Dolphin image credits: Public source Dolphin NASA   **Two toed sloth image credits: Paul Bratescu**

Mammalian keratin associated proteins (KRTAPs) subgenomes: disentangling hair diversity and adaptation to terrestrial and aquatic environments

Khan *et al.*

BMC
Genomics

RESEARCH ARTICLE

Open Access

Mammalian keratin associated proteins (KRTAPs) subgenomes: disentangling hair diversity and adaptation to terrestrial and aquatic environments

Imran Khan[1,2], Emanuel Maldonado[1], Vítor Vasconcelos[1,2], Stephen J O'Brien[3,4], Warren E Johnson[5] and Agostinho Antunes[1,2*]

Abstract

Background: Adaptation of mammals to terrestrial life was facilitated by the unique vertebrate trait of body hair, which occurs in a range of morphological patterns. Keratin associated proteins (KRTAPs), the major structural hair shaft proteins, are largely responsible for hair variation.

Results: We exhaustively characterized the KRTAP gene family in 22 mammalian genomes, confirming the existence of 30 KRTAP subfamilies evolving at different rates with varying degrees of diversification and homogenization. Within the two major classes of KRTAPs, the high cysteine (HS) subfamily experienced strong concerted evolution, high rates of gene conversion/recombination and high GC content. In contrast, high glycine-tyrosine (HGT) KRTAPs showed evidence of positive selection and low rates of gene conversion/recombination. Species with more hair and of higher complexity tended to have more KRATP genes (gene expansion). The sloth, with long and coarse hair, had the most KRTAP genes (175 with 141 being intact). By contrast, the "hairless" dolphin had 35 KRTAPs and the highest pseudogenization rate (74% relative to the 19% mammalian average). Unique hair-related phenotypes, such as scales (armadillo) and spines (hedgehog), were correlated with changes in KRTAPs. Gene expression variation probably also influences hair diversification patterns, for example human have an identical KRTAP repertoire as apes, but much less hair.

Conclusions: We hypothesize that differences in KRTAP gene repertoire and gene expression, together with distinct rates of gene conversion/recombination, pseudogenization and positive selection, are likely responsible for micro and macro-phenotypic hair diversification among mammals in response to adaptations to ecological pressures.

Keywords: Concerted evolution, Gene family, Keratin Associated Proteins, Keratin, Hair, Gene conversion, Recombination, Positive selection

Background

Terrestrial life in extant vertebrates was accompanied by the formation of diverse and rigid body coverings (scales, feathers and hairs), along with other cornified appendages (e.g. horns, hoofs, claws, nails), that evolved in response to strong selective pressures. These body coverings helped protect vertebrates and allowed them to successfully adapt to environmental pressures like heat, ultra violet radiation, water loss, and mechanical forces [1,2]. Since keratinization helps protect the body by forming a barrier between the body and outside world, genes involved in keratinization evolve rapidly in response to changing environment (e.g., KRTAP4-5 showed evidence of positive selection in chimpanzee and hominids) [3]. Changes in gene family (gene gain, and gene loss/pseudogenization) are involved in adaptive evolution and changes in gene family size could affect expression levels [4]. In terrestrial vertebrates, the formation of hard cornified skin appendages involves interactions between fibrous (keratin)

* Correspondence: aantunes@ciimar.up.pt
[1]CIMAR/CIIMAR, Centro Interdisciplinar de Investigação Marinha e Ambiental, Universidade do Porto, Rua dos Bragas 177, 4050-123 Porto, Portugal
[2]Departamento de Biologia, Faculdade de Ciências, Universidade do Porto, Rua do Campo Alegre, 4169-007 Porto, Portugal
Full list of author information is available at the end of the article

and matrix proteins (KRTAPs) [5-7]. The fibrous alpha-keratins, type I and II, appear to have evolved in stem vertebrates [8,9] and recent studies suggest the presence of hair-specific alpha keratins orthologs in amphibians, reptiles, and birds [10-12]. However, there is no evidence of KRTAPs proteins in fishes and amphibians, suggesting that these proteins originated only after the divergence of sauropsids (sKRTAPs/beta-keratins) and mammals (mKRTAPs), leading to the formation and diversification of hard keratin appendages, feather, claw, scale in sauropsids, and hairs in mammals [5,6,11]. The structural and functional conservation of keratin intermediate filaments (KIFs) within mammals contrasts with the large diversity of mammalian hair phenotypes [13-15] and highlights the importance of understanding the molecular diversification of the keratin associated protein (KRATP) multigene family.

Hair is a dynamic mini-organ formed by ectodermal-mesodermal interactions [16-19] and is broadly divided into the root sheath (outer and inner), hair shaft, and matrix zone. Hair has microscopic differences (e.g., cuticular, medullar and cross section), which have long been used as forensic markers for identifying human ethnicity and classifying mammalian species [20-23]. Hair-fiber formation is a cyclical process, which involves growth (anagen), regression (catagen), and resting phases (telogen), followed by the shedding of the hair shaft. The process involves the expression of both hair keratin intermediate filament proteins and their keratin associated proteins [24-28]. This cycle is of particular importance in diverse processes such as determining hair size, shedding fur for body surface cleansing, and changing the body cover to adapt to changing environments, such as from hot summers to cold winters [29].

The current diversity of hair in extant mammals is due to innovations and changes in numerous genes and their corresponding proteins. Humans have 54 functional alpha-keratin genes comprising 28 type I and 26 type II keratins [30,31,13] arranged in two clusters on chromosomes 17q21.2 and 12q13.13 [32,33], which include 11 type I and 6 type II hair keratins [34,35]. Hair keratin types I and II undergo higher-ordered copolymerization-forming keratin intermediate filaments (KIFs) [36-39], which are embedded into a matrix formed by keratin associated proteins (KRTAPs) involved in the formation of hard cornified resilient hair shafts [40-42]. The KRTAP multigene family is divided into two broad groups, high cysteine and high glycine-tyrosine, which together comprise 30 subfamilies based on amino acid composition and phylogenetic relationships [14]. In humans, KRTAPs include approximately 100 gene members that are arranged in tandem and are clustered on chromosomes 11p15.5, 11q13.4, 17q21.2, 21q22.1, and 21q22.3 [43,26,44-46,7,25]. Given the role of the KRTAP multigene family in formation

of hair morphology, we have characterized them in the genomes of 22 diverse mammalian species to provide insights on KRTAP evolution and diversification. We found contrasting KRTAP gene family repertoires among mammals, as well as differences in rates of gene expansion, contraction and pseudogenization. The two major groups of KRTAPs showed distinct evolutionary patterns with high concerted evolution influencing species-specific copy number variation and gene homogenization in high cysteine KRTAPs. In contrast, high glycine-tyrosine genes had more dynamic evolutionary patterns with less gene conversion and recombination, lower GC content, and evidence of positive selection (e.g. subfamily 20), which may also have been an important force of the evolution in subfamilies of high glycine-tyrosine.

Results
Genome scans
Advances in genome sequencing have made it easier to explore multigene families across different genomes. Expansion, contraction and pseudogenization, along with genomic/chromosomal organization (gene clusters) of gene families, are important mechanisms driving genome evolution and influencing fitness within lineages or species [47], as suggested by lineage- or species-specific variations in genes involved in pathogen recognition, stress response and structural proteins [48-51]. Here, we explored the KRTAP multigene family in the genome assemblies of 22 mammalian species: (1) alpaca (*Vicugna pacos*) low-coverage 2.51×, assembly, vicPac1, Jul 2008, (2) armadillo (*Dasypus novemcinctus*) low-coverage 2×, assembly, dasNov2, Jul 2008, (3) bushbaby (*Otolemur garnettii*) low-coverage 1.5×, assembly, otoGar1, May 2006, (4) cow (*Bos taurus*) coverage 7×, assembly Btau_4.0, Oct 2007, (5) dolphin (*Tursiops truncatus*) low-coverage 2.59×, assembly, turTru1, Jul 2008, (6) elephant (*Loxodonta africana*) coverage 7×, assembly, Loxafr3.0, Jul 2009, (7) gibbon (*Nomascus leucogenys*) whole genome coverage 5.6×, assembly, Nleu1.0, Jan 2010, (8) gorilla (*Gorilla gorilla*) gorGor3, Dec 2009, (9) guinea Pig (*Cavia porcellus*) high-coverage 6.79×, assembly, cavPor3, Mar 2008, (10) hedgehog (*Erinaceus europaeus*) low-coverage 1.86×, assembly, eriEur1, Jun 2006, (11) horse (*Equus caballus*) coverage 6.79×, assembly, Equ Cab 2, Sep 2007, (12) marmoset (*Callithrix jacchus*), NCBI build 1.1, (13) megabat (*Pteropus vampyrus*) low-coverage 2.63× assembly, pteVam1, Jul 2008, (14) mouse Lemur (*Microcebus murinus*) low-coverage 1.93×, assembly, micMur1, Jun 2007, (15) orangutan (*Pongo abelii*) NCBI build 1.2, (16) panda (*Ailuropoda melanoleuca*) high-coverage, assembly, ailMel1, Jul 2009, (17) pig (*Sus scrofa*) from NCBI build 3.1, high-coverage, assembly, Sscrofa10, Jun 27, 2011, (18) rabbit (*Oryctolagus cuniculus*) high-coverage, assembly, oryCun2, Nov 2009, (19) sloth (*Choloepus hoffmanni*)

low-coverage 2.05×, assembly, choHof1, Sep 2008, (20) tarsier (*Tarsius syrichta*) low-coverage, 1.82× assembly, tarSyr1, Jul 2008, (21) tree shrew (*Tupaia belangeri*) low-coverage 2×, assembly, tupBel1, Jun 2006, and (22) wallaby (*Macropus eugenii*) coverage 2×, assembly, Meug_1.0, Dec 2008, available at Ensembl and NCBI websites http://www.ensembl.org/ [52,53] and http://www.ncbi.nlm.nih.gov/ [54].

Characterization of KRTAP gene family

The KRTAP multigene family consists of ~100-180 gene members divided into two major classes, High Cystine (HS) and High Glycine/Tyrosine (HGT), which in turn are divided into many subfamilies with unique motifs and sequence repeats. We assigned all KRTAP multigene family members to their respective subfamilies following previously published guidelines [14]. We built species-specific phylogenetic trees to classify the gene subfamilies for each genome (Additional file 1: Figures S1-21), as well as a phylogenetic tree incorporating all members belonging to the high glycine-tyrosine KRTAP multigene family from 22 genomes (Figure 1 and Additional file 1: Figure S24). We also observed that one-to-one orthologous relationships diminished as species diverged over time (Additional file 1: Figures S22 and S23). We used the amino acid composition, unique motifs and sequence repeats, as well as blast results, to classify intact genes, partial genes and pseudogenes. The most-closely-related subfamilies are generally located in close proximity and in tandem arrangements in the genome, as are the members of the same subfamily. Due to the incomplete nature of the genomes analyzed, not all the genes may have been retrieved. Therefore, in the dolphin and other species of low genome coverage, we have assumed that missing genes may be due to low coverage that will be characterized by future research. However, we found a large number of pseudogenes in the dolphin genome compared with other low coverage genomes.

To prove the absence of KRTAP genes in high coverage genomes with intact gene clusters, we performed synteny analysis and searched for human orthologs that should be flanking the missing KRTAP genes. For example, in pig the 5' and 3' human orthologs flanking the KRTAP cluster 5 was missing, indicating that this region has most-likely not been sequenced and that further research and/or higher genomic coverage is needed for confirmation. We verified the synteny of conserved orthologs flanking the missing genes for the subfamily KRTAP25 in the callitrix, cow and elephant, along with KRTAP25, KRTAP19 and KRTAP29 in cavia, and KRTAP12 in rabbit.

Species-specific subfamily differences, changes in the total number of genes, functional genes, pseudogenes, amino acid content (changes in sulfur content are responsible for disulphide bonds, which provide rigidity, strength and flexibility to hair) and size polymorphism in genes within subfamilies may be responsible for the species-specific hair characteristics and the marked variability found in hair patterns among mammalian species.
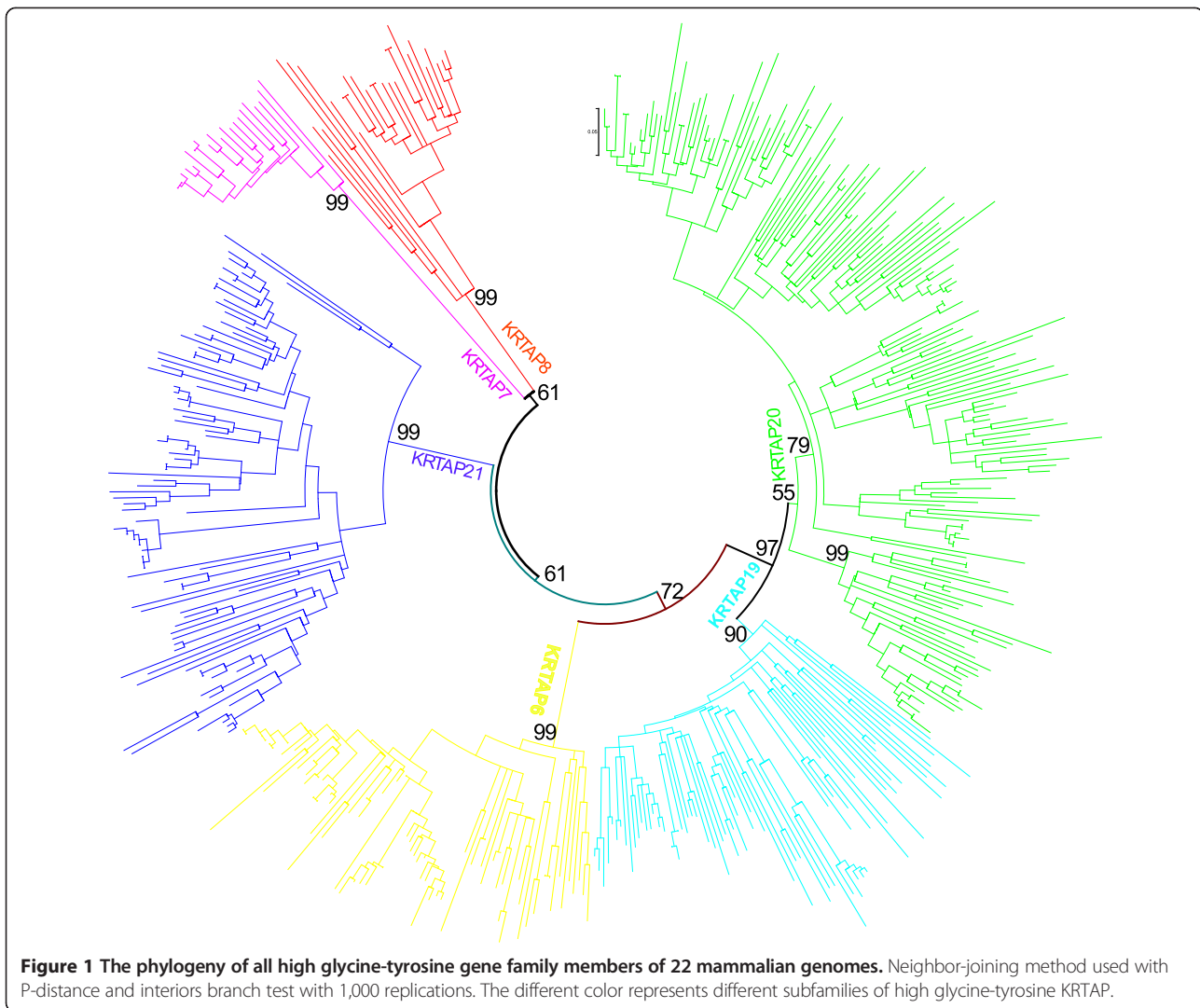
Genomic organization of the KRTAP gene family

The KRTAP gene family consists of 30 subfamilies, 24 of which are high cysteine and six are high glycine-tyrosine. The complete KRTAP gene family is arranged into five clusters at five different genomic locations (Figure 2). Each cluster contains members of one or more subfamilies arranged in a tandem array. The genomic organization of the KRTAP gene family is similar in all species studied, with only slight variations. Subfamilies KRTAP 1, 2, 3, 4, 9, 16, 17 and 29 are present in cluster one. All high glycine-tyrosine (HGT) KRTAP subfamilies, together with KRTAP 11, 13, 24-27 subfamilies, form cluster two. Subfamilies KRTAP10 and KRTAP12 form cluster three, whereas cluster four consists of subfamily KRTAP28 and cluster five of subfamily KRTAP5.

Cluster 5 shows some variation. For example in primates, KRTAP cluster 5 is divided into two paralogous gene clusters, most likely through segmental duplication, with both clusters having members of the KRTAP5 subfamily (Figure 2). In all of the other mammals studied, genes of the KRTAP5 subfamily form a single cluster. The KRTAP subfamilies that are clustered together in the genome (Figure 2) are phylogenetically closely related (Additional file 1: Figure S1) (e.g. all subfamilies of high glycine-tyrosine KRTAPs are located in close proximity in cluster 2 represented by HGT in Figure 2, which supports their functional relatedness and common ancestry arising from duplications and divergence. The conserved genomic organization of the KRTAP gene clusters over more than 166 Myr (i.e. divergence of therian from the monothermes mammals) [55] confirms the strong evolutionary constrain acting on their genomic arrangement [56]. The conserved clustering of KRTAPs seems to be related with its ordered expression in follicle [57].

KRTAP Gene family dynamics and hair characteristics

Previously, the KRTAP gene repertoire had been assessed in eight mammalian species [14], all terrestrial species with few characteristic differences in hair phenotypes. Here we expanded on previous results by analyzing 22 additional mammalian species consisting of a much more diverse group of mammals including species from different mammalian orders with diverse hair characteristics, such as the armadillo (modified scales), hedgehog (spines), alpaca (fiber), sloth (symbionts) and dolphin (mostly hairless and aquatic), along with several more-closely related species, e.g. members of hominidae family in primates.

We identified near complete KRTAP gene repertoires in 22 mammalian genomes, including 11 high-coverage

Figure 1 The phylogeny of all high glycine-tyrosine gene family members of 22 mammalian genomes. Neighbor-joining method used with P-distance and interiors branch test with 1,000 replications. The different color represents different subfamilies of high glycine-tyrosine KRTAP.

genomes (Table 1, Figure 3, and Additional file 2). Our findings suggest that the most recent common ancestor of mammals is supposed to have had 53% (16 of 30) of the known KRTAP subfamilies (1-5, 8, 10, 11, 13, 16, 17, 20, 21, 26, 28, and 29) (Figure 3). Extant monotremes (Platypus) and marsupials (Opossum and Wallaby), have slightly different subfamilies representation (60%, 18 of 30 subfamilies, and 50%, 15 of 30, respectively), while eutherians have up to 93% (28 of 30) of the KRTAP subfamilies. This shows that the diversification of the KRTAP gene family occurred early in mammalian evolution, likely starting after the split of sauropsids (leading to birds and reptiles) and synapsids (leading to mammals-like reptiles) around 350 Myr ago [58,55]. Sauropsids developed KRTAPs (the beta keratins) in hard appendages like feathers, beaks, scales and claw, etc., and synapsids developed mKRTAPs present in hair, nails, hoofs, claws, etc. The presence of glycine rich proteins such as HGT in mammals and HGP (high glycine proline) in reptiles and

birds is evidence of their radiation from a common ancestor [5] and suggest that these changes may have contributed to the successful radiation of mammals, reptiles and birds. Further expansion and diversification of the KRTAP gene family, favored by high rates of concerted evolution in HS-KRTAPs and positive selection in HGT-KRTAPs, led to the species-specific hair characteristics observed in extant mammals. Additional analyses of sauropsid and the mammalian KRTAPs are likely to reveal insights into the patterns of adaptive radiation present in extant reptilian, birds and mammalian KRTAPs. Subfamilies 7 and 12 first appear in therian mammals after their divergence from monothremes around 166 Myr ago [55]. Subfamilies 6, 9, 19, 24, and 27 are specific to placental mammals (eutherians), and thus appeared after their divergence from marsupials around 148 Myr ago [55]. Subfamily 25 is absent in Afrotheria and Xenarthra, which suggests an origin within placental mammals only after the divergence from the atlantogenata clade (Figure 3 and Table 1). Monotremes

Figure 2 Genomic organization of KRTAP gene family in the gorilla genome. The KRTAP gene family is arranged in five different clusters, shown with the size in base pairs (bp) for each cluster with name of cluster and chromosome in which they are present. Each triangle represents a gene member; where p means a pseudogene, same subfamily members are shown with same colors. The triangle points the direction of transcription. The distance between the genes is not to scale.

and marsupials lack subfamily 9, which we observed to have expanded dramatically in the basal placental mammal xenartha (sloth) to 50 members. We noted that the KRTAP gene family shows species-specific variation as expected due to concerted evolution, and some of the subfamilies are restricted to particular species, e.g. subfamilies 30,

31, and 34 are present only in mouse and rat, subfamily 35 in mouse, and subfamilies 32 and 33 in platypus [14] (Figure 3 and Table 1). We also observed remarkable differences among these KRTAP genes (Table 1, Figure 3 and Additional file 2), including a dramatic gene expansion with 175 members (50 in subfamily 9 and 37 genes

Table 1 Number of KRTAP gene present in each subfamily in twenty-two mammalian species
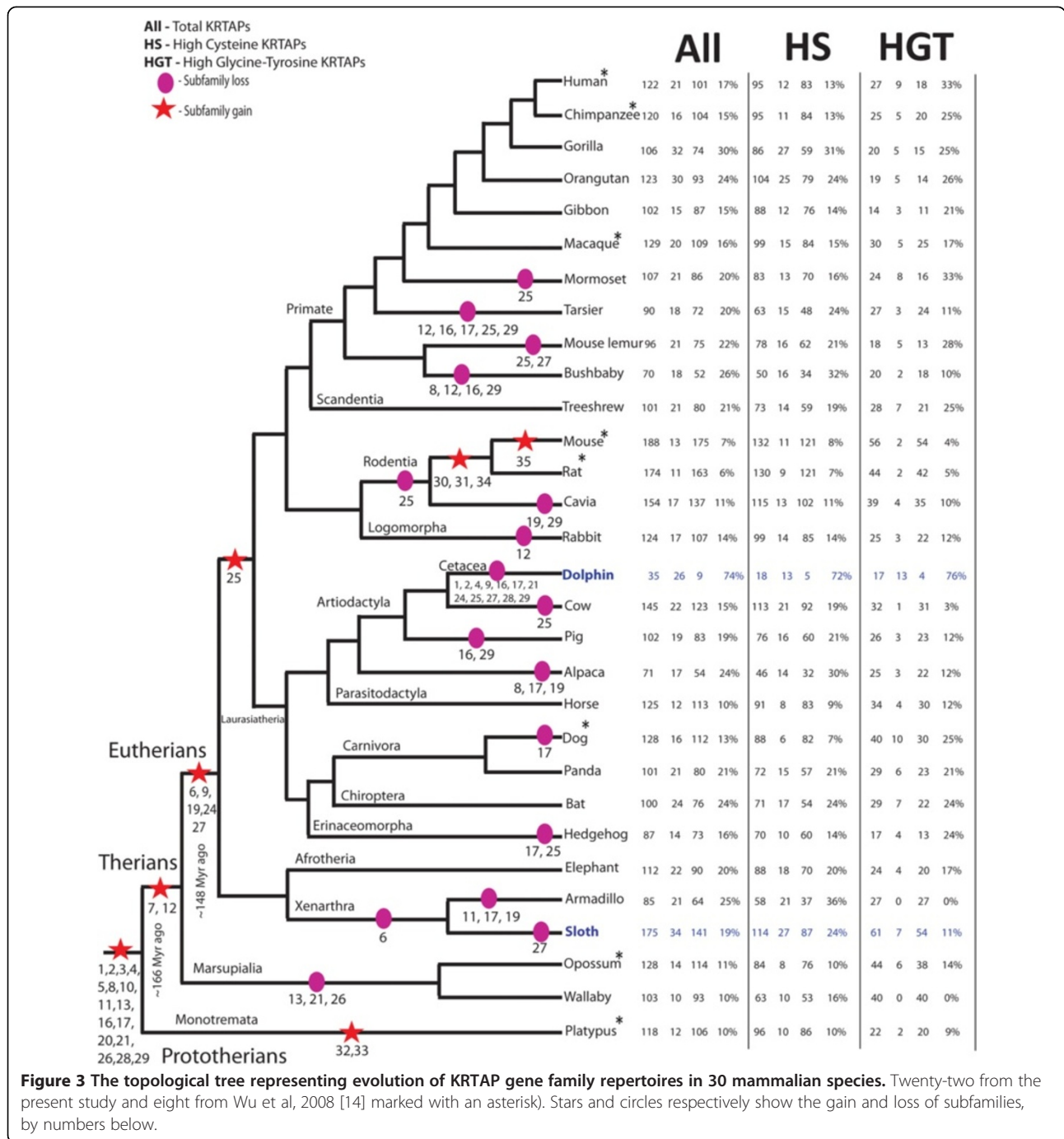
	KRTAP	Gorilla	Pongo	Gibbon	Calllitrix	Tarsius	Mouse lemur	Otolemur	Treeshrew	Cavia	Rabbit	Dolphin
HS-KRTAP	KRTAP1	4(2)	4(0)	3(0)	4(1)	4(0)	4(0)	4(0)	2(1)	4(0)	4(0)	0(0)
	KRTAP2	4(0)	3(0)	4(0)	4(0)	3(0)	4(0)	2(0)	5(1)	4(0)	4(0)	0(0)
	KRTAP3	3(0)	6(0)	4(1)	4(0)	3(0)	2(0)	4(1)	2(1)	4(0)	4(0)	2(2)
	KRTAP4	15(6)	15(4)	10(2)	8(1)	8(1)	9(2)	14(3)	12(1)	16(1)	14(1)	0(0)
	KRTAP5	13(8)	14(6)	15(3)	19(5)	16(8)	8(4)	13(2)	11(4)	18(5)	18(1)	1(1)
	KRTAP9	9(2)	12(2)	9(1)	5(0)	6(1)	4(1)	8(0)	5(0)	12(2)	6(1)	0(0)
	KRTAP10	13(5)	17(2)	13(0)	13(2)	8(2)	5(1)	7(1)	10(2)	12(1)	10(2)	2(2)
	KRTAP11	1(0)	1(0)	1(0)	1(0)	1(0)	1(0)	1(0)	1(0)	1(0)	1(0)	1(1)
	KRTAP12	3(0)	4(2)	2(0)	3(2)	0(0)	0(0)	3(1)	1(0)	12(1)	0(0)	1(0)
	KRTAP13	4(2)	8(4)	7(3)	6(1)	4(2)	4(2)	8(0)	6(1)	8(2)	8(3)	10(6)
	KRTAP16	1(0)	2(1)	1(0)	1(0)	0(0)	0(0)	1(0)	1(0)	1(0)	1(0)	0(0)
	KRTAP17	1(0)	1(0)	1(0)	1(0)	0(0)	1(0)	1(1)	1(0)	1(0)	1(0)	0(0)
	KRTAP24	1(0)	2(0)	1(0)	1(0)	1(0)	1(0)	1(1)	1(0)	1(0)	1(0)	0(0)
	KRTAP25	1(1)	1(1)	1(0)	0(0)	0(0)	1(1)	0(0)	1(1)	0(0)	1(1)	0(0)
	KRTAP26	2(1)	1(0)	2(0)	1(0)	1(0)	1(1)	1(1)	1(0)	1(0)	1(0)	1(1)
	KRTAP27	1(0)	1(0)	1(0)	1(0)	1(0)	1(1)	0(0)	1(0)	1(0)	1(0)	0(0)
	KRTAP28	9(0)	11(1)	12(2)	10(1)	7(1)	4(3)	9(4)	11(3)	19(1)	23(5)	0(0)
	KRTAP29	1(0)	1(1)	1(0)	1(0)	0(0)	0(0)	1(1)	1(0)	0(0)	1(0)	0(0)
	KRTAP30	0	0	0	0	0	0	0	0	0	0	0
	KRTAP31	0	0	0	0	0	0	0	0	0	0	0
	KRTAP32	0	0	0	0	0	0	0	0	0	0	0
	KRTAP33	0	0	0	0	0	0	0	0	0	0	0
	KRTAP34	0	0	0	0	0	0	0	0	0	0	0
	KRTAP35	0	0	0	0	0	0	0	0	0	0	0
	TOTAL-HS	**86(27)**	**104(25)**	**88(12)**	**83(13)**	**63(15)**	**78(16)**	**50(16)**	**73(14)**	**115(13)**	**99(14)**	**18(13)**
HGT-KRTAP	KRTAP6	4(1)	3(1)	3(0)	3(0)	8(0)	4(0)	4(1)	5(0)	7(1)	9(1)	1(1)
	KRTAP7	1(0)	1(0)	1(0)	1(0)	1(0)	2(0)	2(0)	1(0)	1(0)	1(0)	1(1)
	KRTAP8	1(0)	1(0)	1(0)	1(0)	2(0)	0(0)	0(0)	1(1)	1(0)	1(0)	3(0)
	KRTAP19	8(3)	9(4)	5(3)	9(5)	10(3)	5(0)	6(0)	9(2)	0(0)	2(0)	6(5)
	KRTAP20	2(0)	2(0)	1(0)	2(1)	4(0)	6(1)	5(1)	10(3)	27(3)	8(1)	6(6)
	KRTAP21	4(1)	3(0)	3(0)	8(2)	2(0)	3(1)	3(1)	2(1)	3(0)	4(1)	0(0)
	TOTAL-HGT	**20(5)**	**19(5)**	**14(3)**	**24(8)**	**27(3)**	**18(5)**	**20(2)**	**28(7)**	**39(4)**	**25(3)**	**17(13)**
	ALL-KRTAP	**106(32)**	**123(30)**	**102(15)**	**107(21)**	**90(18)**	**96(21)**	**70(18)**	**101(21)**	**154(17)**	**124(17)**	**35(26)**

KRTAP		Cow	Pig	Alpaca	Horse	Panda	Bat	Hedgehog	Elephant	Armadillo	Sloth	Wallaby
HS-KRTAP	KRTAP1	4(0)	4(1)	5(1)	4(0)	3(0)	4(1)	4(1)	4(0)	2(0)	4(1)	4(1)
	KRTAP2	4(0)	4(0)	3(0)	4(0)	5(1)	3(0)	4(1)	4(0)	4(0)	3(0)	5(0)
	KRTAP3	4(0)	4(0)	5(2)	4(0)	4(1)	4(0)	4(0)	4(1)	3(1)	4(0)	4(0)
	KRTAP4	16(5)	7(0)	7(3)	12(0)	4(1)	6(0)	8(1)	13(4)	9(3)	8(0)	17(4)
	KRTAP5	18(5)	0(0)	2(2)	11(2)	13(3)	6(4)	5(0)	12(4)	8(7)	8(1)	7(1)
	KRTAP9	19(3)	5(0)	4(1)	8(2)	2(0)	9(3)	9(1)	4(0)	7(2)	50(5)	0(0)
	KRTAP10	15(3)	14(1)	1(1)	11(0)	8(1)	14(5)	1(1)	8(1)	3(1)	1(0)	5(3)
	KRTAP11	1(0)	1(0)	1(0)	1(0)	1(0)	1(0)	1(1)	1(0)	0(0)	1(0)	1(0)
	KRTAP12	6(0)	4(0)	1(0)	10(1)	5(1)	1(0)	10(1)	11(2)	1(0)	5(2)	7(0)
	KRTAP13	8(3)	14(10)	8(0)	10(2)	8(1)	12(3)	7(1)	11(5)	9(5)	16(13)	0(0)
	KRTAP16	1(0)	0(0)	1(0)	1(0)	1(0)	1(0)	2(1)	1(0)	1(0)	2(1)	1(0)
	KRTAP17	1(0)	1(0)	0(0)	1(0)	1(0)	1(0)	0(0)	1(0)	0(0)	1(0)	1(0)
	KRTAP24	1(0)	1(1)	1(0)	1(0)	1(0)	1(0)	1(0)	1(0)	1(0)	1(0)	0(0)
	KRTAP25	0(0)	1(1)	1(1)	1(0)	1(0)	1(1)	0(0)	0(0)	0(0)	0(0)	0(0)
	KRTAP26	2(0)	1(0)	1(0)	1(0)	1(0)	1(0)	1(0)	3(0)	1(0)	4(1)	0(0)
	KRTAP27	1(0)	1(0)	1(0)	1(0)	1(0)	1(0)	1(1)	1(0)	1(0)	0(0)	0(0)
	KRTAP28	10(2)	14(2)	2(2)	9(1)	12(6)	4(0)	11(0)	8(1)	7(2)	5(2)	10(1)
	KRTAP29	2(0)	0(0)	2(2)	1(0)	1(0)	1(0)	1(0)	1(0)	1(0)	1(1)	1(0)
	KRTAP30	0	0	0	0	0	0	0	0	0	0	0
	KRTAP31	0	0	0	0	0	0	0	0	0	0	0
	KRTAP32	0	0	0	0	0	0	0	0	0	0	0
	KRTAP33	0	0	0	0	0	0	0	0	0	0	0
	KRTAP34	0	0	0	0	0	0	0	0	0	0	0
	KRTAP35	0	0	0	0	0	0	0	0	0	0	0
	TOTAL-HS	**113(21)**	**76(16)**	**46(14)**	**91(8)**	**72(15)**	**71(17)**	**70(10)**	**88(18)**	**58(21)**	**114(27)**	**63(10)**
HGT-KRTAP	KRTAP6	5(0)	4(0)	9(0)	4(0)	7(2)	6(0)	4(1)	5(1)	0(0)	0(0)	0(0)
	KRTAP7	1(0)	1(0)	1(0)	1(0)	1(0)	1(0)	1(0)	1(1)	1(0)	1(0)	1(0)
	KRTAP8	2(0)	4(1)	0(0)	3(1)	1(0)	1(0)	1(0)	1(0)	1(0)	4(2)	1(0)
	KRTAP19	8(1)	3(2)	0(0)	8(2)	8(3)	7(2)	4(0)	6(1)	0(0)	7(1)	0(0)
	KRTAP20	5(0)	5(0)	12(2)	12(0)	4(0)	6(3)	2(2)	6(1)	12(0)	37(3)	38(0)
	KRTAP21	11(0)	9(0)	3(1)	6(1)	8(1)	8(2)	5(1)	5(0)	13(0)	12(1)	0(0)
	TOTAL-HGT	**32(1)**	**26(3)**	**25(3)**	**34(4)**	**29(6)**	**29(7)**	**17(4)**	**24(4)**	**27(0)**	**61(7)**	**40(0)**
	ALL-KRTAP	**145(22)**	**102(19)**	**71(17)**	**125(12)**	**101(21)**	**100(24)**	**87(14)**	**112(22)**	**85(21)**	**175(34)**	**103(10)**

Number of pesudogene is represented in parenthesis.

Figure 3 The topological tree representing evolution of KRTAP gene family repertoires in 30 mammalian species. Twenty-two from the present study and eight from Wu et al, 2008 [14] marked with an asterisk). Stars and circles respectively show the gain and loss of subfamilies, by numbers below.

Species	All (Total KRTAPs)				HS (High Cysteine KRTAPs)				HGT (High Glycine-Tyrosine KRTAPs)			
Human*	122	21	101	17%	95	12	83	13%	27	9	18	33%
Chimpanzee*	120	16	104	15%	95	11	84	13%	25	5	20	25%
Gorilla	106	32	74	30%	86	27	59	31%	20	5	15	25%
Orangutan	123	30	93	24%	104	25	79	24%	19	5	14	26%
Gibbon	102	15	87	15%	88	12	76	14%	14	3	11	21%
Macaque*	129	20	109	16%	99	15	84	15%	30	5	25	17%
Mormoset	107	21	86	20%	83	13	70	16%	24	8	16	33%
Tarsier	90	18	72	20%	63	15	48	24%	27	3	24	11%
Mouse lemur	96	21	75	22%	78	16	62	21%	18	5	13	28%
Bushbaby	70	18	52	26%	50	16	34	32%	20	2	18	10%
Treeshrew	101	21	80	21%	73	14	59	19%	28	7	21	25%
Mouse*	188	13	175	7%	132	11	121	8%	56	2	54	4%
Rat	174	11	163	6%	130	9	121	7%	44	2	42	5%
Cavia	154	17	137	11%	115	13	102	11%	39	4	35	10%
Rabbit	124	17	107	14%	99	14	85	14%	25	3	22	12%
Dolphin	35	26	9	74%	18	13	5	72%	17	13	4	76%
Cow	145	22	123	15%	113	21	92	19%	32	1	31	3%
Pig	102	19	83	19%	76	16	60	21%	26	3	23	12%
Alpaca	71	17	54	24%	46	14	32	30%	25	3	22	12%
Horse	125	12	113	10%	91	8	83	9%	34	4	30	12%
Dog*	128	16	112	13%	88	6	82	7%	40	10	30	25%
Panda	101	21	80	21%	72	15	57	21%	29	6	23	21%
Bat	100	24	76	24%	71	17	54	24%	29	7	22	24%
Hedgehog	87	14	73	16%	70	10	60	14%	17	4	13	24%
Elephant	112	22	90	20%	88	18	70	20%	24	4	20	17%
Armadillo	85	21	64	25%	58	21	37	36%	27	0	27	0%
Sloth	175	34	141	19%	114	27	87	24%	61	7	54	11%
Opossum*	128	14	114	11%	84	8	76	10%	44	6	38	14%
Wallaby	103	10	93	10%	63	10	53	16%	40	0	40	0%
Platypus*	118	12	106	10%	96	10	86	10%	22	2	20	9%

in subfamily 20, respectively) in sloth (*Choloepus hoffmanni*), a nocturnal hairy mammal with long, coarse and shaggy fur that serves as a host for different microorganism [59] (Table 1, Figures 3 and 4). Similarly, we found gene expansion in subfamily 20 (27 genes copies) in the rodent Guinea Pig (*Cavia porcellus*), and 38 genes copies in the marsupial Wallaby (*Macropus eugenii*). Subfamily 28 has expanded in Rabbit (*Oryctolagus cuniculus*) (23 genes copies), which belongs to order lagomorpha. We typically observed functional genes in (HS-KRTAPs) subfamilies 11, 16, 17, 24-27 and 29 varying from a minimum of one to a maximum of three. The subfamilies 11, 16, 17 and 25 have a maximum of one functional gene member, subfamilies 24 and 29 have a maximum of two functional genes (present in orangutan and cow, respectively), and subfamily 26 has a maximum of three members (present in sloth and elephant). Subfamily 7, belonging to the high glycine-tyrosine group, has a maximum of one functional
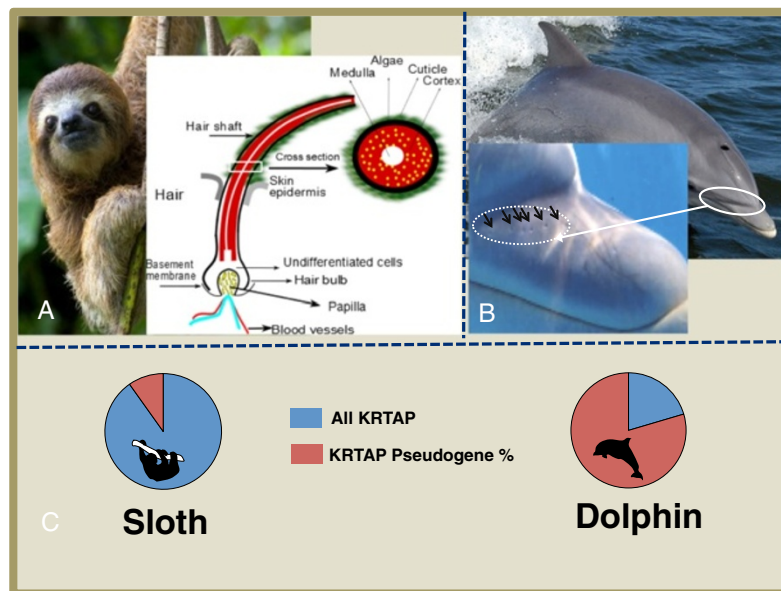
Figure 4 Hair characteristic adaption in terrestrial and aquatic mammal. Sloth an arboreal mammal with high density of hair harboring algae (image credits: firstworldfacts.com). Representation of sloth hair with algal growth and cross section of hair showing the major layers of hair shaft **(A)**. Bottlenose dolphin (image credits: Public source Dolphin NASA) with rostrum selected in circle and detailed in image with arrows point the hairless vibrissae crypts of dolphin (image credits: Élio A. Vicente, Zoomarine) **(B)**. Overall number of KRTAP genes and percentage of pseudogene present in sloth and dolphin **(C)**.

gene member (Table 1). We found that closely related species, e.g. among Hominidae family (Human, chimpanzee, gorilla and orangutan), have very similar gene repertoires with only slight differences (e.g. Humans have highest HGT pseudogenes; Figure 3). We also observed the apparent reduction in the KRTAP gene repertoire in alpaca (fibre), armadillo (modified scales), hedgehog (spines), and dolphin (mostly hairless and aquatic) (Figure 3, Figure 5), probably due to the replacement or modification of hair function or to extensive specialization and subsequent selection and pseudogenization.

For example, we observed high rates of pseudogenization (Figures 3, 4 and 5 and Table 1) (74% compared to the mammalian average of 19%) and only nine intact genes in the dolphin (*Tursiops truncatus*). This aquatic mammal is almost hairless, with only a few hairs (bristles) on the upper lip of the rostrum, which are shed soon after birth, leaving hairless pits on the rostrum of adults that have specialized sensory function [60-65] (Figure 4). The epidermal surface also undergoes high proliferation and sloughing of epidermis cells in order to maintain a smooth skin, a major advantage for swimming [66,67].

Concerted evolution, GC bias and sequence divergence
Tandemly arranged gene members of multigene families often show more similarity among each other than with their counterpart's orthologs in other species, which suggest that they evolved in similar or concerted fashion. This would further lead to species-specific variation, as

observed in KRTAP gene family. Two mechanisms play an important role in concerted evolution. Recombination increases the copy number of gene by providing raw material for further functional innovations and diversification, and gene conversion, which principally homogenizes genes, can help insure the rapid synthesis of a gene product (protein) that may be required during a precise stage of cell cycle [68]. Gene conversion also decreases the evolutionary distance among paralogous members and shifts the substitutions from weak (A or T) to strong (G or C) by increasing GC content through biased gene conversion (gBGC) [69-72]. The negative correlation between evolutionary distance calculated by synonymous substitution rates and GC content provides the level of divergence between the members of a subfamily [73].

Using Geneconv [74] and RDP3 [75] we found higher rates of gene conversion and recombination events in the high cysteine KRTAPs compared with the high glycine-tyrosine KRTAP genes (Additional files 3 and 4). KRTAP subfamilies also displayed different rates of gene conversion in different species. For example, in gorilla we found 44 gene pairs of the KRTAP10 under gene conversion, compared to only 17 gene pairs found for this subfamily in gibbon (Additional file 3). The high level of gene conversion also reduces orthologous relationship between genes of two different species.

Sequences with higher synonymous substitution rates (dS) had higher overall GC content (GC%) and third-codon GC content (GC3%), and lower synonymous substitution

Figure 5 Variation in KRTAP gene family in mammals and relation with hair characteristic features.
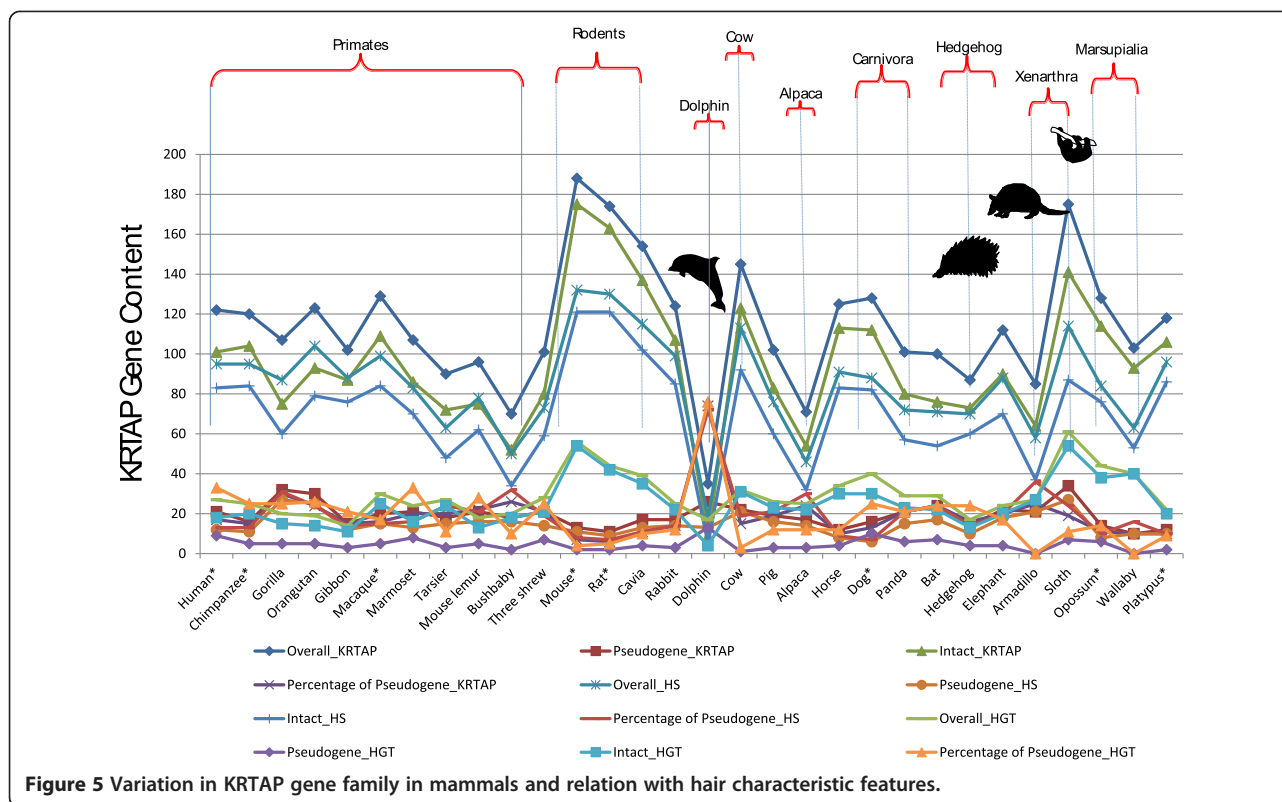
rates in the high-cysteine genes than the high glycine-tyrosine genes. The negative correlation between GC content and synonymous substitution rate (dS) is consistent with the higher rates of concerted evolution observed in high cysteines (Figure 6A and B). The high GC content in the HS –KRTAP gene family compared with the HGT-KRTAP could be a consequence of the high number of gene conversion events.

Adaptive evolution

Gene expansion provides the essential raw material for the positive selection to act [76], which in turn accelerates the diversification of duplicated copies by increasing the number of nonsynonymous substitutions (dN) relative to the synonymous substitutions (dS) through positive selection (dN/dS > 1). The PAML package [77] was used to identify signatures of positive selection. Specifically, we used likelihood ratio test for positive selection [78,79] to test site-specific models comparing twice the difference in log-likelihood between two models to chi square distribution with two degrees of freedom. For expanded subfamilies, such as in the case of the KRTAP20 in wallaby with 38 members, we tested if this species-specific expansion has been influenced by adaptive evolution. We tested two nested pairs of site-specific models (M1a vs. M2a and M7 vs. M8), where M1a and M7 states no positive

selection ($\omega \le 1$) and M2a and M8 states positive selection ($\omega \ge 1$). In both cases the likelihood of positive selection was significantly higher (p < 0.0001), retrieving similar sites under positive selection. The likelihood ratio test is a conservative approach, which can be biased by false positives in the presence of high recombination rates [80]. Thus, we evaluated the possibility of gene conversion/recombination in the KRTAP20 subfamily that has expanded dramatically in the wallaby genome, but did not detect such evidence. The results of positive selection tests are shown in Table 2. The positive selection acting of KRTAP probably favored the diversification and adaptation to different environment.

Differential evolution of the HS and HGT KRTAPs

The KRTAP multigene family has experienced dynamic evolution and diversification within and among genomes as observed in the 30 diverse subfamilies of high cysteine and high glycine-tyrosine subfamilies. These two groups have evolved differently, with the high cysteine group showing high rates of gene conversion within subfamilies, with some exhibiting characteristic differences in copy number, while others have been more conserved. This may be an adaptive mechanism promoting a high order of amplification of similar copies to meet the high demand for the structural proteins required to adapt to changing environmental conditions (e.g. sloth have extensive hairs
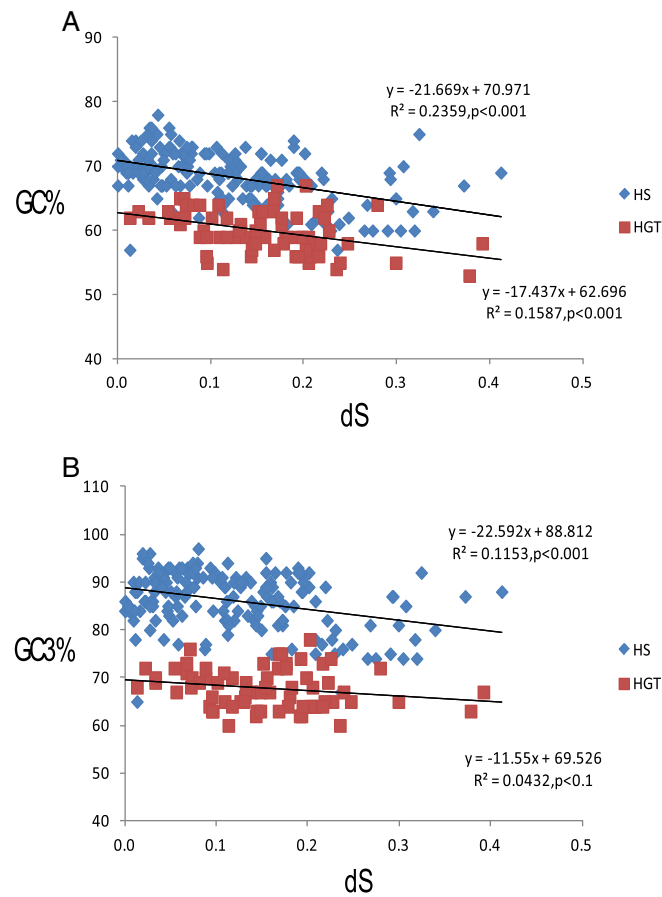
Figure 6 GC-content dynamics. Figure legend text GC-biased gene conversion (gBGC) and evolutionary distance between the KRTAP genes, shown by the correlation between the synonymous substitution rates (dS) and GC content (GC%) among paralogous members of each subfamily **(A)** and third codon GC content (GC3%) **(B)**. Negative correlation points towards the gene conversion. High cyteine KRTAP (HS) and high glycine-tyrosine KRTAP (HGT) are represented by blue and red squares respectively. The linear regression is shown.

that can harbor symbiotic microorganism communities while the dolphin is "hairless" in response to a more-predictable and constant environment and to create less resistance when swimming). We also compared the differential evolutionary patterns between high cysteine and high glycine–tyrosine genes using the Pearson correlation coefficient for the number of genes in each subfamily between species. The coefficient value for high cysteine is significantly higher than for high glycine-tyrosine (Figure 7A) and the coefficient values for the two are positively correlated (p < 0.001) (Figure 7B). The high GC content and negative correlation between GC content and synonymous

Table 2 Likelihood ratio test for PAML site models within Wallaby

Model	Parameters	lnL	2ΔlnL (LRT)
M0	ω: = 0.65464	-1673.250363	NA
M1a	p_0: 0.57003 p_1: 0.42997	-1576.668729	
	$ω_0$: 0.09616 ω1: 1.00000		
M2a	p_0: 0.49130 p_1: 0.31322 p_2: 0.19547	-1556.441999	M1a vs. M2a 40.45346 (p = 1.643E-09)
	$ω_0$: 0.10488 $ω_1$: 1.00000 $ω_2$: 3.48389		
M7	p = 0.33061 q = 0.37447	-1581.277897	
M8	p_0 = 0.74935 p = 0.55372 q = 1.09725	-1560.301816	M7 vs. M8 41.952162 (p = 7.765E-10)
	(p_1 = 0.25065) $ω_1$ = 2.75722		
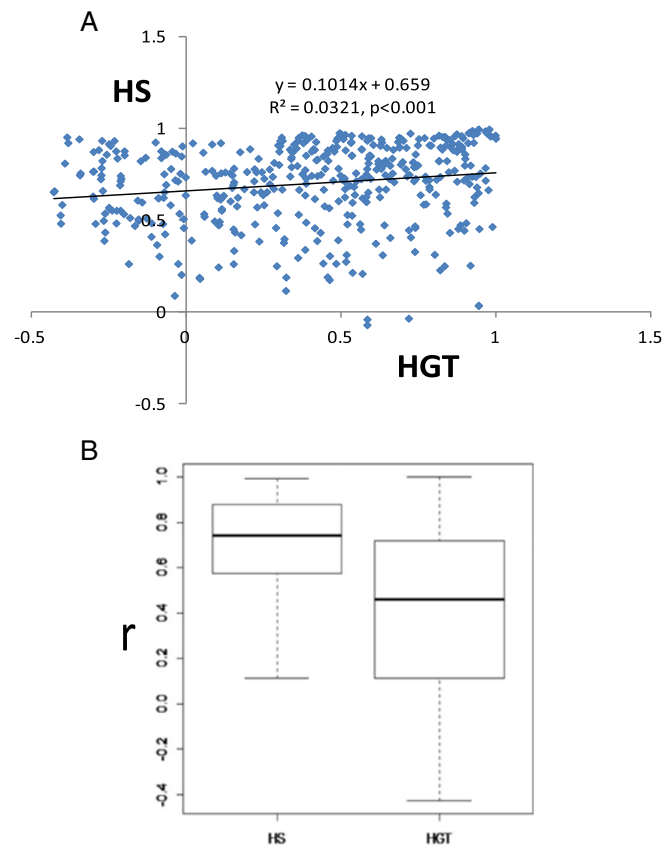
Figure 7 Pearson correlation coefficients (r) show the evolutionary differentiation of KRTAP genes. Pearson correlation coefficients (r) values of the high cysteine and high glycine-tyrosine KRTAP are positively correlated. The linear regression is shown. **(A)** The boxplot for Pearson correlation coefficients (r) of gene numbers of each subfamily between species shows, high cysteine KRTAP genes have higher correlation coefficient than high glycine-tyrosine KRTAP genes **(B)**.

substitution rates also support the higher rates of gene conversion observed in high cysteine genes relative to high glycine-tyrosine KRTAPs, suggesting that high cysteine are under high rates of concerted evolution promoted by gene conversion and recombination events (see Additional files 3 and 4). By contrast, HGT-KRTAP had a more-dynamic evolutionary pattern, with less evidence of gene conversion or recombination, but with signatures of positive selection.

Size Polymorphism and amino acid composition affects KRTAP matrix formation and interactions with hair KIFs

The KRTAP family is widely grouped into three major categories based on amino acid composition: (i) high sulfur (<30% cysteine content), including subfamilies 1, 2, 3, 10-13, 16, 24-27, 29, 31, 34 and 35; (ii) ultrahigh sulfur (>30% cysteine content), including subfamilies 4, 5, 9, 17, 28, 30, 32 and 33 subfamilies; and (iii) high glycine/tyrosine, including subfamilies 6, 7, 8, 19, 20 and 21. The amino acid composition is shown in Table 3. Subfamily gene members also showed size polymorphism [81-83] mostly due to cysteine-rich repeats, which create difference in cysteine content. Cysteine is important for the formation of strong disulphide bonds. Thus, changes in cysteine composition can result in differential interaction among KRTAPs and between KIFs and KRATPs leading to combinatorial complexity and thereby creating morphological differences in hair fiber strength, rigidity and flexibility [40].

Discussion

Gene families are formed by gene duplication, a process that provides important raw material for functional innovation and adaptive selection. Gene families vary in size from a few to thousands of gene members, which makes it difficult to identify and characterize them without sufficient genome sequences. The genome sequencing projects have made it possible to explore complex gene families involved in different phenotypes. Here we explored the mammal-specific KRTAP gene family, which is the major constituent of the hair proteome and plays a primary role in hair formation and thus long been associated with phenotypic differences in hair and wool. This study assessed patterns of variation using comparative genomic approaches in the KRTAP gene

Table 3 Amino acid composition of KRTAPs subfamily genes in mammals

KRTAP	Subfamily	Cysteine	Glycine	Leucine	Proline	Glutamine	Serine	Threonine	Tyrosine
HS-KRTAP	KRTAP1	25.98	9.68	1.83	9.26	7.16	13.89	9.41	1.70
	KRTAP2	28.83	4.35	1.59	14.50	5.29	9.17	9.28	0.52
	KRTAP3	19.65	3.92	8.05	15.43	3.12	8.51	10.97	1.29
	KRTAP4	34.84	3.29	1.15	9.87	6.14	17.60	7.19	0.72
	KRTAP5	33.67	23.00	0.35	5.02	3.74	20.58	0.28	0.36
	KRTAP9	35.30	3.57	1.17	11.64	7.41	12.57	12.79	1.81
	KRTAP10	25.92	1.95	3.66	13.79	5.85	18.86	4.85	0.73
	KRTAP11	12.84	6.86	3.79	8.25	7.16	16.22	10.81	2.64
	KRTAP12	22.54	4.26	4.15	13.85	6.24	17.94	4.27	1.11
	KRTAP13	11.11	10.67	6.62	7.25	3.83	21.93	5.99	7.71
	KRTAP16	19.24	1.89	2.54	14.12	5.33	17.40	5.94	2.24
	KRTAP17	36.09	28.93	0.00	5.29	3.73	10.37	2.83	0.15
	KRTAP24	9.73	5.89	7.90	9.49	4.16	17.18	6.87	6.55
	KRTAP25	8.17	3.92	6.86	8.82	8.17	16.99	1.96	7.52
	KRTAP26	10.38	6.58	9.55	11.13	4.08	18.98	6.24	4.36
	KRTAP27	8.87	4.33	6.69	7.97	8.32	17.15	7.90	2.04
	KRTAP28	38.43	30.92	0.04	1.78	4.19	7.71	2.68	1.20
	KRTAP29	16.25	6.13	3.74	11.10	7.96	16.66	6.92	2.29
HGT-KRTAP	KRTAP6	14.37	39.89	5.40	0.11	0.02	6.96	0.16	22.07
	KRTAP7	7.82	20.26	5.99	6.85	0.29	11.93	5.65	11.76
	KRTAP8	6.15	22.47	4.52	6.64	0.14	8.83	2.69	19.15
	KRTAP19	8.07	35.88	4.51	1.53	0.10	10.39	0.33	18.18
	KRTAP20	10.62	32.08	4.97	2.22	0.13	7.01	0.54	25.05
	KRTAP21	17.88	34.32	0.76	1.76	0.16	11.50	1.77	21.10

Average percentage of eight amino acids.

family. Our study used 22 diverse mammalian genomes that encompassed closely related species, e.g. family Hominidae of primates, comparing apes with dense hair cover with human with much less hair cover, along with species with diverse hair related characteristic, e.g. alpaca (fibre), armadillo (modified scales), hedgehog (spines), sloth (hosting hair symbionts) and dolphin (mostly hairless and aquatic), to obtain greater insights into the KRTAP gene family evolution relative to mammalian hair and phenotypic variations.

We found high molecular diversity within the KRTAP gene family, with 30 subfamilies (24 belonging to high cysteine and six belonging to high glycine-tyrosine KRTAP) (Additional file 2 and Table 1) and approximately 100-180 KRTAP gene members, which are arranged in five clusters at five different chromosome locations in a genome (Figure 2). Most KRTAP subfamilies are found in all mammalian orders, with variations in expansion, contraction, presence/absence, different rates of pseudogenization and sequence variation (length polymorphisms and amino acid changes). For example, we found species-specific differences in the size and compositions of some subfamilies

(e.g. subfamilies 4, 5 and 9) probably caused by unequal crossing over accompanied with high GC content. Moreover, we also found lineage-specific trends, such as in marsupials, where both wallaby and opossum lacked subfamilies 13, 21 and 26 and showed expansion of the KRTAP20 subfamily, which is under positive selection (Table 2). However, highly conserved sequences and the maintenance of the same number of members in subfamilies 1, 2, and 3 suggests that high rates of gene conversion maintain homogeneity and with evolution occurring through a process of punctuated equilibrium [14,84,85]. Similarly, the conserved synteny of KRTAP gene clusters shows that there are also strong constraints acting on this gene family and supports the important role of KRTAP gene family in shaping hair characteristics.

Together with the high molecular diversity of the KTRAP gene family observed in our study, considerable intraspecific diversification has been reported earlier with copy number variations [86-91] in ethnic human populations and allelic variations in sheep. Such sequence polymorphism found in KRTAP gene members

may influence its expression, protein structure, and/or post-translation modifications and consequently effect wool/hair fibre structure and wool/hair quality traits [92,23,93]. For example, evidence of linkage reported between KRTAP6-8 and wool fiber diameter (quantitative trait) in sheep [94] may be related with similar characteristics in alpaca (fiber), which has one of the largest number of KRTAP6 genes (n = 9) in mammals. Further exploration of KRATP gene family in sheep could help shed light on the improvement of hair/wool traits [57,95,94,96].

Interestingly, we found differences in KRTAP gene repertoire related with hair features. A very expanded KRTAP gene family repertoire (175 total genes and 141 intact genes) was found in sloth, which has long, dense and coarse body hair cover, which also serves as a host for symbiotic microorganisms (e.g., cyanobacteria) in this arboreal mammal. By contrast, we have detected a reduced number of functional KRTAP genes and high percentage of KRTAP pseudogenes (74%) in dolphin (aquatic mammal), highlighting the much lower KRTAP gene requirement in this smooth-skinned species that only has a few hairs (bristle) at the rostrum. These are lost soon after birth and in adults the hairless pits are adapted for sensory functions (Figure 4). This example illustrates the adaptive potential of hair follicles to diversify into more specialized sensory organs.

We also observed that several unique hair-related phenotypes in some of the species, such as scales in armadillo, fiber in alpaca and spines in hedgehog, are linked with an inverse correlation between the number of intact KRTAP genes and the number of pseudogenes. Armadillo and hedgehog have specialized hair features, where the pattern in alpaca could be due to inbreeding and genome homogenization during domestication. The "hairless" dolphin had a large number of KRTAP pseudogenes relative to intact genes (Figure 5). In contrast, the sloth showed a high positive correlation with intact KRTAP genes, suggesting that changes in KRTAP can be related to morphological diversity of hair phenotypes (Figure 5). In contrast, we did not find any correlation between the comparatively hairless human and other primates, which favors the hypothesis, that diversification of keratinization structures in mammals may be collectively explained by KRTAP gene number variation together with other biological mechanisms, such as gene expression variation in KRTAPs (which can be further influenced by KRTAP genes polymorphism).

We suggest that the diverse repertoire and variability in KRTAPs (at gene, family and genome level) provides extraordinary combinatorial complexity [97] for interaction between KRTAPs and Keratin intermediate filaments, resulting in a rich diversity of pathways for evolutionary change, which together with differences in higher order expression of KRTAP genes results in the diverse hair morphological characteristic visible in extant mammals. Overall, we conclude that KRTAPs play an important role in evolution and diversification of hair character across mammals and are responsible for unique features of hair.

Conclusions

The present study explored KRTAP gene family evolution in various mammalian species inhabiting diverse terrestrial and aquatic environments. The two groups of the KRTAP gene family, high cysteine and high glycine-tyrosine KRTAP genes, have evolve differently, resulting in species-specific diversification of this multi-gene family and leading wide morphological diversity in hair characteristics in extant mammals. We conclude that differences in KRTAP gene family repertoires, together with changes in expression patterns, are responsible for shaping unique hair characteristics in diverse mammalian species. These differences are more pronounced between aquatic and terrestrial species and demonstrate the important adaptive role of hairs in terrestrial colonization and the radiation of mammals from water to land. Future studies comparing the KRTAP repertoire in key model organisms, e.g. Alpaca and sheep, may provide insights to understanding the role of KRTAP gene variations in hair fibre traits and its use in textile industry.

Methods

Gene identification

All KRTAP genes are relatively small (ca. 1 kb) and generally have single exon [83]. Some KRTAP genes appear to possess small introns. However these are similar to repeat regions present in the gene [98] and can be translated in-frame with the coding exon, leading to the conclusion that all KRTAP are intron-less [14,99]. The presence of KRTAP gene clusters in mammalian genomes makes it easy to identify and fully characterize the gene family in genomes with high coverage, but in low coverage genomes it requires much more manual inspection and in-depth screening to insure an almost complete or maximum possible repertoire of non-redundant KRTAP genes (Additional file 2). In order to identify the complete gene repertoire in the KRTAP gene family, all previously annotated gene sequences were taken and used as query in blast searches against the genomes from Ensembl http://www.ensembl.org/Multi/Tools/Blast and NCBI genome database http://www.ncbi.nlm.nih.gov/blast/Blast.cgi using BLASTN algorithm [100] and E-value cut-off of 10. We retrieved multiple hits for each query and selected all the non-redundant hits by extending 500 bp at both 5' and 3' ends. Non-redundant hits, which were seen to be clustered in the same region (chromosome, contig, genescaffold, scaffold, supercontig) were merged together to form a single extended common

DNA fragment, bearing all these hits and the ends of this fragment were further extended to maximum 0.3 Mbp were ever possible. Finally all the hits were used to identify and annotate KRTAP gene using program BLAST 2 Sequences [101] TFASTX and TFASTY incorporated in Fasta programs [102] and ORF finder from NCBI http://www.ncbi.nlm.nih.gov/gorf/gorf.html and Mobyle [103]. The identified gene were blast searched against non redundant NBCI blast database, all best hits which resulted in KRTAP or KRTAP like sequences were finally taken as KRTAP genes. The KRTAP genes were further classified into intact/complete genes, partial genes and pseudogene with interrupting frame-shift mutations and/or stop codons.

Phylogenetic analysis

We employed phylogenic tree building method to further classify the identified KRTAP gene repertoire to their respective subfamilies. For each species the intact genes were used for building phylogentic tree. All intact KRTAP genes were translated to amino-acid and aligned using ClustalW incorporated in MEGA4.0 [104,105] with Blosum protein weight matrix the manual adjustments were done when ever needed to correct the final alignment. This final protein sequence alignment was used to build the KRTAP gene tree with the Neighbour-Joining method with P-distance and the interior branch test evaluated with 1,000 replications [106] (Additional file 1: Figures S1-24). We make use of unique motifs and repeat sequence structure present in KRTAP subfamilies along with phylogeny and blast results to further help identify and classify partial and pseudogenes to the respective subfamilies.

Gene conversion and Recombination study

We used the program Geneconv http://www.math.wustl.edu/~sawyer/geneconv/ [74] to detect statistically significant events of sequence homogenization on paralogs using Global Bonferroni corrected P values. The lower P values indicate greater support for gene conversion. The multiple sequence alignment of protein were back translated and used as input for the Geneconv to give both global and pairwise fragments involved in gene conversion. We also used the RDP3 software [75] to detect recombination events using RDP, Bootscan, MaxChi and Chimaera with 1,000 permutations and cutoff p value of 0.01 employing Bonferroni correction.

The evolutionary distance between genes can be calculated with synonymous substitution, which are immune to selection and are not decreased by negative selection [50]. The sequence divergence was estimated using approximate synonymous substitution rates (dS) implemented in MEGA using modified Nei-Gojobori (P-distance) method with transition/transversion ratio of 2. GC content was

estimated using MEGA5.0 [107]. More than two sequences are needed to detect the signals of recombation therefore subfamilies having more than three genes were used for studies of gene conversion (Additional files 3 and 4).

Statistical analysis

In order to study the differential evolutionary pattern of high cysteine and high glycine-tyrosine KRTAP genes, we compared the pairwise-pearson correction coefficient (Figure 7) of the number of genes present in each subfamily (Table 1). We also compared the correlation between GC content (GC% and GC3%) and synonymous substitution rates (Figure 6) using the Nei-Gojobori (P-distance) method with transition/transversion ratio of 2 in MEGA4.0 [95].

Availability of supporting data

All the supporting data are included as additional files.

Additional files

Additional file 1: Figure S1-S24. The phylogeny of high cysteine KRTAP genes in 22 mammalian species. Neighbor-joining method with P-distance and interiors branch test with 1,000 replications (shown on the branches) was employed to build the trees. Figures S23 and S24 shows loss of, one to one orthologous relationship between two species due to concerted evolution. The KRTAP members are labeled with species abbreviation, Gene ID and KRTAP subfamily (Additional file 2) Figure S1-21 are in order, Gorilla, Pongo, Gibbon, Mormoset, Tarsies, Mouse lemur, Bushbaby, Treeshrew, Cavia, rabbit, Cow, Pig, Alpaca, Horse, Panda, Bat, Hedgehog, Elephant, Armadillo, Sloth and Wallaby. Figure S22 (Gorilla and Gibbon) and S23 (Gorilla and Cavia) shows reduced orthology with increase in divergence time. Figure S24 shows relationship between all HGT members in 22 genomes.

Additional file 2: Table S2. The excel file shows the genomic coordinates of the KRTAP gene repertoires in 22 mammalian species studied. The Gene ID corresponds to the genomic location.

Additional file 3: Table S3. Gene pairs under significant gene conversion, as detected by GeneConv program.

Additional file 4: Table S4. Results of RDP3 showing unique recombination events with statistical significance P value of less than 0.01 employing Bonferroni correction.

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

Authors' contributions
IK performed all the genomics, phylogenetics, and evolutionary analyses and drafted the manuscript. EM participated in the genome mining analysis and drafting of the study. VV participated in the drafting and coordination of the study. SJOB participated in the drafting and coordination of the study. WEJ participated in the drafting and coordination of the study. AA participated in the design, genetic analyses, drafting and coordination of the study. All authors read and approved the final manuscript.

Author details

[1]CIMAR/CIIMAR, Centro Interdisciplinar de Investigação Marinha e Ambiental, Universidade do Porto, Rua dos Bragas 177, 4050-123 Porto, Portugal. [2]Departamento de Biologia, Faculdade de Ciências, Universidade do Porto, Rua do Campo Alegre, 4169-007 Porto, Portugal. [3]Theodosius Dobzhansky Center for Genome Bioinformatics, St. Petersburg State University, St. Petersburg 199004, Russia. [4]Oceanographic Center 8000 N. Ocean Drive, Nova Southeastern University, Ft Lauderdale, Florida 33004, USA. [5]Smithsonian Conservation Biology Institute, National Zoological Park, 1500 Remount Rd, Front Royal, VA 22630, USA.

References

1. Alibardi L: **Adaptation to the land: the skin of reptiles in comparison to that of amphibians and endotherm amniotes.** *Biologia* 2003, **41**:12–41. doi:10.1002/jez.b.00024.
2. Chuong C-M, Homberger DG: **Development and evolution of the amniote integument: current landscape and future horizon.** *J Exp Zool B Mol Dev Evol* 2003, **298**:1–11. doi:10.1002/jez.b.23.
3. George RD, McVicker G, Diederich R, Ng SB, MacKenzie AP, Swanson WJ, Shendure J, Thomas JH: **Trans genomic capture and sequencing of primate exomes reveals new targets of positive selection.** *Genome Res* 2011, **21**:1686–1694. doi:10.1101/gr.121327.111.
4. Sun Y-B, Zhou W-P, Liu H-Q, Irwin DM, Shen Y-Y, Zhang Y-P: **Genome-Wide Scans for Candidate Genes Involved to the Aquatic Adaptation of Dolphins.** *Genome Biol Evol* 2012. 10.1093/gbe/evs123 Available: http://gbe.oxfordjournals.org/content/5/1/130.
5. Alibardi L, Valle LD, Nardi A, Toni M: **Evolution of hard proteins in the sauropsid integument in relation to the cornification of skin derivatives in amniotes.** *J Anat* 2009, **214**:560–586. doi:10.1111/j.1469-7580.2009.01045.x.
6. Alibardi L: **Embryonic keratinization in vertebrates in relation to land colonization.** *Acta Zoologica* 2009, **90**:1–17. doi:10.1111/j.1463-6395.2008.00327.x.
7. Rogers MA, Winter H, Langbein L, Wollschläger A, Praetzel-Wunder S, Jave-Suarez LF, Schweizer J: **Characterization of human KAP24.1, a cuticular hair keratin-associated protein with unusual amino-acid composition and repeat structure.** *J Invest Dermatol* 2007, **127**:1197–1204. 10.1038/sj.jid.5700702.
8. Zimek A, Weber K: **Terrestrial vertebrates have two keratin gene clusters; striking differences in teleost fish.** *Eur J Cell Biol* 2005, **84**:623–635. doi:10.1016/j.ejcb.2005.01.007.
9. Alibardi L: **Structural and immunocytochemical characterization of keratinization in vertebrate epidermis and epidermal derivatives.** *Int Rev Cytol* 2006, **253**:177–259. doi:10.1016/S0074-7696(06)53005-0.
10. Alibardi L, Jaeger K, Valle LD, Eckhart L: **Ultrastructural localization of hair keratin homologs in the claw of the lizard Anolis carolinensis.** *J Morphol* 2011, **272**:363–370. doi:10.1002/jmor.10920.
11. Vandebergh W, Bossuyt F: **Radiation and functional diversification of alpha keratins during early vertebrate evolution.** *Mol Biol Evol* 2011, **2011**. doi:10.1093/molbev/msr269.
12. Eckhart L, Valle LD, Jaeger K, Ballaun C, Szabo S, Nardi A, Buchberger M, Hermann M, Alibardi L, Tschachler E: **Identification of reptilian genes encoding hair keratin-like proteins suggests a new scenario for the evolutionary origin of hair.** *Proc Natl Acad Sci U S A* 2008, **105**:18419–18423. doi:10.1073/pnas.0805154105.
13. Hesse M, Zimek A, Weber K, Magin TM: **Comprehensive analysis of keratin gene clusters in humans and rodents.** *Eur J Cell Biol* 2004, **83**:19–26.
14. Wu D-D, Irwin DM, Zhang Y-P: **Molecular evolution of the keratin associated protein gene family in mammals, role in the evolution of mammalian hair.** *BMC Evol Biol* 2008, **8**:241. doi:10.1186/1471-2148-8-241.
15. Alibardi L: **Fine structure of marsupial hairs, with emphasis on trichohyalin and the structure of the inner root sheath.** *J Morphol* 2004, **261**:390–402. doi:10.1002/jmor.10257.
16. Botchkarev VA, Paus R: **Molecular biology of hair morphogenesis: development and cycling.** *J Exp Zool B Mol Dev Evol* 2003, **298**:164–180. doi:10.1002/jez.b.33.
17. Millar SE: **Molecular Mechanisms Regulating Hair Follicle Development.** *J Invest Dermatol* 2002, **118**:216–225.
18. Schneider MR, Schmidt-Ullrich R, Paus R: **The hair follicle as a dynamic miniorgan.** *Curr Biol* 2009, **19**:R132–R142. doi:10.1016/j.cub.2008.12.005.
19. Hardy MH: **The secret life of the hair follicle.** *Trends Genet* 1992, **8**:55–61. 16/0168-9525(92)90350-D.
20. Franbourg A, Hallegot P, Baltenneck F, Toutaina C, Leroy F: **Current research on ethnic hair.** *J Am Acad Dermatol* 2003, **48**:S115–S119. doi:10.1067/mjd.2003.277.
21. Sahajpal V, Goyal S, Singh K, Thakur V: **Dealing Wildlife Offences in India: Role of the Hair as Physical Evidence.** *Int J Trichology* 2009, **1**:18–26. doi:10.4103/0974-7753.51928.
22. Bahuguna A, Mukherjee SK: **Use of SEM to recognise Tibetan antelope (Chiru) hair and blending in wool products.** *Sci Justice* 2000, **40**:177–182. doi:10.1016/S1355-0306(00)71973-3.
23. Jenkins BJ, Powell BC: **Differential expression of genes encoding a cysteine-rich keratin family in the hair cuticle.** *J Invest Dermatol* 1994, **103**:310–317.
24. Shimomura Y, Aoki N, Rogers MA, Langbein L, Schweizer J, Ito M: **hKAP1.6 and hKAP1.7, two novel human high sulfur keratin-associated proteins are expressed in the hair follicle cortex.** *J Invest Dermatol* 2002, **118**:226–231.
25. Rogers MA, Langbein L, Praetzel-Wunder S, Giehl K: **Characterization and expression analysis of the hair keratin associated protein KAP26.1.** *Br J Dermatol* 2008, **159**:725–729.
26. Rogers MA, Langbein L, Winter H, Ehmann C, Praetzel S, Schweizer J: **Characterization of a first domain of human high glycine-tyrosine and high sulfur keratin-associated protein (KAP) genes on chromosome 21q22.1.** *J Biol Chem* 2002, **277**:48993–49002.
27. Powell BC, Arthur J, Nesci A: **Characterization of a gene encoding a cysteine-rich keratin associated protein synthesized late in rabbit hair follicle differentiation.** *Differentiation* 1995, **58**:227–232. doi:10.1046/j.1432-0436.1995.5830227.x.
28. Pruett ND, Tkatchenko TV, Jave-Suarez L, Jacobs DF, Potter CS, Tkatchenko AV, Schweizer J, Awgulewitsch A: **Krtap16, characterization of a new hair keratin-associated protein (KAP) gene complex on mouse chromosome 16 and evidence for regulation by Hoxc13.** *J Biol Chem* 2004, **279**:51524–51533. doi:10.1074/jbc.M404331200.
29. Stenn KS, Paus R: **Controls of hair follicle cycling.** *Physiol Rev* 2001, **81**:449–494.
30. Rogers GE: **Hair follicle differentiation and regulation.** *Int J Dev Biol* 2004, **48**:163–170. doi:10.1387/ijdb.021587gr.
31. Schweizer J, Bowden PE, Coulombe PA, Langbein L, Lane EB, Magin TM, Maltais L, Omary MB, Parry DD, Rogers MA, Wright MW: **New consensus nomenclature for mammalian keratins.** *J Cell Biol* 2006, **174**:169–174. doi:10.1083/jcb.200603161.
32. Rogers M, Winter H, Langbein L, Bleiler R: **The human type I keratin gene family: characterization of new hair follicle specific members and evaluation of the chromosome 17q21. 2 gene domain.** *Differentiation* 2004, **72**:527–540. doi:10.1111/j.1432-0436.2004.07209006.x.
33. Rogers MA, Edler L, Winter H, Langbein L, Beckmann I, Schweizer J: **Characterization of new members of the human type II keratin gene family and a general evaluation of the keratin gene domain on chromosome 12q13.13.** *J Invest Dermatol* 2005, **124**:536–544.
34. Rogers MA, Winter H, Wolf C, Heck M, Schweizer J: **Characterization of a 190-kilobase pair domain of human type i hair keratin genes.** *J Biol Chem* 1998, **273**:26683–26691. doi:10.1074/jbc.273.41.26683.
35. Rogers MA, Winter H, Langbein L, Wolf C, Schweizer J: **Characterization of a 300 kbp region of human DNA containing the type II hair keratin gene domain.** *J Invest Dermatol* 2000, **114**:464–472. doi:10.1046/j.1523-1747.2000.00910.x.
36. Steinert PM, North AC, Parry DA: **Structural features of keratin intermediate filaments.** *J Invest Dermatol* 1994, **103**:19S–24S.
37. Powell BC, Nesci A, Rogers GE: **Regulation of keratin gene expression in hair follicle differentiation.** *Ann N Y Acad Sci* 1991, **642**:1–20.
38. Powell BC, Rogers GE: **The role of keratin proteins and their genes in the growth, structure and properties of hair.** *EXS* 1997, **78**:59–148.
39. Fujikawa H, Fujimoto A, Farooq M, Ito M, Shimomura Y: **Characterization of the Human Hair Keratin–Associated Protein 2 (KRTAP2) Gene Family.** *J Invest Dermatol* 2012, **132**:1806–1813. doi:10.1038/jid.2012.73.

40. Shimomura Y, Ito M: **Human hair keratin-associated proteins.** *J Invest Dermatol Symp Proc Soc Invest Dermatol Inc Eur Soc Dermatol Res* 2005, **10**:230–233. doi:10.1111/j.1087-0024.2005.10112.x.

41. Lee YJ, Rice RH, Lee YM: **Proteome analysis of human hair shaft: from protein identification to posttranslational modification.** *Mol Cell Proteomics* 2006, **5**:789–800. doi:10.1074/mcp.M500278-MCP200.

42. Koehn H, Clerens S, Deb-Choudhury S, Morton JD, Dyer JM, Plowman JE: **The proteome of the wool cuticle.** *J Proteome Res* 2010, **9**:2920–2928. doi:10.1021/pr901106m.

43. Rogers MA, Langbein L, Winter H, Ehmann C, Praetzel S, Korn B, Schweizer J: **Characterization of a cluster of human high/ultrahigh sulfur keratin-associated protein genes embedded in the type I keratin gene domain on chromosome 17q12-21.** *J Biol Chem* 2001, **276**:19440–19451. doi:10.1074/jbc.M100657200.

44. Rogers MA, Langbein L, Winter H, Beckmann I, Praetzel S, Schweizer J: **Hair keratin associated proteins: characterization of a second high sulfur KAP gene domain on human chromosome 21.** *J Invest Dermatol* 2004, **122**:147–158. doi:10.1046/j.0022-202x.2003.22128.x.

45. Shibuya K, Obayashi I, Asakawa S, Minoshima S, Kudoh J, Shimizu N: **A cluster of 21 keratin-associated protein genes within introns of another gene on human chromosome 21q22.3.** *Genomics* 2004, **83**:679–693.

46. Yahagi S, Shibuya K, Obayashi I, Masaki H, Kurata Y, Kudoh J, Shimizu N: **Identification of two novel clusters of ultrahigh-sulfur keratin-associated protein genes on human chromosome 11.** *Biochem Biophys Res Commun* 2004, **318**:655–664. doi:10.1016/j.bbrc.2004.04.074.

47. Fumasoni I, Meani N, Rambaldi D, Scafetta G, Alcalay M, Ciccarelli FD: **Family expansion and gene rearrangements contributed to the functional specialization of PRDM genes in vertebrates.** *BMC Evol Biol* 2007, **7**:187. doi:10.1186/1471-2148-7-187.

48. Lespinet O, Wolf YI, Koonin EV, Aravind L: **The role of lineage-specific gene family expansion in the evolution of eukaryotes.** *Genome Res* 2002, 1048–1059. doi:10.1101/gr.174302.eages.

49. Leister D: **Tandem and segmental gene duplication and recombination in the evolution of plant disease resistance gene.** *Trends Genet* 2004, **20**:116–122.

50. Zhang J: **Evolution by gene duplication: an update.** *Trends Ecol Evol* 2003, **18**:292–298. doi:16/S0169-5347(03)00033-8.

51. Lynch M, Force A: **The probability of duplicate gene preservation by subfunctionalization.** *Genetics* 2000, **154**:459–473.

52. **Ensembl Genome Browser.** http://www.ensembl.org/index.html.

53. Flicek P, Amode MR, Barrell D, Beal K, Brent S, Denise C-S, Clapham P, Coates G, Fairley S, Fitzgerald S, Gil L, Gordon L, Hendrix M, Hourlier T, Johnson N, Kähäri AK, Keefe D, Keenan S, Kinsella R, Komorowska M, Koscielny G, Kulesha E, Larsson P, Longden I, McLaren W, Muffato M, Overduin B, Pignatelli M, Pritchard B, Riat HS, *et al*: **Ensembl 2012.** *Nucl Acids Res* 2011. doi:10.1093/nar/gkr991 Available: http://nar.oxfordjournals.org/content/40/D1/D84.

54. **Map Viewer - National Center for Biotechnology Information.** http://www.ncbi.nlm.nih.gov/mapview/.

55. Bininda-Emonds ORP, Cardillo M, Jones KE, MacPhee RDE, Beck RMD, Grenyer R, Price SA, Vos RA, Gittleman JL, Purvis A: **The delayed rise of present-day mammals.** *Nature* 2007, **446**:507–512. doi:10.1038/nature05634.

56. Walsh JB, Stephan W: **Multigene Families: Evolution.** *Encyclopedia Life Sci* 2001, 1–6. http://nitro.biosci.arizona.edu/zdownload/papers/ELSGenFamily.pdf.

57. McLaren RJ, Rogers GR, Davies KP, Maddox JF, Montgomery GW: **Linkage mapping of wool keratin and keratin-associated protein genes in sheep.** *Mamm Genome* 1997, **8**:938–940.

58. Warren WC, Hillier LW, Graves JAM, Birney E, Ponting CP, Grützner F, Belov K, Miller W, Clarke L, Chinwalla AT, Yang S-P, Heger A, Locke DP, Miethke P, Waters PD, Veyrunes F, Fulton L, Fulton B, Graves T, Wallis J, Puente XS, López-Otín C, Ordóñez GR GR, Eichler EE, Chen L, Cheng Z, Deakin JE, Alsop A, Thompson K, Kirby P, *et al*: **Genome analysis of the platypus reveals unique signatures of evolution.** *Nature* 2008, **453**:175–183.

59. Suutari M, Majaneva M, Fewer DP, Voirin B, Aiello A, Friedl T, Chiarello AG, Blomster J: **Molecular evidence for a diverse green algal community growing in the hair of sloths and a specific association with *Trichophilus welckeri* (Chlorophyta, Ulvophyceae).** *BMC Evol Biol* 2010, **10**:86. doi:10.1186/1471-2148-10-86.

60. Palmer E, Weddell G: **The Relationship between structure, innervation and function of the skin of the bottle nose dolphin (*Tursiops truncatus*).** *Proc Zool Soc London* 1964, **143**:553–568. doi:10.1111/j.1469-7998.1964.tb03881.x.

61. Meyer W, Schmidt J, Busche R, Jacob R, Naim HY: **Demonstration of free fatty acids in the integument of semi-aquatic and aquatic mammals.** *Acta Histochem* 2012, **114**:145–150. doi:10.1016/j.acthis.2011.03.011.

62. Czech-Damal NU, Liebschner A, Miersch L, Klauer G, Hanke FD, Marshall C, Dehnhardt G, Hanke W: **Electroreception in the Guiana dolphin (Sotalia guianensis).** *Proc R Soc B* 2012. 2011, doi:10.1098/rspb.2011.1127 Available: http://rspb.royalsocietypublishing.org/content/early/2011/07/21/rspb.2011.1127. Accessed 23 July.

63. Mauck B, Eysel U, Dehnhardt G: **Selective heating of vibrissal follicles in seals (Phoca vitulina) and dolphins (Sotalia fluviatilis guianensis).** *J Exp Biol* 2000, **203**:2125–2131.

64. Jenkins J: **"Tursiops truncatus" (On-line), animal diversity web.** 2009. Accessed July 23, 2012 at http://animaldiversity.ummz.umich.edu/site/accounts/information/Tursiops_truncatus.html.

65. Thewissen JGM, Cooper LN, George JC, Bajpai S: **From land to water: the origin of whales, dolphins, and porpoises.** *Evol Educ Outreach* 2009, **2**:272–288. doi:10.1007/s12052-009-0135-2.

66. Fish FE, Hui CA: **Dolphin swimming–a review.** *Mammal Rev* 1991, **21**:181–195. doi:10.1111/j.1365-2907.1991.tb00292.x.

67. Hicks BD, Aubin DJS, Geraci JR, Brown WR: **Epidermal growth in the Bottlenose Dolphin, Tursiops truncatus.** *J Invest Dermatol* 1985, **85**:60–63. doi:10.1111/1523-1747.ep12275348.

68. Brown TA: **Genomes.** In *How Genomes Evolve.NCBI Bookshelf. Chapter 15.* Oxford: Wiley-Liss. Available: http://www.ncbi.nlm.nih.gov/books/NBK21112/.

69. Kostka D, Hubisz MJ, Siepel A, Pollard KS: **The role of GC-biased gene conversion in shaping the fastest evolving regions of the human genome.** *Mol Biol Evol* 2012, **29**(3):1047–1057. doi:10.1093/molbev/msr279. Epub 2011 Nov 10.

70. Duret L, Arndt PF: **The impact of recombination on nucleotide substitutions in the human genome.** *PLoS Genet* 2008, **4**:e1000071. doi:10.1371/journal.pgen.1000071.

71. Escobar JS, Glémin S, Galtier N: **GC-biased gene conversion impacts ribosomal DNA evolution in vertebrates, angiosperms, and other eukaryotes.** *Mol Biol Evol* 2011, **28**:2561–2575. doi:10.1093/molbev/msr079.

72. Galtier N, Duret L: **Adaptation or biased gene conversion? extending the null hypothesis of molecular evolution.** *Trends Genet* 2007, **23**:273–277. doi:10.1016/j.tig.2007.03.011.

73. Noonan JP, Grimwood J, Schmutz J, Dickson M, Myers RM: **Gene conversion and the evolution of protocadherin gene cluster diversity.** *Genome Res* 2004, **14**:354–366. doi:10.1101/gr.2133704.

74. Sawyer S: **Statistical tests for detecting gene conversion.** *Mol Biol Evol* 1989, **6**:526–538.

75. Martin DP, Lemey P, Lott M, Moulton V, Posada D, Lefeuvre P: **RDP3: a flexible and fast computer program for analyzing recombination.** *Bioinformatics* 2010, **26**:2462–2463. doi:10.1093/bioinformatics/btq467.

76. Han MV, Demuth JP, Mcgrath CL, Casola C, Hahn MW: **Adaptive evolution of young gene duplicates in mammals.** *Genome Res* 2009, **19**(5):859–867. doi: 10.1101/gr.085951.108.

77. Yang Z: **PAML 4: phylogenetic analysis by maximum likelihood.** *Mol Biol Evol* 2007, **24**:1586–1591. doi:10.1093/molbev/msm088.

78. Yang Z: **Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution.** *Mol Biol Evol* 1998, **15**:568–573.

79. Nielsen R, Yang Z: **Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene.** *Genetics* 1998, **148**:929–936.

80. Anisimova M, Nielsen R, Yang Z: **Effect of recombination on the accuracy of the likelihood method for detecting positive selection at amino acid sites.** *Genetics* 2003, **164**:1229–1236.

81. Kariya N, Shimomura Y, Ito M: **Size polymorphisms in the human ultrahigh sulfur hair keratin-associated protein 4, kap4, gene family.** *J Invest Dermatol* 2005, **124**:1111–1118.

82. Parry DD, Smith TA, Rogers MA, Schweizer J: **Human hair keratin-associated proteins: sequence regularities and structural implications.** *J Struct Biol* 2006, **155**:361–369. doi:10.1016/j.jsb.2006.03.018.

83. Rogers M, Schweizer J: **Human KAP genes, only the half of it? extensive size polymorphisms in hair keratin-associated protein genes.** *J Invest Dermatol* 2005, **124**:vii–ix. doi:10.1111/j.0022-202x.2005.23728.x.

84. Gould SJ, Eldredge N: **Punctuated equilibrium comes of age.** *Nature* 1993, **366**:223–227. 10.1038/366223a0.

85. Mattila TM, Bokma F: **Extant mammal body masses suggest punctuated equilibrium.** *Proc Biol Sci* 2008, **275**:2195–2199. doi:10.1098/rspb.2008.0354.

86. Marotta M, Chen X, Inoshita A, Stephens R, Budd GT, Crowe J, Lyons J, Kondratova A, Tubbs R, Tanaka H: **A common copy number breakpoint of ERBB2 amplification in breast cancer co-localizes with a complex block of segmental duplications.** *Breast Cancer Res* 2012, **14**:R150. doi:10.1186/bcr3362.

87. Gautam P, Jha P, Kumar D, Tyagi S, Varma B, Dash D, Mukhopadhyay A, Mukerji M: **Spectrum of large copy number variations in 26 diverse Indian populations: potential involvement in phenotypic diversity.** *Hum Genet* 2012, **131**:131–143. doi:10.1007/s00439-011-1050-5.

88. Gong H, Zhou H, Yu Z, Dyer J, Plowman JE, Hickford J: **Identification of the ovine keratin-associated protein KAP1-2 gene (KRTAP1-2).** *Exp Dermatol* 2011, **20**:815–819. doi:10.1111/j.1600-0625.2011.01333.x.

89. Gong H, Zhou H, Plowman JE, Dyer JM, Hickford JGH: **Analysis of variation in the ovine ultra-high sulphur keratin-associated protein KAP5-4 gene using PCR-SSCP technique.** *Electrophoresis* 2010, **31**:3545–3547. doi:10.1002/elps.201000301.

90. Gong H, Zhou H, Hickford JGH: **Diversity of the glycine/tyrosine-rich keratin-associated protein 6 gene (KAP6) family in sheep.** *Mol Biol Rep* 2011, **38**:31–35. doi:10.1007/s11033-010-0074-6.

91. Zhou H, Gong H, Yan W, Luo Y, Hickford JGH: **Identification and sequence analysis of the keratin-associated protein 24-1 (KAP24-1) gene homologue in sheep.** *Gene* 2012, **511**:62–65. doi:10.1016/j.gene.2012.08.049.

92. Yu Z, Gordon SW, Nixon AJ, Bawden CS, Rogers MA, Wildermoth JE, Maqbool NJ, Pearson AJ: **Expression patterns of keratin intermediate filament and keratin associated protein genes in wool follicles.** *Differentiation* 2009, **77**:307–316. doi:10.1016/j.diff.2008.10.009.

93. Gong H, Zhou H, Dyer JM, Hickford JGH: **Identification of the ovine KAP11-1 gene (KRTAP11-1) and genetic variation in its coding sequence.** *Mol Biol Rep* 2011, **38**:5429–5433. doi:10.1007/s11033-011-0697-2.

94. Parsons YM, Cooper DW, Piper LR: **Evidence of linkage between high-glycine-tyrosine keratin gene loci and wool fibre diameter in a Merino half-sib family.** *Anim Genet* 1994, **25**:105–108.

95. McKenzie GW, Abbott J, Zhou H, Fang Q, Merrick N, Forrest RH, Sedcole JR, Hickford JG: **Genetic diversity of selected genes that are potentially economically important in feral sheep of New Zealand.** *Genet Sel Evol* 2010, **42**:43. doi:10.1186/1297-9686-42-43.

96. Purvis IW, Jeffery N: **Genetics of fibre production in sheep and goats.** *Small Ruminant Res* 2007, **70**:42–47. doi:10.1016/j.smallrumres.2007.01.002.

97. Henikoff S: **Gene families: the taxonomy of protein paralogs and chimeras.** *Science* 1997, **278**:609–614. doi:10.1126/science.278.5338.609.

98. Shibuya K, Kudoh J, Obayashi I, Shimizu A, Sasaki T, Minoshima S, Shimizu N: **Comparative genomics of the keratin-associated protein (KAP) gene clusters in human, chimpanzee, and baboon.** *Mamm Genome* 2004, **15**:179–192. doi:10.1007/s00335-003-2313-9.

99. Wu D-D, Irwin DM, Zhang Y-P: **Correction: molecular evolution of the keratin associated protein gene family in mammals, role in the evolution of mammalian hair.** *BMC Evol Biol* 2009, **9**:213. doi:10.1186/1471-2148-9-213.

100. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool.** *J Mol Biol* 1990, **215**:403–410. doi:10.1016/S0022-2836(05)80360-2.

101. Tatusova TA, Madden TL: **BLAST 2 Sequences, a new tool for comparing protein and nucleotide sequences.** *FEMS Microbiol Lett* 1999, **174**:247–250.

102. Pearson WR, Wood T, Zhang Z, Miller W: **Comparison of DNA sequences with protein sequences.** *Genomics* 1997, **46**:24–36. doi:10.1006/geno.1997.4995.

103. Néron B, Ménager H, Maufrais C, Joly N, Maupetit J, Letort S, Carrere S, Tuffery P, Letondal C: **Mobyle: a new full web bioinformatics framework.** *Bioinformatics* 2009, **25**:3005–3011. doi:10.1093/bioinformatics/btp493.

104. Thompson JD, Higgins DG, Gibson TJ: **CLUSTAL W : improving the sensitivity of progressive multiple ; sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice.** *Nucleic Acids Res* 1994, **22**:4673–4680.

105. Tamura K, Dudley J, Nei M, Kumar S: **MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0.** *Mol Biol Evol* 2007, **24**:1596–1599.

106. Saitou N, Nei M: **The neighbor-joining method: a new method for reconstructing phylogenetic trees.** *Mol Biol Evol* 1987, **4**:406–425.

107. Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S: **MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods.** *Mol Biol Evol* 2011, **28**:2731–2739. doi:10.1093/molbev/msr121.