The importance of annotations (reference genome and parent gene) for the study of circRNAs

Annie Robic, Thomas Faraut, Chloé Cerutti, Julie Demars, Christa Kühn

The importance of annotations (reference genome and parent gene) for the study of circRNAs

Annie Robic[1], Thomas Faraut[1], Chloé Cerutti[1], Julie Demars[1] and Christa Kühn[2,3]

1 GenPhySE, Université de Toulouse, INRAE, ENVT, 31326 Castanet Tolosan, France
2 Institute Genome Biology, Leibniz Institute for Farm Animal Biology (FBN), Dummerstorf, Germany
3 Faculty of Agricultural and Environmental Sciences, University of Rostock, Rostock, Germany

Circular transcripts can be of several types, although the majority of circular RNAs (circRNAs) are generated at the expense of a linear transcript as backsplicing competes with linear splicing. Many pipelines have been developed to identify circRNA in RNA-seq datasets depleted for ribosomal RNA with the core principle of identifying reads including a circular junction.

However, sporadic circularization events should be excluded from circRNA lists for a good compromise between exhaustive identification of circRNAs and false positive data. In our study, we considered two widely used detection pipelines (CIRCexplorer2 (CE2) and CIRI2) as well as an in-house approach, and applied them on bovine, porcine and ovine datasets to understand the differences in their circRNA output lists. Substantial differences in results reflect the alternative circRNA detection strategies: CE2 only retains exonic and intronic lariat circRNAs compatible with an annotated gene, while CIRI2 retains exonic circRNAs due to requiring a junction of two putative exons with canonical splicing site signals. We show that considering only an intersection of circRNA output from these two pipelines is not the final compromise but only an option, which still requires applying a threshold to discard sporadic circulation events. All pipelines provide a list of circRNAs with an associated gene name, but some pipelines proceed to a comprehensive identification of the parent gene (CE2), while others (CIRI2) only propose a gene name based on raw mapping coordinates. We showed that a poor reference genome assembly in a given region can lead to the detection of artifactual circRNAs (possibly detected by CIRI2). In addition, we demonstrated that circRNAs can also originate from incompletely annotated regions (possibly detected by CIRI2) and that all types of genes can produce circRNAs even RNA genes (still incompletely annotated in livestock species). In the three species considered in this study, and with the current state of knowledge of the respective reference genomes and gene annotation, we suggest working with only properly annotated circRNAs.