ROSLIN

Cap analysis gene expression (CAGE) data processing workflow

The single end Illumina CAGE sequence data was processed using in house scripting (bash and R) on University of Edinburgh high performance computing facility (Eddie3). The CAGE tag data statistical analysis was carried out in **R v ≥ 3.5** [1] using **CAGEfightR package v.1.5.1** [2,3] . Briefly the multiplexed FASTQ files were de-multiplexed using **FASTX-Toolkit v0.0.13** (http://hannonlab.cshl.edu/fastx_toolkit) perl script (*fastx_barcode_splitter.pl --bcfile barcode.list --bol --mismatch 0*).

The barcode separated reads were then trimmed in order to remove any CAGCAG finger print or 3' illumina PCR adapter and occasional poly-A tail using **tagDust2 v2.2** [4] and the following read/tag structure (*tagdust INPUT.fastq.gz -t 8 -1 B:3NBARCODE -2 F:CAGNNN -3 R:N -4 P:TCGTATGCCGTCTTCTGCTT -dust 100 -o OUTPUT_trimmed.fq.gz*).

The trimmed read were then mapped to the NCBI's sheep genome (Ovis aries Oar_rambouillet_v1.0 NCBI annotation release 103) using **BowTie2 tool v2.3.5.1** [5,6] and the following flags (*bowtie2 -p 8 --met 1 --very-sensitive -x Oar_ram1 -U $INPUT*). The mapped BAM files were then processed for base pair resolution strand specific read counts using **bedtools v2.29.0** [7] and the following scripts

```
bedtools genomecov -ibam $INPUT -d -strand + | \
awk -v width=1 '!($1~/^NW/)&&($3!=0) {print $1,$2,$2+width,$3}' > ${NAME}.plus.bedGraph &\
bedtools genomecov -ibam $INPUT -d -strand - | \
awk -v width=1 '!($1~/^NW/)&&($3!=0) {print $1,$2,$2+width,$3}' > ${NAME}.minus.bedGraph).
```

The strand specific bedGraph files were converted to bigWig format using UCSCs tool **BedGraphToBigWig** [8] in order to be used in CAGEfightR package. The **CAGEfightR** workflow was used in order to analyse the 56 tissue CAGE-Seq samples [3,9].

NB. The BAM files were also processed using ***CAGEr*** [10] *for quality control of the library size and annotation distribution.*

References:

1. R Core Team. R: A language and environment for statistical computing. *R Found. Stat. Comput.* (2017).
2. Thodberg, M., Thieffry, A., Vitting-Seerup, K., Andersson, R. & Sandelin, A. CAGEfightR: Cap Analysis of Gene Expression (CAGE) in R/Bioconductor. *bioRxiv* 310623 (2018). doi:10.1101/310623
3. Sandelin, A. & Lun, A. A step-by-step guide to analyzing CAGE data using R / Bioconductor [version 1 ; peer review : 1 approved] Malte Thodberg. **886,** 1–44 (2019).
4. Lassmann, T. TagDust2: a generic method to extract reads from sequencing data. *BMC Bioinformatics* **16,** 24 (2015).
5. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9,** 357–9 (2012).
6. Langmead, B., Wilks, C., Antonescu, V. & Charles, R. Scaling read aligners to hundreds of threads on general-purpose processors. *Bioinformatics* **35,** 421–432 (2019).
7. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26,** 841–842 (2010).
8. Kent, W. J., Zweig, A. S., Barber, G., Hinrichs, A. S. & Karolchik, D. BigWig and BigBed: enabling browsing of large distributed datasets. *Bioinformatics* **26,** 2204–2207 (2010).
9. Thodberg, M., Thieffry, A., Vitting-Seerup, K., Andersson, R. & Sandelin, A. CAGEfightR: analysis of 5′-end data using R/Bioconductor. *BMC Bioinformatics* **20,** 487 (2019).
10. Haberle, V., Forrest, A. R. R. R., Hayashizaki, Y., Carninci, P. & Lenhard, B. CAGEr: precise TSS data retrieval and high-resolution promoterome mining for integrative analyses. *Nucleic Acids Res.* **43,** e51–e51 (2015).