Assembly and validation of conserved long non-coding RNAs in the ruminant transcriptome

Stephen J. Bush[1, 2], Charity Muriuki[1], Mary E. B. McCulloch[1], Iseabail L. Farquhar[3], Emily L. Clark[1*], David A. Hume[1, 4 * †]

[1] The Roslin Institute, University of Edinburgh, Easter Bush Campus, Edinburgh, Midlothian, EH25 9RG

[2] Experimental Medicine Division, Nuffield Department of Clinical Medicine, University of Oxford, John Radcliffe Hospital, Headington, Oxford, OX3 9DU

[3] Centre for Synthetic and Systems Biology, CH Waddington Building, Max Borne Crescent, King's Buildings, University of Edinburgh, EH9 3BF

[4] Mater Research-University of Queensland, Translational Research Institute, 37 Kent Street, Woolloongabba, Queensland 4102, Australia

* These authors contributed equally to this work.

[†] Corresponding author: david.hume@roslin.ed.ac.uk / david.hume@uq.edu.au

Email addresses:

Stephen J. Bush	stephen.bush@ndm.ox.ac.uk / stephen.bush@roslin.ed.ac.uk
Charity Muriuki	charity.muriuki@ed.ac.uk
Mary E. B. McCulloch	mary.mcculloch@ed.ac.uk

22 Iseabail L. Farquhar iseabail.farquhar@ed.ac.uk

23 Emily L. Clark emily.clark@roslin.ed.ac.uk

24 David A. Hume david.hume@uq.edu.au / david.hume@roslin.ed.ac.uk

25

Abstract

mRNA-like long non-coding RNAs (lncRNA) are a significant component of mammalian transcriptomes, although most are expressed only at low levels, with high tissue-specificity and/or at specific developmental stages. In many cases, therefore, lncRNA detection by RNA-sequencing (RNA-seq) is compromised by stochastic sampling. To account for this and create a catalogue of ruminant lncRNA, we compared *de novo* assembled lncRNA derived from large RNA-seq datasets in transcriptional atlas projects for sheep and goats with previous lncRNA assembled in cattle and human. Few lncRNA could be reproducibly assembled from a single dataset, even with deep sequencing of the same tissues from multiple animals. Furthermore, there was little sequence overlap between lncRNA assembled from pooled RNA-seq data. We combined positional conservation (synteny) with cross-species mapping of candidate lncRNA to identify a consensus set of ruminant lncRNA and then used the RNA-seq data to demonstrate detectable and reproducible expression in each species. The majority of lncRNA were encoded by single exons, and expressed at < 1 TPM. In sheep, 20-30% of lncRNA had expression profiles significantly correlated with neighbouring protein-coding genes, suggesting association with enhancers. Alongside substantially expanding the ruminant lncRNA repertoire, the outcomes of our analysis demonstrate that stochastic sampling can be partly overcome by combining RNA-seq datasets from related species. This has practical implications for the future discovery of lncRNA in other species.

Introduction

Mammalian transcriptomes include many long non-coding RNAs (lncRNAs), a collective term for transcripts of > 200 nucleotides that resemble mRNAs (many being 3' polyadenylated, 5' capped and spliced) but do not encode a protein product [1]. Proposed functional roles of lncRNAs include transcriptional regulation, epigenetic regulation, intracellular trafficking and chromatin remodelling (see reviews [2-9]). Some view lncRNAs as transcriptional noise [10, 11]. Full length lncRNAs are difficult to assemble: many are expressed at low levels [12], with high tissue-specificity [13, 14], at specific developmental time points (e.g. [15-17]), and with few signs of selective constraint [18, 19]. Many are also expressed transiently, and so may be partly degraded by the exosome complex [20].

The initial recognition of lncRNAs as widespread and *bona fide* outputs of mammalian transcription was based upon the isolation and sequencing of large numbers of mouse and human full-length cDNAs [21-23], many of which were experimentally validated [24] and shown to participate in sense-antisense pairs [25]. They were captured in significant numbers because the cDNA libraries were subtracted to remove abundant transcripts. More recent studies have used RNA-sequencing (RNA-seq) to assemble larger catalogues of lncRNAs [26]. Because of the power-law relationship of individual transcript abundance in mammalian transcriptomes [27], unless sequencing is carried out at massive depth, the exons of lowly-abundant transcripts (such as lncRNAs) are subject to stochastic sampling and are detected inconsistently between technical replicates of the same sample [28]. RNA-seq is also a relatively inaccurate means of reconstructing the 5' ends of transcripts [29]. To overcome this constraint, the FANTOM Consortium supplemented RNA-seq with Cap Analysis of Gene Expression (CAGE) data, characterising – in humans – a 5'-complete lncRNA transcriptome [30].

RNA-seq libraries from multiple tissues, cell types and developmental stages are commonly pooled to maximise the number of lncRNA gene models assembled. Genome-wide surveys have expanded the lncRNA repertoire of livestock species such as cattle (18 tissues, sequenced at approx. 40-100 million reads each) [31], pig (10 tissues, sequenced at approx. 6-40 million reads each) [32], and horse (8 tissues, sequenced at approx. 20-200 million reads each) [33], complementing tissue-specific lncRNA catalogues of, for example, cattle muscle [34, 35] and skin [36], and pig adipose [37, 38], liver [39] and testis [40].

The low level of lncRNA conservation (at some loci, it appears that only the act of transcription, rather than the transcript sequence itself, is functionally relevant [41]) reduces the utility of comparative analysis of the large RNA-seq datasets available from human [30, 42] and mouse [43]. Amongst 200 human and mouse lncRNAs, each characteristic of specific immune cell types, there was <1% sequence conservation [44].

Here we focus on more closely related species. We have generated atlases of gene expression for the domestic sheep, *Ovis aries* [45], and the goat, *Capra hircus* (manuscript in preparation). As the two species are closely related (sharing a common ancestor < 10mya [46]) and their respective RNA-seq datasets contain many of the same tissues, it is possible to use data from one species to infer the presence of lncRNAs in the other. Cattle and humans are more distantly related to small ruminants, but nevertheless are substantially more similar than mice. We extend our approach by utilising existing human and cattle lncRNA datasets to identify a consensus ruminant lncRNA transcriptome, and use the sheep transcriptional atlas to confirm that candidate lncRNA identified by cross-species inference are reproducibly expressed. The lncRNA catalogues we have generated in the sheep and goat are of interest in themselves [47] and contribute valuable information to the Functional Annotation of Animal Genomes (FAANG) project [48, 49].

Results and Discussion

Identifying lncRNAs in the sheep and goat transcriptomes

We have previously created an expression atlas for the domestic sheep [45], using both polyadenylated and rRNA-depleted RNA-seq data collected primarily from three male and three female adult Texel x Scottish Blackface (TxBF) sheep at two years of age: 441 RNA-seq libraries in total, comprising 5 cell types and multiple tissues spanning all major organ systems and several developmental stages, from embryonic to adult. To complement this dataset, we also created a smaller-scale expression atlas – of 54 mRNA-seq libraries – from 6 day old crossbred goats, which will be the subject of a dedicated analysis. For both species, each RNA-seq library was aligned against its reference genome (Oar v3.1 and ARS1, for sheep and goat, respectively) using HISAT2 [50], with transcripts assembled using StringTie [51]. This pipeline produced a non-redundant set of *de novo* gene and transcript models, as previously described [45], and expanded the set of transcripts in each reference genome to include *ab initio* lncRNA predictions and novel protein-coding genes. As the primary purpose of the sheep expression atlas was to improve the functional characterisation of the protein-coding transcriptome, the novel sheep protein-coding transcript models generated by this pipeline have been previously discussed [45] (novel protein-coding transcripts for goats will be discussed in a dedicated analysis of the protein-coding goat transcriptome).

Using similar filter criteria to a previous study [52], the *de novo* gene models were parsed to create longlists of 30,677 (sheep) and 7671 (goat) candidate lncRNAs, each of which was >= 200bp and was not associated, on the same strand, with a known protein-coding locus. The 4-fold difference in the length of each longlist can be attributed to the relative size of each dataset. The sheep atlas contains 8 times as many RNA-seq libraries, spans multiple developmental stages (from embryonic to adult), and has a subset of its samples specifically

prepared to ensure the comprehensive capture of ncRNAs – unlike any sample in the goat dataset, this subset is sequenced at a 4-fold higher depth (>100 million reads, rather than >25 million reads) using a total RNA-seq, rather than mRNA-seq, protocol.

Each model on both longlists was assessed for coding potential using the classification tools CPC [53], CPAT [54] and PLEK [55], alongside homology searches of its longest ORF – with blastp [56] and HMMER [57] – to known protein and domain sequences (within the Swiss-Prot [58, 59] and Pfam-A [60] databases, respectively). Those gene models classified as non-coding by CPC, CPAT and PLEK, and having no detectable blastp and HMMER hits, are considered novel lncRNAs.

This pipeline creates shortlists of 12,296 (sheep) and 2657 (goat) lncRNAs (Tables S1 and S2, respectively), representing approximately 40% (sheep) and 35% (goat) of the gene models on each longlist. The mean gene length is similar in both shortlists – 6.7kb (sheep) and 8.8kb (goat) – as is summed exon length, averaging 1.2kb in each species.

Consistent with previous analysis in several other species [31, 61], 6956 (57%) of the sheep lncRNAs, and 1284 (48%) of the goat, were single-exonic. For sheep, the shortlist contains 11,646 previously unknown lncRNA models and provides additional evidence for 650 existing Oar v3.1 lncRNA models (Table S1). A small proportion of longlisted gene models were considered non-coding by at least one of CPC, CPAT or PLEK, but nevertheless showed some degree of sequence homology to either a known protein or protein domain: for sheep, 226 (including 13 existing Oar v3.1 models) (Table S3), and for goats, 153 (Table S4). The number of novel lncRNAs identified is also given per chromosome (Tables S5 (sheep) and S6 (goat)) and per type (Tables S7 (sheep) and S8 (goat)), the majority of which – in both species – are found in intergenic regions, 10-100kb from the nearest gene. Overall, these lncRNA models increase the number of possible genes in the reference annotation by approximately 30% (sheep) and 12% (goat).

The sets of ab initio sheep and goat lncRNAs only minimally overlap at the sequence level

Even with full length cDNA sequences, comparative analysis revealed that only 27% of the lncRNAs identified in human had mouse counterparts [23]. When comparing the sets of sheep and goat lncRNAs, few predicted transcripts – in either species – show sequence-level similarity either to each other or to other closely or distantly related species (cattle and human, respectively, which shared a common ancestor with sheep and goats approx. 25 and 95mya [46]). Of the 12,296 shortlisted sheep lncRNAs, less than half (n = 5139, i.e. 42%) had any detectable pairwise alignment – of any quality and of any length – to either the shortlisted goat lncRNAs, a set of 9778 cattle lncRNAs from a previous study [31] or two sets of human lncRNAs (Figure 1 and Table S9). In only a small proportion of these alignments can there be high confidence: that is, the alignment has a % identity >= 50% within an alignment >= 50% the length of the target sequence. Of the 5139 sheep lncRNAs that could be aligned to any species, only 293 (5.7%) could be aligned with high confidence to goat and 265 (5.2%) to cattle transcripts. Similarly, of the sheep lncRNAs that could be aligned to either of two human lncRNA databases – NONCODE [62] and lncRNAdb [63] – 68 (1.6% of the total alignable lncRNAs) aligned with high confidence to the NONCODE database, and none to the lncRNAdb. Similar findings are observed with the 2657 shortlisted goat lncRNAs: 1343 (50.5%) had a detectable pairwise alignment, of any quality, to either set of sheep, cattle or human lncRNAs. However, of these 1343 lncRNAs, only 113 (8.4%) aligned with high confidence to sheep, 88 (6.6%) to cattle, 55 (4.1%) to the human NONCODE database, and 1 (0.1%) to the human lncRNAdb database (Figure 1 and Table S10). These observations allow for two possibilities. Firstly, lncRNAs may, in general, be poorly conserved at the sequence level, consistent with previous findings [18, 19] and the observation that only 6% of the sheep/goat alignments have >50% reciprocal identity.

171 However, an alternative is that despite the apparent depth of coverage, we have only
172 assembled a subset of the total lncRNA transcriptome in each species.
173
174 ***lncRNAs not captured by the RNA-seq libraries of one species can be found using data***
175 ***from a related species***
176 A reasonable *a priori* prediction is that lncRNAs – if functionally relevant – are most likely
177 to share expression in a closely related species. Whereas human and mouse lncRNAs
178 identified as full length cDNAs were generally less conserved between species than the 5'
179 and 3'UTRs of protein-coding transcripts, their promoters were more highly conserved than
180 those of protein-coding transcripts, some extending as far as chicken [43, 64]. These findings
181 suggested that the large majority of lncRNAs that were analyzed displayed positional
182 conservation across species. Accordingly, rather than comparing the similarity of two sets of
183 lncRNA transcripts, we mapped the lncRNAs assembled in one species (e.g. sheep) to the
184 genome of another (e.g. goat), deriving confidence in the mapping location from synteny.
185 For each of the pairwise sheep/cattle, sheep/goat, cattle/goat, sheep/human, goat/human, and
186 cattle/human comparisons, we identified sets of syntenic blocks: regions in the genome where
187 gene order is conserved both up- and downstream of a focal gene (see Table 1 and Methods).
188 In the sheep/cattle comparison, approximately 5% of the syntenic blocks contain at least one
189 lncRNA with a relative position conserved in both species, either upstream (n=139 lncRNAs)
190 or downstream (n=141) of the central gene in each block (Table S11). In the sheep/goat and
191 cattle/goat comparisons, respectively, approximately 2 and 3% of the syntenic blocks contain
192 a lncRNA (for sheep/goat, n=42 upstream, 40 downstream; for cattle/goat, 86 upstream, 83
193 downstream) (Tables S12 and S13, respectively). With increased species divergence, far
194 fewer lncRNAs (<1%) have relative positions conserved in either the upstream or
195 downstream positions of the sheep/human, goat/human and cattle/human syntenic blocks

(Tables S14, S15 and S16, respectively). These comparatively small proportions highlight the minimal overlap between each set of assembled transcripts, consistent with stochastic assembly – lncRNAs expected to be present in a particular location are captured in only one species, not both. As such, very few lncRNAs in either of the sheep, goat and cattle subsets have evidence of both shared sequence homology and conserved synteny. When comparing sheep and cattle, 16 unique lncRNAs have high-confidence pairwise alignments within a region of conserved synteny, and when comparing sheep and goat, 6 (Table S17).

In most of the syntenic blocks examined, if a lncRNA was detected in one location in one species (either up- or downstream of a focal gene), no corresponding assembled lncRNA was annotated in the comparison species, even though both species sequenced a similar range of tissues. For example, of the 2927 syntenic blocks in the sheep/cattle comparison, 347 (12%) of the sheep blocks, and 506 (17%) of the cattle blocks, contain a lncRNA in the 'upstream' position (that is, between genes 1 and 2), with little overlap between the two species: in only 139 blocks (5%) is a lncRNA present in this position in both species (Table S11). Similar results are found if considering the 'downstream' position, as well as the sheep/goat, goat/cattle, sheep/human, goat/human and cattle/human comparisons: approximately 2-5 times as many lncRNAs are found in either of the two species than are found in both (Tables S11, S12, S13, S14, S15 and S16).

Each set of syntenic blocks, by definition, represents a set of conserved intergenic regions. Given that the majority of lncRNAs are intergenic (Tables S7 and S8), these regions are reasonable locations for directly mapping candidate transcripts (strictly speaking, concatenated exon sequences) to the genome. For the syntenic blocks in each species comparison, we made global alignments of the lncRNAs in species *x* to the intergenic region of species *y*, and vice versa (see Methods). Retaining only those alignments in which the lncRNA can match the intergenic region with 20 or more consecutive residues (the majority

of these alignments in any case have >= 75% identity across their entire length), we predicted 1077 additional lncRNAs in cattle, 1401 in sheep, and 1735 in goat, although only 44 in human (Table 2 and Table S18). That comparatively few ruminant lncRNAs are recognisable at the sequence level in humans (and vice versa) is consistent with the rapid turnover of the lncRNA repertoire between species [65]. In the case of the goat, the number of new lncRNAs predicted by this approach is > 50% the number captured (and shortlisted) using goat-specific RNA-seq (Figure 2). This suggests that for the purposes of lncRNA detection, datasets from related species can help overcome limitations of sequencing breadth and depth. This is even apparent with comparatively large datasets – the sheep RNA-seq, for instance, spans more tissues and developmental stages than goat, but in absolute terms, it still fails to generate assemblies of many lncRNAs.

Many of the sheep lncRNAs inferred by synteny – which could not be fully assembled from the RNA-seq reads – are nevertheless detectably expressed

To determine the expression level of the sheep lncRNAs, we utilised a subset of 71 high-depth (>100 million reads) RNA-seq libraries from the sheep expression atlas [45]. This subset constitutes a set of 11 transcriptionally rich tissues (bicep muscle, hippocampus, ileum, kidney medulla, left ventricle, liver, ovary, reticulum, spleen, testes, thymus), plus one cell type in two conditions (bone marrow derived macrophages, unstimulated and 7 hours after simulation with lipopolysaccharide), each of which was sequenced in up to 6 individuals (where possible, 3 adult males and 3 adult females).

For each sample, expression was quantified – as transcripts per million (TPM) – using the quantification tool Kallisto [66]. Kallisto quantifies expression by matching k-mers from the RNA-seq reads to a pre-built index of k-mers, derived from a set of reference transcripts. For sheep, we supplemented the complete set of Oar v3.1 reference transcripts (n=28,828

transcripts, representing 26,764 genes) both with the shortlist of 11,646 novel lncRNAs (each of which is a single-transcript gene model) (Table S1), and those lncRNAs assembled from either human, goat and cattle (respectively, 18, 164 and 1219 lncRNAs), whose presence was predicted in sheep by mapping the transcript to a conserved genomic region (Table S18).

Of these 13,047 novel lncRNAs, 8826 were detected at a level of TPM > 1 in at least one of the 71 adult samples, including 14 of the human transcripts (78%), 128 of the goat transcripts (78%), and 772 of the cattle transcripts (63%) (Table S19). At a depth of coverage of 100 million reads, we would expect to detect transcripts reproducibly at between 0.01 and 0.1 TPM if they are expressed in all libraries derived from the same tissue/cell type. Indeed, of the 13,047 total novel lncRNAs, 5353 (41%) were detected with at least one paired-end read in all 6 replicates of the tissue in which it is most highly expressed (Table S19). Those lncRNAs derived from goat and cattle transcripts are similarly reproducible: 83 (51%) of the goat transcripts were detected with at least one paired-end read in all 6 replicates of its most expressed tissue, as were 570 (47%) of the cattle transcripts, and 7 (39%) of the human transcripts (Table S19).

By extension, we can consider sheep, cattle and human lncRNA to be goat lncRNA, and create a Kallisto index containing candidate lncRNAs extracted from the goat genome after mapping sheep and cattle transcripts. Using such a Kallisto index (which contains the 2657 shortlisted goat lncRNAs (Table S2), 507 sheep lncRNAs, 1213 cattle lncRNAs, and 15 human lncRNAs), 1478 (34%) of a total set of 4392 candidate goat lncRNAs were reproducibly detected (> 0.01 TPM) in all 4 of the goats sampled (Table S20). Hence, data from the sheep expression atlas can be used to provide additional functional annotation of the goat genome, despite the much lower number of tissue samples relative to sheep.

In general, lncRNA expression is low: 12,325 sheep lncRNAs (94% of the total) have a mean TPM, across all 71 samples, below 10. The mean and median maximum TPM for each

lncRNA across the total sheep dataset was 18.4 and 2.2 TPM, respectively (Table S19). Other reports have described pervasive, but low-level, mammalian lncRNA transcription [12], and – given the mean TPM exceeds the median – a high degree of lncRNA tissue-specificity [67-69]. Indeed, for those lncRNAs detected at > 1 TPM, the average value of *tau* – a scalar measure of expression breadth bound between 0 (for housekeeping genes) and 1 (for genes expressed in one sample only) [70] (see Methods) – is 0.66. Although most of the lncRNAs (n = 4972, 64% of the 7809 lncRNAs with average TPM > 1 in at least one tissue) have idiosyncratic 'mixed expression' profiles (see Methods), 1339 lncRNAs (17%) are nevertheless detected at an average TPM > 1 in all 13 tissues (Table S19). Many are enriched in specific tissues, with 904 (12%) lncRNAs exhibiting a testes-specific expression pattern, consistent with a previous study identifying numerous lncRNAs involved in ovine testicular development and spermatogenesis [71].

Few lncRNAs are fully captured by biological replicates of the same RNA-seq library

In the largest assembly of predicted lncRNAs, from humans, the transfrags (transcript fragments) assembled from 7256 RNA-seq libraries were consolidated into 58,648 candidate lncRNAs [72]. Before assembling transfrags, machine learning methods were employed to filter, from each library, any library-specific background noise (genomic DNA contamination and incompletely processed RNA). Filtered libraries were then merged before assembling the final gene models, in effect pooling together transfrags (which may be partial or full-length transcripts) from all possible libraries. Consequently, a given set of transfrags can be assembled into a consensus transcript for a lncRNA, but that consensus transcript might not actually exist in any one cellular source. The only unequivocal means to confirm the full length expression would be to clone the full length cDNA. However, additional confidence can be obtained by increasing the depth of coverage in the same tissue/cell type in a technical

296 replicate. In the sheep expression atlas, 31 diverse tissues/cell types were sampled in each of
297 6 individual adults (3 females, 3 males, all unrelated virgin animals approximately 2 years of
298 age). By taking a subset of 31 common tissues per individual, each of the 6 adults was
299 represented by ~0.75 billion reads.
300 In a typical lncRNA assembly pipeline, read alignments from all individuals are merged, to
301 maximise the number of candidate gene models (using, for instance, StringTie --merge; see
302 Methods). With n = 6 adults (and ~0.75 billion reads per adult), there are 2^n-1 = 63 possible
303 combinations of data for which GTFs can be made with StringTie --merge. The
304 reproducibility of each shortlisted lncRNA, in terms of the number of GTFs it is
305 reconstructed in, is shown in Table S21. The GTFs themselves are available as Dataset S1
306 (available via the University of Edinburgh DataShare portal;
307 http://dx.doi.org/10.7488/ds/2284).
308 Only 812 of the 12,296 sheep lncRNRAs (6.6%) could be fully reconstructed by any of the
309 63 GTF combinations (Table S21). One caveat in this assessment is that these sheep libraries
310 are exclusively from adults. Many of the 12,296 lncRNA models may instead be expressed
311 during embryonic development. There is evidence of extensive embryonic lncRNA
312 expression in human [15, 73] and mouse [16, 74]. The lack of embryonic tissues could also
313 explain why fewer lncRNAs were assembled in goat. Nevertheless, when considering all 429
314 RNA-seq libraries in the sheep expression atlas (i.e. including non-adult samples), there are
315 only, on average, 29 libraries (7%) in which any individual lncRNA can be fully
316 reconstructed (Figure 3 and Table S22).
317 In many cases, full-length sheep lncRNAs cannot be reconstructed using all reads sequenced
318 from a given individual. For instance, the known lncRNA ENSOARG00000025201 is
319 reconstructed by 28 of the 63 possible GTFs, but none of these GTFs was built using reads

from only one individual (Table S21). Only 189 lncRNAs (1.5%) were fully reconstructed in all 63 possible GTFs. Notably, 154 of these are known Ensembl lncRNAs (Table S21).

lncRNAs are enriched in the vicinity of co-expressed protein-coding genes

Enhancer sequences positively modulate the transcription of nearby genes (see reviews [75, 76]), and may be the evolutionary origin of a fraction of these lncRNAs (as suggested by [77, 78]), including a novel class of enhancer-transcribed ncRNAs, enhancer (eRNAs), which – although a distinct subset – are arbitrarily classified as lncRNAs [79]. eRNAs are likely to be co-expressed with protein-coding genes in their immediate genomic vicinity.

To identify co-regulated sets of protein-coding and non-coding loci, we performed network cluster analysis of the sheep expression level dataset (Table S19) using the Markov clustering (MCL) algorithm [80], as implemented by Graphia Professional (Kajeka Ltd., Edinburgh, UK) (see Methods) [81, 82]. To reduce noise, only those novel lncRNAs with reproducible expression (that is, having > 0.01 TPM in every replicate of the tissue in which it is most highly expressed) are included in this analysis (n = 5353). The resulting graph contained only genes with tightly correlated expression profiles (Pearson's $r \geq 0.95$) (Figure 4) and was highly structured, organised into clusters of genes with a tissue or cell-type specific expression profile (Table S23).

We expect that for a given cluster of co-expressed genes (which contains *x* lncRNAs and *y* protein-coding genes, each on chromosome *z*), the distance between an enhancer-derived lncRNA and the nearest protein-coding gene should be significantly shorter than the distance between that lncRNA and a random subset of protein-coding genes. For the purposes of this test, each random subset, of size *y*, is drawn from the complete set of protein-coding genes on the same chromosome *z* (that is, the same chromosome as the lncRNA), irrespective of strand

and their degree of co-expression with the lncRNA. The significance of any difference in distance was then assessed using a randomisation test (see Methods).

Of the 5353 lncRNAs included in the analysis, 1351 (25%) were found on the same chromosome as a highly co-expressed protein-coding gene (Table S24), with 252 of these (19%) significantly closer to the co-expressed gene than to randomly selected genes from the same chromosome ($p < 0.05$; Table S25).

Even where the lncRNA is reproducibly expressed in each of 6 animals, there is still substantial noise in the expression estimates with compromises co-expression analysis. We therefore calculated the Pearson's *r* between the expression profile of each reproducibly expressed lncRNA and its nearest protein-coding gene (which may overlap it), located both 5' and 3' on the sheep genome (Table S26). The distance to the nearest gene correlates negatively with the absolute value of Pearson's *r*, both for genes upstream (*rho* = -0.19, $p < 2.2x10^{-16}$) and downstream (*rho* = -0.21, $p < 2.2x10^{-16}$) of the lncRNA (Table S26). This suggests that, in general, the expression profile of a lncRNA is more similar to nearer than more distant protein-coding genes. Using a variant of the above randomisation test, we also tested whether the absolute value of Pearson's *r*, when correlating the expression profiles of the lncRNA and its nearest protein-coding gene, was significantly greater than the value of *r* obtained when correlating the lncRNA with 1000 random protein-coding genes drawn from the same chromosome. For this test, analysis was restricted to those lncRNAs on complete chromosomes rather than the smaller unplaced scaffolds. 27% of lncRNA had a Pearson correlation of > 0.5 with either the nearest upstream or downstream gene, and in around 20% of cases, correlation was significantly different ($p < 0.05$) from the average correlation with the random set (Table S26).

Conclusion

Comparative analysis of lncRNAs assembled using RNA-seq data from several closely related species – sheep, goat and cattle – demonstrates that for the *de novo* assembly of lncRNAs requires very high-depth RNA-seq datasets with a large number of replicates (> 6 replicates per sample, each sequencing >> 100 million reads). The transcription of many lncRNAs identified by this cross-species approach is conserved, effectively validating their existence. We identified a subset of lncRNAs in close proximity to protein-coding genes with which they are strongly co-expressed, consistent with the evolutionary origin of some ncRNAs in enhancer sequences. Conversely, the majority of lncRNA do not share transcriptional regulation with neighbouring protein-coding genes. Overall, alongside substantially expanding the lncRNA repertoire for several livestock species, we demonstrate that the conventional approach to lncRNA detection – that is, species-specific *de novo* assembly – can be reliably supplemented by data from related species.

Materials and Methods

Sheep RNA-sequencing data

We have previously created an expression atlas for the domestic sheep [45], using RNA-seq data largely collected from adult Texel x Scottish Blackface (TxBF) sheep. Experimental protocols for tissue collection, cell isolation, RNA extraction, library preparation, RNA sequencing and quality control are as previously described [45], and independently available on the FAANG Consortium website (http://ftp.faang.ebi.ac.uk/ftp/protocols). All RNA-seq libraries were prepared by Edinburgh Genomics (Edinburgh Genomics, Edinburgh, UK) and sequenced using the Illumina HiSeq 2500 sequencing platform (Illumina, San Diego, USA).

394 The majority of these libraries were sequenced to a depth of >25 million paired-end reads per
395 sample using the Illumina TruSeq mRNA library preparation protocol (polyA-selected)
396 (Illumina; Part: 15031047, Revision E). A subset of 11 transcriptionally rich 'core' tissues
397 (bicep muscle, hippocampus, ileum, kidney medulla, left ventricle, liver, ovary, reticulum,
398 spleen, testes, thymus), plus one cell type in two conditions (bone marrow derived
399 macrophages (BMDMs), unstimulated and 7 hours after simulation with lipopolysaccharide
400 (LPS)), were sequenced to a depth of >100 million paired-end reads per sample using the
401 Illumina TruSeq total RNA library preparation protocol (rRNA-depleted) (Illumina; Part:
402 15031048, Revision E).
403 Sample metadata for all tissue and cell samples are deposited in the EBI BioSamples database
404 under submission identifier GSB-718
405 (https://www.ebi.ac.uk/biosamples/groups/SAMEG317052). The raw read data, as .fastq
406 files, are deposited in the European Nucleotide Archive (ENA) under study accession
407 PRJEB19199 (http://www.ebi.ac.uk/ena/data/view/PRJEB19199).
408
409 ***Goat RNA-sequencing data***
410 All RNA-seq libraries for goat were prepared by Edinburgh Genomics (Edinburgh Genomics,
411 Edinburgh, UK) (as above) and sequenced using the Illumina HiSeq 4000 sequencing
412 platform (Illumina, San Diego, USA). These libraries were sequenced to a depth of >30
413 million paired-end reads per sample using the Illumina TruSeq mRNA library preparation
414 protocol (polyA-selected) (Illumina; Part: 15031047, Revision E). Sample metadata for all
415 tissue and cell samples are deposited in the EBI BioSamples database under submission
416 identifier GSB-2131 (https://www.ebi.ac.uk/biosamples/groups/SAMEG330351). The raw
417 read data, as .fastq files, are deposited in the ENA under study accession PRJEB23196
418 (http://www.ebi.ac.uk/ena/data/view/PRJEB23196).

Identifying candidate lncRNAs in sheep and goats

We have previously described an RNA-seq processing pipeline for sheep [45] – using the HISAT2 aligner [50] and StringTie assembler [51] – for generating a uniform, non-redundant set of *de novo* assembled transcripts. The same pipeline is applied to the goat RNA-seq data. This pipeline culminates in a single file per species, merged.gtf; that is, the output of StringTie --merge, which collates every transcript model from the 54 goat assemblies (each assembly being both individual- and tissue-specific), and 429 of the 441 assemblies within the sheep expression atlas [45] (12 sheep libraries were not used for this purpose as they were replicates of pre-existing bone marrow-derived macrophage libraries, prepared using an mRNA-seq rather than a total RNA-seq protocol). Not all transcript models in either GTF will be stranded. This is because HISAT2 infers the transcription strand of a given transcript by reference to its splice sites; this is not possible for single exon transcripts, which are un-spliced.

The GTF was parsed to distinguish candidate lncRNAs from assembly artefacts, and from other RNAs, by applying the filter criteria of Ilott, *et al.* [52], excluding gene models that (a) were < 200bp in length, (b) overlapped (by >= 1bp on the same strand) any coordinates annotated as 'protein-coding' or 'pseudogene' (this classifications are explicitly stated in the Ensembl-hosted Oar v3.1 annotation and assumed true of all gene models in the ARS1 annotation), or (c) were associated with multiple transcript models (which are more likely to be spurious). For single-exon gene models, we used a more conservative length threshold of 500bp – the lower threshold of 200bp could otherwise be met by a single pair of reads. We further excluded any novel gene model that was previously considered protein-coding in each species' expression atlas (as described in [45]); these models contain an ORF encoding a peptide homologous to a ruminant protein in the NCBI nr database [45]. These criteria establish longlists of 30,677 candidate sheep lncRNAs (14,862 of which are multi-exonic)

and 7671 candidate goat lncRNAs (3289 of which are multi-exonic). The sheep genome, Oar v3.1, already contains 1858 lncRNA models, of which the StringTie assembly precisely reconstructs 1402 (75%). Despite this pre-existing support, these models were included on the sheep longlist for independent verification. The goat genome, by contrast, was annotated with a focus on protein-coding gene models [83], by consolidating protein and cDNA alignments – from exonerate [84] and tblastn [56] – with the annotation tool EVidence Modeller (EVM) [85]. Consequently, there are no unambiguous lncRNAs in the associated GTF (ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/001/704/415/GCF_001704415.1_ARS1/GCF_001704415.1_ARS1_genomic.gff.gz, accessed 23rd October 2017) (unlike the Ensembl-hosted sheep annotation, the goat annotation is currently only available via NCBI).

Each longlist of candidates was assessed for coding potential using three different tools: CPAT v1.2.3 [54], which assigns coding probabilities to a given sequence based on differential hexamer usage [86] and Fickett TESTCODE score [87], PLEK v1.2, a support vector machine classifier utilising k-mer frequencies [55], and CPC v0.9-r2 [53], which was used in conjunction with the non-redundant sequence database, UniRef90 (the Uniref Reference Cluster, a clustered set of sequences from the UniProt KnowledgeBase that constitutes comprehensive coverage of sequence space at a resolution of 90% identity) [88, 89] (ftp://ftp.uniprot.org/pub/databases/uniprot/uniref/uniref90/uniref90.fasta.gz, accessed 18th August 2017). CPC scores putatively coding sequences positively and non-coding sequences negatively. We retained only those sequences with a CPC score < -0.5 (consistent with previous studies [31, 90]) and a CPAT probability < 0.58 (after creating sheep-specific coding and non-coding CPAT training data, from Oar v3.1 CDS and ncRNA, this cut-off is the intersection of two receiver operating characteristic curves, obtained using the R package

468 ROCR [91]; this cut-off is also used for the goat data, as there are insufficient non-coding
469 training data for this species).
470 For each remaining gene model, we concatenated its exon sequence and identified the longest
471 ORF within it. Should CPC, CPAT or PLEK make a false positive classification of 'non-
472 coding', this translated ORF was considered the most likely peptide encoded by the gene.
473 Gene models were further excluded if the translated ORF (a) contained a protein domain,
474 based on a search by HMMER v3.1b2 [57] of the Pfam database of protein families, v31.0
475 [60], with a threshold E-value of $1x10^{-5}$, or (b) shared homology with a known peptide in the
476 Swiss-Prot March 2016 release [58, 59], based on a search with BLAST+ v2.3.0 [56]: blastp
477 with a threshold E-value of $1x10^{-5}$. Shortlists of 12,296 (sheep) and 2657 (goat) candidate
478 lncRNAs – each with three independent 'non-coding' classifications and no detectable blastp
479 and HMMER hits – are given in Tables S1 and S2, respectively.
480
481 ***Classification of lncRNAs***
482 Using the set of Oar v3.1 transcription start sites (TSS), obtained from Ensembl BioMart
483 [92], and the set of ARS1 gene start sites
484 (ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/001/704/415/GCF_001704415.1_ARS1/GCF_0
485 01704415.1_ARS1_genomic.gff.gz, accessed 23rd October 2017), we classified novel
486 candidate lncRNAs for each species in the manner of [93], as either (a) sense or antisense (if
487 the coordinates of the lncRNA overlap, or are encapsulated by, a known gene on the same, or
488 opposite, strand), (b) up- or downstream, and on the same or opposite strand (if < 5kb from
489 the nearest TSS), or (c) intergenic (if ≥ 5kb, 10kb, 20kb, 50kb, 100kb, 500kb or 1 Mb from
490 the nearest TSS, irrespective of strand). The HISAT2/StringTie pipeline, used to generate
491 these transcript models, cannot infer the transcription strand in all cases, particularly for
492 single-exon transcripts. Accordingly, some lncRNAs will overlap the coordinates of a known

493 gene, but its strandedness with respect to that gene – whether it is sense or antisense – will be
494 unknown.
495
496 ***Conservation of lncRNAs in terms of sequence***
497 To assess the sequence-level conservation of sheep and goat lncRNA transcripts, we obtained
498 human lncRNA sequences from two databases, NONCODE v5 [62]
499 (http://www.noncode.org/datadownload/NONCODEv5_human.fa.gz, accessed 27th
500 September 2017) and lncRNAdb v2.0 [63]
501 (http://www.lncrnadb.com/media/cms_page_media/10651/Sequences_lncrnadb_27Jan2015.c
502 sv, accessed 27th September 2017) (which contain 172,216 and 152 lncRNAs, respectively).
503 A previous study of lncRNAs in cattle [31] also generated a conservative set of 9778
504 lncRNAs, all of which were detectably expressed in at least one of 18 tissues (read count >
505 25 in each of three replicates per tissue). These sets of sequences constitute three independent
506 BLAST databases. For each sheep and goat lncRNA, blastn searches [56] were made against
507 each database using an arbitrarily high E-value of 10, as substantial sequence-level
508 conservation was not expected.
509
510 ***Conservation of lncRNAs in terms of synteny***
511 For each of the human (GRCh38.p10), sheep (Oar v3.1), cattle (UMD3.1) and goat (ARS1)
512 reference genomes, we established those regions in each pairwise comparison where gene
513 order is conserved, obtaining reference annotations from Ensembl BioMart v90 [92] (sheep,
514 cattle and human) and NCBI (goat;
515 ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/001/704/415/GCF_001704415.1_ARS1/GCF_00
516 1704415.1_ARS1_genomic.gff.gz, accessed 27th September 2017). By advancing a sliding
517 window across each chromosome gene-by-gene from the 5' end, we identified the first

upstream and first downstream gene of each focal gene, irrespective of strand. For the purpose of this analysis, the first and last genes on each chromosome are excluded, having no upstream or downstream neighbour, respectively. For each pairwise species comparison, we then determined which set of blocks were present in both – that is, where the HGNC symbols for upstream gene/focal gene/downstream gene were identical. These syntenic blocks, of three consecutive genes each, are regions in the genome where gene order is conserved both up- and downstream of a focal gene: between sheep and cattle, there are 2927 regions (comprising 5601 unique genes); sheep and goat, 2038 regions (3883 unique genes); cattle and goat, 2982 regions (5258 unique genes); sheep and human, 380 regions (930 unique genes); goat and human, 527 regions (1262 unique genes); cattle and human, 443 regions (1063 unique genes). If in each syntenic block a lncRNA was found between the upstream and focal gene, or the focal and downstream gene, in only one of the two species, a global alignment was made between the transcript and the intergenic region of the corresponding species. Alignments were made using the Needleman-Wunsch algorithm, as implemented by the 'needle' module of EMBOSS v6.6.0 [94], with default parameters. By effectively treating lncRNA transcripts as if they were CAGE tags (that is, short reads of 20-50 nucleotides [95]), we considered successful alignments to be those containing one or more consecutive runs of 20 identical residues, without gaps (the majority of these alignments in any case have >= 75% identity across the entire length of the transcript (Table S18)). The probability that a transcript randomly matches 20 consecutive residues, within a pre-defined region, is extremely low.

For successful alignments, the target sequence (that is, an extract from the intergenic region) was considered a novel lncRNA. For this analysis, the sheep and goat lncRNAs used are those from their respective shortlists (Tables S1 and S2). lncRNA locations in other species are obtained from previous studies applying similarly conservative classification criteria. For

cattle, 9778 lncRNAs were obtained [31], each of which were >200bp, considered non-coding by the classification tools CPC [53] and CNCI [96], lacked sequence similarity to the NCBI nr [45] and Pfam databases [60], and had a normalised read count > 25 in at least 2 of 3 replicates per tissue for 18 tissues. For human, 17,134 lncRNAs were obtained [72], each of which were assembled from >250bp transfrags, considered non-coding by the classification tool CPAT [54], lacked sequence similarity to the Pfam database [60], and had active transcription confirmed by intersecting intervals surrounding the transcriptional start site with chromatin immunoprecipitation and sequencing (ChIP-seq) data from 13 cell lines.

Expression level quantification

For the 11 'core' tissues of the sheep expression atlas, plus unstimulated and LPS-stimulated BMDMs (detailed in S2 Table of [45] and available under ENA accession PRJEB19199), expression was quantified using Kallisto v0.43.0 [66] with a k-mer index (k=31) derived after supplementing the Oar v3.1 reference transcriptome with the shortlist of 11,646 novel sheep lncRNA models (Table S1) and those lncRNAs assembled in either human (n = 18), goat (n=164), or cattle (n=1219), and which map to a conserved region of the sheep genome (Table S15). Oar v3.1 transcripts were obtained from Ensembl v90 [92] in the form of separate files for 22,823 CDS (ftp://ftp.ensembl.org/pub/release-90/fasta/ovis_aries/cds/Ovis_aries.Oar_v3.1.cds.all.fa.gz, accessed 27th September 2017) and 6005 ncRNAs (ftp://ftp.ensembl.org/pub/release-90/fasta/ovis_aries/ncrna/Ovis_aries.Oar_v3.1.ncrna.fa.gz, accessed 27th September 2017). An equivalent set of expression estimates was made for goat, across the 21 tissues and cell types of the goat expression atlas (i.e., 54 RNA-seq libraries available under ENA accession PRJEB23196). 47,193 transcripts, from assembly ARS1, were obtained from NCBI (ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/001/704/415/GCF_001704415.1_ARS1/GCF_0

568 01704415.1_ARS1_rna.fna.gz, accessed 27th September 2017), and supplemented both with
569 the shortlist of 2657 novel goat lncRNA models (Table S2), and those lncRNAs assembled in
570 human (n = 15), sheep (n = 507), or cattle (n= 1213) (Table S15). After quantification in each
571 species, transcript-level abundances were summarised to the gene-level.
572
573 ***Categorisation of expression profiles***
574 Expression levels were categorised in the manner of the Human Protein Atlas [97], and as
575 previously employed in the Sheep Gene Expression Atlas [45]. Each gene is considered to
576 have either no expression (average TPM < 1, a threshold chosen to minimise the influence of
577 stochastic sampling), low expression (10 > average TPM >= 1), medium expression (50 >
578 average TPM > 10), or high expression (average TPM >= 50). Two sample specificity indices
579 were calculated for each gene, as in [45]: firstly, *tau*, a scalar measure of expression breadth
580 bound between 0 (for housekeeping genes) and 1 (for genes expressed in one sample only)
581 [70], and secondly, the mean TPM (across all samples) divided by the median TPM (across
582 all tissues). Genes with greater sample specificity will have a more strongly skewed
583 distribution (i.e. a higher mean and a lower median), and so the larger the ratio, the more
584 sample-specific the expression. To avoid undefined values, should median TPM be 0, it is
585 considered instead to be 0.01.
586 Each gene is also assigned one or more categories, to allow an at-a-glance overview of its
587 expression profile: (a) 'tissue enriched' (expression in one tissue at least five-fold higher than
588 all other tissues ['tissue specific' if all other tissues have 0 TPM]), (b) 'tissue enhanced'
589 (five-fold higher average TPM in one or more tissues compared to the mean TPM of all
590 tissues with detectable expression [this category is mutually exclusive with 'tissue enriched'),
591 (c) 'group enriched' (five-fold higher average TPM in a group of two or more tissues
592 compared to all other tissues ('groups' are analogous to organ systems, and are as described

593 in the sheep expression atlas [45]), (d) mixed expression (detected in one or more tissues and
594 neither of the previous categories), (e) 'expressed in all' (>= 1 TPM in all tissues), and (f)
595 'not detected' (< 1 TPM in all tissues).
596
597 ***Network analysis***
598 Network analysis of the sheep expression level data was performed using Graphia Professional
599 (Kajeka Ltd, Edinburgh, UK), a commercial version of BioLayout *Express*^{3D} [81, 82]. A correlation
600 matrix was built for each gene-to-gene comparison, which was then filtered by removing all
601 correlations below a given threshold (Pearson's $r < 0.95$). A network graph was then constructed by
602 connecting nodes (genes) with edges (correlations above the threshold). The local structure of the
603 graph – that is, clusters of co-expressed genes (detailed in Table S23) – was interpreted by applying
604 the Markov clustering (MCL) algorithm [80] at an inflation value (which determines cluster
605 granularity) of 2.2.
606
607 ***Enrichment of lncRNAs in the vicinity of protein-coding genes***
608 To test whether lncRNAs co-expressed with protein-coding genes are more likely to be closer
609 to them (from which we can infer they are more likely to have been derived from an enhancer
610 sequence affecting that protein-coding gene), we employed a randomisation test in the
611 manner of [98]. We first obtained clusters of co-expressed genes from a network graph of the
612 sheep expression level dataset (see above). We then calculated q, the number of times the
613 distance between each lncRNA and the nearest protein-coding gene within the same cluster
614 was higher than the distance between each lncRNA and the nearest gene within $s = 1000$
615 randomly selected, equally sized, subsets of protein-coding genes, drawn from the same
616 chromosome as each lncRNA. Letting $r = s-q$, then the p-value of this test is $r+1/s+1$.
617

Declarations

Acknowledgements

The authors would like to thank the farm staff at Dryden farm and members of the sheep tissue collection team from The Roslin Institute and R(D)SVS who were involved in tissue collections for the sheep gene expression atlas project. Rachel Young and Lucas Lefevre isolated the bone marrow derived macrophages and Zofia Lisowski provided technical assistance with collection and post mortem for the goat samples. Technical expertise for dissection of the sheep brain samples was provided by Fiona Houston and heart samples by Kim Summers and Hiu-Gwen Tsang. The authors are also grateful for the support of the FAANG Data Coordination Centre in the upload and archiving of the sample data and metadata.

Funding

This work was supported by a Biotechnology and Biological Sciences Research Council (BBSRC; www.bbsrc.ac.uk) grant BB/L001209/1 ('Functional Annotation of the Sheep Genome') and Institute Strategic Program grants 'Farm Animal Genomics' (BBS/E/D/2021550), 'Blueprints for Healthy Animals' (BB/P013732/1) and 'Transcriptomes, Networks and Systems' (BBS/E/D/20211552). The goat RNA-seq data was funded by the Roslin Foundation (www.roslinfoundation.com) which also supported SJB. CM was supported by a Newton Fund PhD studentship (www.newtonfund.ac.uk). Edinburgh Genomics is partly supported through core grants from the BBSRC (BB/J004243/1), National Research Council (NERC; www.nationalacademies.org.uk/nrc) (R8/H10/56), and Medical Research Council (MRC; www.mrc.ac.uk) (MR/K001744/1). The funders had no role in

study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Ethics approval and consent to participate

Approval was obtained from The Roslin Institute's and the University of Edinburgh's Protocols and Ethics Committees. All animal work was carried out under the regulations of the Animals (Scientific Procedures) Act 1986.

Competing interests

The authors declare they have no competing interests.

Data availability

The raw RNA-sequencing data are deposited in the European Nucleotide Archive (ENA) under study accessions PRJEB19199 (sheep) and PRJEB23196 (goat). Sample metadata for all tissue and cell samples, prepared in accordance with FAANG consortium metadata standards, are deposited in the EBI BioSamples database under group identifiers SAMEG317052 (sheep) and SAMEG330351 (goat). All experimental protocols are available on the FAANG consortium website at http://ftp.faang.ebi.ac.uk/ftp/protocols.

Tables

Species 1	Species 2	No. of syntenic blocks (i.e. three conserved consecutive genes)	No. of unique protein-coding genes in the set of syntenic blocks	Total no. of positionally conserved lncRNAs in the set of syntenic blocks (in either the up- or downstream position)	% of syntenic blocks with at least one positionally conserved lncRNA
sheep	cattle	2927	5601	280	9.57
sheep	goat	2038	3883	82	4.02
sheep	human	380	930	8	2.11
goat	cattle	2982	5258	169	5.67
goat	human	527	1262	2	0.38
cattle	human	443	1063	5	1.13

Table 1. Comparatively few lncRNAs appear positionally conserved, suggesting minimal overlap between each species' set of assembled transcripts. This suggests that those lncRNAs expected to be found at a given genomic location are captured in only one species, not both, consistent with the stochastic sampling of lncRNAs by RNA-seq libraries.

Species 1 (in which lncRNA is captured by RNA-seq libraries)	Species 2 (in which lncRNA can be inferred)	No. of lncRNA models detected within a region of conserved synteny between species 1 and 2, but not captured by the RNA-seq libraries of species 2	No. of lncRNA models from species 1 mapped to the genome of species 2	% of lncRNA models detected by direct genome mapping	Number of intergenic regions in the syntenic blocks conserved between these two species	% of intergenic regions in which a lncRNA from species 1 is inferred in species 2
	goat	2593	1213	46.78	5964	20.34
cattle	human	163	20	12.27	886	2.26
	sheep	2939	1219	41.48	5854	20.82
goat	cattle	2593	286	11.03	5964	4.8

	human	76	9	11.84	1054	0.85
	sheep	991	164	16.55	4076	4.02
	cattle	163	16	9.82	886	1.81
human	goat	76	15	19.74	1054	1.42
	sheep	93	18	19.35	760	2.37
	cattle	2939	775	26.37	5854	13.24
sheep	goat	991	507	51.16	4076	12.44
	human	93	15	16.13	760	1.97

Table 2. lncRNA transcripts assembled using the RNA-seq libraries of only one species can in many cases be directly mapped to the genome of another species, assuming the lncRNA is located within a region of conserved synteny.

References

1. Ponting CP, Oliver PL, Reik W: **Evolution and functions of long noncoding RNAs.** *Cell* 2009, **136:**629-641.
2. Engreitz JM, Ollikainen N, Guttman M: **Long non-coding RNAs: spatial amplifiers that control nuclear structure and gene expression.** *Nat Rev Mol Cell Biol* 2016, **17:**756-770.
3. Rinn JL, Chang HY: **Genome regulation by long noncoding RNAs.** *Annu Rev Biochem* 2012, **81:**145-166.
4. Chen J, Xue Y: **Emerging roles of non-coding RNAs in epigenetic regulation.** *Sci China Life Sci* 2016, **59:**227-235.
5. Kung JT, Colognori D, Lee JT: **Long noncoding RNAs: past, present, and future.** *Genetics* 2013, **193:**651-669.
6. Quinn JJ, Chang HY: **Unique features of long non-coding RNA biogenesis and function.** *Nat Rev Genet* 2016, **17:**47-62.
7. Villegas VE, Zaphiropoulos PG: **Neighboring Gene Regulation by Antisense Long Non-Coding RNAs.** *International Journal of Molecular Sciences* 2015, **16:**3251-3266.
8. Goff LA, Rinn JL: **Linking RNA biology to lncRNAs.** *Genome Research* 2015, **25:**1456-1465.
9. Cech TR, Steitz JA: **The noncoding RNA revolution-trashing old rules to forge new ones.** *Cell* 2014, **157:**77-94.
10. Kapranov P, St Laurent G, Raz T, Ozsolak F, Reynolds CP, Sorensen PHB, Reaman G, Milos P, Arceci RJ, Thompson JF, Triche TJ: **The majority of total nuclear-encoded non-ribosomal RNA in a human cell is 'dark matter' un-annotated RNA.** *BMC Biology* 2010, **8:**149.
11. van Bakel H, Nislow C, Blencowe BJ, Hughes TR: **Most "Dark Matter" Transcripts Are Associated With Known Genes.** *PLOS Biology* 2010, **8:**e1000371.
12. Kornienko AE, Dotter CP, Guenzl PM, Gisslinger H, Gisslinger B, Cleary C, Kralovics R, Pauler FM, Barlow DP: **Long non-coding RNAs display higher natural expression variation than protein-coding genes in healthy humans.** *Genome Biology* 2016, **17:**14.
13. Mercer TR, Dinger ME, Sunkin SM, Mehler MF, Mattick JS: **Specific expression of long noncoding RNAs in the mouse brain.** *Proc Natl Acad Sci U S A* 2008, **105:**716-721.
14. Gloss BS, Dinger ME: **The specificity of long noncoding RNA expression.** *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms* 2016, **1859:**16-22.
15. Qiu JJ, Ren ZR, Yan JB: **Identification and functional analysis of long non-coding RNAs in human and mouse early embryos based on single-cell transcriptome data.** *Oncotarget* 2016, **7:**61215-61228.
16. Zhang K, Huang K, Luo Y, Li S: **Identification and functional analysis of long non-coding RNAs in mouse cleavage stage embryonic development based on single cell transcriptome data.** *BMC Genomics* 2014, **15:**845.
17. Pauli A, Valen E, Lin MF, Garber M, Vastenhouw NL, Levin JZ, Fan L, Sandelin A, Rinn JL, Regev A, Schier AF: **Systematic identification of long noncoding RNAs expressed during zebrafish embryogenesis.** *Genome Res* 2012, **22:**577-591.
18. Cabili MN, Trapnell C, Goff L, Koziol M, Tazon-Vega B, Regev A, Rinn JL: **Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses.** *Genes Dev* 2011, **25:**1915-1927.

19. Johnsson P, Lipovich L, Grander D, Morris KV: **Evolutionary conservation of long non-coding RNAs; sequence, structure, function.** *Biochim Biophys Acta* 2014, **1840:**1063-1071.
20. Andersson R, Refsing Andersen P, Valen E, Core LJ, Bornholdt J, Boyd M, Heick Jensen T, Sandelin A: **Nuclear stability and transcriptional directionality separate functionally distinct RNA species.** *Nat Commun* 2014, **5:**5336.
21. Jia H, Osak M, Bogu GK, Stanton LW, Johnson R, Lipovich L: **Genome-wide computational identification and manual annotation of human long noncoding RNA genes.** *RNA* 2010, **16:**1478-1487.
22. Maeda N, Kasukawa T, Oyama R, Gough J, Frith M, Engstrom PG, Lenhard B, Aturaliya RN, Batalov S, Beisel KW, et al: **Transcript annotation in FANTOM3: mouse gene catalog based on physical cDNAs.** *PLoS Genet* 2006, **2:**e62.
23. Sasaki YT, Sano M, Ideue T, Kin T, Asai K, Hirose T: **Identification and characterization of human non-coding RNAs with tissue-specific expression.** *Biochem Biophys Res Commun* 2007, **357:**991-996.
24. Ravasi T, Suzuki H, Pang KC, Katayama S, Furuno M, Okunishi R, Fukuda S, Ru K, Frith MC, Gongora MM, et al: **Experimental validation of the regulated expression of large numbers of non-coding RNAs from the mouse genome.** *Genome Res* 2006, **16:**11-19.
25. Katayama S, Tomaru Y, Kasukawa T, Waki K, Nakanishi M, Nakamura M, Nishida H, Yap CC, Suzuki M, Kawai J, et al: **Antisense transcription in the mammalian transcriptome.** *Science* 2005, **309:**1564-1566.
26. Mattick JS, Rinn JL: **Discovery and annotation of long noncoding RNAs.** *Nat Struct Mol Biol* 2015, **22:**5-7.
27. Balwierz PJ, Carninci P, Daub CO, Kawai J, Hayashizaki Y, Van Belle W, Beisel C, van Nimwegen E: **Methods for analyzing deep sequencing expression data: constructing the human and mouse promoterome with deepCAGE data.** *Genome Biol* 2009, **10:**R79.
28. McIntyre LM, Lopiano KK, Morse AM, Amin V, Oberg AL, Young LJ, Nuzhdin SV: **RNA-seq: technical variability and sampling.** *BMC Genomics* 2011, **12:**1-13.
29. Steijger T, Abril JF, Engstrom PG, Kokocinski F, Hubbard TJ, Guigo R, Harrow J, Bertone P: **Assessment of transcript reconstruction methods for RNA-seq.** *Nat Methods* 2013, **10:**1177-1184.
30. Hon C-C, Ramilowski JA, Harshbarger J, Bertin N, Rackham OJL, Gough J, Denisenko E, Schmeier S, Poulsen TM, Severin J, et al: **An atlas of human long non-coding RNAs with accurate 5′ ends.** *Nature* 2017, **543:**199-204.
31. Koufariotis LT, Chen YP, Chamberlain A, Vander Jagt C, Hayes BJ: **A catalogue of novel bovine long noncoding RNA across 18 tissues.** *PLoS One* 2015, **10:**e0141225.
32. Zhou ZY, Li AM, Adeola AC, Liu YH, Irwin DM, Xie HB, Zhang YP: **Genome-wide identification of long intergenic noncoding RNA genes and their potential association with domestication in pigs.** *Genome Biol Evol* 2014, **6:**1387-1392.
33. Scott EY, Mansour T, Bellone RR, Brown CT, Mienaltowski MJ, Penedo MC, Ross PJ, Valberg SJ, Murray JD, Finno CJ: **Identification of long non-coding RNA in the horse transcriptome.** *BMC Genomics* 2017, **18:**511.
34. Billerey C, Boussaha M, Esquerré D, Rebours E, Djari A, Meersseman C, Klopp C, Gautheret D, Rocha D: **Identification of large intergenic non-coding RNAs in bovine muscle using next-generation transcriptomic sequencing.** *BMC Genomics* 2014, **15:**499.

35. Liu XF, Ding XB, Li X, Jin CF, Yue YW, Li GP, Guo H: **An atlas and analysis of bovine skeletal muscle long noncoding RNAs.** *Anim Genet* 2017, **48:**278-286.
36. Weikard R, Hadlich F, Kuehn C: **Identification of novel transcripts and noncoding RNAs in bovine skin by deep next generation sequencing.** *BMC Genomics* 2013, **14:**789.
37. Yu L, Tai L, Zhang L, Chu Y, Li Y, Zhou L: **Comparative analyses of long non-coding RNA in lean and obese pig.** *Oncotarget* 2017, **8:**41440-41450.
38. Wang J, Hua L, Chen J, Zhang J, Bai X, Gao B, Li C, Shi Z, Sheng W, Gao Y, Xing B: **Identification and characterization of long non-coding RNAs in subcutaneous adipose tissue from castrated and intact full-sib pair Huainan male pigs.** *BMC Genomics* 2017, **18:**542.
39. Xia J, Xin L, Zhu W, Li L, Li C, Wang Y, Mu Y, Yang S, Li K: **Characterization of long non-coding RNA transcriptome in high-energy diet induced nonalcoholic steatohepatitis minipigs.** *Scientific Reports* 2016, **6:**30709.
40. Esteve-Codina A, Kofler R, Palmieri N, Bussotti G, Notredame C, Pérez-Enciso M: **Exploring the gonad transcriptome of two extreme male pigs with RNA-seq.** *BMC Genomics* 2011, **12:**552.
41. Engreitz JM, Haines JE, Perez EM, Munson G, Chen J, Kane M, McDonel PE, Guttman M, Lander ES: **Local regulation of gene expression by lncRNA promoters, transcription and splicing.** *Nature* 2016, **539:**452-455.
42. Derrien T, Johnson R, Bussotti G, Tanzer A, Djebali S, Tilgner H, Guernec G, Martin D, Merkel A, Knowles DG, et al: **The GENCODE v7 catalog of human long noncoding RNAs: Analysis of their gene structure, evolution, and expression.** *Genome Research* 2012, **22:**1775-1789.
43. Carninci P, Kasukawa T, Katayama S, Gough J, Frith MC, Maeda N, Oyama R, Ravasi T, Lenhard B, Wells C, et al: **The transcriptional landscape of the mammalian genome.** *Science* 2005, **309:**1559-1563.
44. Roux BT, Heward JA, Donnelly LE, Jones SW, Lindsay MA: **Catalog of Differentially Expressed Long Non-Coding RNA following Activation of Human and Mouse Innate Immune Response.** *Frontiers in Immunology* 2017, **8:**1038.
45. Clark EL, Bush SJ, McCulloch MEB, Farquhar IL, Young R, Lefevre L, Pridans C, Tsang H, Wu C, Afrasiabi C, et al: **A high resolution atlas of gene expression in the domestic sheep (Ovis aries).** *PLoS Genet* 2017, **13:**e1006997.
46. Kumar S, Stecher G, Suleski M, Hedges SB: **TimeTree: A Resource for Timelines, Timetrees, and Divergence Times.** *Mol Biol Evol* 2017, **34:**1812-1819.
47. Weikard R, Demasius W, Kuehn C: **Mining long noncoding RNA in livestock.** *Anim Genet* 2017, **48:**3-18.
48. Andersson L, Archibald AL, Bottema CD, Brauning R, Burgess SC, Burt DW, Casas E, Cheng HH, Clarke L, Couldrey C, et al: **Coordinated international action to accelerate genome-to-phenome with FAANG, the Functional Annotation of Animal Genomes project.** *Genome Biology* 2015, **16:**57.
49. Tuggle CK, Giuffra E, White SN, Clarke L, Zhou H, Ross PJ, Acloque H, Reecy JM, Archibald A, Bellone RR, et al: **GO-FAANG meeting: a Gathering On Functional Annotation of Animal Genomes.** *Animal Genetics* 2016, **47:**528-533.
50. Kim D, Langmead B, Salzberg SL: **HISAT: a fast spliced aligner with low memory requirements.** *Nat Meth* 2015, **12:**357-360.
51. Pertea M, Pertea GM, Antonescu CM, Chang T-C, Mendell JT, Salzberg SL: **StringTie enables improved reconstruction of a transcriptome from RNA-seq reads.** *Nat Biotech* 2015, **33:**290-295.

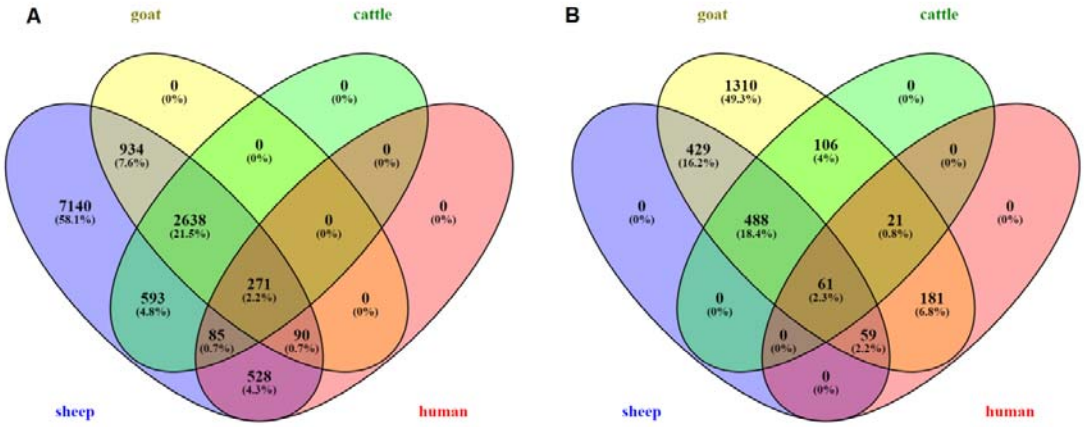
52. Ilott NE, Ponting CP: **Predicting long non-coding RNAs using RNA sequencing.** *Methods* 2013, **63:**50-59.
53. Kong L, Zhang Y, Ye Z-Q, Liu X-Q, Zhao S-Q, Wei L, Gao G: **CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine.** *Nucleic Acids Research* 2007, **35:**W345-W349.
54. Wang L, Park HJ, Dasari S, Wang S, Kocher J-P, Li W: **CPAT: Coding-Potential Assessment Tool using an alignment-free logistic regression model.** *Nucleic Acids Research* 2013.
55. Li A, Zhang J, Zhou Z: **PLEK: a tool for predicting long non-coding RNAs and messenger RNAs based on an improved k-mer scheme.** *BMC Bioinformatics* 2014, **15:**311.
56. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL: **BLAST+: architecture and applications.** *BMC Bioinformatics* 2009, **10:**421.
57. Mistry J, Finn RD, Eddy SR, Bateman A, Punta M: **Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions.** *Nucleic Acids Research* 2013, **41:**e121.
58. Boutet E, Lieberherr D, Tognolli M, Schneider M, Bairoch A: **UniProtKB/Swiss-Prot.** *Methods Mol Biol* 2007, **406:**89-112.
59. The UniProt Consortium: **UniProt: a hub for protein information.** *Nucleic Acids Res* 2015, **43:**D204-212.
60. Finn RD, Coggill P, Eberhardt RY, Eddy SR, Mistry J, Mitchell AL, Potter SC, Punta M, Qureshi M, Sangrador-Vegas A, et al: **The Pfam protein families database: towards a more sustainable future.** *Nucleic Acids Research* 2016, **44:**D279-D285.
61. Liu SJ, Nowakowski TJ, Pollen AA, Lui JH, Horlbeck MA, Attenello FJ, He D, Weissman JS, Kriegstein AR, Diaz AA, Lim DA: **Single-cell analysis of long non-coding RNAs in the developing human neocortex.** *Genome Biology* 2016, **17:**67.
62. Zhao Y, Li H, Fang S, Kang Y, wu W, Hao Y, Li Z, Bu D, Sun N, Zhang MQ, Chen R: **NONCODE 2016: an informative and valuable data source of long non-coding RNAs.** *Nucleic Acids Research* 2016, **44:**D203-D208.
63. Quek XC, Thomson DW, Maag JL, Bartonicek N, Signal B, Clark MB, Gloss BS, Dinger ME: **lncRNAdb v2.0: expanding the reference database for functional long noncoding RNAs.** *Nucleic Acids Res* 2015, **43:**D168-173.
64. Bajic VB, Tan SL, Christoffels A, Schonbach C, Lipovich L, Yang L, Hofmann O, Kruger A, Hide W, Kai C, et al: **Mice and men: their promoter properties.** *PLoS Genet* 2006, **2:**e54.
65. Necsulea A, Soumillon M, Warnefors M, Liechti A, Daish T, Zeller U, Baker JC, Grützner F, Kaessmann H: **The evolution of lncRNA repertoires and expression patterns in tetrapods.** *Nature* 2014, **505:**635.
66. Bray NL, Pimentel H, Melsted P, Pachter L: **Near-optimal probabilistic RNA-seq quantification.** *Nat Biotech* 2016, **34:**525-527.
67. Tsoi LC, Iyer MK, Stuart PE, Swindell WR, Gudjonsson JE, Tejasvi T, Sarkar MK, Li B, Ding J, Voorhees JJ, et al: **Analysis of long non-coding RNAs highlights tissue-specific expression patterns and epigenetic profiles in normal and psoriatic skin.** *Genome Biol* 2015, **16:**24.
68. Jiang C, Li Y, Zhao Z, Lu J, Chen H, Ding N, Wang G, Xu J, Li X: **Identifying and functionally characterizing tissue-specific and ubiquitously expressed human lncRNAs.** *Oncotarget* 2016, **7:**7120-7133.
69. Wu W, Wagner EK, Hao Y, Rao X, Dai H, Han J, Chen J, Storniolo AM, Liu Y, He C: **Tissue-specific Co-expression of Long Non-coding and Coding RNAs Associated with Breast Cancer.** *Sci Rep* 2016, **6:**32731.

70. Yanai I, Benjamin H, Shmoish M, Chalifa-Caspi V, Shklar M, Ophir R, Bar-Even A, Horn-Saban S, Safran M, Domany E, et al: **Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification.** *Bioinformatics* 2005, **21:**650-659.
71. Zhang Y, Yang H, Han L, Li F, Zhang T, Pang J, Feng X, Ren C, Mao S, Wang F: **Long noncoding RNA expression profile changes associated with dietary energy in the sheep testis during sexual maturation.** *Sci Rep* 2017, **7:**5180.
72. Iyer MK, Niknafs YS, Malik R, Singhal U, Sahu A, Hosono Y, Barrette TR, Prensner JR, Evans JR, Zhao S, et al: **The Landscape of Long Noncoding RNAs in the Human Transcriptome.** *Nature genetics* 2015, **47:**199-208.
73. Bouckenheimer J, Assou S, Riquier S, Hou C, Philippe N, Sansac C, Lavabre-Bertrand T, Commes T, Lemaitre JM, Boureux A, De Vos J: **Long non-coding RNAs in human early embryonic development and their potential in ART.** *Hum Reprod Update* 2016, **23:**19-40.
74. Karlic R, Ganesh S, Franke V, Svobodova E, Urbanova J, Suzuki Y, Aoki F, Vlahovicek K, Svoboda P: **Long non-coding RNA exchange during the oocyte-to-embryo transition in mice.** *DNA Res* 2017, **24:**129-141.
75. Li W, Notani D, Rosenfeld MG: **Enhancers as non-coding RNA transcription units: recent insights and future perspectives.** *Nat Rev Genet* 2016, **17:**207-223.
76. Chen H, Du G, Song X, Li L: **Non-coding Transcripts from Enhancers: New Insights into Enhancer Activity and Gene Expression Regulation.** *Genomics, Proteomics & Bioinformatics* 2017, **15:**201-207.
77. De Santa F, Barozzi I, Mietton F, Ghisletti S, Polletti S, Tusi BK, Muller H, Ragoussis J, Wei C-L, Natoli G: **A Large Fraction of Extragenic RNA Pol II Transcription Sites Overlap Enhancers.** *PLOS Biology* 2010, **8:**e1000384.
78. Kim TK, Hemberg M, Gray JM, Costa AM, Bear DM, Wu J, Harmin DA, Laptewicz M, Barbara-Haley K, Kuersten S, et al: **Widespread transcription at neuronal activity-regulated enhancers.** *Nature* 2010, **465:**182-187.
79. Natoli G, Andrau JC: **Noncoding transcription at enhancers: general principles and functional models.** *Annu Rev Genet* 2012, **46:**1-19.
80. van Dongen S, Abreu-Goodger C: **Using MCL to extract clusters from networks.** *Methods Mol Biol* 2012, **804:**281-295.
81. Freeman TC, Goldovsky L, Brosch M, van Dongen S, Maziere P, Grocock RJ, Freilich S, Thornton J, Enright AJ: **Construction, visualisation, and clustering of transcription networks from microarray expression data.** *PLoS Comput Biol* 2007, **3:**2032-2042.
82. Theocharidis A, van Dongen S, Enright AJ, Freeman TC: **Network visualization and analysis of gene expression data using BioLayout Express(3D).** *Nat Protoc* 2009, **4:**1535-1550.
83. Bickhart DM, Rosen BD, Koren S, Sayre BL, Hastie AR, Chan S, Lee J, Lam ET, Liachko I, Sullivan ST, et al: **Single-molecule sequencing and chromatin conformation capture enable de novo reference assembly of the domestic goat genome.** *Nat Genet* 2017, **49:**643-650.
84. Slater GSC, Birney E: **Automated generation of heuristics for biological sequence comparison.** *BMC Bioinformatics* 2005, **6:**31.
85. Haas BJ, Salzberg SL, Zhu W, Pertea M, Allen JE, Orvis J, White O, Buell CR, Wortman JR: **Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments.** *Genome Biol* 2008, **9:**R7.

86. Fickett JW: **Recognition of protein coding regions in DNA sequences.** *Nucleic Acids Research* 1982, **10:**5303-5318.
87. Fickett JW, Tung C-S: **Assessment of protein coding measures.** *Nucleic Acids Research* 1992, **20:**6441-6450.
88. Suzek BE, Huang H, McGarvey P, Mazumder R, Wu CH: **UniRef: comprehensive and non-redundant UniProt reference clusters.** *Bioinformatics* 2007, **23:**1282-1288.
89. Suzek BE, Wang Y, Huang H, McGarvey PB, Wu CH: **UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches.** *Bioinformatics* 2015, **31:**926-932.
90. Weikard R, Hadlich F, Kuehn C: **Identification of novel transcripts and noncoding RNAs in bovine skin by deep next generation sequencing.** *BMC Genomics* 2013, **14:**789.
91. Sing T, Sander O, Beerenwinkel N, Lengauer T: **ROCR: visualizing classifier performance in R.** *Bioinformatics* 2005, **21:**3940-3941.
92. Kinsella RJ, Kahari A, Haider S, Zamora J, Proctor G, Spudich G, Almeida-King J, Staines D, Derwent P, Kerhornou A, et al: **Ensembl BioMarts: a hub for data retrieval across taxonomic space.** *Database (Oxford)* 2011, **2011:**bar030.
93. Ma L, Bajic VB, Zhang Z: **On the classification of long non-coding RNAs.** *RNA Biology* 2013, **10:**924-933.
94. Rice P, Longden I, Bleasby A: **EMBOSS: the European Molecular Biology Open Software Suite.** *Trends Genet* 2000, **16:**276-277.
95. Takahashi H, Kato S, Murata M, Carninci P: **CAGE (cap analysis of gene expression): a protocol for the detection of promoter and transcriptional networks.** *Methods Mol Biol* 2012, **786:**181-200.
96. Sun L, Luo H, Bu D, Zhao G, Yu K, Zhang C, Liu Y, Chen R, Zhao Y: **Utilizing sequence intrinsic composition to classify protein-coding and long non-coding transcripts.** *Nucleic Acids Research* 2013, **41:**e166-e166.
97. Uhlen M, Oksvold P, Fagerberg L, Lundberg E, Jonasson K, Forsberg M, Zwahlen M, Kampf C, Wester K, Hober S, et al: **Towards a knowledge-based Human Protein Atlas.** *Nat Biotechnol* 2010, **28:**1248-1250.
98. Bush SJ, Castillo-Morales A, Tovar-Corona JM, Chen L, Kover PX, Urrutia AO: **Presence–Absence Variation in A. thaliana Is Primarily Associated with Genomic Signatures Consistent with Relaxed Selective Constraints.** *Molecular Biology and Evolution* 2014, **31:**59-69.

956 **Figures**

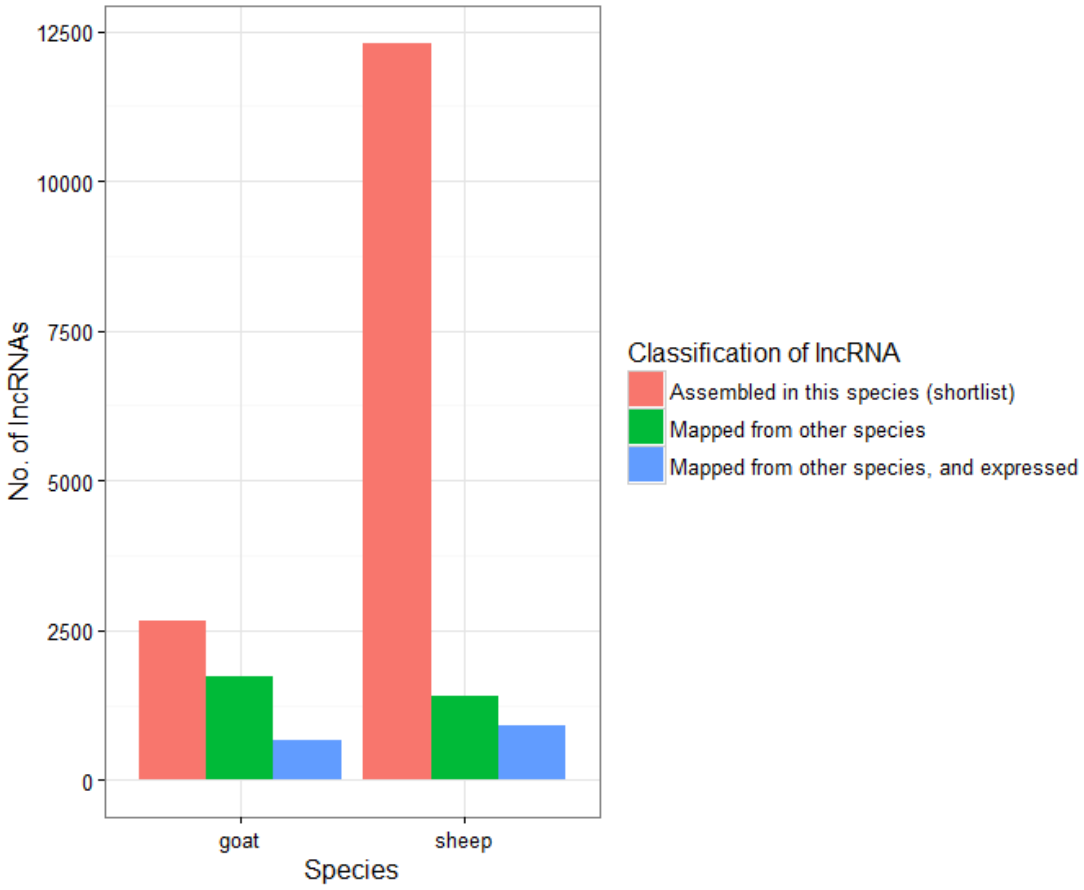
957



958

959 **Figure 1.** Minimal overlap of lncRNAs at the sequence level. Venn diagrams show the
960 number of sheep (A) or goat (B) lncRNAs that can be aligned – with an alignment of any
961 length or quality – to either shortlist of goat (A) or sheep (B) lncRNAs, and to sets of cattle
962 and human lncRNAs from previous studies. The majority (58% of sheep lncRNAs, and 49%
963 of goat lncRNAs) have no associated alignment. Alignments are detailed in Tables S9 (sheep)
964 and S10 (goat).

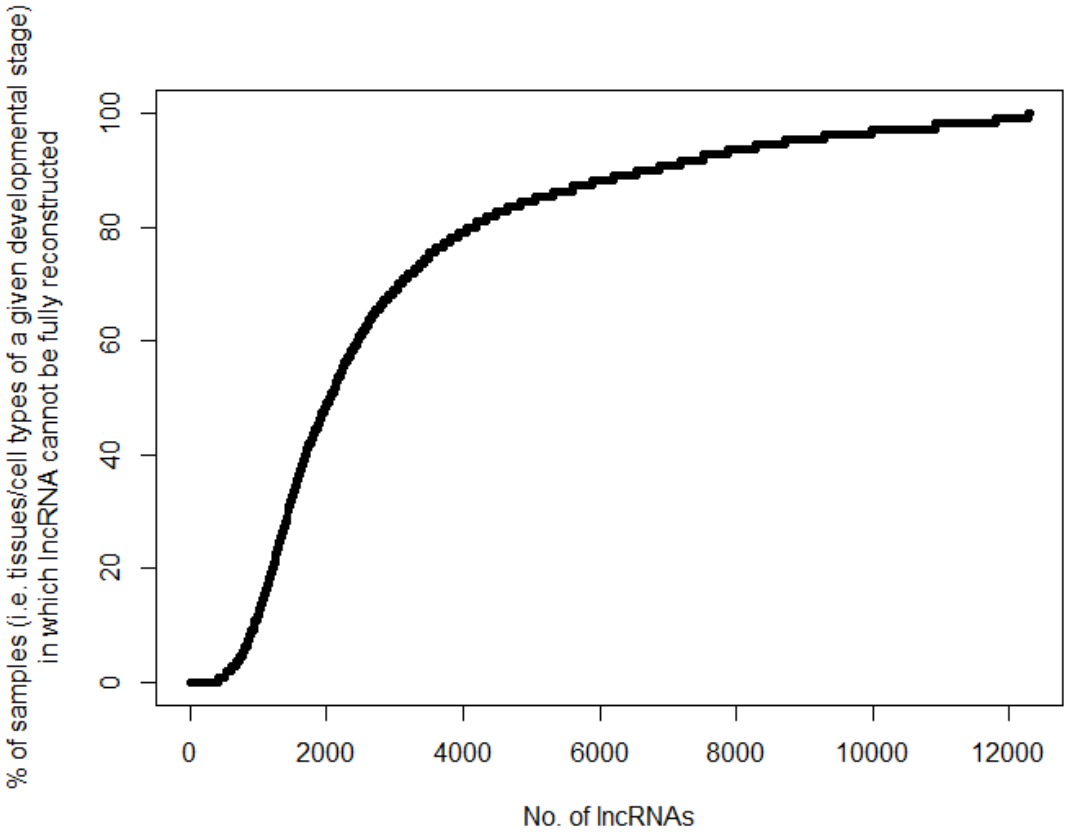
965



966

967 **Figure 2.** The stochastic detection and assembly of lncRNAs by RNA-seq libraries – a
968 consequence of limitations in sequencing breadth and depth – suggests that for a given
969 species, only a subset of the total lncRNRA transcriptome is likely to be captured.
970 Nevertheless, the number of candidate lncRNAs for that species can be increased if directly
971 mapping, to a positionally conserved region of the genome, the lncRNAs from either a related
972 (sheep, goat, cattle) or more distant (human) species. Many of these mapped lncRNAs (which
973 could not be completely reconstructed with the RNA-seq libraries of that species) are
974 nevertheless detectably expressed.

975



976

977 **Figure 3.** Proportion of samples in the sheep expression atlas for which a candidate lncRNA
978 (n = 12,296) cannot be fully reconstructed. The atlas comprises 429 RNA-seq libraries,
979 representing 110 distinct samples; that is, each sample is a tissue/cell type at a given
980 developmental stage, with up to 6 replicates per sample. 22 candidate lncRNAs cannot be
981 reconstructed in any given sample (i.e., the proportion of samples is 100%). These lncRNAs
982 could only be assembled after pooling data from multiple samples. Data for this figure is
983 given in Table S22.

984

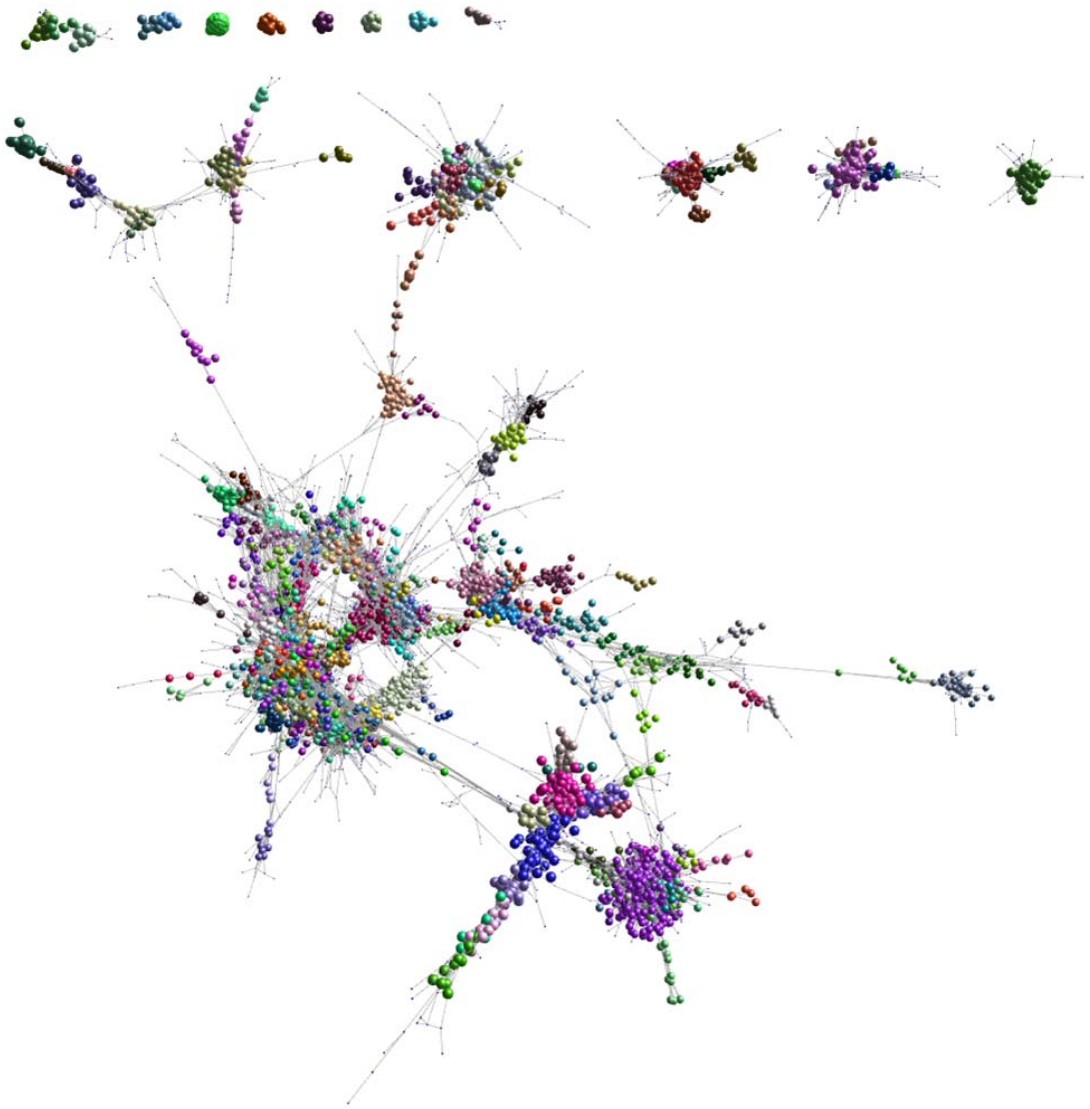


Figure 4. 3D visualisation of a gene-to-gene correlation graph. Each node (sphere) represents a gene. Nodes are connected by edges (lines) that represent Pearson's correlations between the two sets of expression level estimates, at a threshold greater than or equal to 0.95. The graph comprises 11,841 nodes and 2,214,099 edges. Genes cluster together according to the similarity of their expression profiles (i.e. their degree of co-expression), with clusters (coloured sets of nodes) determined using the MCL algorithm. Expression level estimates for the lncRNAs in this graph are given in Table S19. The genes comprising each co-expression

cluster are given in Table S23. Those lncRNAs co-regulated with protein-coding genes will be found within the same co-expression cluster.

Supplementary Material

Dataset S1. 63 sequence assemblies (as GTFs): all possible combinations when merging 6 different sets of RNA-seq reads (available via the University of Edinburgh DataShare portal; http://dx.doi.org/10.7488/ds/2284).

Table S1. Candidate sheep lncRNAs: a shortlist of novel gene models (plus independently confirmed known gene models) assessed for coding potential using CPC, CPAT, PLEK, blastp vs. Swiss-Prot, and HMMER vs. Pfam.

Table S2. Candidate goat lncRNAs: a shortlist of novel gene models assessed for coding potential using CPC, CPAT, PLEK, blastp vs. Swiss-Prot, and HMMER vs. Pfam.

Table S3. Sheep gene models considered non-coding by either CPC, CPAT or PLEK but showing sequence homology to either a known protein (in Swiss-Prot) or protein domain (in Pfam-A).

Table S4. Goat gene models considered non-coding by either CPC, CPAT or PLEK but showing sequence homology to either a known protein (in Swiss-Prot) or protein domain (in Pfam-A).

Table S5. Number of novel sheep lncRNA gene models identified per chromosome.

Table S6. Number of novel goat lncRNA gene models identified per chromosome.

Table S7. Number of novel sheep lncRNA gene models identified, by category.

Table S8. Number of novel goat lncRNA gene models identified, by category.

Table S9. Alignments of novel sheep lncRNA gene models to goat, cattle and human lncRNAs.

Table S10. Alignments of novel goat lncRNA gene models to sheep, cattle and human lncRNAs.

Table S11. Presence of intergenic lncRNAs both in sheep and cattle, in regions of conserved synteny.

Table S12. Presence of intergenic lncRNAs both in sheep and goat, in regions of conserved synteny.

Table S13. Presence of intergenic lncRNAs both in cattle and goat, in regions of conserved synteny.

Table S14. Presence of intergenic lncRNAs both in sheep and human, in regions of conserved synteny.

Table S15. Presence of intergenic lncRNAs both in goat and human, in regions of conserved synteny.

Table S16. Presence of intergenic lncRNAs both in cattle and human, in regions of conserved synteny.

Table S17. High-confidence lncRNA pairs, those conserved across species both sequentially and positionally.

Table S18. lncRNAs inferred in one species by the genomic alignment of a transcript assembled with the RNA-seq libraries from a related species.

Table S19. Expression level estimates for 13,047 novel sheep lncRNAs, as transcripts per million (TPM), assessed using 71 adult RNA-seq libraries (11 tissues plus one cell type in two different conditions, each sequenced in 6 individuals).

Table S20. Expression level estimates for 4392 novel goat lncRNAs, as transcripts per million (TPM), assessed using 54 RNA-seq libraries (20 tissues plus one cell type in two different conditions, each sequenced in 4 individuals).

Table S21. Reproducibility of sheep lncRNA gene models when merging all combinations of data from 6 adults (3 female, 3 male), each individual having sequenced a common set of RNA-seq libraries (comprising 31 tissues/cell types).

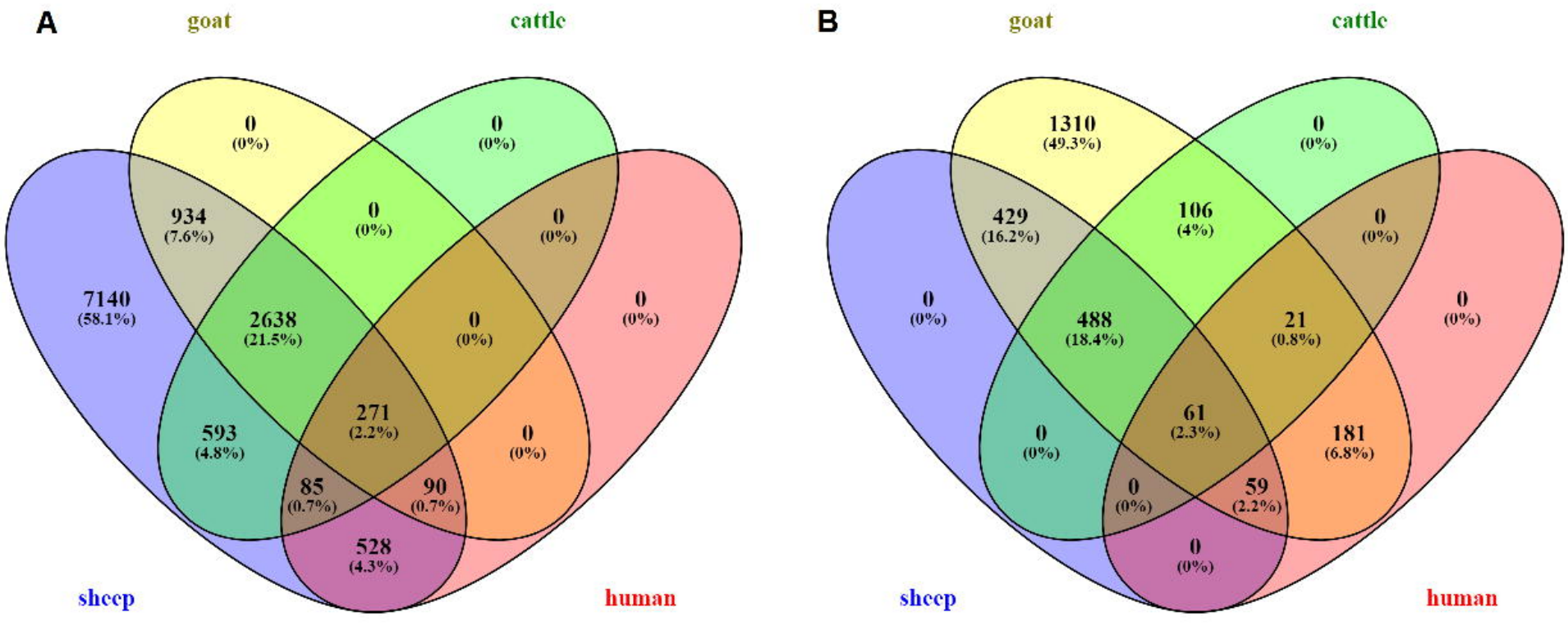
Table S22. Number of sheep expression atlas RNA-seq libraries (out of 429 in total) in which a candidate lncRNA gene model cannot be fully reconstructed.

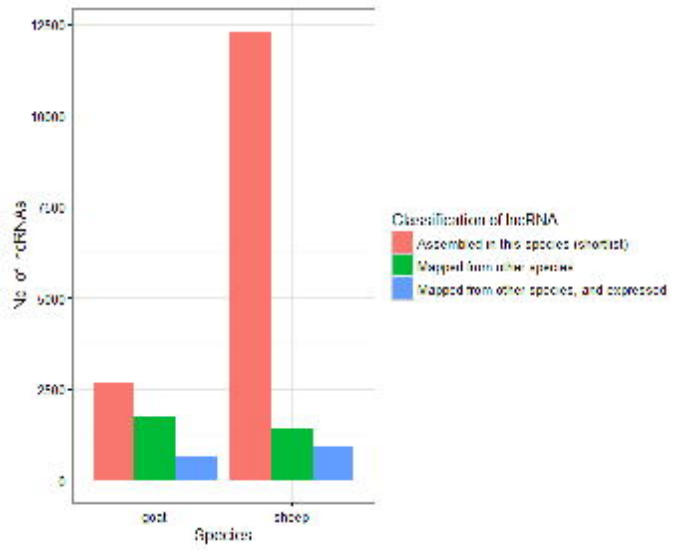
Table S23. Genes within each co-expression cluster, after network analysis of the sheep RNA-seq libraries.

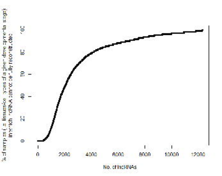
Table S24. No. of lncRNAs co-expressed with protein-coding genes.

Table S25. Distance between lncRNAs and protein-coding genes within the same co-expression cluster, on the same chromosome.

Table S26. Correlation between the expression profile of sheep lncRNAs and their nearest protein-coding genes, both 5' and 3'.


A
goat
cattle
0
(0%)
0
(0%)
934
(7.6%)
0
(0%)
0
(0%)
7140
(58.1%)
2638
(21.5%)
0
(0%)
0
(0%)
271
(2.2%)
593
(4.8%)
0
(0%)
85
(0.7%)
90
(0.7%)
528
(4.3%)
sheep
human
B
goat
cattle
1310
(49.3%)
0
(0%)
429
(16.2%)
106
(4%)
0
(0%)
0
(0%)
488
(18.4%)
21
(0.8%)
0
(0%)
61
(2.3%)
0
(0%)
181
(6.8%)
0
(0%)
59
(2.2%)
0
(0%)
sheep
human


12500
10000
7500
5000
2500
0
goat
sheep
Species
Classification of lncRNA
Assembled in this species (control)
Mapped from other species
Mapped from other species, and expressed



0
20
40
60
80
100
0
2000
4000
6000
8000
10000
12000
No. of lncRNAs