Enrichment of causative variants in tissue-specific and shared ATAC-Seq peaks in cattle

C. Yuan[1], L. Tang[1], T. Lopdell[2], C. Oget-Ebrad[1], G. Costa Monteiro Moreira[1], J.L. Gualdron[1], Z. Cheng[3], M. Salavati[3,4], D.C. Wathes[3], M.A. Crowe[5], GplusE consortium[5], W. Coppieters[6], C. Charlier[1], T. Druet, [1] M. Georges[1] & H. Takeda[1*]

[1] Unit of Animal Genomics, University of Liège, 4000 Liège, Belgium; [2] Research and Development, Livestock Improvement Corporation, 3240 Hamilton, New Zealand; [3] Royal Veterinary College, AL9 7TA Herts, United Kingdom; [4] Roslin Institute, Royal School of Veterinary Studies, Midlothian, United Kingdom; [5] School of Veterinary Medicine, University College Dublin, Dublin 4, Ireland; [6] GIGA Genomics platform, University of Liège, 4000 Liège, Belgium. * htakeda@uliege.be

Abstract

To profile putative regulatory variants in the bovine and test their validity, we first characterized ~1 million putative regulatory elements by 'Assay for Transposase Accessible Chromatin using Sequencing (ATAC-Seq)' in 77 distinct types of tissues and cells. The elements collectively accounted for 10.3% of the genome, from which we identified 938,374 common DNA variants. We then analyzed bovine blood and liver RNA-Sequencing data and identified 5,336 and 3,613 *cis*-expression Quantitative Trait Loci (eQTL), respectively. Credible sets of DNA variants that drive these *cis*-eQTL were enriched in tissue-specific as well as tissue-shared putative regulatory elements identified by ATAC-Seq. These results pave the way to utilizing ATAC-Seq information to improve genomic selection.

Introduction

Genomic selection has had a tremendous impact on livestock selection in the last ten years (e.g., Georges *et al.*, 2019). The accuracy of selection nonetheless remains inferior to what may be expected given the heritability of the selected traits. This could be due to a number of factors including that all variants are generally given an equivalent weight in computing the additive relationship between animals for Restricted Maximum Likelihood analyses or equivalent prior probabilities of variant effects in Bayesian approaches. Only a minority of variants are indeed causative, with the remainder being (at best) passenger variants in linkage disequilibrium (LD) with one or more of the causative variants. It is generally believed that knowing the causative variants, or at least those that are more likely to be, may help further improve the accuracy of genomic selection.

Amongst causative variants, coding variants (including missense, nonsense, frameshift and splice site variants) are easily recognized. However, these account for only a limited part of the genetic variance for complex phenotypes. It is increasingly apparent that most of the genetic variation for complex traits is due to regulatory variants that act either by perturbing the expression profile of genes located in *cis*, or by perturbing the gene regulatory network and affecting the expression profile of a restricted number of core genes in *trans*. Active regulatory elements can be recognized by virtue of epigenetic features including chromatin accessibility, histone marks, transcriptional activity, their participation in loop structures, and transcription factor occupancy. However, regulatory variants remain difficult to identify, as the effect of polymorphisms on the functionality of proximal and distant *cis*-acting regulatory elements is difficult to predict.

In an effort to identify putative regulatory variants in the bovine, we herein report (i) the generation of a comprehensive catalogue of bovine putative regulatory elements identified by Assay for Transposase Accessible Chromatin using Sequencing (ATAC-Seq), (ii) the generation of a catalogue of common DNA variants that map to identified proximal and distal putative regulatory elements, (iii) the demonstration that variants that drive expression Quantitative Trait Loci (eQTL) in liver and blood are more likely to map to regulatory elements that are active in the relevant tissue, and hence that variants in these regulatory elements are more likely to be causative.

Materials & Methods

ATAC-Seq. Libraries were constructed following Corces *et al.* (2017) using 78 tissues (65 tissue-types) collected from two juvenile males and 11 archived frozen samples (4 tissue types). Additionally, fifteen public ATAC-Seq data sets (eight tissue/cell-types) were included (Fang *et al*., 2019, Foissac *et al*., 2019, Halstead *et al*., 2020, Johnston *et al*., 2021). The ATAC-Seq reads were mapped to the bovine reference genome (ARS-UCD1.2) using Bowtie2 (Langmead and Salzberg, 2012) and peaks called with Macs2 (Zhang *et al.*, 2008). Peaks were merged across samples to build a reference chromatin accessibility map following Meuleman *et al*. (2020). The complexity of peak patterning was decomposed using non-negative matrix factorization (NMF) (Meuleman *et al*., 2020). We used transcriptome data from Fang *et al*. (2020) for a correlation analysis.

eQTL analysis. 224 blood and 176 liver RNA-Seq samples collected from Holstein-Friesian cows were analyzed (Wathes *et al*., 2021, Lee *et al*., 2021). The reads were mapped to the reference genome using HISAT2 (Kim *et al.*, 2015). Transcript abundances were estimated using StringTie (Pertea *et al.*, 2015) based on a reference gene annotation (Bos_taurus.ARS-UCD1.2.105.gtf). The data were normalized within sample using DESeq2 (Anders and Huber, 2010) and across samples using inverse normal transformation. Genotypes obtained with 770 K single nucleotide polymorphism (SNP) arrays were further imputed to whole genome using Minimac4 (Das *et al*., 2016) with the Damona population as a reference, yielding 8.4 million SNPs and 1.3 million insertion-deletions (INDELs) with minor allele frequency (MAF) > 0.02. *Cis*-eQTL analyses were conducted using residuals corrected for country and hidden PEER factors (Stegle *et al*., 2010) under an additive model using QTLtools (false discovery rate (FDR) < 0.001) (Delaneau *et al*., 2017). Enrichment analyses of eQTL variants in ATAC-Seq peaks were conducted following Trynka *et al.* (2017).

Results

Generating a catalogue of bovine cis-acting gene regulatory elements. 89 in-house and 15 publicly available ATAC-Seq data sets (including 77 tissue/cell types) were analyzed and identified on average 166,999 ATAC-Seq peaks per sample. To build a chromatin accessibility reference map, overlapping peaks were merged across the samples, yielding 977,261 chromatin accessible sites. 16 components (representing immune, liver, muscle, central nervous system, etc) were identified from the 977,261 chromatin accessibility × 104 sample matrix by NMF (Figure 1). Public RNA-Seq data for 58 tissue types were used to search for correlations between chromatin accessibility and gene expression levels to connect regulatory elements with target genes. Highly significant enrichment of positive correlations for genes ≤ 250 Kb from ATAC-Seq peaks were observed, i.e., potential enhancer effects, without any enrichment for negative correlations, which would be expected for silencers.

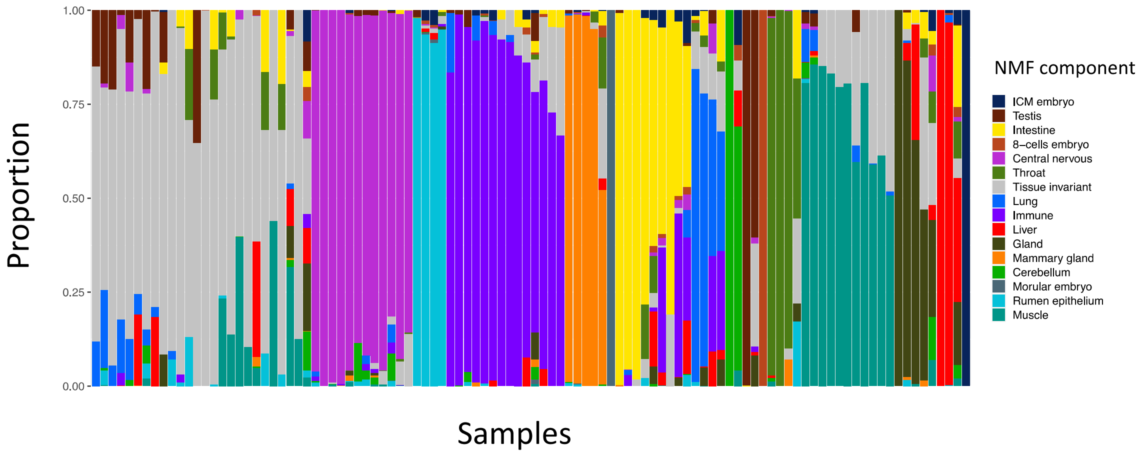

Figure 1: Identification of 16 components by NMF-decomposition of a 977,261 chromatin accessibility × 104 sample matrix and importance of the components in the bovine tissues.

Generating a catalogue of common variants mapping to cis-acting regulatory elements. A catalog of DNA variants called from 264 Holstein-Friesian animals containing ~7.6 million SNPs and ~1.2 million INDELs with MAF > 0.05 were used. Of these 29,494 (0.34%) mapped to proximal and 909,074 (10.4%) to distal putative regulatory elements accounting respectively for 0.37% and 9.9% of sequence space, hence providing no strong evidence for detectable purifying selection.

Identifying bovine cis-eQTL in blood and liver. To identify putative causative variants affecting gene expression, eQTL analyses using 224 blood and 176 liver RNA-Seq data were performed. 5,336 (37% of genes) and 3,613 (23%) significant eQTL in the blood and liver, respectively, were obtained. The proportion of blood-eQTL that would also operate in liver (π_1) was estimated by following Storey & Tibshirani (2003) at 74%, while the proportion of liver-eQTL that would also operate in blood was estimated at 81%. These unexpectedly high values could in part be due to blood irrigating the liver.

eQTL-driving variants are enriched in variants mapping to ATAC-Seq peaks. It was assumed that, if variants mapping to ATAC-Seq peaks were indeed enriched in causative variants, they should be enriched in credible sets of variants driving *cis*-eQTL effects (i.e. the lead *cis*-eQTL SNPs and syntenic variants in LD with it). The significance of the overlap between *cis*-eQTL credible sets and ATAC-Seq peaks was evaluated by NMF components (by assigning each peak to its dominant NMF component). Credible variant sets ($r^2 \geq 0.8$) for 3,336 blood-specific *cis*-eQTL were significantly ($p < 10^{-5}$) enriched in variants overlapping ATAC-Seq peaks that were assigned to the immune (1.21-fold enrichment), tissue-shared (1.17-fold) and pulmonary (1.15-fold) components. Credible sets for 1,613 liver-specific *cis*-eQTL were enriched in variants mapping to ATAC-Seq peaks assigned to the hepatic (1.42-fold), tissue-shared (1.21-fold) and pulmonary (1.15-fold) components. Permutation testing conditional on proximal vs distal status of the ATAC-Seq peaks indicated that the observed enrichment was not due to an enrichment of variants mapping to promotor regions irrespective of ATAC-Seq chromatin accessibility.

Discussion

These results indicate that genetic variants causing *cis*-eQTL are enriched in variants that map to ATAC-Seq peaks marking accessible chromatin regions, and hence likely to active regulatory elements, in the corresponding tissue. This suggests that sets of variants overlapping ATAC-Seq peaks are enriched for regulatory variants, and hence potentially driving organismal phenotypes. Assigning such variants with extra weight in genomic selection may therefore improve its accuracy. However, this putative improvement will dependent on (i) the proportion of variants mapping to ATAC-Seq peaks that are indeed causative for a trait of interest, and (ii) the proportion of causative regulatory variants that map to ATAC-Seq peaks. Further analyses are in progress to estimate these features, and to evaluate the impact of the identified variants sorted by the NMF component on genomic selection.

References

Anders S., and Huber W. (2010) Genome Biol 11(10):R106. https://doi.org/10.1186/gb-2010-11-10-r106

Corces M.R., Trevino A.E., Hamilton E.G., Greenside P.G., Sinnott-Armstrong N.A. *et al.* (2017) Nat Methods 14(10):959-962. https://doi.org/10.1038/nmeth.4396

Das S., Forer L., Schönherr S., Sidore C., Locke A.E. *et al.* (2016) Nat Genet 48(10):1284-1287. https://doi.org/10.1038/ng.3656

Delaneau O., Ongen H., Brown A.A., Fort A., Panousis N.I. *et al.* (2017) Nat Commun 8:15452. https://doi.org/10.1038/ncomms15452

Fang L., Liu S., Liu M., Kang X., Lin S. *et al.* (2019) BMC Biol 17(1):68. https://doi.org/10.1186/s12915-019-0687-8

Fang L., Cai W., Liu S., Canela-Xandri O., Gao Y. *et al.* (2020) Genome Res 30(5):790-801. https://doi.org/10.1101/gr.250704.119

Foissac S., Djebali S., Munyard K., Vialaneix N., Rau A. *et al.* (2019) BMC Biol 17(1):108. https://doi.org/10.1186/s12915-019-0726-5

Georges M., Charlier C., Hayes B. (2019) Nat Rev Genet 20(3):135-156. https://doi.org/10.1038/s41576-018-0082-2

Halstead M.M., Ma X., Zhou C., Schultz R.M., Ross P.J. (2020) Nat Commun 11(1):4654. https://doi.org/10.1038/s41467-020-18508-3

Halstead M.M., Kern C., Saelao P., Wang Y., Chanthavixay G. *et al.* (2020) BMC Genomics 21(1):698. https://doi.org/10.1186/s12864-020-07078-9

Johnston D., Kim J., Taylor J.F., Earley B., McCabe M.S. *et al.* (2021) BMC Genomics 22(1):14. https://doi.org/10.1186/s12864-020-07268-5

Kim D., Langmead B., and Salzberg S.L. (2015) Nat Methods 12:357-360. https://doi.org/10.1038/nmeth.3317

Langmead B., and Salzberg S.L. (2012) Nat Methods 9:357-359. http://doi.org/10.1038/NMETH.1923

Lee Y.L., Takeda H., Costa Monteiro Moreira G., Karim L., Mullaart E. *et al.* (2021) PLoS Genet 17(7):e1009331. http://dx.doi.org/10.1371/journal.pgen.1009331

Meuleman W., Muratov A., Rynes E., Halow J, Lee K. *et al.* (2020). Nature 584:244-251. https://doi.org/10.1038/s41586-020-2559-3

Pertea M., Pertea G.M., Antonescu C.M., Chang T., Mendell J.T. *et al.* (2015) Nat Biotechnol 33(3):290-295. https://doi.org/10.1038/nbt.3122

Stegle O., Parts L., Durbin R., Winn J. (2010) PLoS Comput Biol 6(5):e1000770. https://doi.org/10.1371/journal.pcbi.1000770

Storey J.D. and Tibshirani E. (2003) Proc Natl Acad Sci USA 100(16):9440-9445. https://doi.org/10.1073/pnas.1530509100

Trynka G., Westra H., Slowikowski K., Hu X., Xu H. *et al.* (2015) Am J Hum Genet 97(1):139-152. http://dx.doi.org/10.1016/j.ajhg.2015.05.016

Wathes, D.C., Becker F., Buggiotti L., Crowe M.A., Ferris C. *et al.* (2021) Ruminants 1(2):147-177. https://doi.org/10.3390/ruminants1020012

Zhang Y., Liu T., Meyer C.A., Eeckhoute J., Johnson D. *et al.* (2008) Genome Biol 9: R137. https://doi.org/10.1186/gb-2008-9-9-r137