Scientific Symposium FAIR Data Sciences for Green Life Sciences
12 December 2018 | Wageningen | the Netherlands

Conference Proceedings | DOI: https://doi.org/10.18174/FAIRdata2018.16267

In silico predictions of variant deleteriousness in the genomes of pig species

Gross, C.[1,3], M. Derks[2], D. de Ridder[3], M. Reinders[1]

[1] Delft Bioinformatics Lab, Delft University of Technology, Van Mourik Broekmanweg 6, 2628XE Delft, The Netherlands
[2] Animal Breeding and Genetics, Wageningen University & Research, Droevendaalsesteeg 1, 6708PB Wageningen, The Netherlands
[3] Bioinformatics Group, Wageningen University & Research, Droevendaalsesteeg 1, 6708PB Wageningen, The Netherlands
Corresponding author's e-mail: c.gross@tudelft.nl

Predicting the deleteriousness of observed genomic variants has taken a step forward with the development of the Combined Annotation Dependent Depletion (CADD) [1] methodology, as it allows for comparable evaluations of variants on a genome-wide scale for coding and non-coding variants respectively. The underlying techniques allow to be reproduced for any species. Sets of putative benign and deleterious variants are generated and used to train a classifier which is capable of differentiating between deleterious and benign mutations. The data set is derived by studying the evolutionary past of the species that is investigated. Benign variants are identified as those mutations which have become fixed between the current population and an inferred ancestral genome; expected deleterious variants are simulated on the basis of substitution rates derived from multiple sequence alignments containing species at different evolutionary distances to the species of interest. All variants are annotated with a plethora of annotations that form the data set on which the final model is learned. Previously, such classifiers were only developed for the investigation of human genomes due to the large amount of genomic information necessary. In a feasibility study in mouse [2] we have shown that even with a limited amount of genomic information, meaningful models can be created. Further, we showed the model constraints and under which conditions the model performs best. To follow up on this research we developed a model capable of scoring variations with respect to their deleteriousness in the genomes of pig species. These scores are discriminative on test sets and may help to identify breeding lines suffering from inbreeding depression. This will enable breeders to increase the overall health of their populations by selectively removing those genetic variants. Such scores can be regularly updated when new information is available which provides breeders with a new tool for functional selection.

References

[1] M. Kircher, D.M. Witten, P. Jain, G.M. Cooper B.J. O'Roak, and J. Shendure. "A general framework for estimating the relative pathogenicity of human genetic variants". *Nature Genetics*, 46(3):310–315, 2014.
[2] C. Groß, D. de Ridder, and M.J.T. Reinders. "Predicting variant deleteriousness in non-human species: applying the CADD approach in mouse". *BMC Bioinformatics* 19:373, 2018.