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

Huitong Zhou[*,†,1], Tina Visnovska[‡,1], Hua Gong[*,†], Sebastian Schmeier[‡], Jon Hickford[*,†], and Austen R.D. Ganley[‡,§]

* State Key Laboratory of Sheep Genetic Improvement and Heathy Production, Shihezi 832000, China

† Faculty of Agricultural and Life Sciences, Lincoln University, Lincoln 7647, New Zealand

‡ Institute of Natural and Mathematical Sciences, Massey University Auckland, Auckland 0632, New Zealand

§ School of Biological Sciences, University of Auckland, Auckland 1142, New Zealand

1 These authors contributed equally to this work

18 **Running title:** Contrasting *KRTAP1* evolutionary patterns

19

20 **Key words:**

21 concerted evolution, gene conversion, keratin associated protein, krtap1, tandem repeat

22

23 **To whom correspondence should be addressed:**

24 **Austen Ganley**: School of Biological Sciences, University of Auckland, 3A Symonds St,

25 Building 110N, Auckland 1142, New Zealand; +64 9 923 2906; a.ganley@auckland.ac.nz

26

27 **Huitong Zhou**: Faculty of Agriculture and Life Sciences, Lincoln University,

28 Cnr Springs Road & Ellesmere Junction Road, Lincoln 7647, New Zealand; +64 3 423 0684;

29 zhouh@lincoln.ac.nz

30

31

Abstract

DNA repeats are common elements in eukaryotic genomes, and their multi-copy nature provides the opportunity for genetic exchange. This exchange can produce altered evolutionary patterns, including concerted evolution where within genome repeat copies are more similar to each other than to orthologous repeats in related species. Here we investigated the genetic architecture of the keratin-associated protein (KAP) gene family, *KRTAP1*. This family encodes proteins that are important components of hair and wool in mammals, and the genes are present in tandem copies. Comparison of *KRTAP1* gene repeats from species across the mammalian phylogeny shows strongly contrasting evolutionary patterns between the coding regions, which have a concerted evolution pattern, and the flanking regions, which have a normal, radiating pattern of evolution. This dichotomy in evolutionary pattern transitions abruptly at the start and stop codons, and we show it is not the result of purifying selection acting to maintain species-specific protein sequences, nor of codon adaptation or reverse transcription of *KRTAP1-n* mRNA. Instead, the results are consistent with short-tract gene conversion events coupled with selection for these events in the coding region driving the contrasting evolutionary patterns found in the *KRTAP1* repeats. Our work shows the power that repeat recombination has to complement selection and finely tune the sequences of repetitive genes. Interplay between selection and recombination may be a more common mechanism than currently appreciated for achieving specific adaptive outcomes in the many eukaryotic multi-gene families, and our work argues for greater emphasis on exploring the sequence structures of these families.

54 # INTRODUCTION

55 Repetitive DNA is widespread in most eukaryote genomes (BRITTEN AND KOHNE 1968;
56 RICHARD *et al.* 2008; LOPEZ-FLORES AND GARRIDO-RAMOS 2012). There are two basic repeat
57 DNA types: tandem repeats that are typically arranged in head-to-tail arrays; and dispersed
58 repeats, and these can occur in either coding or non-coding DNA. Repeats are thought to
59 arise from recombination-based duplication/amplification events (STEPHAN 1989). Sequence
60 identity between duplicates will then decay through the diversifying force of mutation, unless
61 counteracting processes operate (BROWN *et al.* 1972; DOVER 1982). The balance between
62 duplication, diversification, selection, and counteracting forces thus dictate the evolutionary
63 dynamics of repeats. Two main paradigms have been proposed to account for the long-term
64 maintenance of repeat identity: concerted evolution and birth-and-death evolution. Concerted
65 evolution describes a pattern of evolution where the repeats within a genome show greater
66 sequence identity to each other than to orthologous repeats in related genomes (ELDER AND
67 TURNER 1995). The pattern of concerted evolution is proposed to result from recombination-
68 based processes, such as gene conversion and unequal cross-over events, that replace the
69 DNA sequence from one repeat with that from another repeat (LIAO 1999). In so doing, these
70 recombination processes maintain sequence identity between repeat copies in the face of
71 mutation, and thus homogenize the repeats (DOVER 1982). 'Birth-and-death' evolution
72 involves purifying selection maintaining sequence identity between repeats that are generated
73 by occasional duplication events (i.e. birth), as well as death, which results from repeat loss
74 or pseudogenization (NEI *et al.* 1997; NEI *et al.* 2000). While there has been debate as to
75 which of these processes best describes the evolutionary dynamics of repetitive DNA (NEI
76 AND ROONEY 2005; ROONEY AND WARD 2005; EIRIN-LOPEZ *et al.* 2012), a basic
77 characterization of the evolutionary dynamics of most repeat families is lacking.

The keratin-associated proteins (KAPs) are a diverse group of proteins, and 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 can be classified into 27 families, with each family comprising 1-12 members that are usually tandemly arranged (ROGERS AND SCHWEIZER 2005; ROGERS *et al.* 2006; GONG *et al.* 2016). The *KRTAP*s are single exon (intron-less) genes, with small coding sequences (less than 1 kb) (ROGERS AND SCHWEIZER 2005), and they have low numbers of pseudogenes. For example, in humans the pseudogene:gene *KRTAP* ratio is approximately 1:5 (GONG *et al.* 2016), while across all human genes the ratio is close to 1:1 (TORRENTS *et al.* 2003; STEIN 2004). In addition, the *KRTAP*s show high levels of population variation, with all known *KRTAP* genes being polymorphic in sheep (GONG *et al.* 2010b; GONG *et al.* 2016; ZHOU *et al.* 2016), where they are well studied because of their roles in determining wool phenotypes (ZHOU *et al.* 2015; LI *et al.* 2017a; LI *et al.* 2017b; LI *et al.* 2017c; TAO *et al.* 2017a; TAO *et al.* 2017b). Despite this variation, it has been reported that at least some *KRTAP* genes show a pattern of concerted evolution between the paraglogous gene copies (ROGERS *et al.* 1994; WU *et al.* 2008; KHAN *et al.* 2014).

The KAP1 proteins form the best characterised KAP family, and they show a high degree of sequence heterogeneity compared to other KAP families. These 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). Their precise role in hair and wool function, 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 copies (**Figure 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) (**Figure 1**).

Here we analyse the *KRTAP1* genes from a number of mammalian species, including four species for which the *KRTAP1-n* loci have not been described. Together with the existing *KRTAP1-n* sequences, we reveal that the *KRTAP1-n* coding regions display a pattern of concerted evolution. In stark contrast to the coding region though, we find that the repeat flanking regions display no evidence of concerted evolution, and instead appear to be evolving by normal vertical or radiating evolution. Surprisingly, we find that this pattern of coding region restricted concerted evolution is not the result of purifying selection, nor does it result from codon adaptation or reverse transcription/reintegration of *KRTAP1-n* mRNA sequences. Instead, the results are best explained by a combination of on-going short-tract gene conversion events between the *KRTAP1-n* copies, and negative selection. We argue that these gene conversion events act as an unusual mechanism of purifying selection to prevent excessive intra-genomic divergence between the four gene copies, while also allowing inter-species diversity. This unusual mode of evolution may apply to other multicopy genes that encode products subject to diversifying selection.

MATERIALS AND METHODS

Sequence Resources and Gene Identification: All genome sequences were sourced from the NCBI GenBank. Previously identified *KRTAP1-n* sequences (ITENGE-MWEZA *et al.* 2007; WU *et al.* 2008; GONG *et al.* 2010a; GONG *et al.* 2011) were used to search the genomes of cattle, horses, rabbits and African elephants using BLAST with default parameters, and the

126 genes retrieved were identified by sequence identity within both the coding and flanking
127 regions (**Table S1**).

128 **Sequence Alignments:** *KRTAP1* nucleotide sequences (**Table S1)** for all four paralogs from
129 the ten species (sheep, cattle, dog, elephant, horse, human, macaque, mouse, rat and rabbit)
130 were separated into 5' flanking regions, coding sequences, and 3' flanking regions. The
131 multiple sequence alignment tool *mafft* (v7.123b) (KATOH AND STANDLEY 2013) was used to
132 separately align the 5' and 3' flanking regions as nucleotide sequences, using the arguments '-
133 -nuc --localpair --maxiterate 1000'. To align the coding sequences at the predicted amino acid
134 level, *mafft* with the arguments '--amino --localpair --maxiterate 1000' was run.

135 The coding sequence alignment was subsequently reverse translated using *revTrans* (v1.4)
136 (WERNERSSON AND PEDERSEN 2003) with two input files: the sequences of all the coding
137 regions, and the amino acid sequence alignments. The sequences in the two files were paired
138 by name using the '-match name' parameter, and default values were used for all other
139 parameters. A number of regions align poorly and have many indels, therefore we used the
140 longest continuous coding sequence block (198 nucleotides; covers on average around 40%
141 of the coding region) where none of the 40 sequences had indels. For the flanking region
142 alignments, we used *Gblocks* (v0.91b) (TALAVERA AND CASTRESANA 2007) to select blocks
143 that cover approximately 40% of the flanking regions having the best alignment. We also
144 used *Gblocks* with less stringent criteria to create multiple sequence alignments of the coding
145 and flanking regions that included more poorly aligning regions.

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

were employed. The Bioconductor package *ggtree* (v1.9.4) (YU *et al.* 2017) was used to plot the phylogenies.

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).

Motifs in the Coding Sequences: We used MEME motif finder (v4.12.0) (BAILEY *et al.* 2006) to explore repetitive elements in the coding sequences. The repetitive structure of the coding regions reported in the Results was obtained with parameters '-dna -oc . -nostatus -time 18000 -maxsize 60000 -mod anr -nmotifs 6 -minw 6 -maxw 30 -minsites 20 -maxsites 600 –revcomp' and all the other parameters set to the default values.

***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.* 2011; GONG *et al.* 2015), 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) with default parameters, and polymorphic sites were identified manually.

Data Availability Statement: Sequence data are available at GenBank and the accession numbers and positions are listed in the **Materials and Methods** (sheep polymorphism data) and Table S1 (*KRTAP1* sequences). Descriptions of the supplemental material are given in **File S1,** and the supplemental material are available on figshare.

RESULTS

Mammalian *KRTAP1-n* repeats show a concerted evolution pattern in the coding but not the flanking regions

To better understand the genetic architecture of the mammalian *KRTAP1* cluster, we selected the *KRTAP1* genomic region from key members of the mammalian phylogeny for analysis. The Basic Local Alignment Search Tool (BLAST) was used to search GenBank with known *KRTAP1-n* sequences to identify and retrieve the *KRTAP1* clusters from the genomes of four species (cattle, horses, rabbits and African elephants) for whom *KRTAP1-n* sequence information has not been reported (**Figure S1**). We then combined these with previously-identified *KRTAP1-n* sequences from other mammalian species to obtain sampling across the mammalian phylogeny (**Figure 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 (ROGERS *et al.* 1994; GONG *et al.* 2011; GONG *et al.* 2016). We used a motif finding tool (MEME; (BAILEY *et al.* 2006) 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 mammalian *KRTAP1-n* genes we obtained (**Figure S2)**, albeit with less amino acid conservation than that observed in sheep. MEME also identified nucleotide level tandem copies of this repeat at the C-terminus of the protein. Furthermore, both the N- and C-terminal repeats vary in copy number, within and between genomes. This copy number variation is responsible for much of the length variation between *KRTAP1-n* sequences.

To determine the genetic relationships between of 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

196 related to each other within a species than to their orthologs in other species, thus exhibiting a
197 concerted evolution pattern. This manifests as clades that group by species, rather than by
198 repeat, in the phylogenetic tree (**Figure 3**). This concerted evolution pattern breaks down
199 between the most closely-related species pairs (cattle/sheep, rat/mouse, human/macaque),
200 presumably because the signal is confounded by these species having more recent shared
201 ancestry. Nevertheless, for most species there is a clear pattern of concerted evolution.

202 For concertedly evolving tandem repeat sequences such as the ribosomal RNA gene repeats,
203 homogenization occurs for the complete repeat unit, including the non-coding regions
204 (GANLEY AND KOBAYASHI 2007). To test whether the *KRTAP1* clusters display a 'whole-
205 unit' pattern of concerted evolution, we generated *KRTAP1* phylogenetic trees from multiple
206 alignments of the 5' and 3' flanking sequences of the mammalian *KRTAP1* genes.
207 Surprisingly, the phylogenies derived from these flanking sequences did not show any pattern
208 of concerted evolution, and in contrast to the coding region phylogeny, the clades in these
209 phylogenetic trees were group by *KRTAP1* repeat number, not by species (**Figure 3**). We
210 note that bootstrap support is not strong for all the clades in these phylogenetic trees, but the
211 contrast between the coding region concerted versus flanking region radiating evolutionary
212 patterns is unmistakable. Furthermore, the topology within many of the *KRTAP1* flanking
213 region clades is consistent with the reported mammalian phylogeny (refer to **Figures 2** and
214 **3**). These phylogenies were generated from multiple sequence alignments that encompass the
215 regions that align well, but phylogenies derived from sequence alignments that include poorly
216 aligned regions give qualitatively similar results (**Figure S3**). Overall, in stark contrast to the
217 coding region, the flanking regions show a phylogenetic pattern expected for normal
218 radiating evolution, and exhibit no evidence of concerted evolution.

219

What is responsible for the different evolutionary patterns of the *KRTAP1* coding and flanking regions?

The difference in evolutionary pattern between the coding and flanking regions is striking, hence we sought to identify the mechanism(s) responsible.

Purifying selection: Previous studies have shown that multi-gene loci undergoing birth-and-death evolution can show high levels of identity within the coding region due to strong purifying selection (NEI *et al.* 2000; PIONTKIVSKA *et al.* 2002). It is possible that purifying selection maintains sequence identity between *KRTAP1-n* copies within a species, whilst diversifying selection results in differences between species. If so, we would predict that while the non-synonymous sites would show a concerted evolution pattern, the synonymous sites would instead show a normal radiating pattern of evolution (resembling the flanking regions).

To investigate this, we looked at the pattern of evolution of the synonymous sites in the coding sequences compared to the non-synonymous sites. 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 third codon positions were used as a proxy for synonymous sites, and first and second codon positions were used as a proxy for non-synonymous sites. We generated phylogenetic trees from multiple sequence alignments of the first-second (which we refer to as "non-synonymous"), and third (which we refer to as "synonymous") codon sites of the *KRTAP1-n* coding regions to test for different evolutionary patterns. Surprisingly, while the non-synonymous sites displayed a pattern of concerted evolution as was expected (**Figure 4A**), the synonymous sites also revealed the same pattern of concerted evolution (**Figure 4B**). The concerted evolution pattern for the synonymous sites seems to be stronger than that of the non-synonymous sites, as they separate sheep and

cattle into separate clades, and also resolve dog, elephant, and rat/mouse into separate clades (**Figure 4**).

Codon adaptation: We considered whether this pattern of concerted evolution amongst the synonymous sites might result from codon adaptation (LIN *et al.* 2006), as a result of synonymous mutations being selected to follow changes in the favoured codons between species. The *KRTAP1-n* genes display strong evidence for codon adaptation (the degree to which the favoured codons for that species are used in a gene). For example, the human *KRTAP1-n* genes collectively show a codon adaptation index (CAI) of 0.91 (out of a maximum of 1), higher than the CAI of randomly permutated human *KRTAP1* sequences (CAI=0.78). Using the *KRTAP1* coding sequence alignment used for the phylogenies presented in **Figure 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 usage preference for that amino acid. Five of these mutations show 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, four of these mutations show the opposite pattern, and most of the codon usage preference changes between human and mouse are small (**Table S2**). These results provide no evidence for adaptation to different codon usage preferences driving the pattern of *KRTAP1* concerted evolution.

Reverse transcription of *KRTAP1* mRNA: Another potential explanation for the incongruence in evolutionary pattern between the *KRTAP1* coding and flanking regions is reverse transcription of *KRTAP1-n* mRNAs, followed by homologous recombination-mediated replacement of a genomic *KRTAP1-n* with the reverse transcribed copy

(COULOMBE-HUNTINGTON AND MAJEWSKI 2007). This is feasible given that *KRTAP1-n* 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. Inspection of the 5' and 3' flanking regions revealed that sequence similarity between *KRTAP1-n* sequences within a genome tends to decay immediately upstream of the ATG codon and downstream of the stop codon (**Figure 5**). This suggests that reverse transcription/integration of *KRTAP1-n* mRNA is unlikely to explain the pattern of *KRTAP1* concerted evolution, as the transcribed flanking regions of the gene would be expected to 'hitch-hike' with the coding regions through such a mechanism.

We also considered whether the *KRTAP1-n* 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, and, independently, that each of these duplications would not involve any flanking sequence (including promoter and terminator sequences) and have inserted into the same site in each species.

Gene conversion: Finally, we considered whether gene conversion could explain the pattern of *KRTAP1* repeat evolution. Gene conversion events within a genome that convert a section of one repeat to the sequence of another can create homogeneity (CHEN *et al.* 2007), and the degree of homogeneity depends on the relative rates of gene conversion and mutation (TESHIMA AND INNAN 2004; HARPAK *et al.* 2017). Our results imply that if gene conversion does occur, it is somehow restricted to the coding region. This pattern could occur if there is selective pressure to maintain a degree of intra-genome homogeneity between the repeat copies. If so, under the assumption that gene conversion occurs in both the coding and flanking regions, those events occurring in the flanking region will not have a selective

advantage, while those occurring in the coding region will. Therefore, the probability of gene conversion events becoming fixed in the population will be greater for events that involve the coding region. There is considerable intra-genomic variation between *KRTAP1* repeats (**Figure 3**), but this incomplete level of homogenization can be explained by relatively infrequent gene conversion events and/or relative infrequent fixation of these events. Therefore, the sequence features of the *KRTAP1* repeats that we document here can all be accounted for by gene conversion coupled with selection.

Evidence for gene conversion events in the *KRTAP1-n* repeats

Inspection of the *KRTAP1* coding region multiple sequence alignment provides evidence for tracts of gene conversion. Specifically, sites where there are mutations that are shared between copies within a species, but that differ between species, are frequently clustered together rather than scattered throughout the gene (**Figure 6**). Such patches of homogeneity are expected if there has been occasional, short-tract gene conversion events. The patches we observe are small, but are within the expected range for mammalian gene conversion events (CHEN *et al.* 2007). In addition, we collected population polymorphism data for *KRTAP1-n* sequences in sheep, as comprehensive sequence variation data are scarce in other species. For many of the sites that are polymorphic, the polymorphism is shared across some, or all, of the *KRTAP1-n* sequences (**Figure 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. Gene conversion has also previously been suggested as an explanation for the pattern of polymorphism in the ovine *KRTAP1* genes (ROGERS *et al.* 1994). Collectively, our results suggest that the unusual evolutionary pattern of the *KRTAP1* repeats, where the coding region evolutionary dynamics

316 are uncoupled from those of the flanking region, is the result of occasional short-tract gene
317 conversion events that are selected for in the coding region but not the flanking regions, and
318 that drive partial homogenization.

319

320

321 **DISCUSSION**

322 Here we have shown that *KRTAP1-n* genes are conserved as a block of four tandem repeats in
323 mammalian species, and this suggests they derive from a relatively ancient gene-
324 amplification event or events that probably pre-date mammalian speciation. These four
325 tandem copies display a strong pattern of concerted evolution in the coding regions, yet the
326 regions flanking show a normal radiating pattern of evolution. We suggest that this
327 dichotomous pattern of evolution is not the result of purifying selection acting to retard
328 changes to the amino acid sequence, but instead results from short gene conversion tracts that
329 periodically homogenize sequences between the four *KRTAP1* genes.

330 The role of gene conversion is supported by two key pieces of evidence: 1) unique amino
331 acid tracts that are shared by KAP1 copies within a species, but are unique to that
332 species/group of related species; and 2) the possession of shared nucleotide variants between
333 *KRTAP1* gene copies in sheep populations. These results extend previous reports of
334 homogenization via ongoing short-tract gene conversion events in other protein coding genes
335 (NOONAN *et al.* 2004; LAMPING *et al.* 2017).

336 We propose that gene conversion is being utilized as an unusual form of purifying selection
337 that prevents accumulation of too much divergence between *KRTAP1* gene copies. We
338 speculate that homogeneity of the *KRTAP1* coding sequences is beneficial as it enables the

production of more homogenous components of the hair and wool fibre matrix, and thus potentially facilitates better associations with the keratin intermediate filaments. We cannot, however, rule out the possibility that individual *KRTAP1* repeats might have functional differences, the signal of which is overwhelmed by the concerted evolution signal from the majority of the gene. However, we note that, particularly in dogs, some of the *KRTAP1-n* genes are very similar in sequence. Therefore, we favour the explanation that *KRTAP1* concerted evolution results from ongoing, stochastic gene conversion events coupled with selection within the coding region against inter-repeat heterogeneity.

Purifying selection is evident in the *KRTAP1-n* coding regions, as the rate of synonymous change is about twice that of the non-synonymous rate (**Figure 4**). While this may seem to contradict the similarity in the synonymous and non-synonymous concerted evolution tree topologies, it can be simply explained by purifying selection acting on residues that are conserved between species, and thus not contributing to the synapomorphies that influence the tree topologies. Any gene conversion events that homogenize unfavourable amino acids will be selected against, thereby preventing deleterious mutations from spreading between copies. However, this same process also allows tolerable and advantageous amino acid changes to sweep through the copies (DOVER 1982). The *KRTAP1-n* sequences from closely related species (i.e. human and macaque, rat and mouse, sheep and cattle) were not separated into different clades for most of the phylogenetic trees we generated (**Figures 3 and 4**). This suggests that the rate of homogenization is relatively slow, and insufficient to drive substantial homogeneity over the evolutionary time frames separating these species pairs. In this context, the shared polymorphisms that we observe in sheep (that are evidence for gene conversion events) are likely intermediate stages in the accumulation of homogenized *KRTAP1-n* sequences.

363 The sharp border between a concerted evolution pattern in the coding region and a radiating
364 evolution pattern in the immediate flanking regions is striking. This can partially be explained
365 by the selection for gene conversion events within the coding region, as we have proposed.
366 However, it is intriguing to speculate that this may also be a consequence of differential
367 expression between the *KRTAP1* genes that is mediated by copy-specific differences in the
368 regulatory regions. Although not direct, some evidence for differential regulation of
369 *KRTAP1-n* gene expression was found in two transcriptome studies looking for differentially
370 expressed genes (FAN *et al.* 2013; CHANG *et al.* 2014). If the *KRTAP1-n* genes do have
371 functionally distinct roles, gene conversion events in the *KRTAP1* regulatory regions that
372 perturb their differential regulation may be maladaptive and therefore selected against. Thus,
373 selective pressure for coding region homogeneity versus regulatory region diversity, coupled
374 with ongoing gene conversion, may be a powerful way to achieve the dichotomy in
375 evolutionary patterns we observe. Clearly, a better understanding of the transcriptional
376 regulation of the *KRTAP1* genes is required to address this hypothesis.

377 Gene conversion is frequently viewed through the lens of impeding sub-functionalization of
378 gene duplicates. This view is consistent with the well characterized case of the opsin gene
379 duplicates in primates, where there is a much stronger signal of gene conversion/concerted
380 evolution in the introns, than in the exons (SHYUE *et al.* 1994; HIWATASHI *et al.* 2011). The
381 interpretation is that selection has largely rejected gene conversion events that include the
382 coding (exon) regions, whilst allowing those occurring in the non-coding (intron) regions to
383 spread in the population (SHYUE *et al.* 1994). This is the opposite of what we observe, and
384 illustrates how gene conversion and selection can intersect to produce a constellation of
385 evolutionary patterns: homogenization of the non-coding but not the coding regions in the
386 opsin paralogs (SHYUE *et al.* 1994); homogenization of the coding but not the non-coding

regions in the *KRTAP1* genes (this study); and homogenisation of both coding and non-coding regions equally in the ribosomal RNA gene repeats (GANLEY AND KOBAYASHI 2007).

The extent to which gene conversion acts to homogenize gene duplicates remains controversial (GAO AND INNAN 2004; CASOLA *et al.* 2012; HARPAK *et al.* 2017). Furthermore, even in examples where recurrent gene conversion events can be detected, they are often not sufficient to produce a strong concerted evolution pattern (PETRONELLA AND DROUIN 2011; PETRONELLA AND DROUIN 2014). There are two potential explanations for why such a strong pattern of concerted evolution is observed in the case the *KRTAP1* genes, despite the relatively high levels of divergence between copies. First, unlike many of the examples that have aroused controversy (GAO AND INNAN 2004; CASOLA *et al.* 2012; HARPAK *et al.* 2017), the *KRTAP1-n* repeats are tandemly-arranged. Proximity effects as a consequence of tandem arrangement may increase the chances of unequal alignment of the repeats during DNA repair-based homologous recombination compared to dispersed repeats, and thus may increase the chances of conversion events. 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). A second explanation relates to the imperfect decapeptide tandem repeat motif found in the coding region. Variation in the copy number of decapeptide repeats between *KRTAP1* genes is possibly the result of unequal recombination (LIAO AND WEINER 1995; GANLEY AND SCOTT 1998; MORRILL *et al.* 2016). If so, the *KRTAP1* genes may harbour a recombination hotspot that drives both decapeptide repeat copy number variation and gene conversion at higher than average levels.

Repeats are ubiquitous denizens of eukaryote genomes, where they exist in different forms (coding, non-coding) and organizations (tandem, dispersed). Our results add to the growing list of examples that illustrate how different molecular and evolutionary processes can

411 impinge on repeats to structure their sequences and create distinctive patterns of evolution
412 (SHYUE *et al.* 1994; NOONAN *et al.* 2004; GANLEY AND KOBAYASHI 2007; STORZ *et al.* 2007;
413 HIWATASHI *et al.* 2011; LAMPING *et al.* 2017). However, it is unclear how widespread these
414 sorts of evolutionary dynamics are for eukaryotic gene repeats, largely because the patterns of
415 evolution have not been investigated for the vast majority of multi-gene families. The
416 increasing availability of high quality genome sequences for a wide range of eukaryotes puts
417 us in an excellent position to determine, on a much more systematic and wide-ranging basis,
418 the patterns of repeat sequence dynamics and evolution. This will, in turn, make it clear
419 whether the impact of recombination on the *KRTAP1*s is unusual, or highlights a common
420 mechanism to finely scale patterns of homogeneity and divergence between repeat copies
421 over time.

422

423

424 **ACKNOWLEDGEMENTS**

425 This work was supported by a Marsden Fund award (14-MAU-053) to ARDG, an
426 AGMARDT Postdoctoral Fellowship to HG, and a Vernon Willey Trust Fellowship to HZ.

427

428

429 **REFERENCES**

430 Bailey, T. L., N. Williams, C. Misleh and W. W. Li, 2006 MEME: discovering and analyzing DNA
431 and protein sequence motifs. Nucleic Acids Res. 34**:** W369-373.
432 Britten, R. J., and D. E. Kohne, 1968 Repeated sequences in DNA. Science 161**:** 529-540.

433 Brown, D. D., P. C. Wensink and E. Jordan, 1972 A comparison of the ribosomal DNA's of
434 *Xenopus laevis* and *Xenopus mulleri*: the evolution of tandem genes. J. Mol. Biol. 63**:**
435 57-73.
436 Casola, C., G. C. Conant and M. W. Hahn, 2012 Very low rate of gene conversion in the yeast
437 genome. Mol. Biol. Evol. 29**:** 3817-3826.
438 Chang, T. H., H. D. Huang, W. K. Ong, Y. J. Fu, O. K. Lee *et al.*, 2014 The effects of actin
439 cytoskeleton perturbation on keratin intermediate filament formation in
440 mesenchymal stem/stromal cells. Biomaterials 35**:** 3934-3944.
441 Chen, J. M., D. N. Cooper, N. Chuzhanova, C. Ferec and G. P. Patrinos, 2007 Gene
442 conversion: mechanisms, evolution and human disease. Nature Reviews Genetics 8**:**
443 762-775.
444 Coulombe-Huntington, J., and J. Majewski, 2007 Characterization of intron loss events in
445 mammals. Genome Res. 17**:** 23-32.
446 Dover, G. A., 1982 Molecular drive: a cohesive mode of species evolution. Nature 299**:** 111-
447 117.
448 Eirin-Lopez, J. M., L. Rebordinos, A. P. Rooney and J. Rozas, 2012 The birth-and-death
449 evolution of multigene families revisited. Genome Dynamics 7**:** 170-196.
450 Elder, J. F., Jr., and B. J. Turner, 1995 Concerted evolution of repetitive DNA sequences in
451 eukaryotes. Q. Rev. Biol. 70**:** 297-320.
452 Fan, R., J. Xie, J. Bai, H. Wang, X. Tian *et al.*, 2013 Skin transcriptome profiles associated with
453 coat color in sheep. BMC Genomics 14**:** 389.
454 Ganley, A. R. D., and T. Kobayashi, 2007 Highly efficient concerted evolution in the
455 ribosomal DNA repeats: total rDNA repeat variation revealed by whole-genome
456 shotgun sequence data. Genome Res. 17**:** 184-191.

457 Ganley, A. R. D., and B. Scott, 1998 Extraordinary ribosomal spacer length heterogeneity in a
458 Neotyphodium endophyte hybrid: implications for concerted evolution. Genetics
459 150**:** 1625-1637.
460 Gao, L.-Z., and H. Innan, 2004 Very low gene duplication rate in the yeast genome. Science
461 306**:** 1367-1370.
462 Gong, H., H. Zhou, R. H. J. Forrest, S. Li, J. Wang *et al.*, 2016 Wool keratin-associated protein
463 genes in sheep—a review. Genes 7**:** 24.
464 Gong, H., H. Zhou and J. G. H. Hickford, 2010a Polymorphism of the ovine keratin-associated
465 protein 1-4 gene (KRTAP1-4). Mol. Biol. Rep. 37**:** 3377-3380.
466 Gong, H., H. Zhou, S. Hodge, J. M. Dyer and J. G. H. Hickford, 2015 Association of wool traits
467 with variation in the ovine KAP1-2 gene in Merino cross lambs. Small Rumin. Res.
468 124**:** 24-29.
469 Gong, H., H. Zhou, G. W. McKenzie, J. G. Hickford, Z. Yu *et al.*, 2010b Emerging issues with
470 the current keratin-associated protein nomenclature. International Journal of
471 Trichology 2**:** 104-105.
472 Gong, H., H. Zhou, G. W. McKenzie, Z. Yu, S. Clerens *et al.*, 2012 An updated nomenclature
473 for keratin-associated proteins (KAPs). Int. J. Biol. Sci. 8**:** 258-264.
474 Gong, H., H. Zhou, Z. Yu, J. Dyer, J. E. Plowman *et al.*, 2011 Identification of the ovine
475 keratin-associated protein KAP1-2 gene (KRTAP1-2). Exp. Dermatol. 20**:** 815-819.
476 Guindon, S., J. F. Dufayard, V. Lefort, M. Anisimova, W. Hordijk *et al.*, 2010 New algorithms
477 and methods to estimate maximum-likelihood phylogenies: assessing the
478 performance of PhyML 3.0. Syst. Biol. 59**:** 307-321.

479 Harpak, A., X. Lan, Z. Gao and J. K. Pritchard, 2017 Frequent nonallelic gene conversion on
480 the human lineage and its effect on the divergence of gene duplicates. PNAS 114**:**
481 12779-12784.
482 Hiwatashi, T., A. Mikami, T. Katsumura, B. Suryobroto, D. Perwitasari-Farajallah *et al.*, 2011
483 Gene conversion and purifying selection shape nucleotide variation in gibbon L/M
484 opsin genes. BMC Evol. Biol. 11**:** 312.
485 Itenge-Mweza, T. O., R. H. Forrest, G. W. McKenzie, A. Hogan, J. Abbott *et al.*, 2007
486 Polymorphism of the KAP1.1, KAP1.3 and K33 genes in Merino sheep. Mol. Cell.
487 Probes 21**:** 338-342.
488 Katoh, K., and D. M. Standley, 2013 MAFFT multiple sequence alignment software version 7:
489 improvements in performance and usability. Mol. Biol. Evol. 30**:** 772-780.
490 Khan, I., E. Maldonado, V. Vasconcelos, J. O. Stephen, W. E. Johnson *et al.*, 2014 Mammalian
491 keratin associated proteins (KRTAPs) subgenomes: disentangling hair diversity and
492 adaptation to terrestrial and aquatic environments. BMC Genomics 15**:** 779.
493 Lamping, E., J. Y. Zhu, M. Niimi and R. D. Cannon, 2017 Role of ectopic gene conversion in
494 the evolution of a *Candida krusei* pleiotropic drug resistance transporter family.
495 Genetics 205**:** 1619-1639.
496 Li, S., H. Zhou, H. Gong, F. Zhao, J. Hu *et al.*, 2017a Identification of the ovine keratin-
497 associated protein 26-1 gene and its association with variation in wool traits. Genes
498 8**:** 225.
499 Li, S., H. Zhou, H. Gong, F. Zhao, J. Wang *et al.*, 2017b Identification of the ovine keratin-
500 associated protein 22-1 (KAP22-1) gene and its effect on wool traits. Genes 8**:** 27.

Li, S., H. Zhou, H. Gong, F. Zhao, J. Wang *et al.*, 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., and A. M. Weiner, 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., J. K. Byrnes, J.-K. Hwang and W.-H. Li, 2006 Codon-usage bias versus gene conversions in the evolution of yeast duplicate genes. PNAS 103**:** 14412-14416.

Lopez-Flores, I., and M. A. Garrido-Ramos, 2012 The repetitive DNA content of eukaryotic genomes. Genome Dynamics 7**:** 1-28.

Morrill, S. A., A. E. Exner, M. Babokhov, B. I. Reinfeld and S. M. Fuchs, 2016 DNA instability maintains the repeat length of the yeast RNA polymerase II C-terminal domain. J. Biol. Chem. 291**:** 11540-11550.

Nakamura, Y., T. Gojobori and T. Ikemura, 2000 Codon usage tabulated from international DNA sequence databases: status for the year 2000. Nucleic Acids Res. 28**:** 292.

Nei, M., X. Gu and T. Sitnikova, 1997 Evolution by the birth-and-death process in multigene families of the vertebrate immune system. PNAS 94**:** 7799-7806.

Nei, M., I. B. Rogozin and H. Piontkivska, 2000 Purifying selection and birth-and-death evolution in the ubiquitin gene family. PNAS 97**:** 10866-10871.

Nei, M., and A. P. Rooney, 2005 Concerted and birth-and-death evolution of multigene families. Annu. Rev. Genet. 39**:** 121-152.

525 Noonan, J. P., J. Grimwood, J. Schmutz, M. Dickson and R. M. Myers, 2004 Gene conversion
526 and the evolution of protocadherin gene cluster diversity. Genome Res. 14: 354-366.
527 Perina, A., D. Seoane, A. M. Gonzalez-Tizon, F. Rodriguez-Farina and A. Martinez-Lage, 2011
528 Molecular organization and phylogenetic analysis of 5S rDNA in crustaceans of the
529 genus Pollicipes reveal birth-and-death evolution and strong purifying selection.
530 BMC Evol. Biol. 11: 304.
531 Petronella, N., and G. Drouin, 2011 Gene conversions in the growth hormone gene family of
532 primates: stronger homogenizing effects in the Hominidae lineage. Genomics 98:
533 173-181.
534 Petronella, N., and G. Drouin, 2014 Purifying selection against gene conversions in the folate
535 receptor genes of primates. Genomics 103: 40-47.
536 Piontkivska, H., A. P. Rooney and M. Nei, 2002 Purifying selection and birth-and-death
537 evolution in the histone H4 gene family. Mol. Biol. Evol. 19: 689-697.
538 Powell, B. C., and G. E. Rogers, 1997 The role of keratin proteins and their genes in the
539 growth, structure and properties of hair, pp. 59-148 in *Formation and structure of*
540 *human hair*. Birkhäuser Verlag.
541 Puigbo, P., I. G. Bravo and S. Garcia-Vallve, 2008 CAIcal: a combined set of tools to assess
542 codon usage adaptation. Biol. Direct 3: 38.
543 Richard, G. F., A. Kerrest and B. Dujon, 2008 Comparative genomics and molecular dynamics
544 of DNA repeats in eukaryotes. Microbiol. Mol. Biol. Rev. 72: 686-727.
545 Rogers, G. R., J. G. H. Hickford and R. Bickerstaffe, 1994 Polymorphism in two genes for B2
546 high sulfur proteins of wool. Anim. Genet. 25: 407-415.
547 Rogers, M. A., L. Langbein, S. Praetzel-Wunder, H. Winter and J. Schweizer, 2006 Human hair
548 keratin-associated proteins (KAPs). Int. Rev. Cytol. 251: 209-263.

549 Rogers, M. A., and J. Schweizer, 2005 Human KAP genes, only the half of it? Extensive size
550 polymorphisms in hair keratin-associated protein genes. J. Invest. Dermatol. 124**:** vii-
551 ix.
552 Rooney, A. P., and T. J. Ward, 2005 Evolution of a large ribosomal RNA multigene family in
553 filamentous fungi: birth and death of a concerted evolution paradigm. PNAS 102**:**
554 5084-5089.
555 Shimomura, Y., N. Aoki, J. Schweizer, L. Langbein, M. A. Rogers *et al.*, 2002 Polymorphisms in
556 the human high sulfur hair keratin-associated protein 1, KAP1, gene family. J. Biol.
557 Chem. 277**:** 45493.
558 Shyue, S. K., L. Li, B. H. Chang and W.-H. Li, 1994 Intronic gene conversion in the evolution of
559 human X-linked color vision genes. Mol. Biol. Evol. 11**:** 548-551.
560 Stein, L. D., 2004 End of the beginning. Nature 431**:** 915-916.
561 Stephan, W., 1989 Tandem-repetitive noncoding DNA: Forms and forces. Mol. Biol. Evol. 6**:**
562 198-212.
563 Storz, J. F., M. Baze, J. L. Waite, F. G. Hoffmann, J. C. Opazo *et al.*, 2007 Complex signatures
564 of selection and gene conversion in the duplicated globin genes of house mice.
565 Genetics 177**:** 481–500.
566 Talavera, G., and J. Castresana, 2007 Improvement of phylogenies after removing divergent
567 and ambiguously aligned blocks from protein sequence alignments. Syst. Biol. 56**:**
568 564-577.
569 Tao, J., H. Zhou, H. Gong, Z. Yang, Q. Ma *et al.*, 2017a Variation in the KAP6-1 gene in
570 Chinese Tan sheep and associations with variation in wool traits. Small Rumin. Res.
571 154**:** 129-132.

572 Tao, J., H. Zhou, Z. Yang, H. Gong, Q. Ma *et al.*, 2017b Variation in the KAP8-2 gene affects
573 wool crimp and growth in Chinese Tan sheep. Small Rumin. Res. 149**:** 77-80.
574 Teshima, K. M., and H. Innan, 2004 The effect of gene conversion on the divergence
575 between duplicated genes. Genetics 166**:** 1553-1560.
576 Torrents, D., M. Suyama, E. Zdobnov and P. Bork, 2003 A genome-wide survey of human
577 pseudogenes. Genome Res. 13**:** 2559-2567.
578 Wernersson, R., and A. G. Pedersen, 2003 RevTrans: Multiple alignment of coding DNA from
579 aligned amino acid sequences. Nucleic Acids Res. 31**:** 3537-3539.
580 Wu, D. D., D. Irwin and Y. P. Zhang, 2008 Molecular evolution of the keratin associated
581 protein gene family in mammals, role in the evolution of mammalian hair. BMC Evol.
582 Biol. 8**:** 241.
583 Yu, G., D. K. Smith, H. Zhu, Y. Guan and T. T.-Y. Lam, 2017 GGTREE: an R package for
584 visualization and annotation of phylogenetic trees with their covariates and other
585 associated data. Methods Ecol. Evol. 8**:** 28–36.
586 Zhou, H., H. Gong, S. Li, Y. Luo and J. Hickford, 2015 A 57-bp deletion in the ovine KAP6-1
587 gene affects wool fibre diameter. Journal of Animal Breeding and Genetics.
588 Zhou, H., H. Gong, J. Wang, J. M. Dyer, Y. Luo *et al.*, 2016 Identification of four new gene
589 members of the KAP6 gene family in sheep. Sci. Rep. 6**:** 24074.

590

591

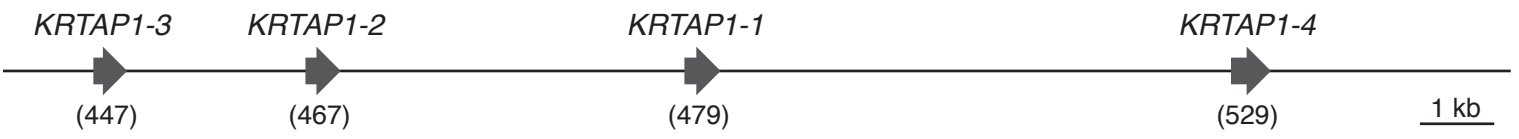

Figure 1. Tandem repeat organization of the keratin associated protein-1 (*KRTAP1*) genes
The organization of mammalian *KRTAP1* genes is illustrated by the arrangement found in sheep. The four *KRTAP1-n* paralogs are represented by arrows that indicate the direction of transcription. Diagram is drawn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

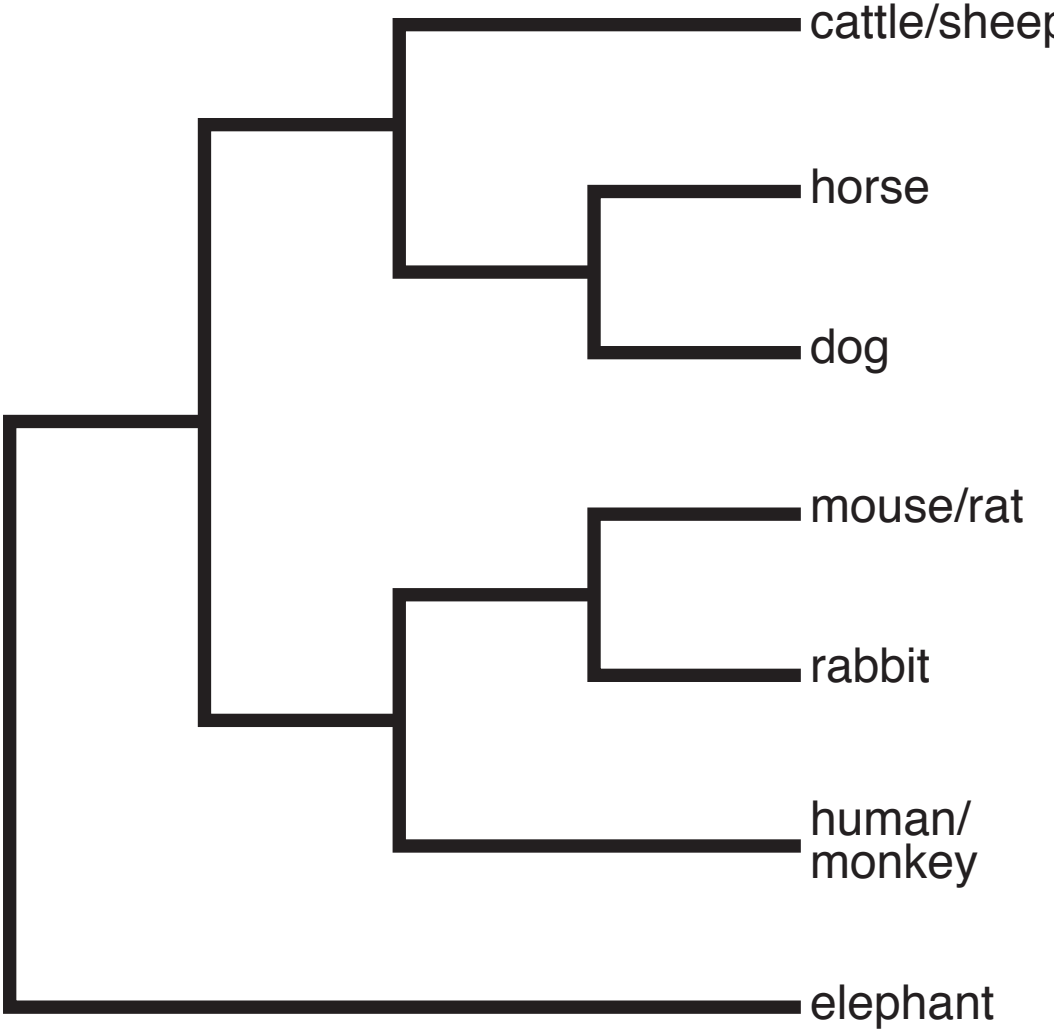

Figure 2. Mammalian *KRTAP1-n* gene phylogenetic relationships
Representative phylogenetic tree illustrating the relationships between the *KRTAP1-n* genes in the species used in this study. Branch lengths are not to scale. The phylogeny is adapted from that presented in McCormack et al. (2012).

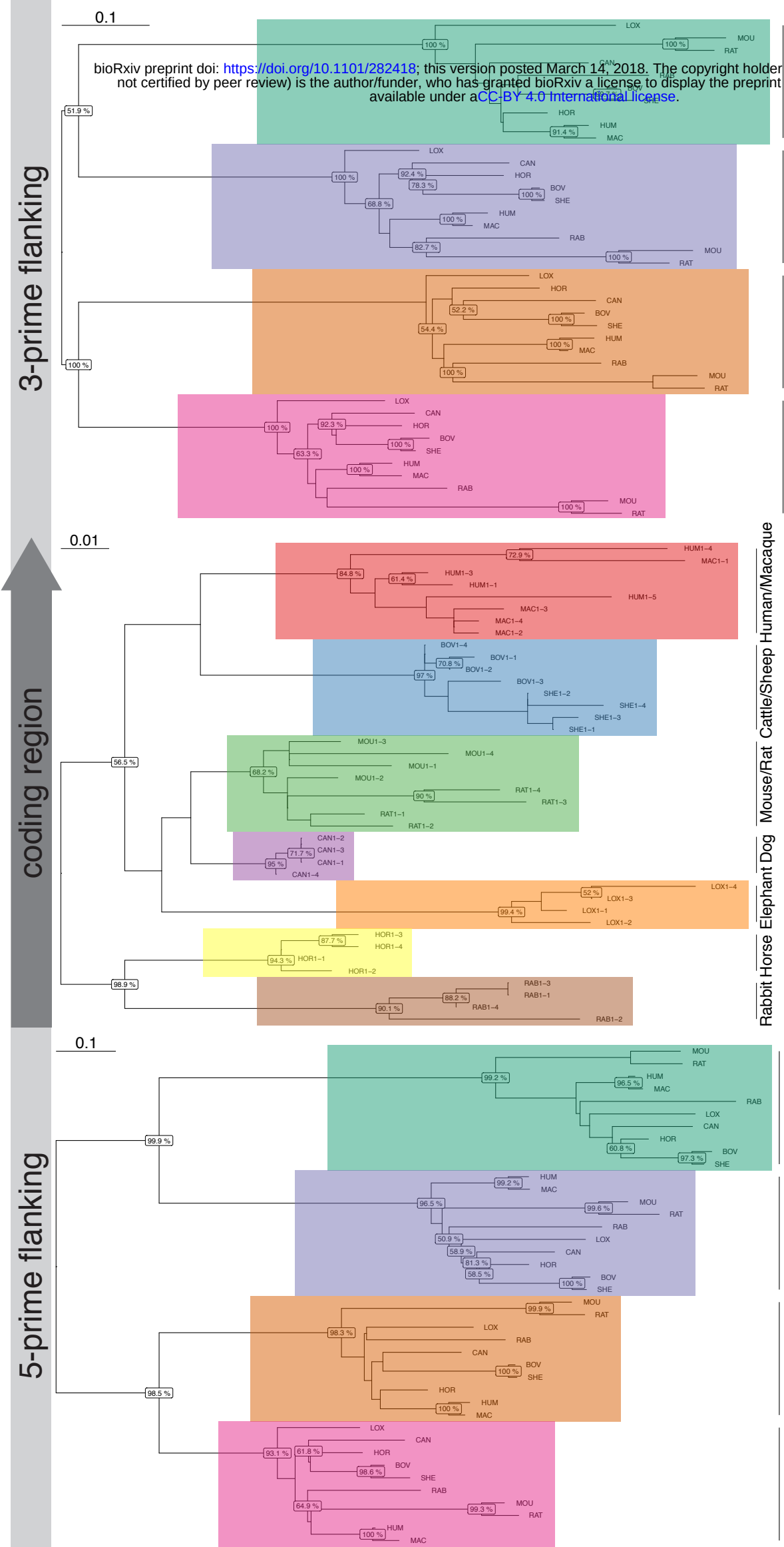

Figure 3. Phylogenetic trees of *KRTAP1-n* coding and flanking region sequences

Phylogenetic trees were constructed for the mammalian *KRTAP1-n* 5' flanking region, coding region, and 3' flanking region using PhyML. The species are indicated by three-letter abbreviations. The number following this for the coding regions indicates the *KRTAP1-n* gene name. The major clades within the trees are indicated by coloured 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 the top left. Human *KRTAP1-n* gene names have been altered for consistency with other species.

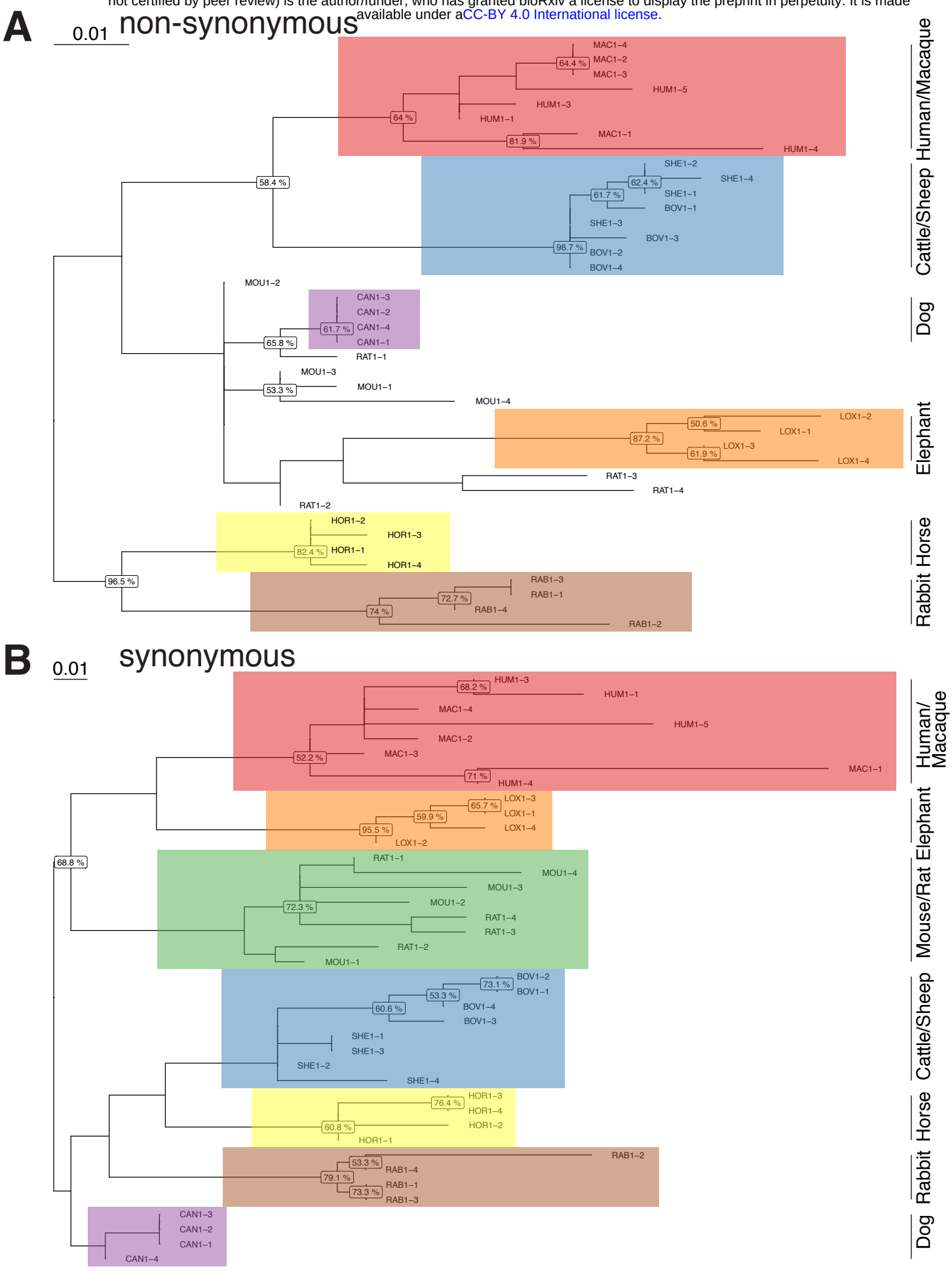

Figure 4. The *KRTAP1-n* concerted evolution pattern is not explained by purifying selection
Phylogenetic trees were constructed for the 1st and 2nd codon positions (”non-synonymous”; **A**), and the 3rd codon position (”synonymous”; **B**), as per **Figure 3**. The major clades in both phylogenies tend to group by species, with this concerted evolution pattern being stronger for the synonymous phylogeny. Numbers on nodes indicate bootstrap supports with values over 50%, and substitution rates are indicated at the top left.

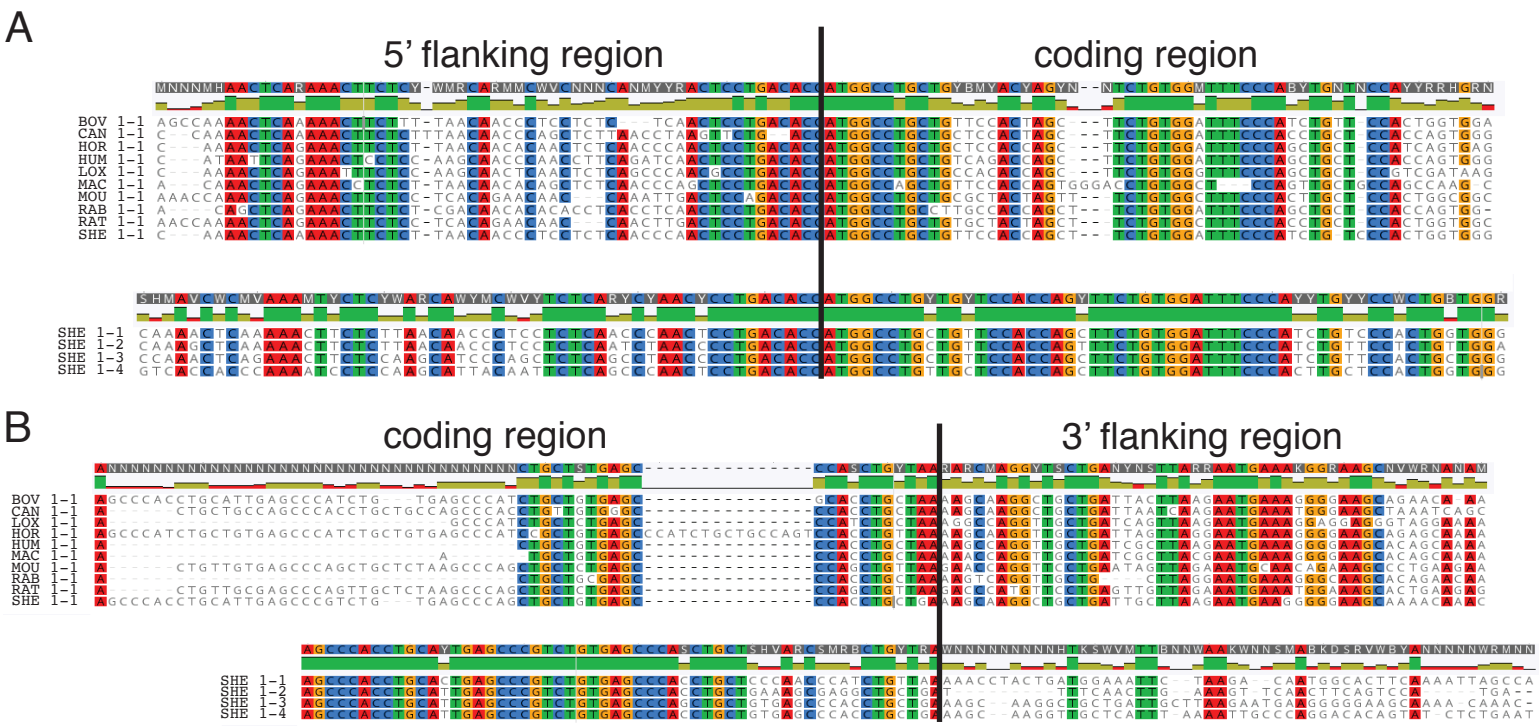

Figure 5. The switch between concerted and radiating evolution patterns is located close to the start/stop sites

A) Alignment of the region flanking the *KRTAP1-1* gene start site. The boundary between the 5’ flanking and coding regions is marked by a vertical line (followed by the ATG). Underneath is an alignment of the same region for all four *KRTAP1-n* sequences from sheep. Mismatches have a white background, conservation is indicated graphically above each alignment, and consensus sequences are shown at the top. **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).

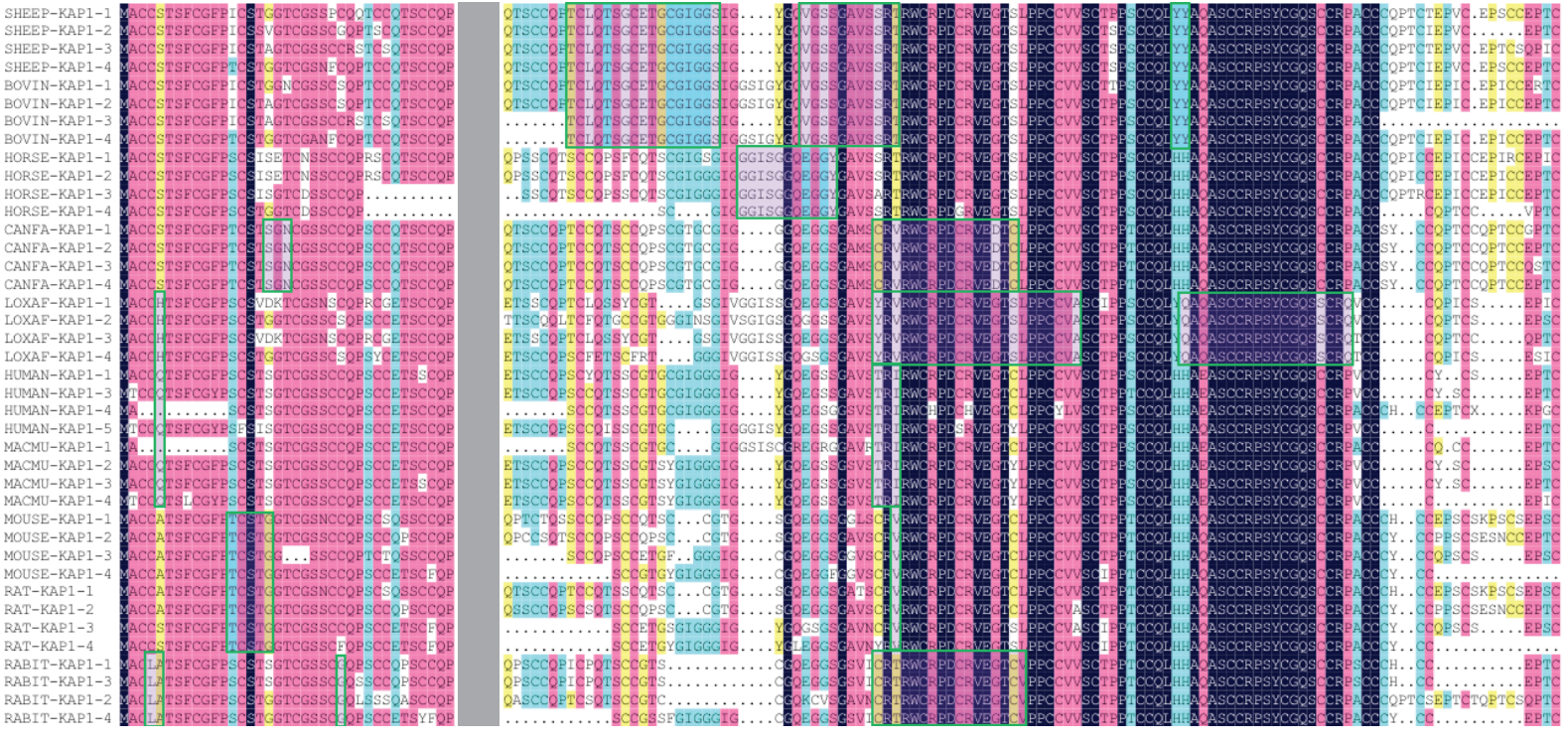

Figure 6. Evidence for short gene conversion tracts between *KRTAP1-n* sequences within species
Alignment of KAP1 amino acid sequences from the ten mammalian species. Amino acid tracts boxed in green represent sequences unique to a species or related species pairs. The grey vertical box represents the conserved decapeptide repeat sequences (which have been removed). Dots represent gaps in the alignment.

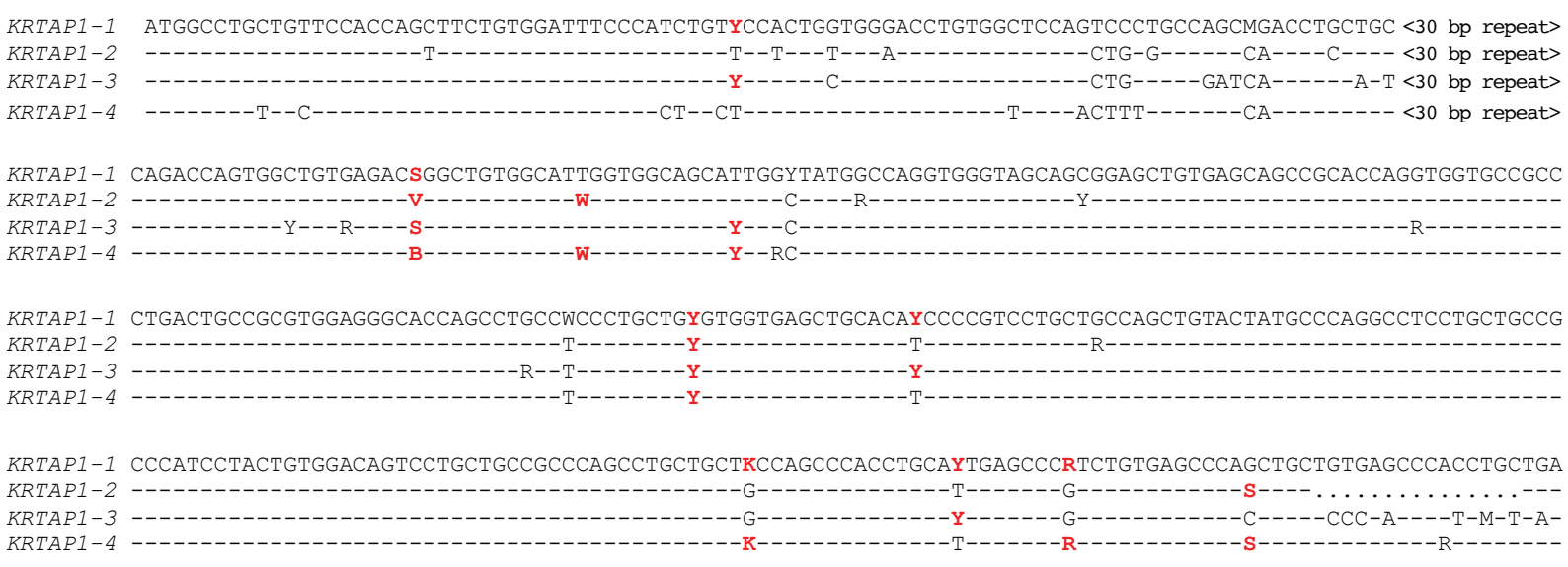

Figure 7. Shared polymorphisms between *KRTAP1-n* sequences in sheep
Alignment of the four sheep *KRTAP1-n* coding region sequences. Dashes represent nucleotides identical to the top sequence, and dots represent gaps. The 30 bp repeats are not shown, as the insertion/deletion positions cannot be precisely determined. Shared nucleotide substitutions between repeat copies are highlighted in red.