095-16729203	GCTTTAGCTATCATTATGAGAAGCAG
105	C2-098-23301831	TTTGGGGTGTGTACCATGAA
106	C2-099-25530118	CCATCACCACCACAACCTGAA
107	C2-100-27119260	TTATCTGTACAATGGGAAGTCAGTA
108	C2-102-28339541	CACATTCCCTGCTCACAATG
109	C2-105-31337265	TCCGTCTCTCTGGGAGCTAA
110	C2-107-32973073	TGGTGTGTGGACAGAGCAT
111	C2-109-37748550	AGCTTGCTGAGGTGCCTTT
112	C2-110-38055606	TGCATTGCATTGAGAAGCAG
113	C2-111-38213363	ATTCCTGCCCTGACACACTT
114	C2-113-41523431	CAAATGCTAATTTGCTTCTGGA
115	C2-118-45107236	GCTCCTGAGCCAGGACTTTA
116	C2-119-48396694	CCACACGGGGAAGTTCTGTA
117	C2-121-49258251	GCTCTGCCATGTGTCTGCTA
118	C2-122-51711902	ATGGTGCATCCATGATGATT
119	C2-124-51996640	GTAACCGTTGCTGTGCTTCC
120	C2-125-53461804	GAGCCTGAAATTCTGCTCCT



121	C20-127-229011	GCGCCTAACCAGTACCAA
122	C3-129-1179687	AGAGGCAGCCAGGATCATT
123	C3-131-2231712	GGAGAAGCGTGCTAGTTTGG
124	C3-132-2850175	TGACACAGCAGAGTAGGTGGA
125	C3-135-10879524	AATGCTTACGGTGTGCAGTG
126	C3-137-11802794	GCAGCAGGAATTTTCAGTCATT
127	C3-138-11954639	TGAAGTGTGGAGTTGGGTGT
128	C3-139-12701177	GCTTGGGCCAGAACAAATACA
129	C3-140-14136961	TACAGCCAGACAAGCCACTG
130	C3-141-15561658	CCCTTTCAACTCTTCCCAGTT
131	C3-142-18834925	GCTTGCAAGAGCCTGAGATT
132	C3-144-22474356	TGAGTGGAACAGCGTTGCTA
133	C3-147-28383035	AAACAAGCTGCTGCTGATCC
134	C3-148-29035641	GCCACACCCCTCCTTTATTC
135	C3-150-34820402	CTTGACCAAGGTCACCCAGT
136	C3-151-40017105	CCCAGCTGGAACAATCTCAT
137	C3-156-50772571	ACGACGTGCTTTCACAGTTG
138	C3-157-52041706	CCCCTAAAAGAGAGAAAACAAATTC
139	C3-159-56100286	TGGGGAGAGAAAAGGGGTAT
140	C3-160-56662301	GGTAGCAGAACCAGGAACCA
141	C4-162-646927	TTTTCTTTGTGCCCTGTATATG
142	C4-163-3592732	ATCCCCAAATAATTGCTCCA
143	C4-166-14414661	GCCAGAATGCCTGCTTATTC
144	C4-167-14681457	CTGCAAAATGTGGGTGTCAG
145	C4-168-15511112	TGCTTTACACACATTAGGCAAT
146	C4-169-19964491	GCTAGTAGCCCGCACAACTG
147	C4-171-25021593	TAActCCCATACCCGTGAGC
148	C4-172-26786718	TCTGGAGGCTAGGAGCTCAG
149	C4-174-31497618	ACTTGCTTGTGCTCCCACT
150	C4-176-34783114	GACTCCCAGTCCCAAGCTCT
151	C4-177-37457154	GCACCAGGAAGCAGCATTAT
152	C4-179-39801240	CCCAAGGTCGCAGAATAGAC
153	C4-180-40623377	TTCTGCTTCTGGCCTTCAAT
154	C4-181-43816528	TGAATGAATGAATGACCTGGTT
155	C4-182-44103178	CATCGCCCTAATTCGTTCTC
156	C4-183-45300849	AGTGCTGGGGTGGATAACAAG
157	C4-187-55917657	CCCCAAGCTCCAATAGTTCC
158	C5-190-1979255	GGTGACCCCGCTATCTAACA
159	C5-191-2196859	TATTTTCTGGCGGTGAGACC
160	C5-192-3460819	CCAATTGTGCACATCTGAGAG
161	C5-193-13126600	GGAGGTGTTGGAGGGAAGTC
162	C5-194-13818053	AGCTCTGGAGTGGCCAGATA
163	C5-195-16154778	GGCTAAATGAATTGCCTGATG
164	C6-197-6420094	TTGCTGTGAACAACATGAGC
165	C7-201-24132518	CCTAGACTCCTGGGCTTGTG
166	C7-202-36243423	TGTGGACATGGAATAGAGAGGA
167	C7-203-36924986	GTAActCCCCAACCCCTGTTT
168	Co-205-10364-444	CCACGTGAGTGAATGAGCTT
169	Co-208-1085-48236	GAGAGCGTGGACAGCAGAG
170	Co-211-11166-2755	TCATTCTGGAGGCCAAAGTC
171	Co-212-1135-98346	TGTGTGCGAGGAAAGATGAA
172	Co-213-11686-8414	TGCCTGTCTATAGTTCGGTCAA
173	Co-215-11964-13354	ATTGTGCTCTCCCAAGCACT
174	Co-216-12092-7011	TGTCAAGCAGCATGGCTTAG
175	Co-221-1248-157557	CTCTTTGCTGGGCACTAAGC
176	Co-223-12536-14823	GAActCCCTTCCCAAGTTCC
177	Co-224-12581-5434	GGGCCAACCAAGGAGAATAGT
178	Co-228-13803-3031	GAAActTCTCCAAAActGTCCA
179	Co-232-14709-7155	TGGACAAGGCAAGCTAGTGA
180	Co-234-15966-17579	CTCCCGTAGTCGAGACTGGT
181	Co-235-16049-11066	GGTGAACCTTAACCTGTGAATC
182	Co-236-16633-14786	TTTTGTTGCATTTCCCTTCC
183	Co-243-26426-5942	TGTCATGGGAATACTGACTCCA

184	Co-244-3438-44400	CTTGAGGAGGGAAACATGGA
185	Co-245-5453-11200	ATCCTTGTCCCACACAGAGC
186	Co-247-9805-1001	GCTCAAAGGCCCCATATACA
187	U-250-514-14164269	CCAACATTCCGGAGGATTTGA
188	C1-001-273788	TCGTCCCAGGTAAGGATCTG
189	C1-004-3989098	TTAGACTGTAATCCCATCACTGG
190	C1-005-4048182	GCAAGGTCACAGGACAGATG
191	C1-006-4210085	GCCTGTGGCTGACCTGATTA
192	C1-007-4558674	TCGAAGTGAATGAGCTGACAA
193	C1-008-4658062	AGGAGAGTGACATGCCCAGA
194	C1-010-5268732	GGGATCAAGTCGACCAACTC
195	C1-012-5770201	TTTATCGAGGGCCTATTCCA
196	C1-013-5858490	CGTGATGACCAGCCCTAACT
197	C1-014-6130962	CCAGTTCATGAGCCCTGATT
198	C1-015-6385496	CAAGAACAATAGAAAAGTAAAATGGT
199	C1-020-7662540	GTTGCTGCAGGTGGGATTT
200	C1-021-8444936	CCCAAGAATAAGGAAGGTCCA
201	C1-024-10636201	GGCAAGTGGCTGTAGACCTG
202	C1-025-10903591	CTGCCTCACATCTTGTTGAGTT
203	C1-026-11240827	TCTGATTGGTGGAGTTTGAGG
204	C1-027-11690228	CTTGGCACTGTCTTCCATCA
205	C1-033-14113846	TCCGCTCCCAGACTTCTCTA
206	C1-038-16022378	TTCCCTAGTTCTAATTCACCACCTA
207	C1-039-16164803	GGGGTTGTGCAAGAACACAG
208	C1-040-17741401	TGTCTGCTGTGACTTGTCTGC
209	C1-041-17787550	TCCCCATTTGACTGAGGAG
210	C1-042-18167364	CCACAGGTTATGGGTAGTGCA
211	C1-045-18601421	CACCACAAAAGGAAAACAGGA
212	C1-046-18683649	TCTAGGACACTTCACTATGTAGGAAAA
213	C1-050-19440122	CAACCAACGGTCAGTCAGTG
214	C1-056-21101084	CCCCTGAGAGGGTAATTTCA
215	C1-058-21346488	CGCTGCCACTGTTTAAGTGC
216	C1-068-24030251	CCCAGAAAGGCGAAGCTACT
217	C1-072-25410495	TCGAGGAGAGGGTTCCTGTA
218	C1-074-25991023	GTGGCCTGAGGGACTTGTTA
219	C1-075-26133201	AAGGCTGTTGCTCAGTGGAT
220	C1-077-26760283	CCCAAGTCTGCACACAGTA
221	C1-080-26990670	TCTAAAGCAACCCATTTGAGC
222	C1-082-27226422	AAGTGAACATTTACTGTGTGCAG
223	C1-085-28140218	GGTCATGGAGCCATTGAATA
224	C1-087-28852714	ATGCCAGCCAACACTTTT
225	C1-090-29468698	TTCCCTCTCTGGGCTACTGA
226	C1-092-30383599	CCCACAGCTCCCCTACTTCT
227	C1-103-32859279	GGGGCAGAGAGATGAGTCAG
228	C1-105-33921897	CTTAGCACATAGCAAGCACTTAAC
229	C1-106-33999676	AGCCAGGGGAGGAACTTTA
230	C1-108-34106118	AAAATGCAGCACCCAATAAA
231	C1-110-34271355	TCAAGCAATGGTATTTGTTGAACT
232	C1-115-36058926	TCCCTCTCTACACCCACGTC
233	C1-116-36141147	CCTTGGCCCTGACATTCTTA
234	C1-117-36313473	TGAACTAATGGTAATCAGGTCAGA
235	C1-118-36613530	CCAAACTCTGGAAGCATGAA
236	C1-119-36679877	AAATCAAACATAACTCATGACTAGAGC
237	C1-121-37444356	CCCGGAAGTGGTGACTAAAC
238	C1-122-37711987	CCCCACATCTGGGATATAACAT
239	C1-123-37811269	CCCAAGTGGGTTGGTCTAGG
240	C1-126-38287004	TGCTTCGTTGAGTATCTTCA
241	C1-127-38388974	TGGACAGTGGCTCGTAAGTG
242	C1-128-38554159	GGTCATGGAATCTCTGGTGGT
243	C1-129-39094211	TCACCCCTGGAGGAAAGTC
244	C1-132-39134001	CCAGCTTCCACTTGCTTTAAC
245	C1-134-41764549	TAAGTGGCAGAGGCAGGATT
246	C1-136-42473515	TTTCAATAAATGCTTTGAATCTACA

247	C1-138-42778489	GCTCACTGTGGACAGAGACTG
248	C1-140-43123545	AGGCTGTGAGCCTCATGTGT
249	C1-141-43411272	AAATTTAGCCATAACCCTCAAAA
250	C1-145-43920663	AATCCATTGCCGGATGAAG
251	C1-146-44231991	CCCAGGTCCCTAGTCTTTCC
252	C1-148-44279609	GGAGCCAGATACCACAAGGA
253	C1-149-44334673	CCAGAACAATTGCAGATGGA
254	C1-150-44464435	GTCATCCCTTCTCCCCACTA
255	C1-153-45346735	CGCTTCAAGTCAGCTCCATT
256	C1-161-46990390	GGTCCTTGGTTTGGGCTCTA
257	C1-162-47442700	CAGGCTCATTTTCTGCACTG
258	C10-163-882524	GACCCTTCACTGGGCAGAT
259	C10-176-6549670	GACCTGGGAGACGAAGCATA
260	C10-177-6898887	AAGCGCTTAACTTCCATCGT
261	C10-180-8553255	GACAGTGAGCCCATGGTTG
262	C10-181-9825511	GCTTGTGGCTCAGTATAGTCTTGA
263	C10-182-10085167	GCTGGACAGCATCTGAATTTT
264	C10-183-10239829	TTGGGACTTGGACTTCATCC
265	C10-186-10492767	TGAAGACGGGGCAAACCT
266	C11-188-87787	TTCCCGAGTTGACAGAGCTT
267	C11-189-599421	GAGGGTAGAAGCGAGATCCA
268	C11-190-845321	AAAACAAACATTACGGGAGAAC
269	C11-196-1422612	CACGCGATTGTTCAAGTTGAG
270	C12-203-149219	CCAGGAGTTAAGGCCCTTTT
271	C12-206-2355559	GCTTGTATCCAACCCAGCAC
272	C12-210-3695162	TCCCAGTTCCACACAGAGC
273	C12-213-5136112	GAGGCCACTGCAAGATCAGT
274	C12-220-6345781	CCTCACCCACAAGGAGCTTA
275	C12-222-6904309	TCCCTTTTCCACGTAGTGCT
276	C12-223-7142683	GTAGACGGTGGGAAGGCTGAG
277	C12-228-7938737	CACCACCACCTCTTTCAATG
278	C12-231-8969858	CAACGTGAAGCCCAAAGAT
279	C12-232-9351104	TTATCACGCCCACTGAAACA
280	C12-236-10542201	AAGCAGGAGGCAGAATGATG
281	C12-241-11159665	CCTGCCACAGTGAGCTTAT
282	C12-242-11175800	ATCAGGAGAGCAGCCACTGT
283	C12-243-11238248	ACTGAATGCTTTCCCATCA
284	C12-248-12786148	AATGTGGTATTTGGTGCTTGG
285	C12-250-13281341	TCAAAAGGATCGCCTTGATT
286	C12-251-13505340	CCATCAGTTCTTGTCCACCA
287	C12-255-14492487	AGGAACATCACAACTTCCA

PCR amplification of each locus was one minute at 94°C initial denaturing, and A final extension time of two minutes at 72° was also used.

Population diversity was classified as: fixed present (FP), low frequency (LF) PO every individual tested had the mobile element homozygous present. Low frequency element is variable as to insertion presence or absence in at least one population heterozygous or absent individuals. N/A indicates non-specific PCR amplification

s and amplicon sizes for platypus Mon1 SINEs and L2 LINES

Reverse Primer (5' to 3')	Filled size	Empty size	March 2007 Assembly Genome Position	Subfamily
TCTTCTTAGCCACAGTCTCTCC	388	197	Ultra454:346998-347619	Mon1f3
CCCGTGCCCTTTCTACTAGG	356	194	Ultra454:385566-386364	Mon1f3
TTGCACTCTACCAATCGCTTA	400	189	Ultra454:343096-343730	Mon1f3
CACACTAATTCAGGGCTTACACT	373	157	Contig141222:180-785	Mon1f4
GCAATATTACTGCTTCCAGGACA	400	185	Ultra454:367414-368044	Mon1f3
GCAAACACGAATCTTGCTCA	389	187	Ultra2:1751038-1751666	Mon1f3
ACGCGTCCACCAACTCTTAC	376	177	Ultra2:1656322-1656948	Mon1f4
GCTTAAGCTGGGACTGCGTA	390	169	Ultra2:1653265-1653860	Mon1f3
CGAAAAAGAATTTAGCAAACACC	381	197	Ultra2:1931295-1931908	Mon1f3
TGGGAATATAACAATAAAAAGAGATG	400	210	Ultra2:1967232-1967844	Mon1f3
CACAGCTCACAGAAGGCAGA	399	171	Ultra2:1848729-1849405	Mon1f3
CATTTGAAAATTGAACAGTTATAGCAA	496	258	Contig190:622744-623391	Mon1f4
GATTCCTGCGTGAGATAGGC	392	183	Contig190:539679-540320	Mon1f2
CAACATTACGGAGGTGGTTG	400	178	Contig190:452697-453343	Mon1f-1
GGCAGGGAATGTGTCTGTTT	398	186	Contig190:352458-353105	Mon1f0
TCTCTGGTGCCCTGTTTTATG	400	182	Contig190:760035-760656	Mon1f3
TAGCCCTGGTTCTCTTCTCG	400	211	Contig190:602116-602715	Mon1f3
CCTTCTCCCCAGAACATCAC	392	208	Contig190:453653-454285	Mon1f3
GAAGCACAGAGCAGCGAAG	489	271	Contig190:797121-797733	Mon1f4
GGGAAATTGAGAGGTGACAG	584	362	Contig190:766226-766865	Mon1f-1
AGCACATCAGCTCCATTTCTC	399	192	Contig190:755126-755763	Mon1f4
CAGTGTATTGAGGGAGGCAAA	387	166	Contig190:632390-633054	Mon1f-1
CCACCTGACCCTCACAGTTT	394	182	Contig190:536509-537128	Mon1f1
TTTTCCCTGAACATACAAGTCC	398	179	Contig190:440462-441130	Mon1f4
AGTGTCTGTCTCCCATCTG	388	158	Contig190:358412-359069	Mon1f4
CTCAGCTCATGGGAAGCACT	400	196	Contig190:324602-325207	Mon1f-1
GGACCCCATAGTCTGAAGCA	393	197	chr5:14630494-14630765	Mon1f3
GGTCGCACCTGGAGAGTTT	388	213	chr5:14614427-14614977	Mon1f3
TTTCCCCATTGATTCCTTTG	395	178	chr5:14593896-14594550	Mon1f3
GCACTGGGGAGAGTACAGTATAA	400	176	Contig219:926274-926814	Mon1f3
TGACTTCCCAAGTTGGCTTT	540	299	?	Mon1f3
TTCACTCCCCAAAGTCCAAG	492	259	Contig294649:1-359	Mon1f4
CCGGAGCTGAGACTGGTAG	400	160	Contig24566:5923-6605	Mon1f3
CTGCTTCTTCAAGGCAAACC	375	175	Ultra479:2831098-2831714	Mon1f3
CGAGAGGAAGCGAACAAGAA	365	195	Ultra479:2370214-2370756	Mon1f3
TCGTATTGGCAACACAGACC	398	160	Ultra479:2952253-2952923	Mon1f3
GAAGAAGCAGGAAGGCACAG	341	122	Ultra479:3265283-3265857	Mon1f2
CAACTGCCACTTTCAGAGCA	440	201	Contig187235:1-256??	Mon1f3
TGAAAATACTATAGGTGGGAAAAAGC	387	178	Ultra479:2203192-2203837	Mon1f3
ACTCTCAAGGCTGGGCTCTT	342	199	Ultra479:2625720-2626287	Mon1f3
CCTCAAGGCGCTATCAGTTT	400	178	Ultra479:2654428-2655086	Mon1f3
TCCAAAAGCGCTGTACCATA	400	197	Ultra479:2593156-2593780	Mon1f3
TGCCATTTGCATGTTTGT	377	197	Contig42163:951-1556	Mon1f4
TTCTGGGCTCCCACTTACAC	485	252	Ultra479:2621960-2623512	Mon1f4
GCGTGTACTACTACCCTCTCC	396	183	chr3:17790226-17790867	Mon1f3
GGAGAGATTGAGGCACAGTGA	398	159	Contig7905:14534-15219	Mon1f3
AATAAGGCGCTGGTTTCGTA	495	252	Ultra336:686908-687579	Mon1f3
CCACCTCACCCGAAATTTT	495	275	Ultra336:744661-745317	Mon1f3
GGAGACAGCGAAAGAAAACG	489	250	Ultra336:875813-876461	Mon1f3
ACACAGCAGAGACGTGCAGA	498	226	chr1_2439995_2441267	Mon1f-1
AGGAGCCTGGAAGCAAGAAT	482	216	chr1_4623259_4624525	Mon1f1
CATGCAAGGGGAAAATGAGA	474	209	chr1_6186795_6188060	Mon1f1
CACAGGACTCCCCTAGAGA	500	234	chr1_7369364_7370630.rev	Mon1f1
GCCATGTGGCTTCTAGGAAA	470	207	chr1_13131414_13132677	Mon1f-1
GCTCCCAGGCCTGTA	500	235	chr1_13439305_13440570	Mon1f-1
GGGAACTGAGGCAAAACAAA	491	227	chr1_13897057_13898321	Mon1f2
CATGTTCCATGACCCTTAGGT	499	235	chr1_16935581_16936845	Mon1f-1

CAGATGTGCCCTTCCATCTT	475	209	chr1_20138644_20139910.rev	Mon1f-1
GTCCTCAGGATGCCAGTCTC	447	185	chr1_25913344_25914606.rev	Mon1f-1
TGGAGACAAGAGGAGATGTGAA	450	185	chr1_27664176_27665441.rev	Mon1f-1
CCCACAAGTGCTCAGTTTCA	442	178	chr1_30282273_30283537.rev	Mon1f1
CACTTTTCATCTATTATCAACAAACCAA	468	202	chr1_30863128_30864394	Mon1f-1
GATCAATATCGGCCATCCAA	489	223	chr1_31191832_31193098.rev	Mon1f1
AAACAACACACACAATTTTCATCA	472	207	chr1_32191881_32193146	Mon1f-1
TGTTCTGAAAACAGTAAGCACTCA	471	210	chr1_32722260_32723521.rev	Mon1f-1
TCAGCTACCTTTGCAACTGAGA	498	234	chr1_33293636_33294900	Mon1f1
CAGGAACTCCAGCATGAACA	475	210	chr1_34717813_34719078.rev	Mon1f1
TGATTGATTAACACATTATAGGCACTT	485	220	chr1_35009282_35010547.rev	Mon1f2
AGCCCTCTGCTTACAAACCA	456	192	chr1_35063550_35064814	Mon1f-1
TGGAAGGAAGAGAGGAATATGG	497	239	chr1_35835046_35836304.rev	Mon1f2
CTTCCTCTGGTCCCTCACAA	495	229	chr1_41231888_41233154.rev	Mon1f0
TGTGCCTTCTGAATTGTTGC	477	214	chr1_45450457_45451720.rev	Mon1f0
TGCATTGCTTCTAGCGAAAC	491	227	chr1_46330980_46332244	Mon1f0
CCCCTCTAGACTGTGAACACG	773	510	chr1_46395538_46396801	Mon1f2
CAGGGACGTGTACCCAGAAT	479	216	chr10_255284_256547	Mon1f-1
GAGAAAGACCCTGTCCACA	460	194	chr10_5494470_5495736.rev	Mon1f2
GCTGTGGTCTGACTGATGGA	473	209	chr10_5574755_5576019	Mon1f0
CGTCGGATTACAGACAAAACA	471	205	chr10_7484318_7485584	Mon1f-1
GTGGTTGTAGCTTGGGAGGA	458	193	chr12_1053642_1054907.rev	Mon1f0
GTGCTAAGCGTTTGGGAGAG	438	172	chr12_2044985_2046251.rev	Mon1f1
CCTCCCTTGCTCAAAGTCAG	467	201	chr12_4932432_4933698	Mon1f-1
ACAAAATTGGCCTGGCTTC	464	198	chr12_5784023_5785289.rev	Mon1f-1
CTCCTGCCCGCAATGTAAT	435	169	chr12_6499232_6500498	Mon1f1
TGGGTCTAACCACTTAACTCAATTT	481	215	chr12_6899752_6901018.rev	Mon1f-2
TGCTCAGAAAGCACTTGATTG	460	197	chr12_7175632_7176895.rev	Mon1f2
TTGTTGCTTTACTTCGTAGATTATTCA	491	225	chr12_7452902_7454168	Mon1f1
TGTTATGGGCAGGAAATGTG	500	234	chr12_7801849_7803115	Mon1f0
CCTTCTATGACCATTTCCCAAG	503	242	chr12_9172697_9173958	Mon1f-1
ATGCCAACACCAAGAAGCTC	471	210	chr14_137579_138840.rev	Mon1f2
AGAGCAGGTTGTGTGGCTTT	426	160	chr14_1307932_1309198.rev	Mon1f-1
GCTGGTTACGAAGGAAGTGG	492	226	chr17_566158_567424	Mon1f-1
GTGGCTTCTGAGGGGTTTT	574	308	chr18_861774_863040.rev	Mon1f-1
TTAGGGGCAAACAGCTGAG	470	201	chr18_2130927_2132196	Mon1f1
TGCAAGATCCATTCGCAATA	476	210	chr18_2256902_2258168.rev	Mon1f-1
TTTACGGGGACGTTTCAGAC	474	208	chr18_2515658_2516924	Mon1f-1
CCATCAACTTGGGGCTGTTA	498	232	chr18_3138592_3139858	Mon1f1
GCTGGTACCCAGTATCTTGGA	469	203	chr18_4359820_4361086	Mon1f0
CAGGAAAAGTGAATGTTTTAGACGA	437	171	chr2_8397014_8398280	Mon1f-1
GGAATCCCTCTCCTCACTGC	438	172	chr2_8660181_8661447.rev	Mon1f1
CATTTCCATCATCCAATCTTGA	443	183	chr2_11466521_11467781	Mon1f2
CGACAGCCATCTCTGTTTCA	443	178	chr2_14029452_14030717	Mon1f1
TGCGCTTAGAAGCACAAACAG	489	225	chr2_14290900_14292164	Mon1f-1
CAACTGGTTTTTCATTTCTGTGG	485	219	chr2_15767493_15768759	Mon1f-1
CAGGTACATATTCGGGCTCTG	470	206	chr2_16729203_16730467.rev	Mon1f-1
TCTTTTCAAGTTGGCCCCTA	495	229	chr2_23301831_23303097.rev	Mon1f1
CAGATGTCTGGACAGGCTGA	452	188	chr2_25530118_25531382.rev	Mon1f-1
GAGGGAAGGAGTGAGCCTTT	483	222	chr2_27119260_27120521.rev	Mon1f1
TCCAAGGTAAAGTAAAATCTGCAA	500	239	chr2_28339541_28340802.rev	Mon1f2
TGGTTCTGGTTTCATGTCTCC	465	201	chr2_31337265_31338529	Mon1f-1
CTGTCCACACTGGACACAT	445	179	chr2_32973073_32974339.rev	Mon1f-1
TCTGGCATATGTTCTAGGGTA	469	203	chr2_37748550_37749816	Mon1f1
TTCCATTTGCACTGTACTGGTC	496	230	chr2_38055606_38056872.rev	Mon1f1
CCACACCGTTTCATTCCTCTAA	498	232	chr2_38213363_38214629	Mon1f2
CTTCCACATTCCACCAGGTT	487	221	chr2_41523431_41524697	Mon1f1
GGTTAGCATTGAGGACCAA	469	203	chr2_45107236_45108502.rev	Mon1f1
CATATTTTCATCAGCTGGCTTTG	446	180	chr2_48396694_48397960.rev	Mon1f2
AAGCCCTGGGGTAGTTACAT	446	181	chr2_49258251_49259516.rev	Mon1f1
GAGAAATGTGGAGCCAGGAA	454	189	chr2_51711902_51713167.rev	Mon1f2
GGTAATCCGGTTTGGACAGA	444	179	chr2_51996640_51997905.rev	Mon1f2
AGCACTGTTGGGCTGTACAAT	500	235	chr2_53461804_53463069.rev	Mon1f1

CATTTTGGGAACCGTAGAGG	489	224	chr20_229011_230276	Mon1f-1
ATACCTCTTCCCGGTCTCGT	484	227	chr3_1179687_1180944	Mon1f1
CCGTGCAAATAAATGAGCA	487	221	chr3_2231712_2232978	Mon1f0
TTGTGAAATGGCTCACAAAGC	477	211	chr3_2850175_2851441.rev	Mon1f1
CTTTATTCATCCCGCCCTTA	451	183	chr3_10879524_10880795.rev	Mon1f1
CTTGGGAGACCAAGGGTTAC	300	35	chr3_11802794_11804059	Mon1f0
TCCCATTCATTTTGGTGTCA	468	204	chr3_11954639_11955903	Mon1f-1
TCTGTGAAAATGAAGGGCAGT	487	224	chr3_12701177_12702440.rev	Mon1f1
AATCCCATCTTGGGCCTTT	458	192	chr3_14136961_14138227	Mon1f2
TAGCCAATCTAGGCCACTCG	454	188	chr3_15561658_15562924	Mon1f1
GTTGGAGGGCTCCACTTGTA	428	166	chr3_18834925_18836187	Mon1f1
AGCTCTGGCACCTCTTGAAG	463	197	chr3_22474356_22475622	Mon1f1
CTGTGCTGAAGTGCCTTGTC	485	224	chr3_28383035_28384296.rev	Mon1f1
CACCTGGAATTGATGACAGC	442	179	chr3_29035641_29036904.rev	Mon1f2
TCAAAGAGCTTTTGGAGGTAATG	452	185	chr3_34820402_34821669.rev	Mon1f1
GAGCACCGGGGTAGATACAA	468	207	chr3_40017105_40018366.rev	Mon1f2
CTCAGGGTGTTCAGGCTTC	468	202	chr3_50772571_50773837.rev	Mon1f1
CCTTTTAGTGAATGCCTCCAA	491	225	chr3_52041706_52042972	Mon1f1
GAGACAGTCCCTGCCTTTTG	467	203	chr3_56100286_56101550	Mon1f1
CTTTGGGGCGACAGTTACAT	584	321	chr3_56662301_56663564	Mon1f1
ATCCATCCATGCATCCACTT	500	234	chr4_646927_648193.rev	Mon1f2
AACCACTCATACCCGAAGTCA	479	212	chr4_3592732_3593999.rev	Mon1f1
GGCAGGTGAAGAGGGAGACT	497	231	chr4_14414661_14415927	Mon1f2
TTGGCAGAGAGGATGCAACT	467	202	chr4_14681457_14682722.rev	Mon1f2
CAGGAGACGTGAGCCTTGTT	462	199	chr4_15511112_15512375.rev	Mon1f0
TTGCTCCTAAGGTGGGACTTT	475	214	chr4_19964491_19965752.rev	Mon1f0
GACCCAGGTCTTCTGACT	498	237	chr4_25021593_25022854	Mon1f2
TGAGGGAGGAAGCAATAGCC	485	220	chr4_26786718_26787983.rev	Mon1f2
CCAGCACTCTCAAATCACA	462	197	chr4_31497618_31498883.rev	Mon1f0
CACTACTGCTTTGGGGAGGA	477	211	chr4_34783114_34784380.rev	Mon1f0
TGCAGAGATGGGTACAAGCA	497	231	chr4_37457154_37458420.rev	Mon1f2
ATTGGGACCTTGTCTTGTGG	425	160	chr4_39801240_39802505	Mon1f0
GCCACAAAGCCATGCTACTT	499	234	chr4_40623377_40624637	Mon1f0
CACCTCCTCTGCTCATGTCA	494	230	chr4_43816528_43817792.rev	Mon1f0
GAGGTAGGCAGGCACCAATA	466	202	chr4_44103178_44104442	Mon1f0
CAGAAGAGGGCCTGAAACTG	492	226	chr4_45300849_45302115.rev	Mon1f0
TCCTTTTACTCGTCCCTCCA	500	236	chr4_55917657_55918921.rev	Mon1f2
AACAGCGTTGGCAGACATTA	483	217	chr5_1979255_1980521	Mon1f2
CCCTTGCCTCTACTGCTCAC	482	217	chr5_2196859_2198124.rev	Mon1f-2
CTCCTTGAGGACAGGGATCA	488	222	chr5_3460819_3462085.rev	Mon1f2
TTGACCTTTTGTGGTTTCA	484	221	chr5_13126600_13127863.rev	Mon1f0
CCCAGGAACCAAGAAGTGTG	496	233	chr5_13818053_13819316.rev	Mon1f2
CTGACACGTACATCGGTCGT	478	213	chr5_16154778_16156043	Mon1f0
CGCTGAGTTTTAGACATTTCC	617	354	chr6_6420094_6421357	Mon1f2
TTCTTCAACATGCCTTGTGC	482	219	chr7_24132518_24133781.rev	Mon1f0
GTTGAGGGCAACCAGAAGAG	472	207	chr7_36243423_36244688	Mon1f2
CTCATTCCCCTTGGGTCATA	455	193	chr7_36924986_36926248	Mon1f2
CTCTCTGCCCTTCTGACTCC	462	198	Contig10364_444_1708.rev	Mon1f0
CCTTCCAGTCATGGAATGT	426	160	Contig1085_48236_49502.rev	Mon1f2
GGGAAGGTCAGGTTGAGTGA	475	209	Contig11166_2755_4021	Mon1f2
CACATTCCCTGCTCACAAAG	494	229	Contig1135_98346_99611	Mon1f2
GGGAGAGACAGATGGCAAAA	435	169	Contig11686_8414_9680	Mon1f0
CGATATCCTTTTCCAGGGTTC	468	211	Contig11964_13354_14611	Mon1f2
CACAAAGTCACACGGAGCTG	600	341	Contig12092_7011_8270.rev	Mon1f2
GGGTCAGAAGGACCTGGATT	462	197	Contig1248_157557_158822.rev	Mon1f0
ATGGATTGGGAGGATGACTG	497	234	Contig12536_14823_16086.rev	Mon1f2
GCCCCTGTTCACTCACTCA	424	157	Contig12581_5434_6701	Mon1f0
TCAAGATGGTCCCAGTCACA	486	218	Contig13803_3031_4299	Mon1f0
GGAGCTGTTCCCAATATGCT	465	199	Contig14709_7155_8421.rev	Mon1f0
CCCAGCGGGTCAGTAGTAAA	473	207	Contig15966_17579_18845	Mon1f0
CATGGCGGAGCTAGGATTAG	426	162	Contig16049_11066_12330.rev	Mon1f0
TAAGCCTGTGGAAAGCCTGT	457	190	Contig16633_14786_16053.rev	Mon1f0
GGGTAAAGCCACAGAAAGCA	470	207	Contig26426_5942_7205	Mon1f-2

AACCCCAAGGACACCTCTTT	456	194	Contig3438_44400_45662.rev	Mon1f-2
CCTCAATGGGCCACTAGAAA	470	201	Contig5453_11200_12469.rev	Mon1f-2
GCCCATGCTCTACCCACTAA	487	221	Contig9805_1001_2267.rev	Mon1f-2
TCTCCCCATCCCTTTAGTCC	473	207	Ultra514_14164269_14165535.r	Mon1f-2
GCAGAGCACTGAGCACTTTCT	432	221	chr1_273788_274999	L2Plat1a
CCACAATCATTCTGCCCTA	282	159	chr1_3989098_3990221.rev	L2Plat1a
AGAGCAGCCTCTTCCCTTTT	657	188	chr1_4048182_4049651.rev	L2Plat1a
TGGCATTATTGAGCACGTT	456	198	chr1_4210085_4211343.rev	L2Plat1a
AATGTGCCCCAGAAGTGTTT	498	233	chr1_4558674_4559939.rev	L2Plat1a
GGCAAATGGAGAGCCAATA	648	183	chr1_4658062_4659527.rev	L2Plat1a
CTCTTGCAAGCAATCAATCA	463	224	chr1_5268732_5269971.rev	L2Plat1a
ATCCCAAAGTCCACTTGAG	372	248	chr1_5770201_5771325.rev	L2Plat1a
TTTCCTAGTTCATGAATCTTTCCTC	284	159	chr1_5858490_5859615.rev	L2Plat1a
GCACTCCTAAAATGCCCAAT	378	193	chr1_6130962_6132147.rev	L2Plat1a
CTGGACATCTTATAAATCCATCTATCT	287	189	chr1_6385496_6386594	L2Plat1a
GGGAGCGGATGTATGTTGAC	285	154	chr1_7662540_7663671.rev	L2Plat1a
CAACTCCTCTACCGAATTGAAA	300	172	chr1_8444936_8446064.rev	L2Plat1a
TGAGTTCGTAATTTTGCACTTTT	399	155	chr1_10636201_10637445.rev	L2Plat1a
CGGTGGCTAAAGAATCCATC	300	179	chr1_10903591_10904712	L2Plat1a
TTGGCCTTCTTTTCTGAGA	346	157	chr1_11240827_11242016	L2Plat1a
GGCCTCTCTCCAGCTTCTTC	396	216	chr1_11690228_11691408.rev	L2Plat1a
CAAGGCTGCTGGACATTGTA	491	193	chr1_14113846_14115144	L2Plat1a
GTTGACCGAGCTGGGATT	295	157	chr1_16022378_16023516	L2Plat1a
AAGCCACCGCACACATTT	288	164	chr1_16164803_16165927.rev	L2Plat1a
AGACACTGTATTAACCGCTGGA	383	163	chr1_17741401_17742621.rev	L2Plat1a
CCATGATCTCCTCTTTGAAGCTA	243	149	chr1_17787550_17788644.rev	L2Plat1a
TTGGCACAAATGAGGCACTTA	355	192	chr1_18167364_18168527	L2Plat1a
GCAGAGCTGGGGTTTGAA	391	152	chr1_18601421_18602660.rev	L2Plat1a
CAAGTGTCAATCCACCCACTT	491	243	chr1_18683649_18684897.rev	L2Plat1a
CAGCCCCACATTACCTGTCT	436	206	chr1_19440122_19441352.rev	L2Plat1a
ATTTGGGCACCTGTGGAATA	471	181	chr1_21101084_21102348	L2Plat1a
TCAGCCTTAAGAATCCCTGA	375	188	chr1_21346488_21347675	L2Plat1a
AAGCCCTCAGAAGCTGAGAA	292	150	chr1_24030251_24031393.rev	L2Plat1a
CCACTCGAGAAGCAAGCA	456	300	chr1_25410495_25411651.rev	L2Plat1a
TCACCCTAGCATTGCATTTG	486	234	chr1_25991023_25992275.rev	L2Plat1a
AAGTCGACCAGTCCATCCTG	497	236	chr1_26133201_26134462	L2Plat1a
GAAGGGCAGATAAACCCACA	308	151	chr1_26760283_26761440	L2Plat1a
TTCTTCCTGAAGCCCAAAG	376	187	chr1_26990670_26991859.rev	L2Plat1a
CCTGTTTTCCAACCTCCTCA	288	177	chr1_27226422_27227533.rev	L2Plat1a
CATGGTAGTGCCTGGGTCTT	377	212	chr1_28140218_28141383	L2Plat1a
CTACGCCAGTGCTCAATCAG	384	238	chr1_28852714_28853860.rev	L2Plat1a
CGAGGGAACAGGTCACAGAG	489	219	chr1_29468698_29469968.rev	L2Plat1a
TGGACCTTTGGGCATTTTAT	600	257	chr1_30383599_30384942	L2Plat1a
TGATCTCCTTGCTTCTAGTCCA	298	181	chr1_32859279_32860396	L2Plat1a
CCTAGGATCATAGAAGCCAGT	399	202	chr1_33921897_33923094.rev	L2Plat1a
TCAAAGACCTGGTATCCTTACCTT	482	256	chr1_33999676_34000902	L2Plat1a
GAACCAACAGAATGCCTTTGA	399	237	chr1_34106118_34107278.rev	L2Plat1a
CCCCTACTCCCTTTGATTGTC	398	203	chr1_34271355_34272550	L2Plat1a
GGGACCACAGAAAGTTCTCC	291	167	chr1_36058926_36060050	L2Plat1a
TCAGCCATCTAGGCCACTAA	579	201	chr1_36141147_36142525	L2Plat1a
CCTGATTTTCTTGATTATCCTAGCAC	280	158	chr1_36313473_36314595	L2Plat1a
GACAGATTGTGGTGCAAACC	278	157	chr1_36613530_36614651.rev	L2Plat1a
GCAAGAGTCATTTCTTGACCTTC	299	178	chr1_36679877_36680998	L2Plat1b
TGACTTGAAAGATTCTGGTGCT	392	261	chr1_37444356_37445487	L2Plat1a
GCAAAAGGTATGACCACAGAAA	296	171	chr1_37711987_37713112.rev	L2Plat1a
CAGGGGTGAACCTGATGATT	379	182	chr1_37811269_37812466	L2Plat1a
TCCACCACATCCTATGCCTAC	595	194	chr1_38287004_38288405	L2Plat1a
CACCTCCATGTGCAGTTGTT	294	196	chr1_38388974_38390072	L2Plat1a
TTCAATTGCCAAGGTAGAAGAA	300	163	chr1_38554159_38555296.rev	L2Plat1a
GGCAGATTTTGCTCTATAAATCATTAG	300	170	chr1_39094211_39095341.rev	L2Plat1a
TCCCAAATACCTTTCTGCATT	286	155	chr1_39134001_39135132.rev	L2Plat1a
TCGTGTGCTCAGTGCGTTAT	681	216	chr1_41764549_41766014	L2Plat1a
CCCCTTGAAACTGTAAGCTCTTT	300	180	chr1_42473515_42474635	L2Plat1a

GCCATTGCCAGAAGATTCAT	297	166	chr1_42778489_42779620	L2Plat1a
GCCCCACTGTGTTCTCAGAT	294	175	chr1_43123545_43124664.rev	L2Plat1a
TAGGGCAGGGAGCTTGTCTA	386	177	chr1_43411272_43412419	L2Plat1a
GGGACTACTCGTCATTGTCCTC	287	148	chr1_43920663_43921802	L2Plat1a
CCCACAAGGAATTTGCAATC	494	279	chr1_44231991_44233206.rev	L2Plat1a
TCCCAGCCCTATAGCACTCA	397	214	chr1_44279609_44280792	L2Plat1a
GATCTTCCCTTGCCCAATCT	288	167	chr1_44334673_44335794.rev	L2Plat1a
TGCTAGGAGTATTAGTAATGGATGTC	586	170	chr1_44464435_44465851.rev	L2Plat1a
GCTGCACACAATCCCTATGA	659	210	chr1_45346735_45348184.rev	L2Plat1a
CCAAATGAGACCAACCAATTCT	397	200	chr1_46990390_46991587	L2Plat1a
GGGATAAGGTCTATGTGCAACC	249	162	chr1_47442700_47443787	L2Plat1a
ACAAGTGGCAGAGCCAGAAT	378	160	chr10_882524_883742.rev	L2Plat1a
CAAAGAAGTAGAACACGAAGAGG	299	144	chr10_6549670_6550825.rev	L2Plat1a
AAGTCGCGGAGTCAGGATTA	636	175	chr10_6898887_6900348.rev	L2Plat1a
AAAATCTACTGCGGGCAAAA	471	168	chr10_8553255_8554558	L2Plat1a
TCAGCTGTGGGAGTGTCTTAAT	400	184	chr10_9825511_9826727.rev	L2Plat1a
TGCTGAATTTTGCCCATATC	297	148	chr10_10085167_10086316.rev	L2Plat1a
CAATCCAAGGGCACCAATAA	369	197	chr10_10239829_10241001.rev	L2Plat1a
GCTCCGTGTCTACCATATCTATTCA	300	161	chr10_10492767_10493906	L2Plat1a
CAGGATGAAGAGGGGAGACA	399	245	chr11_87787_88941.rev	L2Plat1a
TCATTCATTCATTCATTTTCATCG	391	148	chr11_599421_600664	L2Plat1a
CCTTAGACTCCTCCGTGAAGAA	482	155	chr11_845321_846648.rev	L2Plat1a
CAGGGACCACTCCTACCAAC	498	222	chr11_1422612_1423888	L2Plat1a
GATAATCAGGGAAGACACAGTCC	488	216	chr12_149219_150491.rev	L2Plat1a
TTTTAAACCCTCAGCAATATGAA	487	230	chr12_2355559_2356816	L2Plat1a
TGGCTCCTATGGGAATCATAA	300	172	chr12_3695162_3696290.rev	L2Plat1a
GACCAAGTTACTGAAGTCGAGA	296	169	chr12_5136112_5137239.rev	L2Plat1a
TCCCTGTAGCTACGACCAGAA	483	223	chr12_6345781_6347041.rev	L2Plat1a
CCCTCCATTTGCTTTGAGAA	380	177	chr12_6904309_6905512	L2Plat1a
GGTCAGAAACGATCCTCGTC	464	366	chr12_7142683_7143781.rev	L2Plat1a
CCCAGTTAGGGCTCACAGTT	600	159	chr12_7938737_7940178	L2Plat1a
TCCCCTCCACCAAGTCTTC	366	219	chr12_8969858_8971005	L2Plat1a
AGCTTGACCAGTGAGGGCTA	374	172	chr12_9351104_9352306.rev	L2Plat1a
TGCAGATCACCATCAATGAA	354	210	chr12_10542201_10543345.rev	L2Plat1a
AAGCCAGCTCCCTCATCTAA	700	225	chr12_11159665_11161140	L2Plat1a
CCTCATGCAGTATTGGAAAGC	471	202	chr12_11175800_11177069.rev	L2Plat1a
TTGCTAGTCAGTCAGCTCCTTG	347	218	chr12_11238248_11239377.rev	L2Plat1a
CCTGGCACCCCTACAGTTCTT	698	199	chr12_12786148_12787647	L2Plat1a
AGGTAGGAGGGGACAAGGTG	399	184	chr12_13281341_13282556.rev	L2Plat1a
GGGAAATTAAGAGCGGAAGG	399	191	chr12_13505340_13506548.rev	L2Plat1a
CTACCCCGGCACTTAGAACA	461	222	chr12_14492487_14493726	L2Plat1a

32 cycles each for 30 seconds at 94°, 30 seconds at 57° to anneal and 30 seconds at 72° elongation.

LY), intermediate (IF POLY), or high frequency (HF POLY) insertion polymorphism. Fixed present: the absence of the the element in nearly all individuals tested. Intermediate frequency: the element is present in nearly all individuals tested, expect for one or two individuals. High frequency: the element is present in nearly all individuals tested, expect for one or two individuals. Genome position was determined using ornAna1 assembly from UCSC version 5.0.1.



Population Diversity	Echidna
FP	N/A
FP	<b>FP</b>
FP	Empty
FP	N/A
FP	N/A
FP	N/A
FP	N/A
FP	N/A
FP	N/A
FP	N/A
FP	N/A
IF POLY	N/A
FP	Empty
FP	N/A
FP	N/A
FP	N/A
FP	N/A
FP	N/A
FP	N/A
IF POLY	N/A
FP	N/A
FP	N/A
FP	N/A
IF POLY	N/A
FP	N/A
FP	N/A
FP	N/A
N/A	N/A
FP	N/A
FP	N/A
paralog	N/A
FP	N/A
FP	N/A
FP	N/A
FP	N/A
FP	N/A
FP	Empty
FP	N/A
FP	N/A
FP	N/A
FP	N/A
IF POLY	N/A
FP	N/A
FP	Empty
LF POLY	Empty
LF POLY	N/A
N/A	N/A
N/A	N/A
FP	N/A
IF POLY	N/A
FP	N/A
FP	Empty
FP	N/A
N/A	N/A
FP	N/A
FP	N/A

FP	N/A
FP	N/A
FP	N/A
FP	Empty
FP	N/A
FP	N/A
HF POLY	Empty
FP	N/A
IF POLY	N/A
FP	N/A
FP	Empty
FP	N/A
FP	Empty
FP	N/A
FP	N/A
FP	N/A
IF POLY	N/A
FP	Empty
paralog	N/A
FP	Empty
FP	N/A
FP	N/A
FP	<b>FP</b>
FP	Empty
FP	N/A
FP	N/A
FP	N/A
FP	N/A
FP	N/A
FP	N/A
FP	Heterozygous
FP	N/A
FP	N/A
FP	N/A
FP	Empty
N/A	N/A
IF POLY	N/A
FP	N/A
FP	Empty
FP	N/A
FP	N/A
FP	N/A
IF POLY	N/A
FP	N/A
IF POLY	N/A
FP	N/A
FP	N/A
FP	N/A
FP	N/A
FP	Empty
FP	N/A
FP	Empty
FP	N/A
HF POLY	N/A
FP	Empty
FP	N/A
FP	N/A
FP	N/A
FP	N/A
FP	Empty
IF POLY	Empty
FP	N/A
FP	N/A
FP	N/A

FP	N/A
FP	Empty
FP	N/A
FP	Heterozygous
paralog	N/A
FP	Empty
FP	N/A
FP	N/A
FP	N/A
FP	N/A
FP	N/A
FP	N/A
IF POLY	N/A
FP	N/A
FP	Empty
FP	N/A
FP	N/A
FP	N/A
FP	N/A
IF POLY	N/A
FP	Empty
FP	N/A
FP	N/A
FP	Empty
FP	N/A
FP	N/A
FP	N/A
FP	Empty
FP	N/A
IF POLY	N/A
FP	N/A
FP	N/A
N/A	N/A
IF POLY	N/A
IF POLY	N/A
N/A	N/A
FP	N/A
FP	N/A
FP	Empty
FP	N/A
IF POLY	Empty
FP	N/A
FP	N/A
FP	N/A
FP	N/A
IF POLY	N/A
FP	N/A
FP	N/A
FP	N/A
FP	N/A
FP	N/A
IF POLY	N/A
FP	N/A
FP	N/A
paralog	N/A
FP	N/A
FP	N/A
HF POLY	Empty
FP	N/A
FP	N/A
FP	N/A
IF POLY	N/A
IF POLY	Empty

FP	N/A
FP	Empty
FP	N/A
FP	N/A
FP	N/A
FP	N/A
FP	N/A
FP	N/A
FP	N/A
FP	N/A
FP	N/A
FP	N/A
FP	N/A
FP	N/A
FP	Empty
FP	N/A
FP	N/A
FP	N/A
FP	N/A
FP	N/A
FP	N/A
IF POLY	N/A
FP	N/A
FP	N/A
paralog	Empty
FP	N/A
FP	Empty
HF POLY	N/A
FP	N/A
FP	N/A
IF POLY	N/A
FP	Empty
FP	N/A
FP	N/A
FP	N/A
IF POLY	N/A
FP	N/A
FP	N/A
FP	N/A
IF POLY	N/A
IF POLY	Empty
FP	N/A
FP	N/A
IF POLY	N/A
IF POLY	N/A
FP	N/A
N/A	N/A
FP	N/A
FP	N/A
FP	N/A
FP	N/A
IF POLY	N/A
FP	Empty
FP	N/A
FP	N/A
IF POLY	N/A
FP	N/A
paralog	Empty
FP	N/A
FP	N/A
FP	N/A
IF POLY	N/A
FP	N/A



## SUPPLEMENTARY INFORMATION

Sequences and accession codes for platypus gene predictions discussed in the text. Ensembl codes correspond to the January 2007 gene build. Additional accession codes are provided in Table 1.

(1)

Pou2. GenBank entry XP\_001520175.1

Pou5f1. GenBank entry ABU86919.1

(2)

enamelin. GenBank entry XP\_001512296.1

ameloblastin. GenBank XP\_001512321

(3)

Xanthine Dehydrogenase/oxidase (XDH) genes

ENSOANG00000014442

ENSOANG00000015165

ENSOANG00000002004

ENSOANG00000014836

ENSOANG00000014298

ENSOANG00000014299

ENSOANG00000014751

ENSOANG00000003947

ENSOANG00000003948

ENSOANG00000003950

ENSOANG00000012924

(4)

Natural killer receptor protein sequences

&gt;Contig35681

CSSEGWEAFGGSCYLICETSQNCGGKQPKCKECSDCDREKHGTWNCSQRECSGADLVT  
IDSKQELVSAPCLHLWVPLENINKTDRDHHCATIRHGNVSTASCNDTDTNHWICE

&gt;Contig26325

CSLNWLEVLGTCFVSFSAGTADWKESNTSCEVDGAILAMMTPQQMEILKTHVGSFTYWFGL  
NKEGDSSWKWVDNSAFNNEFTIHGTGEFAYLDATGVKSDEWLKWKWICS

&gt;Ultra500.8

CDHNWRFHGGKCYGSFKNNKTWEESKKYCDDRNSTLLKIDTQEAWELNLSVSSLIFLVP  
TLRRLSWILGARKEENFIQGRPDFTRWIGLSRPSSGGRWTWMDHSALTDNLLCSQLDIC  
HRQCKARTECL

&gt;Ultra500.7

CSSLKYHQNCYLLSHRNRTWEDSRTYCASKNYVMLKVDNQEELAYINRNTNKIRWIGLS  
RRAIDSPMWWEDGSVLATDLFQISGDGEANRHCAFFHNGKIQAADQCYESYPTLFP

&gt;Ultra500.6.1

CPADWKLHQGKCYWISKTEKKTWYESKVNCTQKHSNLAMIRDMCDLGFWWSQIPGSTYYW  
VGLHIPSPWNNWTWLDGSALNWNLFQVKGSDKGTCAVLSRDGIYPEDCNTRPWIFL

&gt;Ultra500.6.2

CPKDWWHENSCLERKQQTWEDSSKACNNRNSLVKIDSKEEWKTALSSGGAWGMGAW  
RGYGNDDGNNDGDNDDEHSERALTGSRLGNPGIISGSSNVSPPTQHRHWSEKSQLRSV  
CGIHPTDPATSPSAIRERERSFISLQYNQYIWVGLSRNSSSSQWKWEDGSALSPDLVIGH  
PRRQLGLGGEAGRFRGMVLEEKGNPLYSLQNFISQ

&gt;Ultra437.17

CSDTWHQLGEKCYKIFPNLKTWPGCKDHCAAQGSNLLRLETTEEMPVLFPLSHAASPESA  
PLGLPEDDPTSANRHLSPYCQDCRRTGAWLSTSCWVADIGVHKRADPGQFGHSPPTLVEL  
QFLTVPSSGQESVNMAVSLCNQDERKSWIGLYYDISLETWVWLHGSEFSIHQVYSY

>Ultra271.3  
FIGNWWVQFSMSCYQYSTEKKPWKESQQACKSQNSTLLHIDGLRELIYVHTLFIYINALSG  
MRCTYAHSHVSSNTPSVATSENDDDDNDDDGGLHLRKPDKQLMLYSSRIHVPGRVTSDL  
ICK

>Ultra117  
CLKGTKVHLKCFLAFPEAKTYHEASEDCISRGGTLSTPQTGDENDGLYDYLKRTVGNEGE  
IWLGINDLAAEGSWVDMAGGKIGYKNWETEITTPDGGKVENCAAMSGVAVGKWFDRCK  
DKLPFICQ

>Contig8321  
CWVGWYQYGQCYRIYQKLSWEECRNYCVALNSRLSTLKKKEEELFVIQLTRRQCLLND  
PKYWIGLNYVNRKWIWHDKTEISSDISKYSAWHIEFVIQLTRRQCLLNDPKYWIGLNYV  
RKWIWH

>Contig4675  
CPPNWNRRNGDSCYLFYRITLDNWNRSKMFCEQSQRSHLLRINSHEELVFIQHLSNNSQMSV  
WIDLTSRSDGSMWGDGVSFPHLFDIRQTNLSLNRCAWIHENLVFDALCSSLAYSISS

>Contig16386  
CLEPWVMNGESCYLFFDGGWKNWASSSEFCVQEKSELLKIGSKEELAFINRNIERKKTGSS  
WSYWVGLTQDKCYGDWRWRDSTVPSS

>Contig27609  
DSCYSRERVAQTSDESSTDCTAQNSSLVKFDKSKEEWNFISYFQCNYNIWVGLSRNSSSSQ  
WKWEDGSALSPGLLKFPYALNERKTCAYISAYHLSIDSCTNSHYICE

>Contig6618  
CPDTHWQLGEKCYKIFPNLKTWPCKCYHCATQGSNLLRLETKEEMMRELHRREVQFLTVV  
PSSQESVNQMAASLCNQDERKWCWIGLYNISLETWVWLHDSEFSIHQYISYLLNILMRK  
AFSPKSGFRTPIDYHTACN

>Contig11742  
HPGGDILEAGGDASLKGGEEDRGRDVDLRVICVEMVVEAVRANEFTKGELVIPLSFSKFF  
YSYWTGLSRNGSGQPWWLDGSPQASDLFQVVVSDSPRSRDCVTILNGKTFKDCCKELR  
RCACQ

>Contig1653.3  
LSEGWHSYNQNFYFFSNDSKTWAEAEGICESWDSHLTSVTSEGEQEPPTMSIMAWESNH  
AHGCHSNIQKQRGLVVTRILKSEEKARPEISPVLRKRFWDKGGQPDNWDQGEGLNENCVFFQ  
KITVQGWNDGNCDLKFHWMRE

>Contig1653.2  
FLGGWRLYGGNLYYFSQDKKSWDEAERFCKSWNSHLTSVMSEKEQDYLSKSKGSPHWIG  
LRERRTEHQWSWSDGTPFQEAESKWFWEEGQPDNWRQEEECVAMRTKERSWNDGNCQR  
LHWIYCK

>Contig1653.1  
LSKGWEYYGGHLYYFSKGKKSWDNAEKFCAAQNSHLASVTSVEEQEFLFKYTSGIYQWIG  
LTDKGKEGTWHWIDGTRYNEAENRRFWVDGQPDNWNQGLDLQEDCVHFQSEHTKSWNDGN  
CNHKYNWICK

>Contig8989  
CLRGTKIHKKCYFAPDGSKNFHEANEDCIAKGGTLAVPRNSEETNALRDYGGKSMPIGIND  
LWLGINDMIDEGKFVDVNGVSIHFFNWDGQPNGGKRENCALFSQSAQGWSEVCRSVK  
RYICE

>Contig24609  
YPERWICARDNCYYSMTKKWGESRKSCEALNSTLVKIDNKEELDFVSSNLQFCHWVGL  
LRKKDSNQWKWVDGSTLSPEL

>Contig14318  
CPENWHHRDSCYWKTSVLNLDSESRKVCAERNSSLVKIENKEELVYVASKLKAYHWLGLT  
RNISSDQVWVEDGSTLSPDLDLVASQKHESDMTLKSRSRVPRGPLFTRGYIRRFVSW

>Contig12615  
CPEHWYWHGNNCYKIMMENLTWSESREACAFQNSSLVKIDNKEEWYFVTAKIERYHWWGL  
SRNARDLPLKWEDGSEVNSE

>Contig33372  
CPPGWKRFEDSCYFFSVTTLWQEAKDHCIEQGAHLVIINNQQEQNFLTQNDGKGYWIGL  
TDMDTGTHQWIDGTDLIFX

>Contig5962.1  
CSPSWKAFEGSCYYSRRTALWHDVKKCAEKEAHLVIINSREEQNFLISNSDDRITYWIG  
LSSVRDGSQKVKHHRWIDETDLTFTYWSPGEPNDAGDQEDCVVMLSNGRWNTPCHTDLE  
KWVCE

>Contig5962.2  
CPIDWKTFFEGSCYFFSPTKSSWHSKSKCLSEGSHLVIINDQQEQNFLTQNTNNGYWIG  
LSDTEVEGKHKWIDGSDITFVYWNRGEPNDSYGREDCVMMLSHGHWNDA PCSSELDNWI  
S  
P

>Contig6405  
CPQDWRPFQTSCYFFPRDVL PWNEAQSCKLQQAHLVVINTEAEKNFITQGKPLNSSIHL  
GLKKRKKERQWRWEDKTPYNPDX

>Contig50169  
CPILARLSWTRKRLIQLLFYASLALCFILIIVLLAQGHLVIINSPEEQNFLSQQLNNIQA  
WIGLSDQRTEGVVHWVDGVALTLQFWDTGEPNNSGNEDCAEMKDKGRWNDVTCQKESVW  
I  
CE

>Contig19895\_CLECT.1  
CPGDWIGFRNSCYLFSNDRKNWRDSKSACAALNSSLLWIDTQEEQAFLAFFSLYAWIGLSRNEPGNSSK  
WEDGTDFSNHSDSVKKEGTPRCDR

>Contig19895\_CLECT.2  
CPGDWIGFRNSCYLFSNDRKNWRRESKSACAALNSSLLWIDTQEEQAFLAFFSLYAWIGLSCNEPGNSSK  
WEDGTDFSNHQLPVMKTQKGVNCFEVHAI TMKVNIMDENEIVNDINS

>Contig19862\_CLECT.1  
CPEDWIQLRKSCYLFSDKKNWHDSSKKTCAALKSSLLWIDSKEEEDFIVLFSIFAWIGLTCIGPGSSWK  
WESGTAFSSHX

>Contig9506\_CLECT.1  
CPEDWIGFRNGCDYYSNETMSWEQSKMACAALNSSLLWIDSQKEQDFLELFSLSAWIGLSRNRSGSSWQ  
WEKRAAFSN

>Contig30336\_CLECT.1  
CSEDWIQLRKSCYLFSDKKNWRNSKKTCAALNSSLLWIDSKEEEDFVELFSSFAWIGLTRIGPGSPWK  
WESGTVFSSHX

>Contig10595\_CLECT.1  
CPDDWIGFREKCYFFSEDTRNWTSSQSFCSSRSASLTGIDTQREMDLLLRHKGPSDHWIGLSREPGQDW  
KWTNGTKFTGWFEVKGGRGECAYLNDERVSSARSYTD RRWICS

>Contig12025\_CLECT.1  
CPQGWLHRHSCYFSEERKSWEESVRNCDARRSOLLVLKDEAEMDFVSNRKEEAFYFWLGLHFQEKEG  
KWLWLGHPPELHGYSRSGPEFRVHLEQCSAHSKRLTNTNII GNPEGEQARGGLNSSRDQQNCAAFRNKEEI  
HPDRCLSSYKWICK

>Contig9655\_CLECT.1  
CPENWFLYRNCCYLFSSKEKRKWNESRAICRSHNSGLLK IENKGELDFLSLFNININGWTGLSRSGRNQPWK  
WSDGSDLSKDLPRRLKEQQMGRRGGGRQLKRRRVKSCSFLVHEGDKEDWFLSATKSTEFYHGK

>Contig17262\_CLECT.1  
CPENWFLYRNCCYLFSSKEKRKWNESRAICRSHNSGLLK IENKGELDFLSLFSIIIGWTGLFRSGRNQPWK  
WSDGSDLSKDLYLICKMGIKTVSPTWDNLIPLCL

>Contig10412\_CLECT.1  
CPEKWIWHRDNCYLLSRQEKWNHSTLLRAYSVQNI TKITGNMSPWFGCGSTQDFLKLKSYWTGLSRN  
GSDGPWLWEDGSLSHNLLQGSQTQQFMKSYAPIEKEHGQV NKAQVWVPIPVSPLVCCGTLRCYSG



>Contig10412\_CLECT.2  
CPENWAWNRGNCYYIISTEAKTWLESQTACRSMNSSLLKIENREELQRITGNVSSWFGCGSIQDFLKLK  
SFYWTGLSQRGPDEPWLWLDRLSGLSHDLGRKADWLPQLLIQPTVSRFSSAIDTGKSLGSDSYQALESP  
FSVLNFASQ

>Contig10412\_CLECT.3  
CPQNWIWNRGNCYYIISTEAKTWPGSQTACKNMNSSLLKMEDREDLKRITGNVSSWFGCSSIQDFLKLK  
CYYWTGLSWRGPDEPWRWEDGSGLSHNX

>Contig14255\_CLECT.1  
CPEKWFLHRKSCYFFSDEKRTWSQSRTACRTQNSSLLKIEDQEMLDFLSRFYLFGWIGLSRPGVTRAW  
WTDNSKLPN

>Contig17200\_CLECT.1  
CPENWIWNRENCYYIISTQVKTWPESQTACQSMNSSLLKIDDREELQHLEQCLAHSKRLTNTITITITSSR  
FCCGSTQDFLKLKLLASYWIGLSLSNSSSGPWLWEDGSGLSHDLVQDQFSRGSMMYYGLKDKFSSEKCS  
HETFYICE

>Contig30657\_CLECT.1  
CPENWIWNRENCYYIISTQVKTWPESQTACQSMNSSLLKIDDREELQHLEQCLAHSKRLTNTITITITSSR  
FCCGSIQDFLKLKLLASYWIGLSLSNSSSGPWLWEDGSGLSHDLGSCMYGLEDKFSSEKCSHETFYICE

>Contig28730\_CLECT.1  
CPEKWLHRKSCYFFSSEKGTWSQSEAACQTQNSSLLKIENQKMVDFLFRFSLYGWIGLSRHKASGRWK  
WTDSDLLNLSKVVLKGEKDKCAKYTSTYPVFAENCTSTLKYICR

>Contig18019\_CLECT.1  
CPKKWVFRHNCYFFSSEKRTWYNSTAATAQNSSLLKIDNKEELIFLNQLPIYFWTGLYHLENNGFQ  
WADGSALSSKLLILISAMVLLKSDRHITVEGVGLRKLLELEKRPIGAVSE

>Contig37996\_CLECT.1  
CPENWIWNRGNCYYLSREVKTWPGSQNAQRSMNSSLLKIEDREELQDFLKLKLSYYWTGLSRTGSDGPW  
LWEDGAGLSHDLVQDIFSNERCANYHPQKDEFSSSECLRSFSY

>Contig36753\_CLECT.1  
CPENWIWNRGNCYYLSREVKTWPGSQNAQRSMNSSLLKIEDREELQDFLKLKLSYYWTGLSRTGSDGPW  
LWEDGAGLSHDL

>Contig6664.1\_CLECT.1  
CPENWAWNRGNCYYIISTEAKTWLESQTACRSMNSSLLKIEDREELQRITGNVSSWFGCGSIQDFLKLK  
SFYWTGLSRRGPDEPWLWLDGSGLSHDLFHVYSLEAPGWDSVSPGPGPHGIRALTGGSHGKPRIDDSRI  
AERRLNWQLQSSTCK

>Contig6664.2\_CLECT.1  
CSENWIWSRGSCYYISKESKPWPNSQAACKKKNSSLLKIDSREELEILRGRGCLTLHQHREVMHVTTSI  
QIVFQ

>Contig29323\_CLECT.1  
CPENWAWNRGNCYYIISTEAKTWLESQTACRSMNSSLLKIEDREELQRITGNVSSWFGCGSIQDFLKLK  
SFYWTGLSRRGPDEPWLWLDGSGLSHDX

>Contig6191\_CLECT.1  
CPVNWNVYRNSCYLISMERKTWQDSHMICVAQNSSLKIDSMEELDFSSLSLFWHMGLNWSQNTGWSQ  
WTDGSPSSSLLFCRTDSHFGWENTFTILRSVHENRNTSHRSERKGLQPTFSEQLVRCQA

>Contig12161\_CLECT.1  
CSENWIWSRGSCYYISKESKPWQDSQAACKKKNSSLLKIDSREELESFLKYLKLFYWIGLSRNSTTGPS  
LWEDASALSQDQLTVINPSHTGNCAFYFRGHYVFYCYVASNTYICE

>Contig16183\_CLECT.1  
CSENWIWSRGSCYYISKESKPWQDSQAACKKMNSSLKIDSREELESFLKYLKLFYWIGLSRNSTTGPS  
LWEDASALSQDQLTVINPSHTGNCAFYFRGHYVFNEYCVASNTYICE

>Contig66858\_CLECT.1  
CPVNWNVYRNSCYLISMERKTWQDSHMICVSNSSLKIDSMEELAVMLLKFFPDYGFLLKSLKSVQFRI  
RPTRDQSFPGYGCNFMLEMTLSDQARERRQEVFVSHLVLLCKFSC

>Contig7665\_CLECT.1  
CPKHWIGFRNSCYLFSNNRKDWRDSKIACATLNSLLWLDNREELDFLGMFYSYSAWTGLFHNGPRSSWQ  
WEDGTAFSIHGITFLKQSGGNCVYQQTWYLYQENCEALNWFICK

>Contig4625\_CLECT.1  
CPKHWIGFRNSCYLFSNNRKDWRDSKIACATLNSLLWLDNREELDFLGMFYSYSAWTGLFHNGPRSSWQ  
WEDGTAFSIHX

>Contig18593\_CLECT.1  
CPQNWIWNRGNCYYIISTEAKTWPGSQTACKNMNSSLLKMEDREDLKRITGNVSSWFGCRSIQDFLKLK  
CYYWTGLSRRGPDEPWRWEDGSGLSHNLSEQEHFFRGRCVHYGFSNQFSSDCHHSTTYICP

>Contig74026\_CLECT.1

CPVDWMNYRNSCYLFSMENTTWQDSHEACVSNSSLLKIDSKEELDFWQFSLFYWIGLNHSQNAGSWQ  
WTDGSPLSHGX  
>Contig16585\_CLECT.1  
CPENWILYRNNCYLFSKEKRNWNESLAICRSHNSGLLK IENKGELDVLFSFKINGWTGLSRSGRNQPWK  
WSDGSDLKDLLTITLTKNGKDKCAKYIPGYPAYAEQCESLQMYICK  
>Contig23054\_CLECT.1  
CPENWILYRNNCYLFSKEKRNWNESLAICRSHNSGLLK IENKGELDFMFLFKINGWTGLSRSGRNQPWK  
WSDGSDLKDLLTITLTKNGKDKCAKYIPGIPAYADQCESLQKYICK  
>Contig68432\_CLECT.1  
CPENWFLYRNNCYLFSKEKRKWNESRAICRSHNSGLLK IENKGELSVTDNNPKVSLFNDVVVRNLVTSR  
LVEARDTLSQSSPTTQPFPTQ  
>Contig15207\_CLECT.1  
CPPQWKHFHQGKCYLFSRGEKNWNESKADCTKKESSMLVIQDVRDLEFVQNTIFEGALFWIGLRAASLET  
AWTWVDHSPYNQLLFKITGNRNGEVCVQLSKMAYSENCESEINWICQ  
>Contig17956.1\_CLECT.1  
CPEHWISAREKCFYFSTTIRNWTSSQHFCAFHGATLAKIEDQMELDFLKEHEGHSDDHWIGLRRSESNQA  
WEWTDGAKFNGLFAVRGEGEYSYLNENGISSARQFIEKSCPPGEVVLQ  
>Contig7463\_CLECT.1  
CPADWKWHQGKCYWISQTEKKTWNESKANCTQKDSNLTIKDMCDLGFWSQMSASTYYWIGLHIPNPG  
SNWTWPDGSGLDWNLFQVKGSDKGICAVLSRDGIYPEDCNNTRPWICE  
>Contig34561\_CLECT.1  
CPDNWVQFEKTCYFFNGKKNWTRSQEFCKSQDAELVVIKNQQELSLVRKFQDSWIGLYRKNGELYWVNG  
VPLNESLFEVTDPGDCAYVSANKIATNACSTPKFYICS  
>Contig4063\_CLECT.1  
CPDDWIGFRGKCYLLSNIPDTWNASKRFCSNLNATLTKVESEKDLDFLKRLTGTSEHWIGSISGTIQGL  
SWTTSNLLSNSRESEAAVTESTASLLYVVAHFNGTTPMILSCP  
>Contig3438.1\_CLECT.1  
CSSAWLEVNETCFFLSTKTRDWEKSKTFCEQEATLAMMTIVQLEYLKTQVGTSDYWIGLTKKSSSDWR  
WLNGSALNNLNFNIRGAGDCAYLDSSSVNSAGCSQPRRWICS  
>Contig3438.2\_CLECT.1  
CPVDWKMYRGQCYWFSKNNEKKTWNQSTAFCAEKKSNNLPKIQNICDLGFWSQIPPSTFFRIGLHILRP  
GANWTWLDGSALDWSLQVKGSTGMNTCAVLSRDGIYSENCDAENPWICQ  
>Contig8580\_CLECT.1  
CPETWFGNGRSCYVVTQOKASWHTSKTICASLKASLLQISSREELDYLKAVPIIGWMGLSREGPGAPWK  
WLNSSDEPSNVVEIHNLDQKGDCTVFRSRDRIFYDDCGNNYPYICE  
>Contig2984\_CLECT.1  
VDGNWVQFSTSCYQYSAEKKPWKESQQACKSQNSTLLHIDGLTELNFFKFFTLSGWIRLSHTGPGSSLK  
WEDETTYTRNFDLICKMGMKTVSVTWDNLI TRYLPQHLEQCSAHKNGEDCQ  
>Contig23122\_CLECT.1  
CPLGWLNNSSRCYHFSEEQRSWEASARNCADRKSWLLVLEDEAEADWVHAGKQRDEYFWIGYKYNIIEE  
QWTWLDNSGFSGYRIKVN GDANGNDCASFKRKNEFGFVKCEDLLHWICK  
>Contig26044\_CLECT.1  
CPDGWLQLRNKCYSYNNGNINWPKSKEVCKSQGANLVVIESHQELSFVQKFKGHWIGLFRENGKLFVWN  
GVPLNESLFPVTDPGNCAYISTNKIATNGCNESKPFVCS  
>Contig49354\_CLECT.1  
KDGNNWVQFDTSCYRYSAELKPWQESQQVCKSQNSSLLHINRREMDFFKLFRLSWVWGLSRKGPTSPWQW  
EDGTIYSRK  
>Contig8101\_CLECT.1  
CPGKWFYRKMSCYHYSTERKSWKESKQACASQKSNLLHIDSTDELDLFLKLLTLRGWIGLIHNESRASWQ  
WENGTEFSSKILSFVDLEYGGSCVLYGSDLAYLADCAKDRHYIFS  
>Contig8101\_CLECT.2  
CSVGWYQYGQKCYRIYCKLNSWEECQNYWRSSVAQWKEHGLWRVCDLTRRQCLENDPKFWIGLKYVNQK  
WIWPKSVISSDMFQQFQFTHQTNDKSCVTLQKRQLQAETCSAHGRCLCQ  
>Contig2815.1\_CLECT.1  
CPGKWFYRKMSCYHYSTERKSWKESKQACASQKSNLLHIDSTEELDFLKLTLHGWI GLIRNESRASWQ  
WENGTEFSIDLGYGGSCVLYGSGDSAYVGD CATDRHYICL  
>Contig43923\_CLECT.1  
STGNWVQFSMSCYQYSTEKKPWKESQQACKSQNSTLLHIDGLRELNFFKFFKLSGWIGLSRTGPGSSLK  
WEDETTYTRN  
>Contig22326\_CLECT.1

QGVSTNVFLNPSPSNIKETDKKNWRDSKKTCAALNSSLLWIDSKEEEDFVALFSIFAWIGLTRIGPASP  
 WKWESGTVFSSHRLTFIKIQTGGNCIYQSGNYVFE  
 >Contig20594\_CLECT.1  
 CRENWVAYRNRCLYFSKESASWEGSRTAWAAQNSSLQIDSREERDFLKLKLSLYHWIGLSLSENEKSWQ  
 WTDGSPLSSDX  
 >Contig18885\_CLECT.1  
 LHTGSCYQYGEKYRVFQQLKTWNDSRDFCIARNSTLLHVENMEELEFIQTKLQEFWSWIGLTRQRPTSP  
 WMWVNGSALSSALNILEPNQKLENSCGILSNDSLQAKACQDISHFMPD  
 >Contig3953\_CLECT.1  
 CSGQQGWLNC SRCY YFSEERRSWEAGATDCANRKS KLWVLENEPEVDCITEMKGRDEYYWIRYNYNTI  
 QGQWTWLDNTGFSGYRVKEIKNFTGKDCASFKKRNSFAPENCNP SHYWICK  
 >Contig21884\_CLECT.1  
 CPPGWQLSNSRCY YFSGECSRWDTSERDCTDKKSHLLVLEDEAEADWISKMRPKDEYFWIGYKYNITQE  
 QMMWLDNTGFSGYREGLTEVLGEDVYFNIQS  
 >Contig10996\_CLECT.1  
 SQADWKLNHGKCYWISNSNHTKNWGSNAFCVENKSNLAKIQDLCDLGF IWSQIPASKYYWIGLHIPKS  
 RGNWTWLDGSALDWS  
 >Contig28883\_CLECT.1  
 CPPGWRLSNSKCY YFSGEKQSWEASARNCANRKSRLLVLEDEAEAGR VQEMRPRDEYFWIGYKYNITTQG  
 QMMWLD DDPGFSRYRWAPLPVQE IPLTQSMQK  
 >Contig14764\_CLECT.1  
 CPADWKLNHGKCYWISNSNHTKNWSR SNAFCVENKSNLAKIQDLCDLGF IWSQIPASKYWIGLHIPKYR  
 GNWTWLDGSALDWSL FHVNPPGGMDTCAALS RDGIYQSNCRGENPWICQ  
 >Contig13620\_CLECT.1  
 CPPGWLWLNSSRCY YFSGEERSWEASAKDCADRKSRLLVLEDEAEAVEVPRMGPRDEYFWIGYKYNMTQR  
 AWTWLDNAGFSGY  
 >Ultra271.1\_CLECT.1  
 CLGKWFICYKTSYHYSTERKSWKESKQACASQKSNLLHIDSTEELDFLKLTLHEWIGLICNESRASWQ  
 WENSTEF SIDLGYGGSCVLYGLRDLAYLEDCATDRHYICL  
 >Ultra271.2\_CLECT.1  
 CPEDWIEFRNSCYLFSNDTKNWRDSKIACATLNSSLWLDNREDLDFLGIFSHIVVTGLYRNGPHSSWW  
 WEDGTVFSIHGNKSRKAGSLAFTSTLTQSGGQTAILRASLRGRKDESKPFVSR  
 >Ultra271.4\_CLECT.1  
 CPGKWFYRYKTSYHYSTERKSWKESKHACASQKSNLLHIDSTEELDFLKLTPHGWIGLIRNGSRASWQ  
 WENGTEFSIDLGYGGSCVLYGLRDLAYLEDCATDRHYICE  
 >Ultra271.6\_CLECT.1  
 ETEDWIKLRKSCYLF SKDKNNWGD SKKTCAALNSSLLWIDSKEEELIIVSCARRPSKSCIRVLSPNSL  
 AVLEIGAKNRIWSSASWRYRLKSLISLMSMIASNAEDTEPWPMGYMLQNLLIHLTGVIDCLLCP  
 >Ultra345.1\_CLECT.1  
 CPEDWLQLRKKCYFSSSEHESKRNWSDSQEYCRSQGGDLVVIENKQELVFIQQFKNFWIGLHRTNEGFF  
 WVTGAPLNQSLFNHYHLFADDTQIYISPVLSPSLQARISSCLQDVSTWMSA  
 >Ultra437.1\_CLECT.1  
 CPEDWIGFRNGCY YFSNETMSWEQSKMACAALNSSLLWIDSQKEQDFLALFSFSAWIGLSHNRSGSSWQ  
 WEKRSFAFSLNLFMQYYAMSRSPPKIIVMEENEIVNDNNSQIIDTSIFVQS  
 >Ultra437.2\_CLECT.1  
 CPGDWIGFRNSCY YFSSNDRKNWGD SKSACAAQNSSLWLDNQE QVFLAFFSLYAWIGLSHNEPGNSWK  
 WEDGTD FSNHQLTLIKIQTEGNCVFRSARYDYEGNCEDLKFVCK  
 >Ultra437.3\_CLECT.1  
 FWRNWVQFDTSYRYSAELKPWQESQVCKSQNSSLLHINRREMDFFKLFRLSWWVGLSRKGP TSPWQW  
 EDGTIYSRKNLDTYEGKGRDCALFSSERSIHIENCASARWYICE  
 >Ultra437.4\_CLECT.1  
 CPVNWMMYRNTCYFVPVENKTWQDSHKACISLNSSLVKIDSKEELDFNSLYLYHWIGLNRSQNAESWK  
 STDGSRFSSGLFNILKSPGHAENCTTAKNCICE  
 >Ultra437.5\_CLECT.1  
 CPHELWFGNGKNCYQISLEKKSWFESKIACASRN TSLQIDSKEELSLCHSPTQLTILRTQFSPYLTRMK  
 PGPHLILIQ TSSVHTYQALKNFQWLPTYICDEQKLLAIGFNDFLKLLPMMVWIGLSRHSNHEPWKWDG  
 STLYTDLINLQNSQKNQ TGNVFRSIDRTFHDNCETKNLYICE  
 >Ultra437.6\_CLECT.1  
 CPENWIWRNGNCY YLSREDKNWSESQTSCRSMNSSLLKIDGNKEVDDFLKLLTFHYWIGLSRNLS DGPW  
 LWEDRSEL SHDLTFYKVFVKPPNYSYYYQWFEEVLKERFQQSTTGMSSWKSMKQKEQPADQG  
 >Ultra437.7\_CLECT.1

PRENWIRNRGNCYYIISIDVKTWRESQTSCQSMNSSLLKIEDRKELQHLEQCLAHSKCLNTIFIIVTSSW  
 FGCGSIQDFLKLKLVSDYIWIGMSLSNSSGPWLWEDGSGLSHDLGIDVDCLEATQRAPKNDEDLSGEGEWR  
 REEEAPASIFISW  
 >Ultra437.8\_CLECT.1  
 CPQNWIIWRGNCYYIISTEAKTWPQSQTACKNMNSSFLKIEDRKDLKDFLKLKLSYYWTGLSRRGPDEPW  
 PWEDGSGLSHDLKRKYFVSVDLLQGKAPLMFLIFCH  
 >Ultra437.9\_CLECT.1  
 CPENWAWNRGNCYYIISTEAKTWLESQTACRSMNSSLLKIEDREELQRITGNVSSWFGCGSIQDFLKLK  
 SFYWTGLSRRGPDEPWLWLDGSGLSHDLNGMQTPRSSYGLVDEAQAWKSKKDLRSNPGSTTCL  
 >Ultra437.10\_CLECT.1  
 CPVNWMMYRNTCYFASVENKTWQDSHKACVSLNSSLVKIDSKEELDFFNLSLYLHWIGLNRSQNAGSWQ  
 WTDGSLSSGLFDINQKSSGGDCIAYGSRSTSGHAENCTTVKNYICE  
 >Ultra437.10\_CLECT.2  
 CPGNWVQFNMSCYRYSAELKLWKESQQACKSQNSSLLHINNTGEMDFFRLFRSLSCWVGLSRKGPSTPWQ  
 WEDGTTYSSKNLDTYEGKGRDCALFSSQRSIHIENCASARCYICE  
 >Ultra437.11\_CLECT.1  
 CPGDWIGFRDSCYLFSDNRKNWRDSKSAARNSSLLWIDNQEEQVFFLEFFSLYAWIGLSRNEPGNSWK  
 WEDGTDFSNHQFTI IKIQTEGNCVFSARYYYEGNCEDEFFVCK  
 >Ultra437.12\_CLECT.1  
 CRENWVAYRNRNCYLFSKESASWEGSRTTCAAQNSSLLQIDSREERGPGLQNQGGDSLGGIPHELVQGO  
 DFLKLLSLHYWIGLSLPENAKSWQWTDGSPLSSDLQIPMTDNDAQNVNTPQVGFSEKLGKIPDLNHP  
 SFVCF  
 >Ultra437.13\_CLECT.1  
 CPGDWIGFRNSCYLFSNDTKNWRDSKIACATLNSSLLWLDNQEELDFLGLFSPAWTGLYRNGPHSSWQ  
 WEDGTAFSIHGFPKRWVMKPLRVAFTLEMSERQVFMELKK  
 >Ultra437.14\_CLECT.1  
 SKGNWVQFSTGICYQYSAEKKPWKESQQACKSQNATLLHIDGLRELIVYVHTLFDINSFFKFFKLSGWIG  
 LFHTGPGSSLKWEDETTYTRNLEEGERQVWNPWFSDDEETETQRSSGDFPKVTQQCLEQCSAHSVCS  
 >Ultra437.15\_CLECT.1  
 FKRNWVQFMSCYQYSAEKEPWKESQQACKSQNSTLLHIDGLRELNFFKFFQLSGWIGLSRTGPDSSLK  
 WEDETTYTRNFLSFTKERKGGDCALYLNEKSISMEDCAIGKLHICE  
 >Ultra437.15\_CLECT.2  
 CPVNWVNYRNSCYLISMERKTWQDSHMICVSNSSLLKIDSMEELDFSSLSLFWIGLNSQNTGWSQ  
 WTDSSPLSSSLFRFTPKATVACIVYSSRTSLFTENCAVLKSYICE  
 >Ultra437.16\_CLECT.1  
 CPGKWFYRYKTSYHYSTERKSWKESKQACASQKSNLLHIDSREELGFLKLLILHGWIGLIRNESHASWQ  
 WENGTEFSSKLISLVDLEYGGICVLYGSGDSAYVGDCAKDRHYICE  
 >Ultra437.18\_CLECT.1  
 CSENWIWSRGSCYYISKESKPWPNSQAACKKKNSSLLKIDSREELESFLKYLKASYWIGLSRNGTAGPL  
 LLTVINPSHKGNCAYFGMGLYVFYEHCDASNTYICE  
 >Ultra437.20\_CLECT.1  
 CPENWIWRGNCYYLSRENKNWSESQTACRSMNSSLLMIEDREELQDFLKLKLSYYWIGLSRTGSAGPW  
 LWEDGSGLSHDLSDYIYTEMGEREVIPMEERLRGPAKLPTAGVILQ  
 >Ultra437.20\_CLECT.2  
 CPENWIRNRGKCYIISTEVRTWQESQTACQSMNSSLLKIEDREELQHLEQCLAHSKPLNTIIIIITSS  
 WFGCGSIQDFLKLKLVVSYWIGLSLSNSSGPWLFTRSWRLLAGILGILGLMLVAALALGVLIANQSYE  
 GSERPPSHVNTTSERVTLSLHPR  
 >Ultra437.20\_CLECT.3  
 CPENWIWRGNCYYLSREDKNWSESQTACRSMNSSLLKIDGSKEVDDFLKLLSSFYWIGLSRNGSDGPW  
 LWEDRSELSHDLMNKLNVAFTTEKKMSHSTQKQKGRMHGTREDSLSAPWKLLAGILGVLVSLVLM  
 >Ultra437.20\_CLECT.4  
 CPEKWIWSRGNCYYISNDIMDWQGSRTDCMKRNSTLLKIESREELNFNLLSAMYWTGLSHNVTASSW  
 MWDDGAALSQELGIDVDCPEATQRAPKNDEDLSGEGEWRREEEAPASIFISW  
 >Ultra437.20\_CLECT.5  
 CPAKWIWNSGNCYYISTKIKTWPESTACRSMNSSLLKLENREELQDFLKLKLFYYWIGLSQNGSDGRW  
 LWEDGSGLSYDLTCFLCQKPKIRSREQIISDPWTSVLCK  
 >Ultra437.20\_CLECT.6  
 CPENWIWRGNCYYLSREDRNWSESQTACRSMNSSLLKIEDREELQDFLKHLLKCYWIGLSRNGSDGPW  
 LWEDGSGLSHDLPLVQDRGSCVFFYGLKDKFSSSEDCFHSTLYICE  
 >Ultra437.21\_CLECT.1

CPENWAWNRCNYIISTEAKPWLESQTACRSMNSSLLKIEGREELQDFLKRRLKTYWTGLSRVGPDEPW  
LWLDGSGLSHDLGCSLVLQNGPLRFDSFGVCR  
>Ultra437.22\_CLECT.1  
CPEHEAEDTAKSYPRSNVCTSPPTASTSAPGMFPRDWKLHEGKCYWFSNETERKLGTKIPPSKYWIGL  
HILSSRNWTWLNGLDWSLFPVKASEAQSKCAVLSRDGIFPDKCENGYSWICE  
>Ultra437.23\_CLECT.1  
CPEPWFNGKSCYQISGEKLTWEKSREACTSKNASLLQIDNRNELEEKQTKLFCDSKQDSIKILPMMGW  
MGIFRESPVKPWRWLNSELSLGLHCPEVGPKGPLLSTAWSLSIKLFTYACREEKQTKLFC  
>Ultra437.23\_CLECT.2  
CPEPWFNGKSCYQISGEKLTWEKSREACTSKNVSLQIDNHKELRSSVAQWKKPGLESQRIQFSNTQG  
KGNCVVFRRSDVFPIDECSEFTYICK  
>Ultra437.24\_CLECT.1  
CPENWIWFRHSCYRFFSESKTWRESRSSCEAQNSSLLTLKSQEELELFTLQLFHWIGVYLKEETGTWIW  
EDGSEVSSKLLSFPTNQKDRTCGIYKSRHSAFSENCTIKNPYIYS  
>Ultra437.25\_CLECT.1  
FPRDWKLHEGKCYWFSNMTEKKTWKKSKANCTHRESNLMIRDNYDKDFLLSQIPPSKYWIGLHIPSS  
RNNWTWLNGLDWSLFPVKASEAQSKCAVLSRDGIFPDKCENEYSWICE  
>Ultra437.26\_CLECT.1  
CPVPWFGDGKNCYQISSEKKSWEESKNDCKSWNATLLQIYNKEELDFLKLPMMGWGLSRQSADDAWK  
WTDGSILQKDLIQINDVQEPGNCVIFRTIDVPYANCKKKITFICK  
>Ultra500.1\_CLECT.1  
CPKSWLRFHKNCFYLSTEARTEGSKSFCASLEATLSKITTKQELDLLKKQVGLFNFWIGLSRKKDSIW  
RWTDDNVFNSLFIITGGGECAYLDASGVKTASCNQPRRHLC  
>Ultra500.2\_CLECT.1  
CSEHWKLVEKCFSLSSETRDWDSSQENCKGHKATLAVVSQKQLEILTKGVGNSDYWIGLKKKFDQWKW  
DDNTTFNKEFNIRGAGNCAAYLDSSSVNSAGCSQPRRSKLT  
>Ultra500.3\_CLECT.1  
CPADWKLNHGKCYWISNSNHTKNWSRSNAFCVENKSNLAKIQDLCDLGFISQIPASKYYWIGLHIPKS  
RGNWTWLDGSAALDWSLHFVVKPPGGMDTCAALSRDGIYQNNCRGENPWICQ  
>Ultra500.4\_CLECT.1  
CPADWKSDBGKCYWFSNVTDKKSWNESIEFCAEMKSNLTTIQEMCDLGFISQIPLSKYYWIGLNLHP  
GGNWTWLDGSAALNVRQIFQITHSPGKDTGALSARDGIYPENCESLNPWICQ  
>Ultra558\_CLECT.1  
CPPGWIEVDGNCFFLSTGTKDWEGSKKACEEHNALLTKLSVLQMEVLKSLVENSNYWIGLKKSGDSGWV  
WNDNSAFHEEFSIRGAGDCAYLDSTSVNSASCSQHRRWICCK  
>Ultra76.1\_CLECT.1  
STGNWVQFMSCYQYSTEKKPWKESQACKSQNSTLLHIDGLRELNFFKFFKLSGWIGLSRTGPGSSLK  
WEDETTYTRNLEEGEQVWNLHFSDEETETQRSSGDFPK  
>Ultra76.1\_CLECT.2  
CPRDWIGFRNSCYLFSNDTKYWRDSKIDCATLNSLLWLDNREELVNYSSILLSRFQLEKAKDSIMPFW  
FSCDSQQDFLALFSPATAWTGLYRNGPHSSWRWEDGTVFSIHGILDVNCTDFSVIRMAPLFLSARCFSE  
LDIKD  
>Ultra76.2\_CLECT.1  
CPDAWHQLGEKCYKIFPNLKTWPCKDHCAAQGSNLLRLETKEEMESVNRMAVSLCNQDERKSWIGLYY  
DISLETWVWLHGSEFSIHQATAAMSKADKRRSLCIDVGHLEVSIQYFKNHLPQTMLK  
>Ultra76.4\_CLECT.1  
CPVDWIGFRNSCYFFNETMSWGQSKMACVALNSSLLWIDSQKEQDFLALFSLSAWIGLSRNISASSWQ  
WEKRAAFSDLPYLIYKMGIKTVSLMWDNLITLYLPQLLE  
>Ultra76.4\_CLECT.2  
SKGNWVQFDTSCFRYSAELKPWKESQACKSQKPSLLNINNTREMDFFKLLRLSWWVGLSRKGPTSPWL  
WEDAKSYSRKNLNTYEGKGRDCALFSSQRSIHIENCAMARSSLIGK  
>Ultra76.4\_CLECT.3  
CPGDWIGFRNSCYLFSNDRKNWGGSKSACAAQNSSLLWIDNQEEQVFLAFFSLYAWIGLSRNEPGNSWK  
WEDGTDFSNHQ  
>Ultra76.5\_CLECT.1  
CPEDWIEFRNSCYLFSNDTKNWRDSKIACATLNSLLWLDNREEDLEYRDRTCYDGTWQDFLGIFSHIVW  
TGLYRNGPHSSWWEDGTVFSIHGSLICKMGIKTVSPPWDNLIPLCL  
>Ultra76.6\_CLECT.1  
CPENWIWDRGNCYIISTEFRSWPESQTACRSMNSSLLKIENREELQNFLLASYWIGLSRKGSDAPW  
LWEDRSDLSCDLSLICK  
>Ultra76.7\_CLECT.2

STGNWVQFSMSCYQYSTEKKPWKESQQACKSQNSTLLHIDGLRELNFFKFFKLSGWIGLSRTGPGSSLK  
 WEDETTYTRNFLKLNMSKTELLIFPPKPCPRPDFPITVDGTTILPVSQAHNLGVI FD  
 >Ultra76.7\_CLECT.3  
 CSENWIWSRGCYIISKQSKPWPNSSQAACKKKNSSLLKIDSREELESFLKYLKASYWVGLSRNGTAGPS  
 LWEDASALSQDELTVINPSHKGNCAYFGMGLYVFEQCDASNTYICE  
 >Ultra76.8\_CLECT.1  
 CPEDWIQLRKSCYLFSDKKNWHDSKKTCAALKSSLLWIDSEEEEWESGTAFSSHRLTFVKIQTGGNCI  
 YQSGKYVFESNCLDLKKNQKDDAHGMGLKGPDPQGTVELPVPFIHLDRHQS IKVPSEKAREIMIPCPGGP  
 QWVWQTTNQQOQRQLPVQCQ  
 >Ultra76.9\_CLECT.1  
 CPGKWFICYKTSYHYSTERKSWKESKQACASQKSNLLHIDSREELGFLKLLILHGWIGLIRNESHASWQ  
 WENGTEFSSKLLISLVDLGYGGSCVLYGRWDSAYLEDCATDRHYICE  
 >Ultra76.11\_CLECT.1  
 PMPRYLASAWREMLQDFSNLKTWPECKDHCATQGSNLLRLETKVEMESVHRMAASLCNQDERKCKWIGWY  
 YNISLKTWVWLYGSEFSIHQFHVYFLAASGRDSDGHPGGLSICASTGDPHNKLLFWPRIREPFFSSQH  
 HNPKSYLIWKMEIQTVSPDFLKLKLLISDYWIGLSRSGSDGPWLWEDGSGLSHDLVQDIFPNGWCVNYP  
 RQNEFSSEDCLRSFSYIFV  
 >Ultra76.12\_CLECT.1  
 CSENWIWSRGCYIISKESKPWPNSQAACKKKNSSLLKIDSREELESFLKYLKASYWVGLSRNGTAGPS  
 LWEDASALSQDELTVINPSHKGKCAAYFGMGLYACYEHCASNTYICE  
 >Ultra76.13\_CLECT.1  
 CPGWFRYKTSYHYSTERKSWKESKQACASRKSNNLLHIDSTDELDFLKLTLHGWIGLIHNESRASWQ  
 WENGTEFSRLSRGPSWVRIPSVSVSSESSRVVSGQEAELMQSLLLLGADLPGRGLKLLLQNTQDMPTV  
 KK  
 >Ultra76.13\_CLECT.4  
 ETEDWIQLRKSCYLFSDKKNWHDSKKTCAALNSSLLWIDSKEEEDFVELFSSFAWIGLTRIGPGSPWK  
 WESGTVFSSHRHKPGPLDQLSPECIFQAEVGEQQVFYTELKQKSSQSHTKRPKINK  
 >Ultra76.14\_CLECT.1  
 CSGNWWQFSTSCYQYSAEKKPWKESQQACKSQNSTLLHIDGLRGLNFFKFFTLSGWIGLSRTGPGSSLK  
 WEEETATYTRNFDLICKMRMKTVSVTWDNLITRYLPQHLEQCSAHKNGEDCQ  
 >Ultra76.15\_CLECT.1  
 CRENWVAYRNCRYLFSKESASWEGSRTAWAAQNSSLLQIDSREERDFLKLKLLSLYHWIGLSLSENEKSWQ  
 WTDGSLSSDLLPYRAQKQEPQYVGTSLFPSTVDSHSWSF ILPLGEQQDLEQCSVHRDMATNSVVLV  
 >Ultra76.16\_CLECT.1  
 CPENWFLYRNCCYLFSEKRNWNEIRAICRSHNSGLLK IENKGELDFLSLFSINRWTGLSRSGRNQPWK  
 WSDGSDLSKDLPRRLKEQQMGRRGGRQLKRRRVKSCSFLVHEGDKEDWFLSATKSTEFYHGK  
 >Ultra76.16\_CLECT.2  
 CPENWILYRNCCYLFSEKRNWNEIRAICRSHNSGLLK IENKGELDFMFLFKINGWTGLSRSGRNQPWK  
 WSDGSDLSKDLLSGRGENVKVPVFGKFGSESDYRSFSPHNVHSTIQV  
 >Ultra76.18\_CLECT.1  
 FRRNWIQFGMNCYHYSAELKPWKESQQACKSQNSSLLHINNSREKDFFKLFRLLSLWIGLSRDIPSTSWQ  
 WEDGMAYSRQKQVSPQYCS  
 >Ultra76.19\_CLECT.1  
 CPVNWMMYRNNTCYFASVENKTWQDSHKACFSLNSSLVKMDSKEDLDFNSLYLYHWIGLNRSQNAGSWQ  
 WTEGSLFSSGLFNKLSPPGGDCIAYGSRTSGHAENCTTAKNYICE  
 >Ultra76.20\_CLECT.1  
 CPKKWVSFRNSCYFFSEEKRTWYNSTTACTAQNSSLLKIDNKEELNFLNQLPIYFWIGLYRHENNGFWQ  
 WTDGSLVSSKLSNTKITRQDIAGWSAVQKFSFRDSEGSTFGLLSVFL  
 >Ultra76.21\_CLECT.1  
 ACQTQNSSLLKIENQEKVKWLLYRKSCYFFSEEKGTWSQSEAACQTQNSSLLKIENQEKVEGISQIGCV  
 FSGKSDVSCNPEDFAWN  
 >Ultra76.22\_CLECT.1  
 CSGNWWQFSTSCYQYSAEKKPWKESQQACKSQNSTLLHIDGLRGLNFFKFFTLSEWIGLSRTGPGSSLK  
 WEDEATYTRNFLSFTNERKGGDCALYLNEKSISMEDCAIGKPHICN  
 >Ultra76.25\_CLECT.1  
 CPGKWFYKMSCYHYSTERKSWKESKQACASQKSNLLHIDSTEELDFLKLTLHGWIGLIRNESRASWQ  
 WENGTEFSRKLKLLSVDLGYGGSCVLYGSGDSAYVGDACATDRHYICE  
 >Ultra76.27\_CLECT.1  
 CSVGWYQYGQKCYSMYRKLSSWEECRNYCISLNSLSTLKTKEEELFVIQLTRRQCLENDPKFWIGLKY  
 VNRKWIWPKSVISSDIAQYSAYSQHLINTVNE  
 >Ultra76.27\_CLECT.2

CPEKWL LYRKSCYFFSEERKTWSQSEAACQTQNSSLLK IENQEKVEGISQIGCVFSGKSALS CNPEDFA  
 WNLFGPQEVLQGERLNNMGVTA  
 >Ultra78\_CLECT.1  
 CPENWFLYGNNCYLFSKEERKWDESQAICRSQNSTLLK IENMGELDFLSL FALYRWTGLYRSGRNQTKW  
 WSDGSNLSIDLHITLDES GKYCAKYHPSNPVSP EECESLQKYICY  
 >Contig12036\_CLECT.1  
 CPEDWMQKGDKCYHISVKQVPWSNCSAQCVLLGSTFLQSGNGELLD FLMPLATHCHWIGLFYKDSSKKW  
 TWIDGSPASSDLNMKESRPTPQSSCAYVTSTGLV SQSCIIPYSCVCE  
 >Contig25135\_CLECT.1  
 CSGKWFHYKMSYHYSTERKSWKESKQACAAQKSNLLHIDSREELN FLKLLTLQGWIGVSRNESCASWQW  
 EKGTEFSSKLI PKFQITTREQFKGEYCSNAEKDFLLD  
 >Contig25622\_CLECT.1  
 CPLGWHMSGKKQSWEAGASDCDDRKFQLLVLEDEAEADRISWIRL TDEYFWIGYNYIKTHGRWMLHDS  
 EFSGFRIVIDNYTGKDCV SFKGRPNII PEDCRTSRGWI  
 >Contig30556\_CLECT.1  
 CPGKWHYKMIYHYSSERKSWKESRQACAAQKSNFLHIDSGEELYFLKLL TLQGWIGLIRNESRASWQW  
 ENGTEFSRSHSGPSEKGLEPLDRVMTETQGNVSSGSGKK  
 >Contig3450\_CLECT.1  
 CPLGGWLNSSRCYFFSGQKHSWEAGVTD CANRKS WLLVLENESEEDWLTRIRPADEYYWIGYKYNTIQ  
 GQAWALGNTEFSGYRWAPLGQVNC SRLPLHKSEKQESNIGLQPE TNQEGQPAQGS SKKPVVPG  
 >Contig3534\_CLECT.1  
 CEKWFVDIEKIYCFHLSHYSWNE SRKFCEDKAFSLLQINN KDDLDFMKNVYH THHWGLSYARISRSWT  
 WEDGSPLSPDX  
 >Contig9884\_CLECT.1  
 CPDSWHQHGEKCYRLMLNMKA WTDCKNHCGSQSSNF SVLENKELDFVNQVIMNKCGKKRICWIGLSFDT  
 NRRKWTWIDGSEL SHK  
 >Ultra271.8\_CLECT.1  
 CPENWILCRNNCYLFSKEERMWNE SQAICRSHNSGLLK IENKGE LDFLSLLEINGWTGLSRSGRNQPWK  
 WSDGSDL SKGLHI INHQPT EIKGAADGEMGWRETEMSE QPVVYS  
 >Ultra271.8\_CLECT.2  
 CPEKWF LHRKSCYFFSEERKTWNQ SRAACGTQNSSLLK IEDQEMLVGF FFSNVPLTGLVTPVAKEDSTD  
 GVTQVKTDNSQDFLSRFYLLGWIGLLWPGVTRAWI WTDNSNLPRNL FYVYKQKGEALEGRRQTL PRAWN  
 PAAFLT LGITSY  
 >Ultra271.8\_CLECT.3  
 CPENWFLYGNNLYLFSKEERKWDESQAICRSQNSTLLK IENKRELD FLSL FALYGTGLYRSGRNQTKW  
 WSDGSNLSIGLLHTTLDES GKDKAYRPSNPASPEECESLQKYICL  
 >Ultra271.9\_CLECT.1  
 CPDSWHQHGEKCYRLMLNMKA WTNCKNHCGSQSSNF SVLENKELDFVNQVIMNKCGKKRICWIGLSFDT  
 NRRKWTWIDGSEL SHK LASPPASRTSPPGCLPT  
 >Chromosome7  
 PCPDNGNGFQ GKCYFFSEDKRNWTSSQS FCTSHAAALIVIDTSWVMDLLMSDKGPCDYWTGLSREAGQD  
 WKWTNGTKFTGWFEIKGRGECAYL NDERVSRARNY TNRRWICSKSNFYTKRRREAFPD

(5)

**Cathelicidins**

&gt;oaCATH1

METT WKLLLLLAAVATVATAQSL SYEEAVVA AVTEYNRES DADALFWLLELKPQERPLESRKAI PVNFTI  
 QETVCPKVENRSPDQCFDFKEGGVVRNCSGTVLPTQEAPVVTLTCDRPI RTKRRIKLIKNGVKKVKDILK  
 NNNIIILPGSNEK

&gt;oaCATH2

MASGAWRVLLLVSVAALTPARGLSHREAVNLALD TYNRGSDVDRAFRVFGSVPRPRPAPSPTTIPLNF  
 TVKETVCFKREKFDLDQCFDRDRGLVRDCTGFVSTAQRPPVTI SCEGAARPWAGNGSVHRYTVLSPRL  
 KTQ

&gt;oaCATH3

MEKRQLG LLLLLGLLLAVEALAGPVPPEGLDFQPPLDADL GTPQVLNFTIQETVCPKAESPTPERCDFK  
 ADG

LVKDCLGSISTGGPPAVVVTCEPKDGLSTVSGT

&gt;oaCATH4

MGGTGRLLLLLGLAAVATAGPPEAEKKYREAVSLALDAYNRGLTGSSAFRLLEPSQLKPDLP RGHKFK  
 IKETVCPVAEKQORECDFKEDGLERDCSGIFLAPSDTRIMVLS CRPAGGGNGQVQIARSASQPYLLSA  
 YCAPSTEGRRLLRSSGSEEPDWSKVPSKYREMYENAKYDI IANILRNF

>oaCATH5  
MAPGIWRVLLLLVSVAAALTPALGLSHGEAVNLALDSYNRGPQHQTRLPGGGAEGGPDEGDAPSPTAIPL  
NFTVKETVCFKKEKLDLEQCAFRDNGLVRDCTGFVSTAQRPPPTVTISCEGPIRTKRFLFSLISAGLGIG  
GAIANLVRRSRG

>oaCATH6  
MASGVWVLLLLVSVAAALTPARGLSHREALNLALDITYNKGSVDVRAFVWFVGPVVKPGPAPSPTTIPLKF  
TVKETVCFKREKFDPEQSAVPRRQSGPPEETNASWAYGPNLKPFSRASPPKTSSEAPQKASPGARPQL  
P

>oaCATH7  
LLLAAVATLATAQSLSYEEALARTIASYNKASDSNDLFRLELKLHEGTVSPGPGEGRNNANAFVNFN  
TIQETVCTKEEMRPLDQCDFKEGGLVRECSGTISMDKDDPSVNVTCDGVSVPGTTRAQSGTTSPPGSGSRR  
VARSRG

>oaCATH8  
EEDGESRMAEGSWRVLLLLVSLAAALATGRGLSQRRRAVKIALDSYKNGSDTDNTFRLVHYIKKSGPVDPS  
KPIPVFTTIQETVCRKSDNPVDPDRCDFKKGGLVKICTGTISSATGSPSASITCDGVSVRRRH

(6)

## CD163 family

>mm\_oanatinus2\_13032.19145  
VRLVGSSEHPCAGRLEVRGLEWGTVCCHSDLNLAHAHVVCQELRCGAAVSMPPGGPLWTQETFRRCRGNESLLFHCPRGPVTELCSSH  
RAAAVRCSEYRLVNGSSRCEGRVELQVGESWGPLCAVHWGLADAHVLCRQLDCRAALTTVGGASFGEGDPAWTDFAHCVGMPEHL  
WNCPRASALGASACLPHNVATAICSGNETRLTQCRISIPPPSPVGMTWDGGVICSEFTDMRLTGGSWTCTGRLEVFYNGTWGSLCHN  
SLSDVSLTVICKQLQCGARGTLLDTPEMDAESKFTWVDRIECREKHHTSLWQCPAPWHPHSCPRMEEAWISCGQGTCKCSVFNNC  
TAQEKVVRVRRGGKDCSGRVEVWVHGGSWGTVCDDAWDLVDAHVVCRQLGCGAALVAPGHAAGFPGMGTIWLDEVGCLGSEASLWDCP  
AAPWGRDTCGHKEDAANVCSGQ

>mm\_oanatinus2\_14563.21452  
HLLMPPQPRLVASAGHREVRLVGGDHRACAGRLEVRGLEWGTVCQSDLDLATAHVCRELHCGAAVSMRLDRTGFGQGPGLWTQETFR  
CQGNESLLFHCPRGPTRMEPCSHSRDATVRCSEYRLVNGSSRCKGRVELQVGGSWGSLCATHWDLADAHVLCRQLGCGPALATTG  
RASFGVGDGPAWTDFAHFCAGTEPHLWNCPSALGAAACPPNHVATAICSGECQPRAGSHHGSVPTSEGLRLAGDLGRACGRVEVL  
HDGRWGTVCDDSWDLADAVVCRQLSCGVALGLASAGHVGVTGPIWLDELGCAGNESRLWQCPMSMGWRHDCCHHKEDAGVLCSSG

>mm\_oanatinus2\_1569.2309  
VKLFSTTSPCVGIVLVKHQDQWGSVCSHSWNMEEASVVCQKLNCGTAFGAPKVFVFNKQYPPPELSTPLPGLHGVSCRGNESLQNCNL  
GEWSWKDCSDDFNSIVCSKGYSAVMLAKGNSPCAGFPQGLVPSKDLMTCDFRGKEATVLCRELGCSSALQPSDLDLHLKEDMRN  
ISRAVTCQGTSTVNLNCFIQEKLKLDLNCNPDLDQVICSGHREVRLVGGHEPCTGRLEVRGLEWGTVCQSDLDLATAHVCRELRC  
GAALSVPEGARFRSRTGTLWTQETFRRCQGNESLLFHCPRGPARTPEPCSHRQEAARIRCSEYRLVNGSSRCEGRVELQVGESWGPLCA  
AHWDLADAHVLCRQLGCGPALATAGGASFGVGDGPAWTDFAHFCAGTEPHLWNCPSALGAAACPPNHVATAICSGEAEALRLQGG  
QSRCDGRVEISLRGWSRSPAWGLAEAGVVCHQLRCGEARRAYSPRSTARDQSHPPGLGLSAVSWCRGGWVNLHNGSWGTVCDSSW  
DLADAHVCRQLSCGVALWAPTSSGPGPTGPIWLDELGCAGNESRLWQCPAGWRHRDCRHKEDVGVLCSEFTDVLQSGSSWACT  
GRLEVFYNGTWGVSVCQYSLPAVALTVICKQLQCGSRGEILNRSETDAESRLSWLNDIRCREQHHTSLWQCPAPWHPHSCSLSTEEA  
WIHCRG

>mm\_oanatinus2\_16550.24447  
YTSAVGFAWEVWGDGGSIVVSIILNTLLELPIISNYCSFPPPPAPEYRLVNGSSRCEGRVELQVGESWGLCASHWDLADAHVLCRQL  
GCGVALETVGRASFGAGDGPWTDFAHFCAGTEPHLWNCPRSGLGAACPPDHVAYAVCSVLPSPQPISLSSLTTPRQQLSPGESA  
EALRLQGGTTRCDGRVEISLRGVWSRIVAEAWGLAEAGVVCRQLRCGEARRAYSLDPTAHGPEATAMGLSDVTCAGNETHLTQCRV  
SISPLSPVGMADWVSVICSGEFLPDPCAPPNLLILPRPSPSSLVQVQCASWESERPASGREAQGAVPGRVEVLHDGRLGTFCDDSS  
WLEDAQVVCRCQCGEALBAPASTGQGAGSGPIWLDELGCAGNESRLWQCPAGWQHDCCHHKEDVGVLCSDVVDGRMSLEKEGSI  
SLQMRKRLRLREKFTDVRTLGGSWACTGHLEVFYNGTWGVSCHNSLSDVSLTVICKQLQCGSRGRLDTPETESRIQDPESRFTWVD  
HIECREKHHTSLWQCPAPWHPHSCPGTEEAWISCGQD

>mm\_oanatinus2\_17198.25406  
VKLIQSRSHCSGIVLVNHQDQWGLVCSHTWNMEEAAVICKQKLNCGTAFVGPVPLKFNKQYPPPELSTFPGLHGISCRGNESLRLHCNQ  
GEWSWKNCSDDFNSVICSKSPYSTAMLVKGNSPCAGFPQGVIANSLDLITCGFQSQEATVLCGELGCGSALQPSVDTQLREDLGN  
TSRAVTCQGTSTVNLNCFIQEKLKLDLNCNPDLDQVICSGHREVRLVGGHEPCTGRLEVRGLEWGTVCQSDLDLATAHVCRELHC  
GSAVSIIPGAASFLGPGQLWTQDTRFCRGNESLLFHCPRGPARTPEPCSHRQDTAVRCSEYRLVNGSSRCEGRVELQVGESWVPLCA  
AHWDLVDAHVLCHQLSCGVALATAGGASFGVGDGPAWTDFAHFCAGTEPHLWNCPRASALGATACQPNHVATAICSGELAEALRLQGG  
QSRCDGRVEISLQAGWSRIVAEAWGLTEARVVCHQLRCGEARRAYSPLYMAHSPPEATVPLNAINSCVGNERTLTCRVSITPPFPV  
GTAWDVVICSESEGLRLVGGPGRACGRVEVLHDGWRWTVCDSSWDLADAHVVCRQLSCGVALGLDASPVHEPGTGPWLDELGCA  
GNESHLWKCPAGWGRNDCRHKEDVGVLCSEFTDLRLAGGSWACTGRLEVFYNGTWGVSVCQNSLMDVTLTVICKQLQCGPWGQVVG  
TPETDAESRLTDLHIECQEKHHTSLWQCPAPWHPHSCPRTEEAWISCGDEKGEKRRPGEILKCSVFNNCTAQEKVVRVRRGGEDGC  
SGRVEVWVHGESWGTVCDDAWDLADAHVVCRQLGCGAALAAPGRAVFGSGTGTIWLDEVRCRGEASLWDCPAPWGRDTCDHKEDA  
AVNCSG

>mm\_oanatinus2\_3158.4622  
KKFNFSLPWDGLGLACGFMKKGCAVEEVKLLHSSDSCSGIVLVNYQDQWGFVNCNHFVNLEEATVVCQKLNCGTGIAPNLIIPMTQ  
SLQELSTFGWLHGVSCSGNESSIQNCNLEQWSWKNSHDFNAIILCSGKVIIVSNFKDCRCDPTMLRAEILSTNQIVLYCTNRREE  
ATVLCRELGCSSALQPSDLDLATAHVCRELRCGAAVSTLGAARFQGMGLLWTQETFRRCRGNESLLFHCPRGPARTKPCSHSQDAAIRCS  
RRGLEWGTVCCHSDLDLATAHVCRELRCGAAVSTLGAARFQGMGLLWTQETFRRCRGNESLLFHCPRGPARTKPCSHSQDAAIRCS  
GESPRAI

>mm\_oanatinus2\_4806.7036  
TGEAAEGLKLEFRKSPCDGVVQVQHEGEGWGSVCSDDWTLAEASVVCQMDCGAAVGPVYVPLPGENVQVPLHGVSCRGNESSEFWN  
CSLGEWSRKNCSFPWVVVALCSKGTFFVEIRLVKPKSPCAGIPEIRNKDGRDLRCLGLYMEEATVFCQELRCGPAPIQAPIEGLGVNPK  
YMTCKGTEPTIRNCRNLNNFRGTGCHREAEVIECAGHREVRLVGGHEPCTGRLEVRGLEWGTVCCHSDLDLATAHVCRELRCGAAVS  
TQAVAGFGRGTGPGTQETFRRCRGNESLLFHCPRGPARTPEPCSHSRDAARIRCSEYRLTNGSTRCEGRVELQVGGSWGSLCAAHWDL



ADAHVLCRQLDCGPALMITSGASFGMGDPTLTDAFHCMGTEPHLWNCPRALGAAACPPDHMAYAVCSGESAEALRLQGGQSRCD  
 GRVEISLRGWSRIVAEAWGLAEAGVVCRQLRCGEARRAYSLLDRTARGPEVTPVGLTTVTCAGNETRLTQCHVSSLPPLPPVMAWD  
 VGVVCSSEGLRLAGGPGRCAGRVEVLHDGHGWTVCDDSWDLADAQVVCRLSCGVALGAPASTGHRAGKGP IWLDDLGC TGNESL  
 LWQCP SAGWRHDCRHKEDVGVLCSEFTEVRLAGGSWACTGRLEVFYNGTWGSVCHNSLTDVSLTVICKQLQCGARGQVLPKND  
 EYNTLWVHDHIECREQHHSI SLWQCPSTPWNEESCNTTEKAWISQEEERDEKNEKLVETLKCSEFNNTAQEKVVRVQGGEDGCSGRV  
 EVVHGGSGWGTVCDDAWDLVDAYVVCRLGCGEALAAPGQAAGFGPTGTVWLDEVQCLGSETSLWDCPAAPWGHTDCGHKEDAAVNC  
 SGKAEPLL  
 >mm\_oanatinus2\_4809.7038  
 RMSALLTLGLPLLATWAGLVTGPCTWEGGYSSSESPCSGIVLVNHQDQWGLVCSHTWSTEEATVVYKQLNCGTAFGAPKISVMT  
 QYLQDFSEVPWLHGVSCRGNESLRNCLGEWSWKDCSHDFWNCICTESLYSKIMLAKRNSPCAGFLGQLVTSKDLRITCGFQSK  
 EATVICRKLGCSSALQPSMDSHLREDLANTSRAVTCQGTTEPTVLNCLHQQLLGHCLNDSEFEVICSDDELGF  
 >mm\_oanatinus2\_8323.12286  
 GRYGGWINVIDCSLPAPEYRLVNGSSRCEGRVELQVGVESWGLCASHWDLADAHVLCRQLGCGVALETVGRASFGAGDGPATWDAF  
 HCTGTTEPHLWNCPRSGLGAACPPDHVAAYVCSESEGLRLAGGPGRCAGRVEVLHDGRWGTVCDDSWDLEDAQVVCRLSCGVAL  
 EAPASTGQAGSGPIWLDELGCVGNESHVWQCP SAGWGQHDCHHKEDVGVLCGSKFCGP  
 >mm\_oanatinus2\_8596.12731  
 GCABEEVKLIHSRSPCIGIVLVKHQDRWGLVCSHSWNVEASVVKQLNCGTAFGAPKAFKPNQYPPPELSTFPGLHGVSCRGNES  
 LRNCLGEWSWKDCSDDWFNSIVCSSACHREVRLVGGEPCTGRLEVRGGLKWTVCQSDLDLATAHLVCRELHCGAAVSMPPGGAC  
 FDQPGQLWTQETFRCRGNESLLIHCPRHLKYASGFSYQSSIVLASCADLI VPRAEYRLVNGSSRCEGRVELQVGVESWGLCAVH  
 WDLADAHVLCRQLGCGPALGTVGRESFGAGDGPATWDAFHCAGTEPHLWNCPRALGAAACQPNHVATAVCSGESAEALRLQGGQ  
 RCDGRVEISLRGWSRIMAEAWGLAEAGVVCHQLRCGEARRAYSLLDPTARGPEATPLGLSAVSCAGNETRLTQCRFISIPSPVGT  
 AWDVGVVCSSEGLRLADGPGRCAGRVEVLHDGRWGTVCDDSWDLEDAQVVCRLSCGVALGAPASSGPGPTGP IWLDELGCAGN  
 ESRWLQCPSTGWRHDCRHKEDVGVLCSEFTEVRLAGGSWACTGRLEVFYNGTWGSVCHNSLTDVSLTVICKQLQCGARGEMLARPE  
 TDAESSPSWVD  
 DIRCRERHHTSLWQCP SAPPWHPHSCLSTEAEAWIRCRAQEKVRLRDGEDDCSGRVEVWHDGSGWGTVCDDAWDLADAHVVCRLGCGV  
 ALAALGLAAFPGKGTIWLDEVRCQGGEASLRDCPAAPWGRTDCGHKEDAAVNC

(7)

Vitellogenin II (Table 2). GenBank entry XP\_001513824.1.

(8)

Organic solute transporter alpha genes (Table 2).

&gt;Contig159089

VFGVTAUVVGIYIPRSSFVFCNFVAALYHSITLWKFLLELVRDFAGGEAQLLRQLHGSHVAPNPPCCCCC  
 CPNITITRDQLRWMSSLAVALQLSLVVRTALFLISLVLWTDEQYDYGDVGSNPNSYFNALIGASTFLSFY  
 YLLFYTASRRALPGFRLRAKFCVIALVLVVCGLQSSVLETMGALGVIPCSPPFSARTRSQIIYYYSITV  
 EMFSIGLLARSCFRRLPAL

&gt;Ultra462

MEDARNATRESACQARQAPFSGTLLQSLDLKGRYLFATLTLMTLVAYLVFLEEYFYIYRKIPPSKRSIF  
 IWINASSPVIALTSLGMWIPRATMFTDFAAGAFFAIVIHKFLMMIKECGGQKMFLLRRFPGRFKIST  
 GPCCCCFCPLPHITSGSLFWLKVGTFLALLQPVLLVLLAIVLWNTGTYSLS

(9)

Short wavelength sensitive 2. GenBank entry ABN43075.1.

(10)

Indigoidine synthase A. GenBank entry XP\_001518909.1.

(11)

Epididymal specific lipocalin 8

GTSARMNSALLISILGLIWANHAQSNVPFKKDFDLNKFSGFWYVIGMATDSEQLMAPKGEKKMAGSIVSVENGO  
 IRLKFSYHQTRCQVENLVGKEIISGKYDFPGLGLGKRVVYISD TDYQDYAVLYITAYLDGENHDGLMFYKSSP  
 LENAHEHINAF