Alignment-based RNA-Seq processing to generate the sheep gene expression atlas BAM files

To complement the alignment-free gene expression atlas for the domestic sheep [1], we generated a parallel dataset using an alignment-based processing pipeline. This data can be used both as a confirmatory validation of Kallisto's expression estimates, and because alignment-based – unlike alignment-free – methods can be used to identify novel, and revise existing, gene and transcript models.

After screening with FastQC v0.11.2 [2], all raw reads were cleaned using Trimmomatic v0.35 [3] with parameters 'TRAILING:20 SLIDINGWINDOW:4:20 MINLEN:100.' These parameters, respectively, remove bases from the end of a read if they are below a Phred score of 20, clip the read if the average Phred score within a 4bp sliding window advanced from the 5' end falls below 20, and specifies a minimum read length of 100bp (Phred scores are the logarithm of the probability that a base was called incorrectly, i.e., a score of 20 is equivalent to 99% accuracy). The parameter 'HEADCROP:8' was also used for the blastocyst samples as these were generated using the NuGen Ovation Single Cell RNA-Seq System (http://www.nugen.com/sites/default/files/M01363_v10 - User Guide, Ovation Single Cell RNA-Seq System.pdf).

These cleaned reads were then aligned against the reference genome (Oar v3.1) using HISAT2 v2.0.4 [4] with the parameter --dta (optimise for downstream transcriptome assembly) and default alignment scoring parameters. In brief, HISAT2 assigns scores to alignments equal to the sum of the scores for individual mates (i.e. two scores for paired-end alignments, one for single-end [unpaired] alignments). Reads are required to align in full and are scored according to successful matching and penalised for mismatching: +2 for each position where a base in the read exactly matches that of the reference, -1 for any ambiguous base (N) on either the read or the reference, -(5+3*n*) for any gap opening or extension (of length *n*) on either the read or reference, -(2 + floor(4 x min(*Q*,40)/40)) where *Q* is the Phred quality score for any non-N mismatch between the read and reference. The minimum alignment score for reporting is -18. If there are a set of multiple valid alignments, the primary alignment is considered the one whose score is greater than or equal to any other member of the set. In the case of equal scores, this primary alignment is assigned arbitrarily.

Using SAMtools view v1.2 [5], the set of primary (uniquely highest scoring) alignments was obtained using parameters -F 256 (which removes non primary alignments) and -F 12 (which removes all reads that are not mapped and whose mate is not mapped; this primarily – but not exclusively – retains those reads mapping in a proper pair, i.e. those located on the same chromosome, one on either strand, orientating towards each other and spanning a reasonable insert size). The set of singleton reads (which map but have an unmapped mate) was obtained using SAMtools view with parameters -F 4 -f 8 -F 256. Finally, these two subsets were merged using Picard Tools [6] to create a file of uniquely mapped reads.

The tissue-specific transcriptome was assembled for this set of mapped reads using StringTie v1.2.3 [7] with default parameters, generating a corresponding GTF. In order to create a uniform, non-redundant set of transcripts for comparative purposes, these individual

GTFs were then merged using StringTie --merge. StringTie was then re-run for each sample with parameters -G, -b and -e, now specifying not the reference annotation (ftp://ftp.ensembl.org/pub/release-81/gff3/ovis_aries/Ovis_aries.Oar_v3.1.81.gff3.gz, downloaded 18th August 2015) but the merged GTF. Finally, gene-level expression estimates – comparable across samples produced by the same experimental protocol – were calculated using the R/Bioconductor package Ballgown [8].

By default, genes are assigned an 'mstrg' ID, a unique, StringTie-specific identifier. However, adjacent genes in the reference annotation may otherwise share an mstrg ID if reads map between them. This is not necessarily incorrect as multiple genes may be transcribed as a single operon, but it does introduce ambiguity into per-gene TPM estimates as there is not always a one-to-one correspondence of mstrg to Ensembl gene IDs. In these cases, we consider gene-level TPM to be identical for each gene assigned a single mstrg ID. That this occurs is because we retain the default StringTie parameter of -g 50 (minimum gap locus separation value = 50bp). In this case, StringTie will merge reads that map closer than 50bp in the same processing bundle, closing coverage gaps so as to increase the number of full-length structures possible for lowly expressed genes. This is a trade-off between sensitivity and specificity and is most pronounced in gene-dense regions: if coverage gaps were not filled, the assembly will be more fragmented (although there will be fewer erroneous merges).

The StringTie assembly is highly accurate with respect to the existing (Oar v3.1) annotation, successfully reconstructing almost all exon (96%), transcript (98%) and gene (99%) models. Nevertheless, StringTie also predicts many novel models, although in the absence of experimental verification, it is not easy to predict which are genuine, as opposed to stochastic noise in RNA processing or assembly artefacts [7]. The number of false positives

is also likely exacerbated by the merger of both mRNA-Seq and total RNA-Seq data. The latter measures nascent (ongoing) transcription [9] and consequently has a larger proportion of retained introns arising from incompletely spliced pre-mature (nuclear) mRNA [10]. In any case, in the context of transcript annotation, false positives are easier to identify and correct than false negatives.

References

1. Clark EL, Bush SJ, McCulloch MEB, Farquhar IL, Young R, Lefevre L, et al. A high resolution atlas of gene expression in the domestic sheep (Ovis aries). PLOS Genetics. 2017;13(9):e1006997. doi: 10.1371/journal.pgen.1006997.
2. Andrews S. FastQC: a quality control tool for high throughput sequence data 2010. Available from: http://www.bioinformatics.babraham.ac.uk/projects/fastqc.
3. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30(15):2114-20. doi: 10.1093/bioinformatics/btu170.
4. Kim D, Langmead B, Salzberg SL. HISAT: a fast spliced aligner with low memory requirements. Nat Meth. 2015;12(4):357-60. doi: doi.org/10.1038/nmeth.3317.
5. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25(16):2078-9. Epub 2009/06/10. doi: 10.1093/bioinformatics/btp352.
6. The Broad Institute. Picard: a set of tools (in Java) for working with next generation sequencing data in the BAM format 2014. Available from: http://broadinstitute.github.io/picard/.
7. 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(3):290-5. doi: doi.org/10.1038/nbt.3122.
8. Frazee AC, Pertea G, Jaffe AE, Langmead B, Salzberg SL, Leek JT. Ballgown bridges the gap between transcriptome assembly and expression analysis. Nat Biotech. 2015;33(3):243-6. doi: 10.1038/nbt.3172 http://www.nature.com/nbt/journal/v33/n3/abs/nbt.3172.html#supplementary-information.
9. Ameur A, Zaghlool A, Halvardson J, Wetterbom A, Gyllensten U, Cavelier L, et al. Total RNA sequencing reveals nascent transcription and widespread co-transcriptional splicing in the human brain. Nat Struct Mol Biol. 2011;18(12):1435-40. doi: http://www.nature.com/nsmb/journal/v18/n12/abs/nsmb.2143.html#supplementary-information.
10. Zhang X, Rosen BD, Tang H, Krishnakumar V, Town CD. Polyribosomal RNA-Seq reveals the decreased complexity and diversity of the Arabidopsis translatome. PLoS One. 2015;10(2):e0117699. Epub 2015/02/24. doi: 10.1371/journal.pone.0117699.