Minireview

# Marsupials and monotremes sort genome treasures from junk

Matthew J Wakefield* and Jennifer AM Graves†

Addresses: *Division of Immunology and Genetics, John Curtin School of Medical Research and †ARC Centre for Kangaroo Genomics, Research School of Biological Science, The Australian National University, Canberra 0200, Australia.

Correspondence: Jennifer AM Graves. E-mail: Jenny.Graves@anu.edu.au

Published: 28 April 2005

*Genome* ***Biology*** 2005, **6:**218 (doi:10.1186/gb-2005-6-5-218)

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2005/6/5/218



## Abstract

A recent landmark paper demonstrates the unique contribution of marsupials and monotremes to comparative genome analysis, filling an evolutionary gap between the eutherian mammals (including humans) and more distant vertebrate species.

Sequencing of a variety of mammalian and other vertebrate genomes is now proceeding apace, and one major goal of this work is to interpret the massive amounts of data from the Human Genome Project by aligning sequence and distinguishing conserved elements from the background of variable sequence ('phylogenetic footprinting'). Sequence data from mammals that are more or less closely related to humans (chimpanzee, mouse, dog) and more distantly related vertebrates (birds, fish) span 450 million years of evolution. But there is still an awkward gap, precisely in the region of the tree from which genomic data are most needed: species that are not so close that sequence comparison gives false-positive signals and not so far that the sequences are unalignable. Marsupials and monotremes, the earliest groups of mammals to diverge, fill this gap (Figure 1). All mammals produce milk and suckle their young, but marsupials and monotremes are distinguished from eutherian ('placental') mammals by differences in reproduction. Marsupials such as kangaroos and wallabies give birth to highly underdeveloped young and much of their development occurs while suckling in the pouch (including of the hindlimbs, eyes, gonads and a significant portion of the brain). Monotremes such as platypus lay eggs that are incubated in a burrow, where the young hatch and suckle from milk patches until they mature.

In a recent landmark paper in *Proceedings of the National Academy of Sciences USA*, Margulies *et al.* [1] present the sequencing and comparative analysis of a 1.9 megabase (Mb) region from three marsupials (the North American opossum *Didelphis virginiana*, the Brazilian opossum *Monodelphis domestica* and the tammar wallaby *Macropus eugenii*) and a monotreme (the platypus *Ornythorhynchus anatinus*). Although previous studies [2] have clearly demonstrated the utility of marsupial sequences in comparative analysis, Margulies *et al.* [1] have analyzed a significantly larger region, looked at multiple marsupial species for the first time and made the first large-scale comparison with a monotreme. This has enabled the identification of sequences that are conserved between multiple species, and which may therefore have functional significance.

The results [1] clearly confirm the prediction [3] that non-eutherian mammals make a unique contribution to the power of comparative analysis. Because marsupials and monotremes diverged from eutherian mammals 180 and 210 million years ago, respectively, non-functional sequences are expected to have diverged beyond recognition, making conserved sequences easier to spot. Approximately 34% of the marsupial sequence and 14% of the platypus sequence was alignable with the human genome, compared with 45%-75% for eutherian mammals. This smaller proportion of alignable sequence improves the selectivity of the analysis, resulting in rapid identification of the most conserved (and by inference the most important) functional non-coding regions. Non-eutherian sequence can therefore make a strong contribution to comprehensive functional annotation of non-coding DNA, such as is being undertaken by the ENCODE project [4].

The 1.26 Mb of contiguous gap-free sequence obtained from the platypus [1] is the largest sample of high-quality

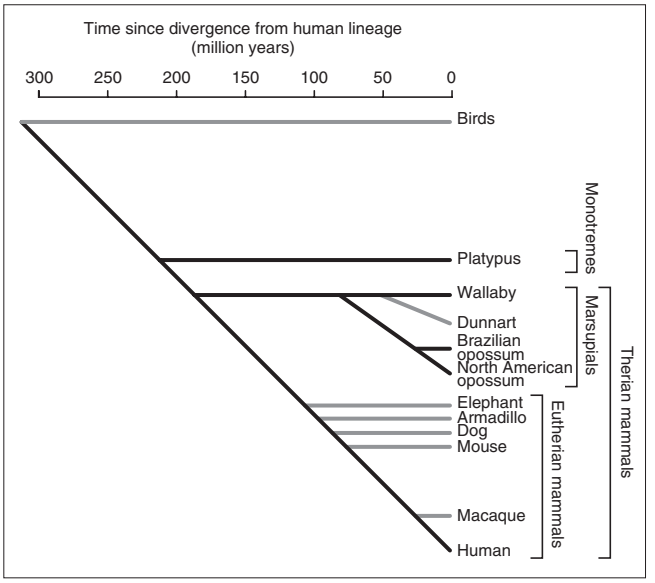


**Figure 1**
The phylogenetic relationships of species discussed in this article [1,14]. Species used in the comparison of Margulies *et al.* [1] are indicated with black lines.

sequence available so far for a monotreme, and it provides some early pointers to what we can expect from the platypus genome project. As found in previous molecular studies [5,6], a large number of core short interspersed nucleotide elements (SINEs) were present in the platypus sequence. The ubiquity of these small repeated elements will make the assembly of whole genome sequence quite a challenge. Another important observation made by Margulies *et al.* [1] is that the GC content of the platypus genome is significantly higher in than that of other mammals. Models of molecular evolution often assume that the sequences being compared have the same frequency of short motifs; we must allow in such models for the difference in composition, which is apparent at 'neutral' four-fold degenerate sites in codons. The higher GC content of the platypus genome may reflect differences in mutational processes, such as decreased rates of CpG methylation, or different selective pressures owing to fundamental differences in physiology between monotremes and other mammals.

As we define the appropriate species to compare and the techniques for identifying sequences conserved between multiple species, we can build a comprehensive list of conserved regions that are likely to be functional. The next challenge will be to work out how these regions function and why they evolved to the form we see today. Valuable insights may be provided by categorizing conserved regions by their different patterns of molecular evolution, and thereby inferring the type of functional constraint. For example, about half of the multispecies conserved sequences that are not in known exons were found to be conserved in all of the species examined [1]. Of those not present in all mammals, some are specific to individual clades (such as eutherians); these could be significant in evolution, as changes in gene regulation, rather than in the protein products of genes, are likely to be the major contributor to the phenotypic diversity of life. In order to start to unravel the role of individual clade-specific conserved noncoding sequences we will need taxonomically rich datasets within all clades, to compensate for the reduced discriminatory power caused by the short evolutionary distances involved and to allow reliable identification of conserved sequences. Such analyses will therefore become more tractable as more genome sequences become available. The interpretation of changes in regulatory elements will also require other genomic data, such as expression information from a representative range of species.

The power of comparative genomics goes beyond the discovery of regulatory regions in non-transcribed DNA highlighted by Margulies *et al.* [1]. We have previously proposed [3] that the increased evolutionary distance from humans to marsupials and monotremes will be particularly valuable in studies of untranslated regions, as the constraints of transcription increase the level of conservation (and therefore of noise) in comparisons between more closely related species. A recent study using four eutherian mammals by Xie *et al.* [7] demonstrates the utility of comparative studies to define regulatory elements in 3′ untranslated regions; this approach can easily be extended by the addition of marsupial and monotreme sequence. Coding regions are also a rich area for comparative genomics to which marsupials and monotreme sequence can contribute. For instance, comparison of cDNA sequence of the large and complex gene α-thalassemia and mental retardation on the X (*ATRX*) between human and tammar wallaby has identified conserved protein-binding sites [8].

Another potential application of marsupial and monotreme sequences is shown by a study sharing the same journal issue with Margulies *et al.* [1]. Sawyer *et al.* [9] present an analysis of adaptive evolution of the primate *TRIM5α* gene, which encodes a protein that limits retrovirus infection by an unknown mechanism. By using evolutionary analysis they show that *TRIM5α* is engaged in an antagonistic conflict between the immune system and retroviruses (including human immunodeficiency virus, HIV) that is at least as old as the primate lineage. Adaptive evolution can be inferred by comparing the ratio of mutations that change an amino-acid sequence ($K_a$) to those that are in degenerate codon positions and therefore silent ($K_s$; for details see the commentary by Yang in the same issue [10]). Genes that are undergoing adaptive evolution are interesting from the point of view of understanding the process of evolution, and they may also be important in human disease [11,12]. Any gene that is evolving new functions or a different mode of regulation is likely to be more prone to error, and there is a correlation between genes evolving adaptively and disease genes that

are catalogued in the Online Mendelian Inheritance in Man (OMIM) database [12]. The addition of marsupial and monotreme sequences will provide evolutionary depth to whole-genome screens for adaptively evolving genes using $K_a/K_s$ ratios - in the style of a recent study comparing human, chimpanzee and mouse sequences [12] - improving their power to detect genes under positive selection. Also, marsupials and monotremes are uniquely positioned to illuminate the genes and domains that were under selection in the early mammalian radiation and that were critical in mammalian evolution.

Mammalian comparative genomics has itself evolved from a data-poor science, in which most effort went into the collection of data, to a science of the genomic age in which large amounts of high-quality data are widely available. The South American opossum genome sequence is currently in assembly, with draft sequence now available. Platypus genome sequencing is underway, with approximately three-fold whole-genome shotgun coverage completed. Two-fold shotgun sequencing of the genomes of tammar wallaby and eight eutherian mammals, including important representatives of the more distantly related eutherian lineages (shrew, tenrec, armadillo and elephant), is starting this year [13]. The escalating comparative-genomics firepower arising from these datasets, along with easy-to-use tools integrated into genome browsers, is available to all researchers to analyse their region of interest. The scale and power of mammalian comparative genome analysis is set to take a big leap forward.

## References


1. Margulies EH, Maduro VV, Thomas PJ, Tomkins JP, Amemiya CT, Luo M, Green ED; NISC Comparative Sequencing Program: **Comparative sequencing provides insights about the structure and conservation of marsupial and monotreme genomes.** *Proc Natl Acad Sci USA* 2005, **102:**3354-3359.
2. Chapman MA, Charchar FJ, Kinston S, Bird CP, Grafham D, Rogers J, Grutzner F, Graves JA, Green AR, Gottgens B: **Comparative and functional analyses of *LYL1* loci establish marsupial sequences as a model for phylogenetic footprinting.** *Genomics* 2003, **81:**249-259.
3. Wakefield MJ, Graves JA**: The kangaroo genome. Leaps and bounds in comparative genomics**. *EMBO Rep* 2003, **4:**143-147.
4. ENCODE Project Consortium: **The ENCODE (ENCyclopedia Of DNA Elements) Project.** *Science* 2004, **306:**636-640.
5. Kirby PJ: **Investigations of the Monotreme Genome.** *PhD thesis.* La Trobe University, Department of Genetics, Bundoora, Australia; 2002.
6. Gilbert N, Labuda D: **Evolutionary inventions and continuity of CORE-SINEs in mammals.** *J Mol Biol* 2000, **298:**365-377.
7. Xie X, Lu J, Kulbokas EJ, Golub TR, Mootha V, Lindblad-Toh K, Lander ES, Kellis M: **Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals.** *Nature* 2005, **434:**338-345.
8. Park DJ, Pask AJ, Huynh K, Renfree MB, Harley VR, Graves JA: **Comparative analysis of ATRX, a chromatin remodeling protein.** *Gene* 2004, **339:**39-48.
9. Sawyer SL, Wu LI, Emerman M, Malik HS: **Positive selection of primate TRIM5alpha identifies a critical species-specific retroviral restriction domain.** *Proc Natl Acad Sci USA* 2005, **102:**2832-2837.
10. Yang Z: **The power of phylogenetic comparison in revealing protein function.** *Proc Natl Acad Sci USA* 2005, **102:**3179-3180.
11. Huttley GA, Easteal S, Southey MC, Tesoriero A, Giles GG, McCredie MR, Hopper JL, Venter DJ: **Adaptive evolution of the tumour suppressor BRCA1 in humans and chimpanzees. Australian Breast Cancer Family Study.** *Nat Genet* 2000, **25:**410-413.
12. Clark AG, Glanowski S, Nielsen R, Thomas PD, Kejariwal A, Todd MA, Tanenbaum DM, Civello D, Lu F, Murphy B, *et al.*: **Inferring nonneutral evolution from human-chimp-mouse orthologous gene trios.** *Science* 2003, **302:**1960-1963.
13. **New Genomic Sequencing Targets** [http://www.genome.gov/11007951]
14. Murphy WJ, Eizirik E, O'Brien SJ, Madsen O, Scally M, Douady CJ, Teeling E, Ryder OA, Stanhope MJ, de Jong WW, Springer MS: **Resolution of the early placental mammal radiation using Bayesian phylogenetics.** *Science* 2001, **294:**2348-2351.