Evolution of coding and non-coding genes in HOX clusters of a marsupial

Hongshi Yu1,2, James Lindsay3, Zhi-Ping Feng4,5, Stephen Frankenberg1,2, Yanqiu Hu1,2, Dawn Carone3, Geoff Shaw1,2, Andrew J Pask1,2,3, Rachel O’Neill3, Anthony T Papenfuss4,6 and Marilyn B Renfree1,2*

Abstract

Background: The HOX gene clusters are thought to be highly conserved amongst mammals and other vertebrates, but the long non-coding RNAs have only been studied in detail in human and mouse. The sequencing of the kangaroo genome provides an opportunity to use comparative analyses to compare the HOX clusters of a mammal with a distinct body plan to those of other mammals.

Results: Here we report a comparative analysis of HOX gene clusters between an Australian marsupial of the kangaroo family and the eutherians. There was a strikingly high level of conservation of HOX gene sequence and structure and non-protein coding genes including the microRNAs miR-196a, miR-196b, miR-10a and miR-10b and the long non-coding RNAs HOTAIR, HOTAIRM1 and HOXA11AS that play critical roles in regulating gene expression and controlling development. By microRNA deep sequencing and comparative genomic analyses, two conserved microRNAs (miR-10a and miR-10b) were identified and one new candidate microRNA with typical hairpin precursor structure that is expressed in both fibroblasts and testes was found. The prediction of microRNA target analysis showed that several known microRNA targets, such as miR-10, miR-414 and miR-464, were found in the tammar HOX clusters. In addition, several novel and putative miRNAs were identified that originated from elsewhere in the tammar genome and that target the tammar HOXB and HOXD clusters.

Conclusions: This study confirms that the emergence of known long non-coding RNAs in the HOX clusters clearly predate the marsupial-eutherian divergence 160 Ma ago. It also identified a new potentially functional microRNA as well as conserved miRNAs. These non-coding RNAs may participate in the regulation of HOX genes to influence the body plan of this marsupial.

Keywords: Marsupial, HOX cluster, MicroRNAs, Long non-coding RNAs

Background

The origin, evolution, function and regulation of HOX genes are amongst the most intriguing questions in developmental biology and evolutionary genetics. Their highly conserved clustered arrangement on chromosomes, their spatio-temporal expression and their patterning results in each distinctive body plan during embryogenesis and organogenesis in bilaterian animals [1,2]. HOX genes are expressed as early as the pre-somite stage of gastrulation in the posterior primitive streak of the epiblast, a region that gives rise mainly to the lateral plate and extraembryonic mesoderm in chicken and mouse embryos [3-5]. The dynamic expression of HOX genes in the ectoderm, mesoderm and endoderm during gastrulation suggests that HOX genes are key regulators of regional patterning along the antero-posterior (A-P) axis [2-4.6]. HOX genes confer positional information for proper organ development and are expressed in ordered patterns that control the segmentation of the hindbrain and axial skeleton along the A-P axis, while mis-expression or mutation leads to the conversion of one structure into another, (homeotic transformation) [2]. Limb development and regeneration depends on patterning formation along three axes: A-P, dorsal ventral (D-V), and proximal distal (P-D) axes [7], where HOX A and HOX D, especially groups 9–13, are responsible for positional information along the A-P and P-
D axes [8,9]. De-regulation of the HOX network results in cancers including breast, bladder, prostate and kidney, as well as abnormal expression during proliferation, differentiation and apoptosis and signal transduction [1,10].

In all vertebrates, HOX genes are comprised of two exons, in which exon 2 includes the highly conserved 180 bp of homeobox region, and a variable length of intron, from less than 200 bp to several kilobase pairs. The homeodomain encoded by a homeobox consists of 60 highly conserved amino acids and forms an N-terminal extended structure followed by three alpha helices. The homeodomain binds target DNA sequences at its N-terminal arm and the third helix from the minor and major groove of DNA, respectively. Orthologues of every HOX gene, including the homeodomain and flanking regions, are highly conserved among species. However, within species, the most conserved region between paralogues is restricted to the homeodomain. HOX genes are clustered on different chromosomes and are believed to have evolved from a single ancestral HOX gene by tandem duplications and sequence divergence [1,11]. There are four HOX clusters, denoted A, B, C and D, produced by two successive whole genome duplication events followed by subsequent divergence [12,13]. Paralogues within each cluster are designated 13 to 1 based on gene 5′-3′ transcribing orientation although there are only 11 paralogues at most found so far in vertebrates.

The low density of interspersed repeats in the human HOX clusters suggests that cis-regulatory elements are important in the tight control of HOX gene expression [14]. Global enhancer sequences located outside the clusters regulate HOXD temporal co-linearity [15]. Non-coding RNAs known to be involved in regulation of HOX gene expression [16,17], include the highly conserved micro-RNAs [18], such as miR-196 [19] and miR-10 [20]. The long non-coding RNAs HOTAIR [21,22] and HOTAIRM1 [23] are known only in the mouse and human.

The comparison of HOX genes between vertebrates and invertebrates has highlighted conserved features of HOX gene expression regulation and evolution. Comparisons of DNA sequences between evolutionarily distantly-related genomes are highly efficient ways to identify conserved (and novel) functional regions, especially non-coding RNAs, and to discover how they regulate HOX gene expression [24,25]. However, some conserved functional features show lineage-specific distributions and will be missed if the taxa chosen are too distant in evolutionary terms. Similarly, if they are too close, differences can be missed. Marsupials fill the mammalian “gap” because they are a distinct lineage that diverged from eutherian mammals 130–160 Ma ago [26-29], but they are still mammals. There is a high ratio of conservation signal to random noise in comparisons between therian mammal (marsupial and eutherian) genomes, suggesting that there are localized regions under evolutionary constraint [30]. The divergence time between these groups is sufficient for non-functional sequences to have diverged while important genes are sufficiently conserved to enable their clear identification. Comparative genomics between eutherians and marsupials is therefore invaluable for predicting new and novel mammalian-specific motifs participating in HOX gene expression and regulation during mammalian evolution.

In this study, we used the tammar wallaby (Macropus eugenii), a macropodid marsupial of the kangaroo family, as our model. We screened BAC clones and further characterized all 39 tammar HOX genes as well as genome mapping and deep sequencing. Comparative genomic analyses identified the known HOX coding genes and non-coding regulatory regions including regulatory elements and non-coding RNAs. Importantly, we uncovered a new potential microRNA in the tammar HOX cluster.

Results

Sequencing and assembly

To map HOX clusters on the tammar chromosomes (Figure 1), partial sequences of 34 tammar HOX genes were retrieved using the assembled tammar genome (assembly 1.0) [31] and the trace archives in GenBank. The HOX genes and clusters were highly fragmented in the genome assembly, so we used these sequences to screen a BAC library (Me_KBa; Arizona Genomics Institute, Tucson, AZ, USA) and then utilized a shotgun sequencing approach. Five BAC clones covering the HOXA to HOXD clusters were pulled down, sequenced using the Roche 454 platform and de novo assembled (Genbank: JN378718, JN378719, JN378720 and JN378721). Contigs were aligned to the genomic sequences of HOX clusters from opossum (Oct. 2006, MonDom5), platypus (Mar. 2007, WUGSC5.0.1/ornAna1) and human (Feb. 2009, GRCh37/hg19). These alignments verified that there were 37 HOX genes contained in the five BAC clones (see Methods), with HOXA1 and HOXD13 missing. Therefore, cross-species primers were designed to obtain full-length sequences for HOXA1 and HOXD13. In addition, to confirm that HOXA1 and other HOX genes were clustered together, we screened a different tammar BAC DNA library (MEB1 library constructed at RIKEN, Japan). The newly obtained clones containing HOXA1 also included HOX genes identified by PCR. The same strategy was also used to confirm that HOXD13 was clustered with other HOXD genes.

Annotation of HOX clusters

Tammar HOX genes were clustered at four different loci with an arrangement of HOX13 to –1 from 5′ to 3′ (Figures 1, 2), showing a similar and highly conserved
relative order and orientation of HOX genes in each cluster. There are 11 HOXA genes in the HOXA cluster, 10 HOXB genes in the HOXB cluster, 9 HOXC genes in the HOXC cluster and 9 HOXD genes in the HOXD cluster, showing a conserved distribution of homologues across each HOX cluster. Gene sequences have been submitted to Genbank with the BAC sequences. Tammar HOX genes each consist of two exons (detailed in Additional files 1, 2) encoding a highly conserved homeodomain, as is found in other vertebrates, including humans. Each HOX gene showed a similar and highly conserved overall exon length as well as sequence. Although the intron of each HOX orthologue varied significantly in sequence, the length of each intron is also conserved (Additional file 1), suggesting a high level of conservation across HOX clusters during evolution.

The abundance of repetitive DNA elements is extremely low in the core of tammar HOX clusters, in agreement with the previous findings in gnathostome HOX clusters [32]. Utilizing RepeatMasker (http://www.repeatmasker.org/cgi-bin/WEBRepeatMasker), repeat elements including short interspersed repeat elements (SINEs), long interspersed repeat elements (LINEs), long terminal repeats (LTRs) and other DNA elements were investigated in each tammar HOX cluster (Additional file 3). Strikingly, there were no Alu (short interspersed repeat element of about 300 bp, comprising 10.75% of the human genome), ERVL (long terminal repeats), TcMar-Tigger and satellite sequences found in any tammar HOX locus, resembling the human HOX clusters [14].

Tammar HOX gene expression in adult tissues
The expression patterns of all 39 HOX genes were analysed in 23 adult tissues by RT-PCR, including brain, gastrointestinal tract, circulatory system, digestive system and reproductive system of the tammar wallaby (Figure 3).
Some anterior HOX genes (HOX1 to −3) were expressed in the forebrain, midbrain and hindbrain in tammar, similar to the expression patterns of human HOXA genes [33], but very few HOX genes were expressed in hypothalamus, pituitary and pancreas. Interestingly, almost all HOX genes were expressed in cerebellum, suggesting that HOX genes continue to participate in coordinating motor activity and communication as they do during development [34,35]. Anterior (1–3) and central (4–8) HOX genes of cluster A/B/D were expressed in the spleen and carry important roles in replenishing red blood cells and in activating the immune response. In the tammar gastrointestinal tract, weak expression was found in intestine while much stronger expression was observed in stomach and caecum, showing tissue-specific expression patterns. Anterior and central HOX genes of clusters A and B, but not C or D, were expressed in liver and heart. In tammar lung tissue, almost no posterior HOX genes were expressed. Skeletal muscle had broad expression of HOX genes (HOX1-11). HOX gene expression in reproductive tissues was similar to those in the developing tissues, displaying ongoing proliferation, differentiation, and degeneration of multiple cell types. HOX genes were strongly expressed in the mammary gland, kidney, adrenal, testis and ovary, but had a restricted expression in epididymis and uterus. Overall, HOX genes had tissue-specific expression patterns, maintaining high expression in some tissues, while in other tissues they were down-regulated or switched off.

Functional and conserved non-coding sequences in the kangaroo HOX clusters

Comparative genomic analysis between tammar, human, mouse and a non-mammalian vertebrate, frog (Additional files 4, 5, 6, 7) using mVISTA [36], showed that the coding regions of each cluster were highly conserved, whereas the non-coding regions including untranslated regions (UTRs), intergenic regions and introns shared a comparatively low sequence similarity but were conserved in length. Furthermore, there was higher conservation in the 3′ UTR of each HOX gene than in the 5′ UTR, similar to previous findings [24] (Figures 2, 4, 5 and Additional files 4, 5, 6, 7). This provided a platform to identify whether these conserved non-coding sequences function as conserved transcription factor binding sites or non-coding RNAs participating in gene expression regulation/RNA processing, or whether they just act as non-functional and randomly conserved elements, maintaining high sequence identity for about 500 Ma of evolution for vertebrates or up to 160 Ma of evolution for mammals [26,28].

Known long non-coding RNAs are conserved in the kangaroo HOX clusters

Long non-coding RNAs (lncRNAs) play critical roles in transcription regulation, epigenetic gene regulation and diseases. They are rapidly evolving genes, and are expected to be poorly conserved at the sequence level [37-39]. However, we found conserved orthologues of all three known mammalian lncRNAs—HOTAIRM1, HOXA11AS and
HOTAIR (sequences provided in Additional file 8)—by comparative genomic analysis and RT-PCR amplification. HOX antisense intergenic RNA myeloid 1 (HOTAIRM1) was located between HOXA1 and HOXA2, and we demonstrated that it was restricted to mammals (Figures 4, 6 and Additional file 4). The tammar HOTAIRM1 has three exons according to RT-PCR size. Exon 1 was highly conserved across all mammals. Exon 2 could not be detected in tammar and opossum using the “Infernal” (v1.0.2) program (http://infernal.janelia.org/), which employs both RNA secondary structure and sequence to search the genomic sequence, but using RT-PCR, we were able to find exon 2. The conservation of the secondary structure of exon 3 is lower than that of exon1, but is much higher than that of exon 2, which can be clearly observed by the phylogenetic trees in the right bottom panels (Figure 6). HOTAIRM1 was expressed in bone marrow as expected (Figure 4), suggesting that this IncRNA has had conserved roles in myelopoiesis across all mammals for up to 160 Ma.
Tammar HOXA11 antisense (HOXA11AS), located between HOXA13 and HOXA11, has two exons similar to that in human (Figures 4, 7 and Additional file 4). HOXA11AS was highly conserved in eutherian mammals, but had a very low conservation in marsupial species, while in the frog it had less than 50% identity. Additionally, using the “Infernal” program we found that exon 1 had a conserved RNA secondary structure, but it failed to predict exon 2. Although there was a low conservation between tammar and eutherian mammals, tammar

Figure 4 Conserved miRNA and long non-coding RNAs analysis in the HOX A cluster. The conserved long non-coding RNAs, HOXA11AS and HOTAIRM1, and microRNA miR-196b were shown by mVISTA with comparison of mouse, tammar and frog against human HOX A cluster genomic sequence. The coding genes HOXA13, HOXA10, HOXA9, HOXA2 and HOXA1 are highly conserved in all species. Expression of tammar long non-coding RNAs in bone marrow and endometrium were confirmed by RT-PCR. The blue stands for coding regions, and the green for non-coding RNA regions whilst the pink represents conserved coding sequences. F, forward primer, R, reverse primer.

Figure 5 Sequence conservation in the HOX C cluster in tammar, human, mouse and frog. mVISTA plot of HOX C genomic sequences from tammar, human (chr12:54332691–54396455), mouse (chr15:102751619–102814560) and frog (scaffold_226:281324–390491). The sequence similarity (50–100%) (vertical axis) is shown in the coordinates of the genomic sequence (horizontal axis) from human, mouse and frog. Genes and their orientation are indicated by grey arrowed line. Exons of genes are indicated by blue solid boxes. Conserved regions above the level of 70%/100 bp are highlighted under the curve, with red indicating conserved non-coding regions, blue representing conserved coding-protein exons, and turquoise representing microRNAs or long non-coding protein exons. The long non-coding RNA HOTAIR located between HOXC12 and HOXC11 was conserved in all mammals and had a much lower conservation in frog. MicroRNA miR-196a2 is extremely highly conserved in all examined species. RT-PCR performed in the tammar with a single band at 81 bp confirmed the presence of the long non-coding RNA HOTAIR providing further evidence of the conservation. In addition, both microRNA miR-196a2 was expressed in tammar cells, verifying the existence of this microRNAs in tammar.
HOXA11AS was expressed in the endometrium during pregnancy (Figure 4) as it is in humans, suggesting a conserved role in mammalian reproduction.

HOX antisense intergenic RNA (HOTAIR) is a trans-regulatory gene, unlike the other lncRNAs that are cis-regulatory. It plays an important role in epigenetics and tumorigenesis. In the tammar, it was located between HOXC11 and HOXC12 as in human and mouse. Exon 1–4 are very short exons of about 100 bp, and exon 5 is just 53 bp in human, but there was only low conservation seen in the mVISTA plot (Figure 8 and Additional file 6). Exon 6 is the longest exon, showing some regions that are highly conserved (Figures 5, 8 and Additional file 6). In the tammar, exon 1–3 could not be identified with the “Infernal” program using human RNA secondary structure and sequence alignment. Phylogenetic trees showing that exon 1 is highly conserved with short genetic distance between them compared to exon 2 and exon 3 consistent with the concept of rapid evolution of non-coding RNAs (lower right).

The kangaroo HOX clusters encode conserved microRNAs. mVISTA plots showed numerous non-coding regions, possibly representing microRNAs, were highly conserved (Additional files 4, 5, 6, 7). We examined the presence of
known microRNAs, miR-196a1, miR-196a2, miR-196b, miR-10a and miR-10b, previously described in the human, mouse and zebrafish HOX clusters. As expected, we found 5 known conserved miRNAs in tammar HOX clusters (summary in Figure 2 and the sequences provided in Additional file 8, genomic sequence alignment referred to Additional files 4, 5, 6, 7). We examined tammar microRNA deep sequencing libraries from different tissues and cells to determine the expression profile of each of these miRNAs. We found that miR-10a and miR-10b were strongly expressed in the testis. They are also expressed in fibroblast cells of the tammar.

In order to computationally explore new or novel miRNAs and their targets in the HOX cluster of the

Figure 7 Evolutionary relationships of HOXA11AS orthologues. The genes flanking HOXA11AS, HOXA11 and HOXA13, in the human (chr7:27,220,777–27,239,725; http://genome.ucsc.edu) are shown along with their conservation score (phylop). HOXA11AS gene structure consists of two exons in eutherian mammals except mouse, but one exon in tammar (lower left), based on predicted with RNA secondary structure and sequence alignment. Phylogenetic trees showing exon 2 is highly conserved in eutherian mammals whilst exon 1 is more divergent with the full predicted exon 1 sequence, consistent with the mammalian consensus sequences in the top panel (lower right).

Figure 8 Evolutionary relationships of HOTAIR orthologues. The genes flanking HOTAIR, HOXC11 and HOXC12, in the human genome (chr12:54,348,714–54,370,201; http://genome.ucsc.edu) are shown along with their conservation score (phylop). HOTAIR gene structure consists of 6 exons in the eutherian mammals, except mouse and rat, which have 5 exons. In contrast, only 3 exons were found in tammar. Phylogenetic trees based on exons 4–6 (lower right).
Discussion

Comparative genomic analysis of the marsupial HOX clusters uncovered a new microRNA and confirmed the presence of numerous known mammalian RNAs. There was a strikingly high level of conservation of coding sequences between this member of the kangaroo family and that of eutherian mammals.

Marsupial HOX gene clusters are compact and uninterrupted by large repeat domains. In the tammar, the length of all clusters were remarkably similar to that found in human (tammar HOXA-D: 113 kb, 207 kb, 144 kb and 110 kb; human HOXA-D: 112 kb, 205 kb, 137 kb and 112 kb retrieved from the UCSC genome browser GRCh37/hg19). Similar patterns are also found in frog, chicken and mouse (Additional files 4, 5, 6, 7), demonstrating that the HOX gene clusters are highly conserved and compact across vertebrate lineages. However, *Amphioxus*, which is viewed as an “archetypal” genus in the chordate lineage, carries a HOX cluster length of about 448 kb [42]. In invertebrates, HOX clusters are often more than 1 Mb, as is found in the sea urchin [43]. Thus the vertebrate HOX clusters are more compact than the ancient and invertebrate HOX clusters [42].

All 39 tammar HOX genes had conserved gene structures (Additional file 11) and chromosomal arrangement (Figure 2), consistent with the theory that two rounds of genome duplications occurred after the vertebrate–invertebrate divergence but before bony fishes and tetrapods split [12,13,44]. In adults, HOX genes continue to be expressed and thereby retain developmental plasticity in certain tissues or maintain homeostasis. However, there has been much less work on gene expression in adult tissues compared to developing tissues [45,46]. We showed that HOX gene expression in adult marsupial tissues was tissue-specific and differentially expressed (Figure 3). Interestingly, almost all HOX genes were expressed in the cerebellum, suggesting that HOX genes continue to participate in coordinating motor activity and communication in adults, as they do during development [2].

Using the tammar HOX genomic sequences as a reference for phylogenetic footprinting, we were able to identify a large number of conserved non-coding genomic sequences which may act as transcription factor binding sites in promoters, regulatory motifs involved in chromatin remodeling or non-coding RNAs that modulate gene expression post-transcriptionally [25,47]. Long non-coding RNAs play diverse roles in biological processes but are thought to be under different evolutionary constraints and are expected to have low sequence conservation compared to protein-coding sequences [38], which has hampered the study of long non-coding RNA in vertebrates. We not only found these IncRNAs

![Figure 9](http://www.biomedcentral.com/1471-2164/13/251) Newly discovered miRNA *meu-miR-6313* in tammar. A) Centroid secondary structure with a minimum free energy of −43.66 kcal/mol; the bar from blue to red represents base-pair probabilities from low (0) to high (1). B) the reads, precursor and secondary structure of new miRNA; C) sequence alignment of miRNA and precursors in tammar, human, mouse and frog.
orthologues in the tammar HOX genome, but also confirmed that they were expressed in certain tissues. For example, human HOTAIRM1 is expressed specifically in myeloid cells to regulate HOXA1 and HOXA4 expression in NB4 cells (an acute promyelocytic leukaemia cell line) [23]. Tammar HOTAIRM1 was also expressed in bone marrow, suggesting it has a conserved role in myelo- 
ihesis across all mammals. In addition, HOTAIRM1 appears to be restricted to mammals and so must have evolved during the mammalian radiation. A recently discovered long non-coding RNA, HOTAIR [21,22], acts as a trans-regulator to regulate HOXD but not HOXC gene expression during limb development [22] and participates in reprogramming chromatin states to promote cancer metastasis [21]. Tammar HOTAIR was also found in the tammar HOX genomic sequence [31], and was expressed at the early head-fold stage of the tammar embryo at the time just before limb buds develop, sug- gesting that it may have a role in the regulation of limb development—especially important structures for the kangaroos [40]. In addition, the 5’ flanking sequence of HOTAIR was conserved, suggesting that it has the same or similar transcriptional regulation mechanism (Figure 5 and Additional file 6). Thus, contrary to expectation, mammalian IncRNAs do show a reasonable level of se- quence conservation.

Micro-RNAs are highly conserved, in contrast to long non-coding RNAs, and play important roles in animal development by controlling translation or stability of mRNAs [48]. They are normally 22 nucleotide RNA that binds to complementary sequences in the 3’UTR to repress gene activities [49]. Using the tammar as a reference and searching the microRNA database we were able to identify four known HOX microRNAs (miR-196a, miR-196b, miR-10a and miR-10b), and most sig- nificantly, we uncovered one new potential microRNA, meu-miR-6313 in the tammar which was expressed in testis and fibroblasts. The precursor sequence was used to search the human, mouse, and frog genomes and was not present (Figure 9). We also searched the opossum and Tasmanian devil genome sequences using the pre- 

Conclusions
Annotation and comparative genomic analysis of tam- 
mar HOX genes demonstrated a high degree of evolu- 

tionary conservation. As expected, 39 HOX marsupial genes were mapped to four different chromosomal loci. The tammar HOX clusters had a low concentration of repetitive elements and were compact as in other verte- 
brates HOX clusters. The protein-coding regions and their UTRs also showed high conservation but there was a novel potentially functional miRNA meu-miR-6313 within a HOX cluster. Interestingly, the long-coding RNAs (HOTAIR, HOTAIRM1 and HOXAI1AS) and microRNAs (miR-196a2, miR-196b, miR-10a and miR-10b) were highly conserved in this marsupial. These IncRNAs and miRNAs may control the HOX genes to influence phenotypic differences in the body plan, as they do in other mammals. This study confirms that the emergence of known long non-coding RNAs in the HOX clusters clearly predates the marsupial-eutherian divergence up to 160 Ma ago.

Methods
Animals, tissues and cells
Tammar wallabies originating from Kangaroo Island, South Australia, were held in the University of Mel- 
bourne marsupial breeding colony in Melbourne, Vic- 
toria. All sampling techniques and collection of tissues 

Tissues (forebrain, midbrain, hindbrain, cerebellum, hypothalamus, pituitary, pancreas, spleen, stomach, intestine, caecum, heart, liver, lung, muscle, kidney and ad- 
renal) were collected from five adults. Bone marrow, 
mammary glands, uterus and ovary were collected from three adult females. Prostates, epididymides and testes 
were collected from two adult males. HOX gene expres- 
sion was examined using all tissues listed above except 
bone marrow. Bone marrow, whole embryos (day 20 of 
the 26.5 day gestation, n = 2) and endometrium (col- 
lected from three additional pregnant females) were used 
to examine IncRNA expression. All tissues were
collected under RNase-free conditions. All collected tissues for molecular analysis were snap frozen in liquid nitrogen and stored at −80°C until use.

Tammar primary cells were prepared from a day 10 post partum pouch young testis. Briefly, the primary cells were cultivated in 50% DMEM (containing 10% fetal bovine serum) (Invitrogen, Melbourne, Australia) and 50% AminoMax (Gibco, Carlsbad, USA) containing 15% fetal calf serum.

**Probe preparation and BAC library screening**

The six frame translation of the tammar genome (assembly 1.0) was searched for homeobox domains using a profile hidden Markov model (Pfam accession PF00046.21) and the HMMer software (version 2.3.2) [52]. An E-value threshold of $10^{-4}$ was used. Predicted homeobox domain sequences of at least 80aa and related DNA were extracted from the tammar genome. The domain sequences of these sequences were then classified using HOXPred [53]. At the same time, tammar HOX partial sequences were also obtained by searching the tammar trace archives with human exon 1 and exon 2 of 39 HOX genes using BLASTN. Gene specific primers were designed to amplify probes and to confirm identity of isolated BACs. All primers and their annealing temperatures as well as the product size are listed in Additional file 12.

When screening the BAC library, at least two probes from the 5′ end and 3′ end were selected and 5 positive clones were identified: 205I5, 9G11, 168N24, 6P18 and 214D22. BAC clone 205I5 covered HOX cluster genes (HOXA2 to HOXA13); BAC clone 9G11 covered the HOXB cluster (HOXB1 to HOXB9); BAC clone 168N24 covered the HOXB cluster (HOXB4 to HOXB13); BAC clone 6P18 contained all HOXC cluster genes and clone 214D22 covered the HOXD cluster (HOXD1 to HOXD12).

**BAC DNA preparation, sequencing and assembly**

Positive BAC bacteria were cultured overnight in LB medium containing 12 μg/ml chloramphenicol at 37°C. BAC DNA was extracted according to manufacturer’s instructions of Maxipreps DNA purification system (Promega, Sydney, Australia). The quality was assessed by gel electrophoresis in 0.8% agarose gel and NanoDrop ND-1000 Spectrophotometer (Wilmington, USA) with the ratio of A260/A280 at over 1.8. The amount of DNA was also measured by NanoDrop ND-1000 Spectrophotometer. BAC samples were sequenced with GS-FLX method at the Australian Genome Research Facility Ltd (AGRF, Brisbane, Australia).

The Roche 454 reads of the tammar were extracted and de novo assembled with the program CAP3 [54]. There are 202 contigs from BAC 205I5 in HOXA cluster, 85 contigs from 168N24 and 2613 contigs from 9G11 in HOXB cluster, 405 contigs from 6P18 in HOXC cluster and 89 contigs from 214D22 in HOXD cluster. The contigs were then aligned to the genome sequence of human, tammar, opossum and platypus and any gaps between the new contigs from the BAC sequencing filled where sequence was available using the tammar genome sequence. Based on these genomic sequences, gene structures of all HOX genes and full HOX scaffolds were identified.

**microRNA sequencing and in silico analysis**

The recently published marsupial genome paper provided deep sequencing information [31] and additional sequencing of the tammar microRNAs was performed on an Illumina GAII platform. Briefly, 40 μg Trizol extracted total RNA from tammar brain, liver, testis, and pouch young fibroblast cells grown in culture was electrophoresed on a 15% denaturing polyacrylamide gel with γ-[32P]-ATP end labeled 19-mer, 24-mer and 33-mer oligonucleotides. The bands corresponding to the miRNA fraction (19–24nt) were excised and ligated to an adenylated 3′ adapter (IDT, Inc.). The 3′ ligated RNA was electrophoresed on a 15% polyacrylamide gel and the bands corresponding to miRNA were excised. A 5′ ligation reaction and subsequent polyacrylamide gel purification followed by reverse transcription and PCR was performed in preparation for Illumina sequencing. Sequencing was performed on an Illumina GAII according to the manufacturer’s protocol.

miRNAs mapped to HOX genome were performed using Bowtie [55], allowing for at most 1 mismatch. Potential hairpin locations were first identified using the SRNALOOP program (http://arep.med.harvard.edu/ miRNA/pgmlicense.html). They were further refined by manual inspection of the hairpin loop using an interactive instance of RNAfold program (http://rna.tbi.univie.ac.at/cgi-bin/RNAfold.cgi). Target prediction was done using the miRanda tool [56] with default parameters. The novel microRNAs and the complete HOX genes were used as the query and target sequences, respectively.

**Phylogenetic footprinting analyses**

For interspecies DNA sequence comparison, tammar or human genomic sequence acted as a reference in four species (Human, Mouse, Tammar and Frog). Genomic sequences containing HOXA, HOXB, HOXC and HOXD clusters from Human (HOXA, chr7: 27098056–27210689;
To perform comparative analyses of long non-coding RNAs, we retrieved the genomic sequences upstream from the nearest HOX gene and the corresponding downstream HOX gene in multiple eutherian mammals including chimpanzee, rhesus, mouse, rat, dog and elephant. The “Infernal” program (http://infernal.janelia.org/) was employed to blast each genome sequence with default parameters. Briefly, we used the secondary RNA structure of each exon in human lincRNAs to produce *.sto file. The secondary structure was predicted by online program RNAfold WebServer (http://rna.tbi.univie.ac.at/cgi-bin/RNAfold.cgi). Cmsearch of “Infernal” program was then used to build a model from above secondary structure. Cmcalibrate of “Infernal” program was used to determine expectation value scores (E-values) for more sensitive searches and appropriate HMM filter score cutoffs for faster searches. Cmsearch was used to blast genomic sequences downloaded from NCBI or Ensembl. Using cmsearch, the lowest E-value with less than 0.01 has the priority.

### miRNA pipeline, miRNA and hairpin annotation

In order to computationally explore the cause and effects of miRNA in the HOX cluster of the tammar wallaby we followed a processes inspired by [61]. Our miRNA has three main goals; separating valid sequences from noise and degradation product, identifying miRNA targets and genes. The targets and genes of our pipeline can then be compared against known features from miRBase (http://
www.mirbase.org/) to determine which are confirmed and which are novel.

Each sequenced library is pre-processed to remove both 3’ and 5’ prime adapters and is then size selected to remove reads with less than 15 or more than 32 bases. Next the reads were aligned against the HOX cluster allowing for no mismatches, all valid alignments for each read were reported. The same reads are aligned against the genome, except one mismatch is allowed to compensate for the draft nature of the tammar genome.

To separate between valid miRNA and degraded product/sequencing noise it is required that each read must align at least once within an annotated miRNA gene or hairpin region. The construction of this annotation is detailed in a later section. The novel miRNA gene in HOX was identified by during the annotation stage detailed in a later section. The novel miRNA targets required to meet the following conditions: 1) a valid read aligned to the HOX cluster, 2) the location of the aligned read did not overlap with a previously annotated target.

The main requirement of the miRNA pipeline previously presented is that each read must have aligned within an annotated miRNA gene or hairpin at least once in the genome. The miRNA gene annotations generally come from an external gene annotation pipeline such as ENSEMBL (http://asia.ensembl.org/info/docs/genebuild/genomeAnnotation.html). Since the tammar genome is quite new, and highly fragmented this annotation is incomplete. To augment it, the hairpin sequences in miRBase [62] are aligned to the genome using BLAST. The locations where the known hairpins align are considered equivalent to a miRNA gene.

To capture novel miRNA genes and hairpins, a simple pipeline of commonly available tools was created. Many published tools which identify new micro RNA genes use sequence and structure based alignments to find the best candidates [63]. Unfortunately these tools do not scale well and are too slow to use on large genomes and large micro RNA datasets. Therefore we implemented a custom version of the strategy mentioned above. First, all miRNAs were mapped to the genome. Next, each aligned sequence plus 100 bp flanking windows were put into SRNALOOP a hairpin prediction tool [64]. Regions containing valid hairpins which did not overlap with a previously known miRNA gene or miRBase annotation were recorded.

miRNA target annotation

miRNA targets were annotated in a two-step process. First the valid miRNA were mapped against the HOX cluster allowing for no mismatches. Then the mature miRNA from miRBase release 18 were mapped against the HOX cluster, allowing for 1 mismatch. A target was considered confirmed if a valid miRNA from our pool co-located with a miRNA from miRBase. Otherwise the aligned sequence was considered to be novel.

Our definition of a valid miRNA required each sequence to be associated with at least one miRNA gene, or hairpin structure somewhere in the genome. All of the putative novel miRNA targets in HOX were associated with a hairpin [table XYZ]. However, none of these hairpins were found within an annotated gene. This could be due to a poor annotation, the draft status of the genome, or it is simply a false signal. Each of these will be further validated in future research.

Additional files

- Additional file 1: The length of exon and intron of tammar 39 HOX genes (bp).
- Additional file 2: The sequences of 39 tammar HOX genes.
- Additional file 3: Repetitive elements in tammar HOX clusters.
- Additional file 4: Phylogenetic footprinting analyses of HOX A cluster with mVISTA. mVISTA plot generated with HOXA genomic sequences from tammar, human (chr7:27131531–27244164), mouse (chr6:52104079–52216539) and frog (scaffold_56:1381000–1485000) with tammar as a reference. Conserved regions above the level of 70%/100 bp are highlighted under the curve, with red indicating conserved non-coding regions, blue representing conserved coding-protein exons, and turquoise representing microRNAs or long non-coding protein exons. HOX4 and HOX11/11:1 representing the long non-coding RNAs are conserved in all mammals and have much lower similarity in frog microRNA miR-165 is highly conserved in all species. Arrow stands for the transcription orientation.
- Additional file 5: Phylogenetic footprinting analyses of HOXB cluster with mVISTA. mVISTA plot generated with HOXB genomic sequences from tammar, human (chr17: 43960868–44165742), mouse (chr11: 96029412–96229585) and frog (scaffold_334: 483000–52244164), mouse (chr15: chr15:102750000–102892969) and frog (scaffold_226: 269568–557892) with tammar as a reference. The information of long lncRNAs and microRNAs is same as in Figure 3. Other details as in figure Additional file 4.
- Additional file 6: Phylogenetic footprinting analyses of HOXC cluster with mVISTA. mVISTA plot generated with HOXC genomic sequences from tammar, human (chr12: 52605461–52742874), mouse (chr15: chr15:102750000–102892969) and frog (scaffold_226: 269568–557892) with tammar as a reference. The information of long lncRNAs and microRNAs is same as in Figure 3. Other details as in figure Additional file 4.
- Additional file 7: Phylogenetic footprinting analyses of HOXD cluster with mVISTA. mVISTA plot generated with HOXD genomic sequences from tammar, human (chr2: 175656359–176766195), mouse (chr2: 74497085–74613489) and frog (scaffold_163: 534804–660354) with tammar as a reference. Other details as in figure Additional file 4.
- Additional file 8: The sequences of IncRNAs and microRNAs in tammar.
- Additional file 9: Valid target positions of known miRNAs in tammar HOX clusters.
- Additional file 10: Newly discovered target positions of putative miRNAs in tammar HOX clusters.
- Additional file 11: Phylogenetic relationships and high-order grouping of HOX families from tammar and human. A representative an unrooted tree with rooting that should be considered arbitrary. Phylogenetic analysis was based on the homeodomain regions with an extension of extra 20 amino acids on both sides from human and tammar. The phylogenetic tree was constructed using neighbor-joining method with 100 bootstrap replicates showing bootstrap support values.
on the nodes. 13 monophyletic groups were shown to form HOX11 to HOX13. Three big branches according to their functions during the developmental events are shown: anterior, central and posterior.

Additional file 12: Primers for RT-PCR.

Abbreviations
A-P: Antero-posterior; D-V: Dorsal ventral; ERVL: Endogenous retrovirus L; HOXAT1R: HOX antero-posterior intergenic RNA; HOXAT1A: HOX antero-posterior intergenic RNA myeloid 1; HOXA11AS: HOXA11 antisense; LINEs: Long interspersed repeat elements; IncRNAs: Long non-coding RNAs; LTRs: Long terminal repeats; MaLR: Mammalian LTR; MIR: Mammalian-wide interspersed repeats; P-D: Proximal distal; RTE: Retrotransposable element; SINE: Short interspersed repeat elements; UTRs: Untranslated regions.

Competing interests
The authors declare that they have no competing interests.

Authors’ contributions
The author(s) have made the following declarations about their contributions: conceived the study and designed the experiments: MHR, MB. HOX. Collected the tissues: HY, GS, MBR. Performed the experiments: HY, DC, YH, JL. Analysed the data: HY, JL, Z-PF, MBR, RON, AJP, ATP, SF. Wrote the paper: HY, MBR, AJP, ATP. All authors edited and approved of the manuscript.

Authors information
Anthony T Papenfuss and Marilyn B Renfree are joint senior authors

Acknowledgements
We thank members of the tammar research team for assistance with the collection of the samples. We thank Ms. Bonnie Dopheide for assistance with the FISH experiment. We also thank Prof A Fujimura and Dr. Y. Kuroki for kindly providing the BAC library for screening the incomplete HOX1A and HOX13 genes. This study was supported by the Australian Research Council (ARC) Centre of Excellence in Kangaroo Genomics; an ARC Federation Fellowship to MBR, a National Health and Medical Research Council (NHMRC) R.D. Wright Fellowship to AJP and an NHMRC Career Development Fellowship to ATP.

Author details
1 ARC Centre of Excellence in Kangaroo Genomics, The University of Melbourne, Melbourne, Victoria 3010, Australia. 2 Department of Zoology, The University of Melbourne, Melbourne, Victoria 3010, Australia. 3 Department of Molecular and Cell Biology, College of Liberal Arts and Sciences, University of Connecticut, Connecticut, CT 06269, USA. 4 Bioinformatics Division, The Walter and Eliza Hall Institute of Medical Research, Parkville, VIC 3052, Australia. 5 Department of Medical Biology, The University of Melbourne, Melbourne, Victoria 3010, Australia. 6 Department of Mathematics and Statistics, The University of Melbourne, Melbourne, Victoria 3010, Australia.

Received: 30 November 2011 Accepted: 22 May 2012
Published: 18 June 2012

References
1. Papageorgiou S: HOX Gene Expression. Texas: Landes Bioscience; 2007.
2. Pourqu"e O: HOX Genes, Volume 88. San Diego: Elsevier; 2007.
3. Iimura T, Pourqu"e O: Collinear activation of Hoxb genes during gastrulation is linked to mesoderm cell ingestion. Nature 2006, 442(7102):568-571.
4. Forl"ar"a S, Lawson KA, Deschamps J: Acquisition of Hox codes during gastrulation and axial elongation in the mouse embryo. Development 2003, 130(14):3807-3819.
5. Gaunt SJ, Strachan L: Temporal colinearity in expression of anterior Hox genes in developing chick embryos. Dev Dyn 1996, 207(3):270-280.
6. Deschamps J, van Nes J: Developmental regulation of the Hox genes during axial morphogenesis in the mouse. Development 2005, 132(12):2931-2942.
7. Capdevila J, Iripisa Belmonte JC: Patterning mechanisms controlling vertebrate limb development. Annu Rev Cell Dev Biol 2001, 17(8):137-149.
8. Ohgo S, Itoh A, Suzuki M, Satoh A, Yokoyama H, Tamura K: Analysis of hoxa11 and hoxa13 expression during patternless limb regeneration in Xenopus. Dev Biol 2010, 338(2):148-157.
9. Kmita M, Tarchini B, Zalany J, Logan M, Talbin CJ, Duboule D: Early developmental arrest of mammalian limbs lacking Hoxa/Hoxd gene function. Nature 2005, 435(7045):1113-1116.
10. Shah N, Sukumar S: The Hox genes and their roles in oncogenesis. Nat Rev Cancer 2010, 10(5):361-371.
11. Ferrier DE, Holland PW: Ancient origin of the Hox gene cluster. Nat Rev Genet 2001, 2(1):133-38.
12. Panopoulos G, Poustia A: Timing and mechanism of ancient vertebrate genome duplications—the adventure of a hypothesis. Trends Genet 2005, 21(10):559-567.
13. Van de Peer Y, Maere S, Meyer A: The evolutionarily significant of ancient genome duplications. Nat Rev Genet 2009, 10(10):725-732.
14. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, et al: Initial sequencing and analysis of the human genome. Nature 2001, 409(6822):860-921.
15. Duboule D: The rise and fall of Hox gene clusters. Development 2007, 134(14):2549-2560.
16. Lempradl A, Ringrose L: How does noncoding transcription regulate Hox genes?. BioEssays 2008, 30(2):110-121.
17. Petkou S, Sedkov Y, Brock FW, Mazo A: A model for initiation of mosaical HOX gene expression patterns by non-coding RNAs in early embryos. RNA Biol 2007, 4(1):3-6.
18. Yekta S, Talbin CJ, Bartel DP: MicroRNAs in the Hox network: an apparent link to posterior prevalence. Nat Rev Genet 2008, 9(10):789-796.
19. Yekta S, Shih IH, Bartel DP: MicroRNA-directed cleavage of HOX8B mRNA. Science 2004, 304(5670):594-596.
20. Woltering JM, Dunston AJ: Mir-10 represses Hoxb1a and Hoxb3a in zebrafish. PLoS One 2008, 3(1):e1396.
21. Gupta RA, Shah N, Wang KC, Kim J, Horlings HM, Wong DJ, Tsai MC, Hung T, Arigae P, Rinn JL, et al: Long non-coding RNA HOTAIR reprograms chromatin state to promote cancer metastasis. Nature 2010, 464(7291):1071-1076.
22. Rinn JL, Kertesz M, Wang JK, Squezalo SL, Xu X, Brugmann SA, Goodnough LH, Helms JA, Fanharn PJ, Segal E, et al: Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs. Cell 2007, 129(7):1311-1323.
23. Zhang X, Lian Z, Padden C, Gerstein MB, Rozovskey J, Snyder M, Gingeras TR, Kapranov P, Weissman SM, Newburger PE: A myelopoesis-associated regulatory intergenic noncoding RNA transcript within the human HOXA cluster. Blood 2009, 114(19):4326-4334.
24. Santini S, Boore JL, Meyer A: Evolutionary conservation of regulatory elements in vertebrate Hox gene clusters. Genome Res 2003, 13(6A):1111-1122.
25. Kuntz SG, Schwarz EM, Demodeana JA, De Buyscher T, Trout D, Shizuya H, Sternberg PW, Wold BJ: Multigene DNA sequence conservation identifies Hox cis-regulatory elements. Genome Res 2008, 18(12):1905-1908.
26. Brindza-Emonds OR, Cardillo M, Jones KE, MacPhee RD, Beck RM, Grenyer R, Price SA, Vos RA, Gittleman JL, Purvis A: The delayed rise of present-day mammals. Nature 2007, 446(7135):507-512.
27. Luo ZX: Transformation and diversification in early mammal evolution. Nature 2007, 450(7172):1011-1019.
28. Luo ZX, Yuan CX, Meng QJ, Ji Q: A Jurassic eutherian mammal and divergence of marsupials and placental. Nature 2011, 476(7361):442-445.
29. Luo ZX, Yuan CX, Meng QJ, Ji Q: A Jurassic eutherian mammal and divergence of marsupials and placental. Nature 2011, 476(7361):442-445.
30. Wakefield MJ, Graves JA: The kangaroo genome. Leaps and bounds in comparative genomics. EMBO Rep 2003, 4(2):143-147.
31. Renfree MB, Papenfuss AT, Deakin JE, Lindsay J, Heider T, Belov K, Rens P, Doherty T, Price SA, Vos RA, Gittleman JL, Purvis A: The delayed rise of present-day mammals. Nature 2007, 446(7135):507-512.
32. Luo ZX: Transformation and diversification in early mammal evolution. Nature 2007, 450(7172):1011-1019.
33. Luo ZX, Yuan CX, Meng QJ, Ji Q: A Jurassic eutherian mammal and divergence of marsupials and placental. Nature 2011, 476(7361):442-445.
