Molecular Phylogenetics and Evolution 133 (2019) 352–361

Contents lists available at ScienceDirect

Molecular Phylogenetics and Evolution

journal homepage: www.elsevier.com/locate/ympev

# Contrasting patterns of coding and flanking region evolution in mammalian keratin associated protein-1 genes

Huitong Zhou[a,b,1,*], Tina Visnovska[c,1], Hua Gong[a,b], Sebastian Schmeier[c], Jon Hickford[a,b], Austen R.D. Ganley[c,d,*]

[a] International Wool Research Institute, Faculty of Animal Science and Technology, Gansu Agricultural University, Lanzhou 730070, China
[b] Faculty of Agricultural and Life Sciences, Lincoln University, Lincoln 7647, New Zealand
[c] Institute of Natural and Mathematical Sciences, Massey University Auckland, Auckland 0632, New Zealand
[d] School of Biological Sciences, University of Auckland, Auckland 1142, New Zealand

ARTICLE INFO

*Keywords:*
Concerted evolution
Gene conversion
Keratin associated protein
Krtap1
Tandem repeat
Recombination

ABSTRACT

Mammalian genomes contain a number of duplicated genes, and sequence identity between these duplicates can be maintained by purifying selection. However, between-duplicate recombination can also maintain sequence identity between copies, resulting in a pattern known as concerted evolution where within-genome repeats are more similar to each other than to orthologous repeats in related species. Here we investigated the tandemly-repeated keratin-associated protein 1 (KAP1) gene family, *KRTAP1*, which encodes proteins that are important components of hair and wool in mammals. Comparison of eutherian mammal *KRTAP1* gene repeats within and between species shows a strong pattern of concerted evolution. However, in striking contrast to the coding regions of these genes, we find that the flanking regions have a divergent pattern of evolution. This contrast in evolutionary pattern transitions abruptly near the start and stop codons of the *KRTAP1* genes. We reveal that this difference in evolutionary patterns is not explained by conventional purifying selection, nor is it likely a consequence of codon adaptation or reverse transcription of *KRTAP1-n* mRNA. Instead, the evidence suggests that these contrasting patterns result from short-tract gene conversion events that are biased to the *KRTAP1* coding region by selection and/or differential sequence divergence. This work demonstrates the power that gene conversion has to finely shape the evolution of repetitive genes, and provides another distinctive pattern of contrasting evolutionary outcomes that results from gene conversion. A greater emphasis on exploring the evolution of multi-gene eukaryotic families will reveal how common different contrasting evolutionary patterns are in gene duplicates.

## 1. Introduction

Most eukaryote genomes contain repetitive DNA sequences (Britten and Kohne, 1968; Lopez-Flores and Garrido-Ramos, 2012; Richard et al., 2008). There are two basic repeat DNA types: tandem repeats that are typically arranged in head-to-tail arrays, and dispersed repeats, both of which can include either coding or non-coding DNA. Repeats are thought to arise from recombination-based duplication/amplification events (Stephan, 1989), and sequence identity between duplicates will then decay through the diversifying force of mutation, unless counteracting processes operate (Brown et al., 1972; Dover, 1982). Two main paradigms have been proposed to account for the long-term maintenance of identity between repeat copies: concerted evolution and 'birth-and-death' evolution.

Concerted evolution describes a pattern of evolution where the repeats within a genome show greater sequence identity to each other than to orthologous repeats in related genomes (Elder and Turner, 1995). It is thought that this pattern results from recombination-based processes such as gene conversion and unequal crossing-over, which replace DNA sequence from one repeat with that from another repeat (Liao, 1999). In so doing, these recombination processes maintain sequence identity between repeat copies in the face of mutation, and thus they homogenize the repeats (Dover, 1982). Birth-and-death evolution (Nei et al., 1997, 2000) involves purifying selection maintaining

* Corresponding authors at: Faculty of Agriculture and Life Sciences, Lincoln University, Cnr Springs Road & Ellesmere Junction Road, Lincoln 7647, New Zealand (H. Zhou). School of Biological Sciences, University of Auckland, 3A Symonds St, Building 110N, Auckland 1142, New Zealand (A. Ganley).
*E-mail addresses:* zhouh@lincoln.ac.nz (H. Zhou), a.ganley@auckland.ac.nz (A.R.D. Ganley).
[1] Huitong Zhou and Tina Visnovska should be considered joint first authors.

https://doi.org/10.1016/j.ympev.2018.12.031
Received 9 October 2018; Received in revised form 15 December 2018; Accepted 26 December 2018
Available online 30 December 2018

sequence identity between repeats that are generated by occasional duplication events (i.e. birth) and lost through repeat deletion or pseudogenization (i.e. death). While there has been debate as to which of these patterns best fits the evolutionary dynamics of repetitive DNA families (Eirin-Lopez et al., 2012; Nei and Rooney, 2005; Rooney and Ward, 2005), a basic characterization of the evolutionary dynamics for many repeat families is lacking.

The keratin-associated proteins (KAPs) are a diverse group of proteins that are rich in either sulphur, or glycine and tyrosine. They are important structural components of hair and wool fibres, and form a matrix that cross-links the keratin intermediate filaments. The genes encoding the KAPs are called *KRTAP*s (Gong et al., 2012), and these can be classified into 27 families which each comprise 1-12 copies that are usually tandemly arranged (Gong et al., 2016; Rogers et al., 2006; Rogers and Schweizer, 2005). The *KRTAP*s are intron-less genes, with small coding sequences of less than 1 kb in length) (Rogers and Schweizer, 2005; Gong et al., 2016; Stein, 2004; Torrents et al., 2003). In addition, the *KRTAP*s show high levels of population variation, with all known *KRTAP* genes being polymorphic in sheep (Gong et al., 2016, 2010b; Zhou et al., 2016), where they are well studied because of their roles in determining wool phenotypes (Li et al., 2017a, 2017b, 2017c; Tao et al., 2017a, 2017b; Zhou et al., 2015). Despite this variation, it has been reported that some *KRTAP* genes show a pattern of concerted evolution between the paralogous gene copies (Khan et al., 2014; Rogers et al., 1994; Wu et al., 2008).

The KAP1 proteins form the best characterised KAP family, and the *KRTAP1* genes show a high degree of sequence heterogeneity compared to other KAP family members. KAP1 proteins appear to be restricted in expression to the middle to upper cortex region of the hair and wool follicle, and are absent in the cuticle (Powell and Rogers, 1997; Shimomura et al., 2002). However, their precise role in wool and hair fibres has yet to be determined. The genes encoding the KAP1 proteins (*KRTAP1-n*) have been characterized in a number of mammalian species, where they are usually arranged as four tandem repeat copies (Fig. 1) (Khan et al., 2014). The coding regions of the *KRTAP1-n* genes vary in length within species, predominantly as a consequence of variation in the number of imperfect tandem decapeptide repeat units (Gong et al., 2016) (Fig. 1).

The goal of this study was to determine the evolutionary dynamics of the *KRTAP1* gene repeats, including to what extent this repeat family has been shaped by concerted evolution and/or birth-and-death evolution. To achieve this, we analysed the phylogenetic relationships of *KRTAP1* genes from a ten mammalian species, including four species for which the *KRTAP1-n* loci have not been described. We reveal that the *KRTAP1-n* coding regions display a pattern of concerted evolution, but in stark contrast to the coding regions, we find that the flanking regions of these genes display no evidence of concerted evolution, and instead appear to be evolving by divergent evolutionary processes. We show that this pattern of coding region-restricted concerted evolution is unlikely to result from purifying selection, codon adaptation, or the reverse transcription/reintegration of *KRTAP1-n* mRNA sequences. Instead, the concerted evolution pattern is best explained by infrequent short-tract gene conversion events that precise pattern the *KRTAP1-n* paralogs at the nucleotide level through a bias to the coding regions that results from the effects of selection against intra-genomic divergence, and/or differing levels of sequence divergence.

## 2. Materials and methods

### 2.1. Sequence resources and gene identification

Species were chosen based on prior identification of the *KRTAP1* genes, public availability of genome sequences, representation across the eutherian mammal phylogeny, and their anthropological, ecological, experimental, and/or commercial interest. All genome sequences were sourced from the NCBI GenBank. Previously identified *KRTAP1-n* sequences (Gong et al., 2010a, 2011; Itenge-Mweza et al., 2007; Wu et al., 2008) were used to search the genomes of cattle, horse, rabbit and African elephant using BLAST, and *KRTAP1* genes and flanking regions were identified based on a low e-value and alignment across the input sequences, with four *KRTAP1* genes that fulfil these criteria being found in each genome (Table S1).

### 2.2. Sequence alignments

*KRTAP1* nucleotide sequences (Table S1**)** for all four paralogs (*KRTAP1-1* to *KRTAP1-4*) from the ten species (sheep, cattle, dog, elephant, horse, human, macaque, mouse, rat and rabbit) were separated into 5′ flanking, coding, and 3′ flanking regions. The multiple sequence alignment tool *mafft* (v7.123b; (Katoh and Standley, 2013) was used to independently align the 5′ and 3′ flanking regions as nucleotide sequences, using the arguments '–nuc –localpair –maxiterate 1000'. To align the coding sequences at the predicted amino acid level, *mafft* with the arguments '–amino –localpair –maxiterate 1000' was used.

The coding sequence alignment was subsequently reverse translated using *revTrans* (v1.4; Wernersson and Pedersen, 2003) with two input files: the sequences of all the coding regions, and the amino acid sequence alignments. The sequences in the two files were paired by name using the '-match name' parameter, and default values were used for all other parameters. A number of regions aligned poorly and contained indels, therefore the longest continuous coding sequence block (198 nucleotides; covers on average around 40% of the coding region) where none of the 40 sequences had indels, were aligned. For the flanking region alignments, we used *Gblocks* (v0.91b; Talavera and Castresana, 2007) to select blocks that cover approximately 40% of the flanking regions having the best alignment (alignment lengths of 473 bp and 386 bp for the 5′ and 3′ flanking regions, respectively). We also used *Gblocks* with less stringent criteria to create multiple sequence alignments of the coding and flanking regions that included more poorly aligning regions (alignment lengths of 759 bp, 455 bp, and 709 bp for the 5′ flanking, coding, and 3′ flanking regions, respectively). To generate the synonymous and non-synonymous alignments, a custom script was used to divide the 198 bp long conserved coding region alignment into two alignments, one consisting of the third codon positions and the other of the first and second codon positions.

### 2.3. Phylogenetic trees

*PhyML* (v3.1; Guindon et al., 2010) was used to construct phylogenies based on the coding and flanking region sequences. The number of resampled bootstrap data sets was set to 1000 (parameter '-b 1000'), and the additional arguments '-q -s BEST -o tlr' were employed. The

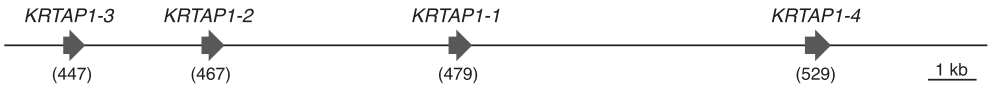


**Fig. 1. Tandem repeat organization of the keratin associated protein-1 (*KRTAP1*) genes.** The general organization of mammalian *KRTAP1* genes is illustrated by the arrangement found in sheep (*Ovis aries*), on chromosome 11. The four *KRTAP1-n* paralogs are represented by arrows that indicate the direction of transcription. Diagram is drawn to scale, with *KRTAP1-n* nucleotide lengths bracketed below the genes. The respective repeats are numbered *KRTAP1-1, 3, 4*, and *5* in human.

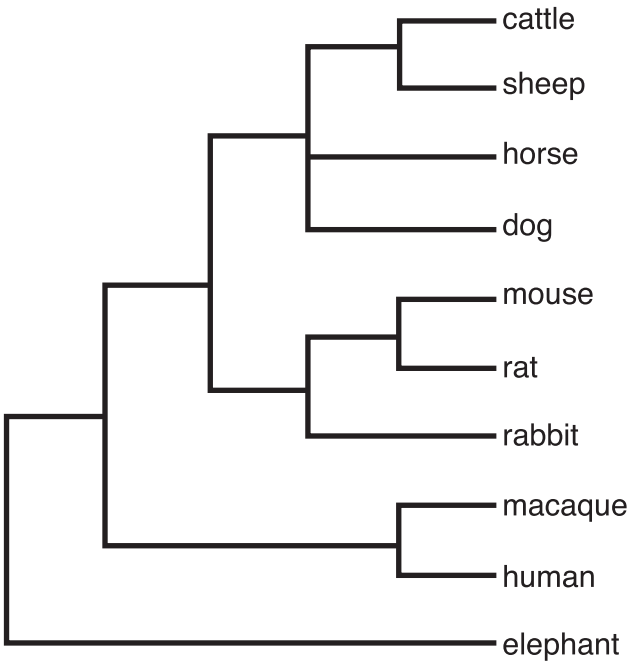


**Fig. 2. Eutherian mammal phylogeny.** Representative phylogeny illustrating the relationships between the species used in this study. Branch lengths are not to scale. Phylogeny is adapted from those presented in McCormack et al. (2012) and Esselstyn et al. (2017), with members of the Laurasiatheria shown as a polytomy.

Bioconductor package *ggtree* (v1.9.4; Yu et al., 2017) was used to plot the phylogenies.

### 2.4. Concerted versus divergent evolution metric

Nucleotide alignments of the last ~54 bp of the 5′ flanking region and the first ~54 bp of the coding region, as well the last ~56 bp of the coding region and the first ~48 bp of the 3′ flanking region were constructed using *mafft* for the *KRTAP1-1* gene across all species from this study, and for all *KRTAP1* genes from sheep. Each position was manually scored for whether the alignment better supported concerted evolution or divergent evolution by determining whether the *KRTAP1-1* alignment or the sheep *KRTAP1* alignment had the highest proportion of sequences with the same base. To allow comparison of each position between the gene and sheep alignments, only the positions in sheep *KRTAP1-1* were scored. The scoring system was as follows:

For each position,

most frequent base proportion in sheep *KRTAP1* alignment > most frequent base proportion in *KRTAP1-1* alignment = 1
most frequent base proportion in sheep *KRTAP1* alignment < most frequent base proportion in *KRTAP1-1* alignment = −1
most frequent base proportion in sheep *KRTAP1* alignment = most frequent base proportion in *KRTAP1-1* alignment = 0

Culmulative sums of these scores were then plotted for the 5′ flanking/coding region and the coding/3′ flanking region alignments using Prism (v.7.0b, Graphpad Software).

### 2.5. Codon adaptation index

The CAIcal server (http://genomes.urv.es/CAIcal; Puigbo et al., 2008) was used to calculate CAI values for the *KRTAP1*s, as well as expected CAI values from permutated sequences using default parameters and published codon usage data (Nakamura et al., 2000).

### 2.6. Motifs in the coding sequences

MEME motif finder (v4.12.0; Bailey et al., 2006) was used to explore repetitive elements in the coding sequences. The repetitive structure of the coding regions was obtained with parameters '-dna -oc . -nostatus -time 18,000 -maxsize 60,000 -mod anr -nmotifs 6 -minw 6 -maxw 30 -minsites 20 -maxsites 600 –revcomp' and all other parameters set to default values.

### 2.7. KRTAP1-n polymorphism in sheep

Intra-specific variation was assessed using three sequences for *KRTAP1-1* (Itenge-Mweza et al., 2007), eleven sequences for *KRTAP1-2* (Gong et al., 2015; Gong et al., 2011), nine sequences for *KRTAP1-3* (Itenge-Mweza et al., 2007), and nine sequences for *KRTAP1-4* (Gong et al., 2010a). These were aligned using DNAMAN (v5.2.10; Lynnon BioSoft, Canada), and the polymorphic sites identified.

### 2.8. Data availability

Sequence data are available at GenBank, with accession numbers and positions listed in Table S1. Sequence data, alignments and phylogenetic tree data together with scripts that can be used to reproduce the results are available as a repository on GitLab (https://gitlab.com/tina_visnovska/concerted_kap1) with a Zenodo DOI assigned (https://zenodo.org/record/1445772).

## 3. Results

### 3.1. Mammalian KRTAP1-n repeats show a concerted pattern of evolution in the coding but not flanking regions

To better understand the genetic architecture of the eutherian mammal *KRTAP1* cluster, we selected the *KRTAP1* genomic region from key members of the eutherian mammal phylogeny for analysis. *KRTAP1* clusters from the genomes of four species (cattle, horses, rabbits and African elephants) for whom *KRTAP1-n* sequence information was not reported (Fig. S1) were identified by querying GenBank with known *KRTAP1-n* sequences using BLAST. We then combined these with previously-identified *KRTAP1-n* sequences from other mammalian species to provide sampling across the mammalian phylogeny (Fig. 2).

Previously, the *KRTAP1* genes of sheep were shown to contain a variable number of occurrences of a QTSCCQPXXX decapeptide tandem repeat in the N-terminal region of the protein (Gong et al., 2016, 2011; Rogers et al., 1994). We used the motif finding tool MEME to search for repetitive motifs in the coding regions of all the mammalian *KRTAP1-n* sequences. This revealed that the decapeptide repeat is present at the N-terminus in all the mammalian *KRTAP1-n* genes studied (Fig. S2). MEME also identified tandem copies of this repeat at the C-terminus of the protein in the nucleotide sequences. Both these N- and C-terminal repeats vary in copy number within and between genomes, and this copy number variation is responsible for much of the *KRTAP1-n* length variation.

To determine the genetic relationships between the mammalian *KRTAP1-n* genes, we generated a *KRTAP1* phylogenetic tree from an alignment of our mammalian *KRTAP1-n* coding region sequences. This revealed that, in most cases, the *KRTAP1* genes are more related to each other within a species than to their orthologs in other species: i.e. they exhibit a concerted evolution pattern. This manifests as clades that group by species, rather than by paralog, in the phylogenetic tree (Fig. 3). This concerted evolution pattern breaks down between the most closely-related species pairs (cattle/sheep, rat/mouse, human/macaque), presumably because the signal is confounded by these species having more recent shared ancestry.

For concertedly evolving tandem repeat sequences such as the ribosomal RNA gene repeats, homogenization occurs for the complete

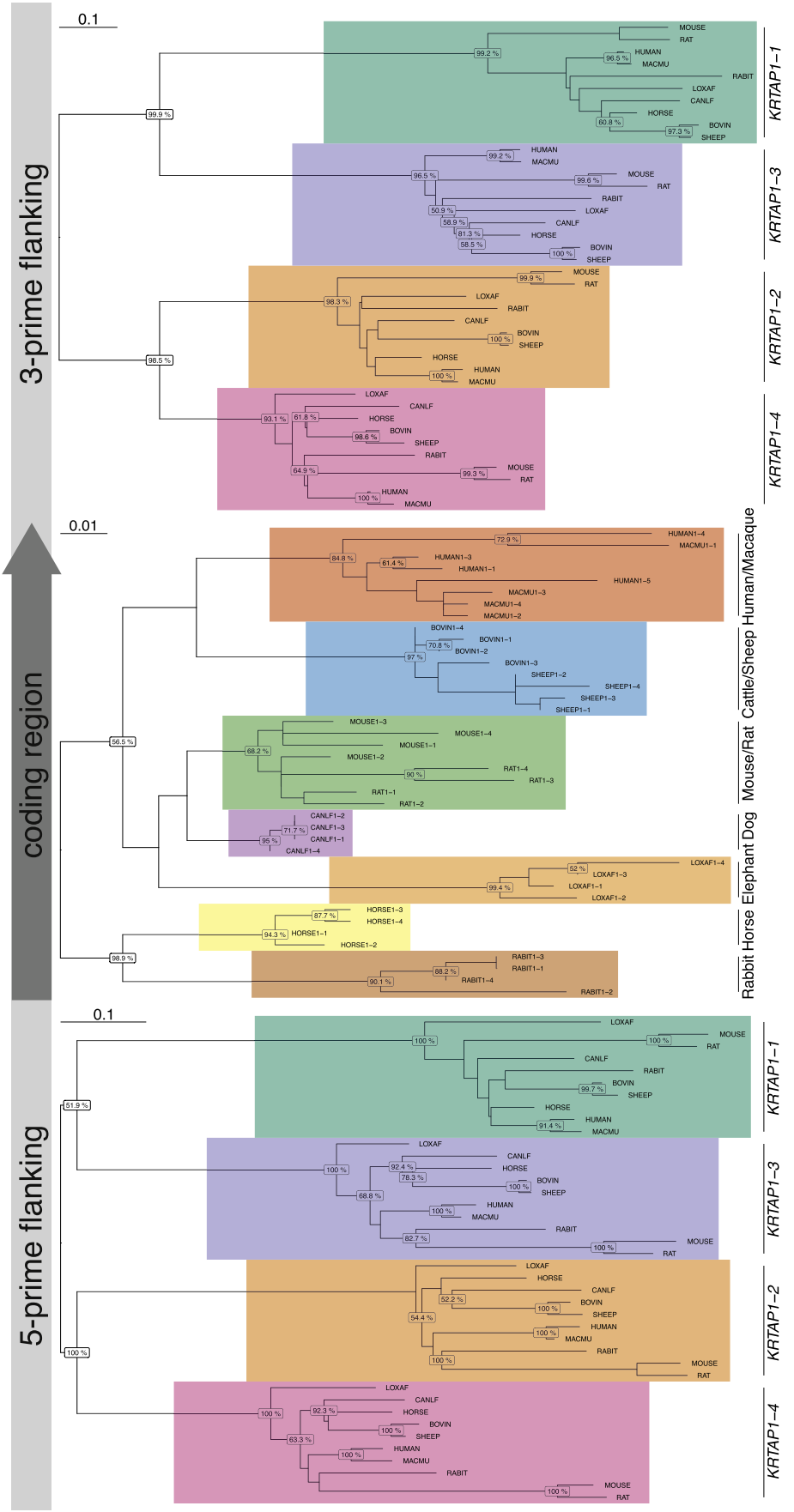


**Fig. 3. Phylogenetic trees of *KRTAP1-n* coding and flanking region sequences.** Phylogenetic trees were constructed for the mammalian *KRTAP1-n* 5′ flanking, coding, and 3′ flanking regions using PhyML. Species are indicated by Uniprot names, and numbers following this for the coding regions indicate *KRTAP1-n* gene name. Major clades in the trees are indicated by colored boxes. The 5′ and 3′ flanking region phylogenies group by repeat number, while the coding region phylogeny tends to group by species. Numbers on nodes indicate bootstrap supports over 50%, and substitution rates are indicated at top left. Human *KRTAP1-n* gene names have been altered for consistency with other species.

repeat unit, including the non-coding regions (Ganley and Kobayashi, 2007), as recombination does not mechanistically distinguish different parts of repeats. To test whether the *KRTAP1* clusters display a 'whole-unit' pattern of concerted evolution, we generated *KRTAP1* phylogenetic trees from multiple alignments of the 5′ and 3′ flanking sequences of the mammalian *KRTAP1* genes. Surprisingly, the phylogenies derived from these flanking sequences did not show any pattern of concerted evolution and, in clear contrast to the coding region phylogeny, the clades in these phylogenetic trees were grouped by *KRTAP1* repeat number, not by species (Fig. 3). We note that bootstrap support is not strong for all the clades in these phylogenetic trees, but the contrast between the coding region and flanking region evolutionary patterns is unmistakable. Furthermore, the topologies of the *KRTAP1-4* 3′ flanking region and *KRTAP1-2, 3*, and *4* 5′ flanking region are largely consistent with the reported mammalian phylogeny (refer to Figs. 2 and 3). The phylogenies were generated from multiple sequence alignments that encompass the *KRTAP1* regions that align well, but phylogenies derived from sequence alignments that include poorly aligned regions give qualitatively similar results (Fig. S3). Overall, in stark contrast to the coding region, the flanking regions show a phylogenetic pattern expected for normal divergent evolution, and they exhibit no evidence of concerted evolution.

### 3.2. What is responsible for the concerted evolution pattern of the KRTAP1 coding region?

We next looked for where the transition point between the concerted and divergent evolutionary patterns appears in the *KRTAP1* sequences. Inspection of the 5′ and 3′ flanking regions revealed that sequence similarity between *KRTAP1-n* sequences within a genome tends to decay around the ATG and stop codons (Fig. 4). We developed a metric that scores whether each nucleotide position is more consistent with a concerted or a divergent evolution pattern (or neither). Cumulative plots of these scores show that at the 5′ end of the gene there is clear transition from a divergent to a concerted evolution pattern about 20 bp upstream of the start codon, and the opposite transition about 15 bp upstream of the stop codon (Fig. 4). This switch in evolutionary patterns close to the boundaries of the coding region suggests that the mechanism responsible for the concerted evolution pattern is able to distinguish the coding region from the flanking region. We therefore sought to identify what mechanism(s) is responsible for the *KRTAP1* coding region-specific concerted evolution pattern.

#### 3.2.1. Purifying selection

The most obvious candidate mechanism is purifying selection, as it is expected to operate more strongly on coding regions, and previous studies have shown high levels of identity within the coding regions of multi-gene loci undergoing birth-and-death evolution due to strong purifying selection (Nei et al., 2000; Piontkivska et al., 2002). A concerted evolution pattern could be generated by purifying selection if sequence identity is maintained between *KRTAP1-n* copies within a species, and diversifying selection drives differences between species. Indeed, our phylogenies clearly show that purifying selection is acting on the *KRTAP1* coding regions, as the level of divergence between coding regions is much lower than that between the flanking regions (Fig. 3). If purifying selection is responsible for the concerted evolution pattern, non-synonymous sites are predicted to show a concerted evolution pattern, while the synonymous sites would instead show a divergent evolution pattern (resembling the flanking regions), as purifying selection is expected to operate primarily on non-synonymous sites.

To investigate this, we looked at the evolutionary patterns of the synonymous and non-synonymous sites in the coding sequences. The number of KAP1 amino acid changes present within and between species makes it difficult to consistently call sites as synonymous or non-synonymous, so we used third codon positions as a proxy for synonymous sites, and first and second codon positions as a proxy for non-synonymous sites. Surprisingly, while the phylogenetic tree generated from *KRTAP1-n* coding region non-synonymous sites displayed a pattern of concerted evolution as expected (Fig. 5A), the tree generated from synonymous sites also revealed the same pattern of concerted evolution (Fig. 5B). Indeed, the concerted evolution pattern for the synonymous sites appears to be stronger than that of the non-synonymous sites, as the synonymous site phylogeny separates sheep and cattle, and dogs, elephants and rat/mouse into separate clades (Fig. 5).

#### 3.2.2. Codon adaptation

Next we considered whether the synonymous site concerted evolution pattern might result from codon adaptation (Lin et al., 2006). This could occur if synonymous mutations that follow changes in the favoured codons between species are beneficial and selected for. The *KRTAP1* genes collectively show a codon adaptation index (CAI; the degree to which the favoured codons for that species are used in a gene) of 0.91 (out of a maximum of 1), higher than the CAI of randomly permutated human *KRTAP1* sequences (0.78). Using the *KRTAP1* coding sequence alignment used for the phylogenies presented in Fig. 3, we identified nine synonymous differences between human and mouse that exhibit a concerted evolution pattern (similarity within species versus difference between species). If codon adaptation can explain this pattern, these synonymous mutations should change in a manner consistent with a change in codon preference for that amino acid. Five of these mutations demonstrate the pattern expected, given the change in codon usage between human and mouse (synonymous change creates the more favoured codon in the species it is found in). However, the other four have the opposite pattern. Furthermore, most of the codon usage preference changes between human and mouse are small in magnitude (Table S2). Thus, there is no evidence that adaptation to different codon usage preferences is driving the pattern of *KRTAP1* concerted evolution.

#### 3.2.3. Reverse transcription of KRTAP1 mRNA

Another mechanism that can distinguish coding and flanking regions and produce a concerted evolution pattern in the coding region is reverse transcription of *KRTAP1* mRNAs, followed by homologous recombination-mediated replacement of a genomic *KRTAP1* with the reverse transcribed copy (Coulombe-Huntington and Majewski, 2007). This is feasible given that the *KRTAP1s* are single-exon genes. If reverse transcription events occur, the 5′ and particularly 3′ flanking regions should show a concerted evolution pattern that is similar to the coding region. However, the transition from concerted to divergent evolution occurs near the start/stop sites of the *KRTAP1* gene, rather than at the 3′/5′ UTRs (Fig. 4), suggesting that reverse transcription/integration of *KRTAP1* mRNA is unlikely to explain the pattern of concerted evolution.

We also considered whether the *KRTAP1* sequences might have arisen through a pure birth-and-death process by independent gene duplication events. However, we think this is improbable as it would require the same number of duplications to occur in at least seven of the species we examined, and, independently, that each of these duplications would not involve any flanking sequence (including promoter and terminator sequences) and would have inserted into the same site in each species. Our results could also formally be explained by purifying selection occurring at the synonymous sites. However, we think this is also unlikely, as it would require the selection to be acting across the coding region specifically in the *KRTAP1* genes but differentially in different species, and it is not clear what the selection would be purifying for.

#### 3.2.4. Gene conversion

Finally, we considered whether gene conversion could explain the *KRTAP1* concerted evolution pattern by examining the *KRTAP1* coding region multiple sequence alignment for gene conversion tracts.

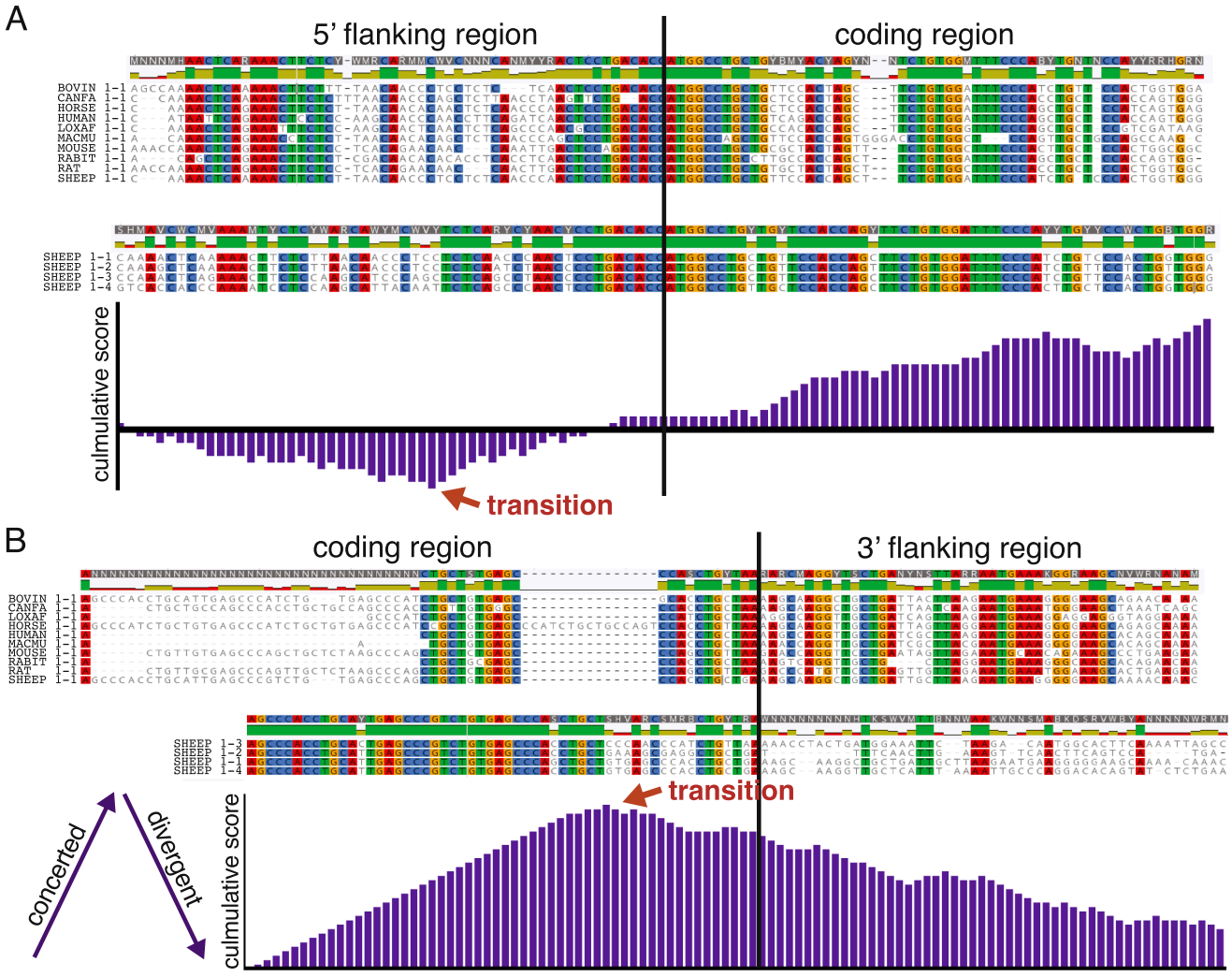


**Fig. 4. Concerted and divergent evolution patterns switch close to the *KRTAP1* start/stop sites.** (A) Alignments of the region flanking the *KRTAP1* gene start site from the *KRTAP1-1* genes and the four *KRTAP1-n* paralogs from sheep (middle). Culmulative plot (bottom) of metric scoring whether each position is more consistent with a concerted evolution (positive score) or divergent evolution (negative score) pattern shows the transition between evolutionary patterns (stationary point; orange arrowhead). Boundary between the 5′ flanking and coding regions is marked by vertical line (followed by the ATG). Positions in the alignments with 90% or greater identity have a colored background, conservation is indicated graphically above each alignment, and consensus sequences are shown above. (B) As in (A), except the region flanking the stop site is shown, with the vertical line marking the boundary between the coding and 3′ flanking regions (preceded by the stop codon). A schematic of the expected cumulative plot slope direction depending on whether the sequence shows a concerted or divergent evolutionary pattern is shown at bottom left. The central part of the coding region is not shown. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Consistent with short-tract gene conversion events, we found several instances of clusters of mutations that are shared between copies within a species, but differ between species (Fig. 6). These patches of homogeneity are small, but within the expected range (3 bp to 1.5 kb) for mammalian gene conversion events (Chen et al., 2007). In addition, we analyzed published population polymorphism data for *KRTAP1* sequences in sheep, as comprehensive sequence variation data are scarce for other species. For many of the polymorphic sites, the polymorphism is shared across some or all of the *KRTAP1* sequences (Fig. 7). While we cannot rule out independent mutation events in each *KRTAP1* copy, we think that gene conversion is a more parsimonious explanation for this observation, particularly for the polymorphisms at synonymous sites. Collectively, our results suggest that the unusual evolutionary pattern of the *KRTAP1* repeats, where the coding region evolutionary dynamics are uncoupled from those of the flanking regions, is the result of coding region-specific short-tract gene conversion events.

## 4. Discussion

Here we have characterized the evolution of the *KRTAP1* gene repeats. The four *KRTAP1-n* copies display a strong pattern of concerted evolution in the coding regions, yet the regions flanking show a normal divergent pattern of evolution. The boundary between these two patterns of evolution is near the start/stop codons of the genes, but our evidence suggests that this dichotomous pattern of evolution is not the result of purifying selection simply acting to reject changes to the amino acid sequence. Instead, we provide evidence that short gene conversion tracts periodically homogenize sequences between the four *KRTAP1* repeats within a genome. The role of gene conversion is supported by two key pieces of evidence: (1) unique amino acid tracts that are shared by KAP1 copies within a species but are unique to that species/group of related species; and (2) the possession of shared nucleotide variants between *KRTAP1* gene copies in sheep populations.

Gene conversion is well-characterized as a mechanism of

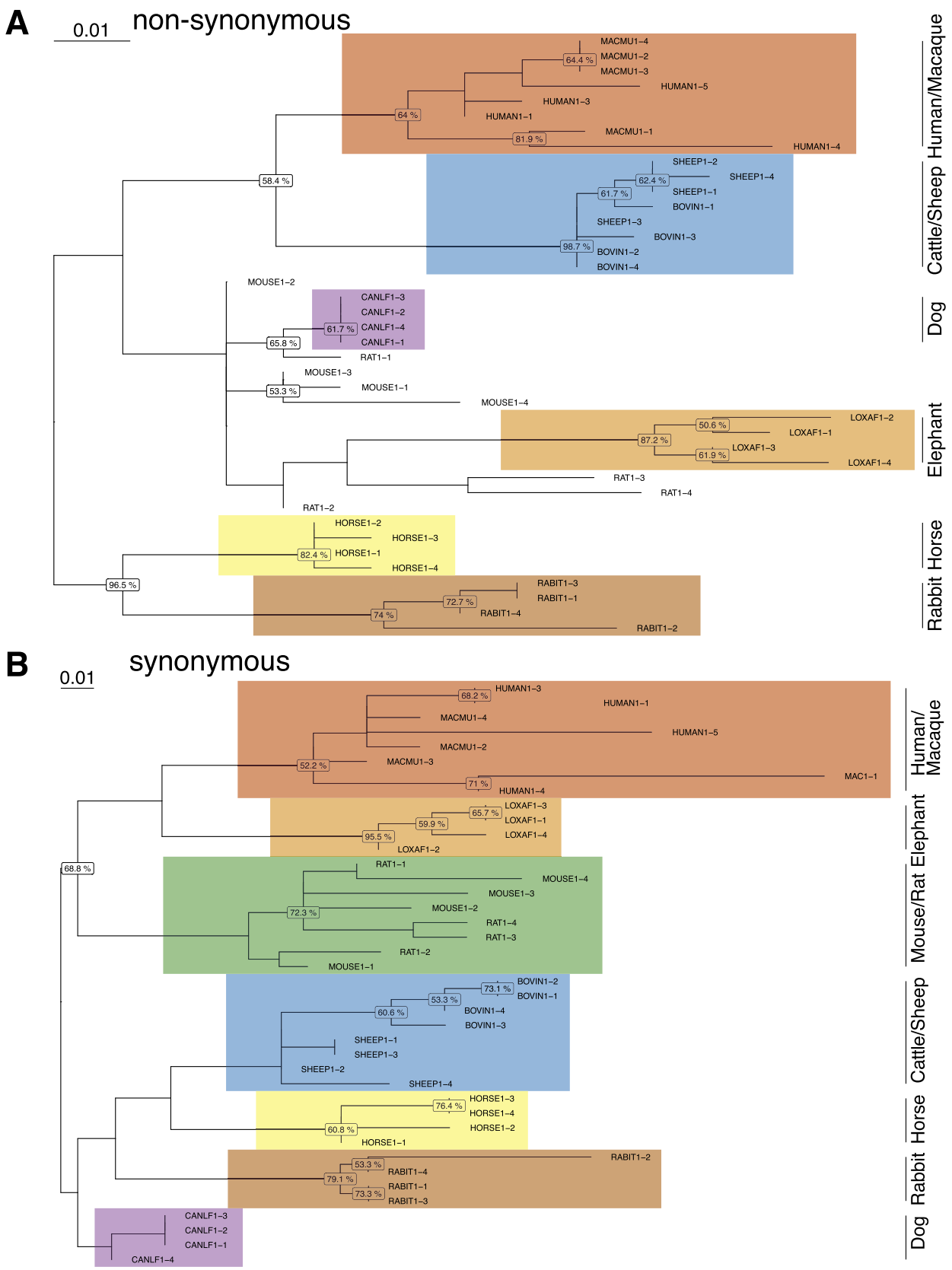


**Fig. 5. *KRTAP1-n* concerted evolution pattern is not explained by purifying selection.** Phylogenetic trees were constructed for the 1st and 2nd codon positions ("non-synonymous"; A), and the 3rd codon position ("synonymous"; B), as per Fig. 3. The major clades (colored boxes) in both phylogenies tend to group by species, with this concerted evolution pattern being stronger in the synonymous phylogeny. Numbers on nodes indicate bootstrap supports with values over 50%, and substitution rates are indicated at top left.

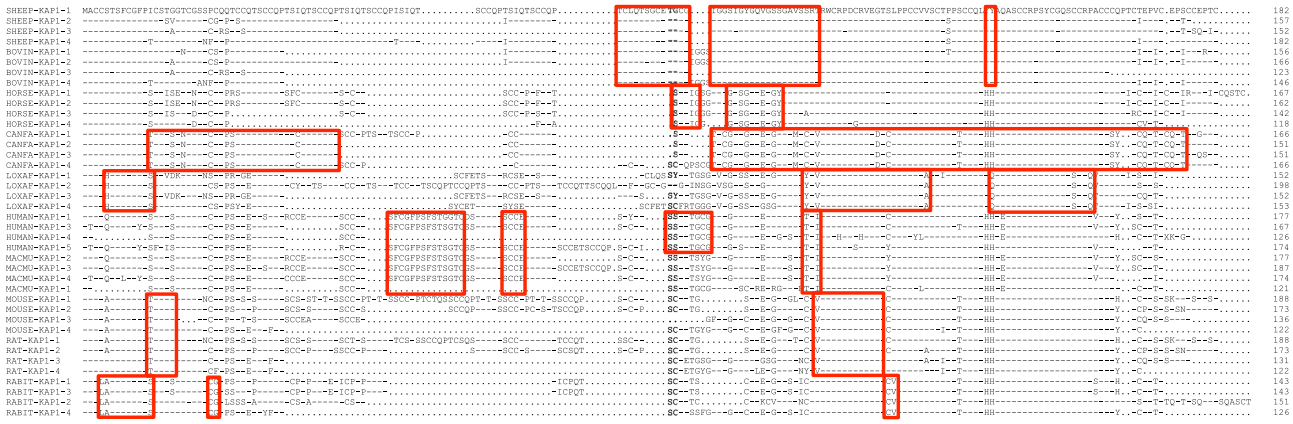

**Fig. 6. Evidence for short gene conversion tracts between *KRTAP1-n* sequences within species.** Alignment of KAP1 amino acid sequences from the ten mammalian species. Amino acid tracts boxed in red represent sequences unique to a species or related species pairs. Dots represent gaps in the alignment, dashes residues identical to the top sequence. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

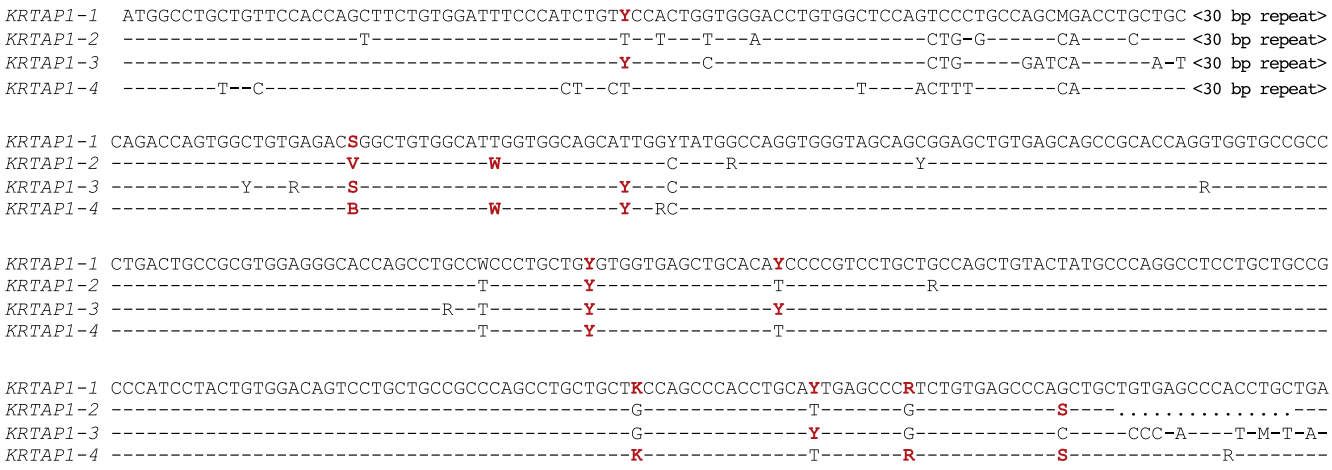

**Fig. 7. Shared polymorphisms between *KRTAP1-n* sequences in sheep.** Alignment of the four sheep *KRTAP1-n* coding region sequences. Dashes represent nucleotides identical to the top sequence, and dots represent gaps. The 30 bp repeats are not shown, as the insertion/deletion positions cannot be precisely determined. Shared nucleotide substitutions between repeat copies are highlighted in red. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

homogenization across all three domains of life (Liao, 1999; Nei and Rooney, 2005; Wang and Chen, 2018), but the striking aspect of our results is the transition from a concerted to a divergent evolutionary pattern that occurs near the coding and flanking region boundaries. What can account for gene conversion events only manifesting in the *KRTAP1* repeat coding regions? We think there are two major possible explanations. The first is a selective argument. Here, gene conversion acts as an unusual form of purifying selection (Innan and Kondrashov, 2010) that prevents accumulation of too much divergence between *KRTAP1* gene copies. Homogeneity of the *KRTAP1* coding sequences may be beneficial by enabling the production of more homogenous components of the hair and wool fibre matrix, thus facilitating better associations with the keratin intermediate filaments. In addition, it is possible that the different *KRTAP1* copies have differential expression that is mediated by copy-specific differences in the flanking regions. If so, gene conversion events in the coding regions would be selected as they remove deleterious heterogeneity, while those in the regulatory flanking regions may perturb differential regulation and thus be selected against. Some evidence for differential regulation of *KRTAP1* gene expression has been found (Chang et al., 2014; Fan et al., 2013). The ability of gene conversion to increase the effective population size of a repeat family (Mano and Innan, 2008) may facilitate more efficient selection against deleterious *KRTAP1* mutations and for the spread of advantageous mutations between *KRTAP1* copies (Dover, 1982). Therefore, selective pressure for coding region homogeneity and possibly regulatory region diversity, coupled with ongoing gene conversion, may be a powerful way to achieve the dichotomy in evolutionary patterns we observe.

The second explanation for the transition in gene conversion between the coding and flanking regions does not invoke selection. Instead, this transition may simply be a consequence of the requirement that gene conversion has for high sequence identity (Chen et al., 2007). Under this explanation, gene conversion promotes sequence homogeneity, which in turn increases the probability that another gene conversion event will occur in that region (Ezawa et al., 2010; Wang et al., 2007). Conversely, regions that diverge in sequence are less likely to undergo gene conversion events. Thus, in a scenario where gene conversion events occur infrequently, the events will be biased to regions undergoing purifying selection, such as the coding region. If true, this explanation suggests that concerted evolution patterns between duplicates should be widely observed. However, the extent to which duplicates undergo concerted evolution is controversial (Casola et al., 2012; Gao and Innan, 2004; Harpak et al., 2017), and even examples where recurrent gene conversion events are detected do not always show a concerted evolution pattern (Petronella and Drouin, 2011, 2014). The *KRTAP1-n* repeats may be exceptions because of their tandem arrangement, with the resulting proximity between copies facilitating unequal alignment and thus inter-repeat gene conversion

during recombination-based DNA damage repair (Ezawa et al., 2010). However, this does not explain examples where tandemly repeated paralogs do not show a strong concerted evolution pattern (Nei et al., 2000; Perina et al., 2011). Instead, a recombination hotspot in the *KRTAP1* genes that drives gene conversion at higher than normal levels could help account for our observations. The presence of a recombination hotspot is supported by *KRTAP1* decapeptide repeat copy number variation, as this may result from unequal recombination between the decapeptide repeats (Liao and Weiner, 1995; Morrill et al., 2016). Such a recombination hotspot could drive both decapeptide repeat copy number variation and gene conversion between *KRTAP1* copies in regions with high sequence similarity, similar to what has been proposed for the ribosomal RNA gene repeats (Ganley and Scott, 1998).

Our results complete a diverse set of evolutionary patterns that can be produced by gene conversion: homogenization of the non-coding but not the coding regions, such as is seen in opsin paralogs (Shyue et al., 1994); homogenization of the coding but not the non-coding regions that we have documented here in the *KRTAP1* genes; selective homogenization of certain regions of the gene, such as has been found in the protocadherin genes in vertebrates (Noonan et al., 2004); and homogenization of both coding and non-coding regions equally, such as has been found in the ribosomal RNA gene repeats (Ganley and Kobayashi, 2007). The pattern we observe is opposite to that found for the opsin gene duplicates in primates, where a much stronger signal of gene conversion/concerted evolution in the introns than the exons has been interpreted as selection largely rejecting coding (exon) region gene conversion events (Hiwatashi et al., 2011; Shyue et al., 1994). It has been suggested that strong selection is required to drive divergence between duplicates undergoing gene conversion (Innan, 2003; Lamping et al., 2017; Storz et al., 2007). However, we think that non-selective mechanisms could also explain the level of divergence observed between *KRTAP1* copies despite the strong concerted evolution pattern. Looking at the evolutionary patterns of a wider range of multi-gene families than have been investigated to date may clarify the extent to which selective versus non-selective forces are responsible for shaping the various evolutionary dynamics they display. The increasing availability of high quality eukaryote genome sequences puts us in an excellent position to achieve this, and to determine whether the impact of gene conversion on the *KRTAP1s* is unusual, or highlights a common mechanism to finely scale patterns of homogeneity and divergence between repeat copies over time.

## 5. Declarations of interest

None.

## Acknowledgements

Funding: This work was supported by a Marsden Fund award (14-MAU-053) to ARDG, an AGMARDT Postdoctoral Fellowship to HG, and a Vernon Willey Trust Fellowship to HZ. The funders had no involvement in the study design; the collection, analysis and interpretation of data; the writing of the report; or the decision to submit the article for publication.

## Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ympev.2018.12.031.

## References

Bailey, T.L., Williams, N., Misleh, C., Li, W.W., 2006. MEME: discovering and analyzing DNA and protein sequence motifs. Nucl. Acids Res. 34, W369–W373.
Britten, R.J., Kohne, D.E., 1968. Repeated sequences in DNA. Science 161, 529–540.
Brown, D.D., Wensink, P.C., Jordan, E., 1972. A comparison of the ribosomal DNA's of *Xenopus laevis* and *Xenopus mulleri*: the evolution of tandem genes. J. Mol. Biol. 63, 57–73.
Casola, C., Conant, G.C., Hahn, M.W., 2012. Very low rate of gene conversion in the yeast genome. Mol. Biol. Evol. 29, 3817–3826.
Chang, T.H., Huang, H.D., Ong, W.K., Fu, Y.J., Lee, O.K., Chien, S., Ho, J.H., 2014. The effects of actin cytoskeleton perturbation on keratin intermediate filament formation in mesenchymal stem/stromal cells. Biomaterials 35, 3934–3944.
Chen, J.M., Cooper, D.N., Chuzhanova, N., Ferec, C., Patrinos, G.P., 2007. Gene conversion: mechanisms, evolution and human disease. Nat. Rev. Genet. 8, 762–775.
Coulombe-Huntington, J., Majewski, J., 2007. Characterization of intron loss events in mammals. Genome Res. 17, 23–32.
Dover, G.A., 1982. Molecular drive: a cohesive mode of species evolution. Nature 299, 111–117.
Eirin-Lopez, J.M., Rebordinos, L., Rooney, A.P., Rozas, J., 2012. The birth-and-death evolution of multigene families revisited. Genome Dyn. 7, 170–196.
Elder Jr., J.F., Turner, B.J., 1995. Concerted evolution of repetitive DNA sequences in eukaryotes. Q. Rev. Biol. 70, 297–320.
Esselstyn, J.A., Oliveros, C.H., Swanson, M.T., Faircloth, B.C., 2017. Investigating difficult nodes in the placental mammal tree with expanded taxon sampling and thousands of ultraconserved elements. Genome Biol. Evol. 9, 2308–2321.
Ezawa, K., Ikeo, K., Gojobori, T., Saitou, N., 2010. Evolutionary pattern of gene homogenization between primate-specific paralogs after human and macaque speciation using the 4-2-4 method. Mol. Biol. Evol. 27, 2152–2171.
Fan, R., Xie, J., Bai, J., Wang, H., Tian, X., Bai, R., Jia, X., Yang, L., Song, Y., Herrid, M., Gao, W., He, X., Yao, J., Smith, G.W., Dong, C., 2013. Skin transcriptome profiles associated with coat color in sheep. BMC Genom. 14, 389.
Ganley, A.R.D., Kobayashi, T., 2007. Highly efficient concerted evolution in the ribosomal DNA repeats: total rDNA repeat variation revealed by whole-genome shotgun sequence data. Genome Res. 17, 184–191.
Ganley, A.R.D., Scott, B., 1998. Extraordinary ribosomal spacer length heterogeneity in a Neotyphodium endophyte hybrid: implications for concerted evolution. Genetics 150, 1625–1637.
Gao, L.-Z., Innan, H., 2004. Very low gene duplication rate in the yeast genome. Science 306, 1367–1370.
Gong, H., Zhou, H., Forrest, R.H.J., Li, S., Wang, J., Dyer, J.M., Luo, Y., Hickford, J.G.H., 2016. Wool keratin-associated protein genes in sheep—a review. Genes 7, 24.
Gong, H., Zhou, H., Hickford, J.G.H., 2010a. Polymorphism of the ovine keratin-associated protein 1–4 gene (KRTAP1-4). Mol. Biol. Rep. 37, 3377–3380.
Gong, H., Zhou, H., Hodge, S., Dyer, J.M., Hickford, J.G.H., 2015. Association of wool traits with variation in the ovine KAP1-2 gene in Merino cross lambs. Small Rumin. Res. 124, 24–29.
Gong, H., Zhou, H., McKenzie, G.W., Hickford, J.G., Yu, Z., Clerens, S., Dyer, J.M., Plowman, J.E., 2010b. Emerging issues with the current keratin-associated protein nomenclature. Int. J. Trichol. 2, 104–105.
Gong, H., Zhou, H., McKenzie, G.W., Yu, Z., Clerens, S., Dyer, J.M., Plowman, J.E., Wright, M.W., Arora, R., Bawden, C.S., 2012. An updated nomenclature for keratin-associated proteins (KAPs). Int. J. Biol. Sci. 8, 258–264.
Gong, H., Zhou, H., Yu, Z., Dyer, J., Plowman, J.E., Hickford, J., 2011. Identification of the ovine keratin-associated protein KAP1-2 gene (KRTAP1-2). Exp. Dermatol. 20, 815–819.
Guindon, S., Dufayard, J.F., Lefort, V., Anisimova, M., Hordijk, W., Gascuel, O., 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst. Biol. 59, 307–321.
Harpak, A., Lan, X., Gao, Z., Pritchard, J.K., 2017. Frequent nonallelic gene conversion on the human lineage and its effect on the divergence of gene duplicates. PNAS 114, 12779–12784.
Hiwatashi, T., Mikami, A., Katsumura, T., Suryobroto, B., Perwitasari-Farajallah, D., Malaivijitnond, S., Siriaroonrat, B., Oota, H., Goto, S., Kawamura, S., 2011. Gene conversion and purifying selection shape nucleotide variation in gibbon L/M opsin genes. BMC Evol. Biol. 11, 312.
Innan, H., 2003. A two-locus gene conversion model with selection and its application to the human RHCE and RHD genes. PNAS 100, 8793–8798.
Innan, H., Kondrashov, F., 2010. The evolution of gene duplications: classifying and distinguishing between models. Nat. Rev. Genet. 11, 97–108.
Itenge-Mweza, T.O., Forrest, R.H., McKenzie, G.W., Hogan, A., Abbott, J., Amoafo, O., Hickford, J.G., 2007. Polymorphism of the KAP1.1, KAP1.3 and K33 genes in Merino sheep. Mol. Cell. Probes 21, 338–342.
Katoh, K., Standley, D.M., 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol. Biol. Evol. 30, 772–780.
Khan, I., Maldonado, E., Vasconcelos, V., Stephen, J.O., Johnson, W.E., Antunes, A., 2014. Mammalian keratin associated proteins (KRTAPs) subgenomes: disentangling hair diversity and adaptation to terrestrial and aquatic environments. BMC Genom. 15, 779.
Lamping, E., Zhu, J.Y., Niimi, M., Cannon, R.D., 2017. Role of ectopic gene conversion in the evolution of a *Candida krusei* pleiotropic drug resistance transporter family. Genetics 205, 1619–1639.
Li, S., Zhou, H., Gong, H., Zhao, F., Hu, J., Luo, Y., Hickford, J.G.H., 2017a. Identification of the ovine keratin-associated protein 26–1 gene and its association with variation in wool traits. Genes 8, 225.
Li, S., Zhou, H., Gong, H., Zhao, F., Wang, J., Liu, X., Luo, Y., Hickford, J.G.H., 2017b. Identification of the ovine keratin-associated protein 22–1 (KAP22-1) gene and its effect on wool traits. Genes 8, 27.
Li, S., Zhou, H., Gong, H., Zhao, F., Wang, J., Luo, Y., Hickford, J.G.H., 2017c. Variation in the ovine KAP6-3 gene (KRTAP6-3) is associated with variation in mean fibre diameter-associated wool traits. Genes 8, 204.

Liao, D., 1999. Concerted evolution: molecular mechanism and biological implications. Am. J. Hum. Genet. 64, 24–30.
Liao, D., Weiner, A.M., 1995. Concerted evolution of the tandemly repeated genes encoding primate U2 small nuclear RNA (the RNU2 locus) does not prevent rapid diversification of the (CT)n. (GA)n microsatellite embedded within the U2 repeat unit. Genomics 30, 583–593.
Lin, Y.-S., Byrnes, J.K., Hwang, J.-K., Li, W.-H., 2006. Codon-usage bias versus gene conversions in the evolution of yeast duplicate genes. PNAS 103, 14412–14416.
Lopez-Flores, I., Garrido-Ramos, M.A., 2012. The repetitive DNA content of eukaryotic genomes. Genome Dyn. 7, 1–28.
Mano, S., Innan, H., 2008. The evolutionary rate of duplicated genes under concerted evolution. Genetics 180, 493–505.
McCormack, J.E., Faircloth, B.C., Crawford, N.G., Gowaty, P.A., Brumfield, R.T., Glenn, T.C., 2012. Ultraconserved elements are novel phylogenomic markers that resolve placental mammal phylogeny when combined with species-tree analysis. Genome Res. 22, 746–754.
Morrill, S.A., Exner, A.E., Babokhov, M., Reinfeld, B.I., Fuchs, S.M., 2016. DNA instability maintains the repeat length of the yeast RNA polymerase II C-terminal domain. J. Biol. Chem. 291, 11540–11550.
Nakamura, Y., Gojobori, T., Ikemura, T., 2000. Codon usage tabulated from international DNA sequence databases: status for the year 2000. Nucl. Acids Res. 28, 292.
Nei, M., Gu, X., Sitnikova, T., 1997. Evolution by the birth-and-death process in multigene families of the vertebrate immune system. PNAS 94, 7799–7806.
Nei, M., Rogozin, I.B., Piontkivska, H., 2000. Purifying selection and birth-and-death evolution in the ubiquitin gene family. PNAS 97, 10866–10871.
Nei, M., Rooney, A.P., 2005. Concerted and birth-and-death evolution of multigene families. Annu. Rev. Genet. 39, 121–152.
Noonan, J.P., Grimwood, J., Schmutz, J., Dickson, M., Myers, R.M., 2004. Gene conversion and the evolution of protocadherin gene cluster diversity. Genome Res. 14, 354–366.
Perina, A., Seoane, D., Gonzalez-Tizon, A.M., Rodriguez-Farina, F., Martinez-Lage, A., 2011. Molecular organization and phylogenetic analysis of 5S rDNA in crustaceans of the genus Pollicipes reveal birth-and-death evolution and strong purifying selection. BMC Evol. Biol. 11, 304.
Petronella, N., Drouin, G., 2011. Gene conversions in the growth hormone gene family of primates: stronger homogenizing effects in the Hominidae lineage. Genomics 98, 173–181.
Petronella, N., Drouin, G., 2014. Purifying selection against gene conversions in the folate receptor genes of primates. Genomics 103, 40–47.
Piontkivska, H., Rooney, A.P., Nei, M., 2002. Purifying selection and birth-and-death evolution in the histone H4 gene family. Mol. Biol. Evol. 19, 689–697.
Powell, B.C., Rogers, G.E., 1997. The Role of Keratin Proteins and Their Genes in the Growth, Structure and Properties of Hair. Formation and Structure of Human Hair. Birkhäuser Verlag, pp. 59–148.
Puigbo, P., Bravo, I.G., Garcia-Vallve, S., 2008. CAIcal: a combined set of tools to assess codon usage adaptation. Biol. Direct 3, 38.
Richard, G.F., Kerrest, A., Dujon, B., 2008. Comparative genomics and molecular dynamics of DNA repeats in eukaryotes. Microbiol. Mol. Biol. Rev. 72, 686–727.
Rogers, G.R., Hickford, J.G.H., Bickerstaffe, R., 1994. Polymorphism in two genes for B2 high sulfur proteins of wool. Anim. Genet. 25, 407–415.
Rogers, M.A., Langbein, L., Praetzel-Wunder, S., Winter, H., Schweizer, J., 2006. Human hair keratin-associated proteins (KAPs). Int. Rev. Cytol. 251, 209–263.
Rogers, M.A., Schweizer, J., 2005. Human KAP genes, only the half of it? Extensive size polymorphisms in hair keratin-associated protein genes. J. Invest. Dermatol. 124, vii–ix.
Rooney, A.P., Ward, T.J., 2005. Evolution of a large ribosomal RNA multigene family in filamentous fungi: birth and death of a concerted evolution paradigm. PNAS 102, 5084–5089.
Shimomura, Y., Aoki, N., Schweizer, J., Langbein, L., Rogers, M.A., Winter, H., Ito, M., 2002. Polymorphisms in the human high sulfur hair keratin-associated protein 1, KAP1, gene family. J. Biol. Chem. 277, 45493–45501.
Shyue, S.K., Li, L., Chang, B.H., Li, W.-H., 1994. Intronic gene conversion in the evolution of human X-linked color vision genes. Mol. Biol. Evol. 11, 548–551.
Stein, L.D., 2004. End of the beginning. Nature 431, 915–916.
Stephan, W., 1989. Tandem-repetitive noncoding DNA: forms and forces. Mol. Biol. Evol. 6, 198–212.
Storz, J.F., Baze, M., Waite, J.L., Hoffmann, F.G., Opazo, J.C., Hayes, J.P., 2007. Complex signatures of selection and gene conversion in the duplicated globin genes of house mice. Genetics 177, 481–500.
Talavera, G., Castresana, J., 2007. Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. Syst. Biol. 56, 564–577.
Tao, J., Zhou, H., Gong, H., Yang, Z., Ma, Q., Cheng, L., Ding, W., Li, Y., Hickford, J.G.H., 2017a. Variation in the KAP6-1 gene in Chinese Tan sheep and associations with variation in wool traits. Small Rumin. Res. 154, 129–132.
Tao, J., Zhou, H., Yang, Z., Gong, H., Ma, Q., Ding, W., Li, Y., Hickford, J.G.H., 2017b. Variation in the KAP8-2 gene affects wool crimp and growth in Chinese Tan sheep. Small Rumin. Res. 149, 77–80.
Torrents, D., Suyama, M., Zdobnov, E., Bork, P., 2003. A genome-wide survey of human pseudogenes. Genome Res. 13, 2559–2567.
Wang, S., Chen, Y., 2018. Phylogenomic analysis demonstrates a pattern of rare and long-lasting concerted evolution in prokaryotes. Commun. Biol. 1, 12.
Wang, X.Y., Tang, H.B., Bowers, J.E., Feltus, F.A., Paterson, A.H., 2007. Extensive concerted evolution of rice paralogs and the road to regaining independence. Genetics 177, 1753–1763.
Wernersson, R., Pedersen, A.G., 2003. RevTrans: multiple alignment of coding DNA from aligned amino acid sequences. Nucl. Acids Res. 31, 3537–3539.
Wu, D.D., Irwin, D., Zhang, Y.P., 2008. Molecular evolution of the keratin associated protein gene family in mammals, role in the evolution of mammalian hair. BMC Evol. Biol. 8, 241.
Yu, G., Smith, D.K., Zhu, H., Guan, Y., Lam, T.T.-Y., 2017. GGTREE: an R package for visualization and annotation of phylogenetic trees with their covariates and other associated data. Methods Ecol. Evol. 8, 28–36.
Zhou, H., Gong, H., Li, S., Luo, Y., Hickford, J., 2015. A 57-bp deletion in the ovine KAP6-1 gene affects wool fibre diameter. J. Animal Breed. Genet. 132, 301–307.
Zhou, H., Gong, H., Wang, J., Dyer, J.M., Luo, Y., Hickford, J.G.H., 2016. Identification of four new gene members of the KAP6 gene family in sheep. Sci. Rep. 6, 24074.